

Kleptoplasty and Plastid-Targeted Gene Diversity in Complex Algae: A Quantitative Phylogenomic Approach

Study of the relative contribution of endosymbiosis and kleptoplasty
in the evolution of complex algae by automatic phylogenetic tree
parsing and gene transfer inference

Mick VAN VLIERBERGHE

A thesis submitted to University of Liège in partial fulfilment of the requirements
of the degree of Doctor of Philosophy in Science

Academic year
2025-2026

Jury:

Pierre Cardol (President)

Nicolas Magain (Secretary)

Denis Baurain (Promoter)

Sébastien Massart

Olivier De Clerck (UGent)

Richard G. Dorrell (Sorbonne Université)

*To my three beloved little monsters,
Raphaël, Coline, and Léa.
You make life so much fun!*

Acknowledgements

I would like to thank all the people who made this possible. This PhD has truly been a journey, and I couldn't have done it alone. In many ways, I am a mosaic of all the people who played their part in it, whether small or large. Even the smallest push can help make a greater leap possible.

I would like to express my gratitude to my parents, Faïza and Elie, and to my stepfather, Georges, for giving me the opportunity to study; this is a precious gift for an open mind. To my mother, for her kindness. To my father, for his out-of-the-box thinking and for bringing me back to Belgium. To my grandfather, Roland, for sharing his appetite for learning and his rigor. To my grandmother, Germaine, for sharing so much with me; she taught me so much about life, love, nature, and art. You will forever be my favourite painter!

To Marc and his family, for their ever-warm welcome; their home has been an oasis in a desert.

To Pauline, for her limitless generosity. More than a push, this was clearly a pull. I would not be on this path without her. Thank you from the bottom of my heart!

To the one without whom none of this would have been possible: Rose. My formidable partner and the exceptional mother of our beloved children, Raphaël, Coline and Léa. Thank you for your love, patience, strength, and support throughout this journey. What a team we make!

To my stepfamily, Mémé, Pépé, Carine and José who helped two young, happy, but slightly overwhelmed parents.

To my childhood friends, my second family: it has been a blast growing up with you guys!

To my colleagues and friends from the Eukaryotic Phylogenomic Unit, that shared this PhD journey from the very beginning: Rosa, Damien, Luc, Raphaël, Loïc, Manu and Ngoc. To those who joined along the way: Catherine, Valérian; and to the new ones: Louise and Marie. To our beloved interns; there were many and they made the lab lively! And let's not forget my former teachers and squash partners: Pierre and Patrick!

A special thought to Loïc, Manu and Ben: from start to finish together, brothers!

To my colleagues and some of whom, former teachers: it's a pleasure to teach with you.

To Dr. Hervé Philippe and Dr. Henner Brinkmann for their precious and insightful discussions.

To Denis, my teacher and promoter. You contaminated me with the puzzle of programming, after all, there's nothing a Perl one-liner can't solve. But more importantly, you contaminated me with the puzzle of plastid evolution, the one that brought us here, to the end of this journey, and the beginning of a new one. Thank you for your tireless support! We also make a great team!

And finally, to myself, for not giving up.

Mick Van Vlierberghe

Abstract

The origin of plastids in complex algae remains one of the most debated questions in eukaryotic evolution. Although plastid acquisition by polyphyletic lineages is classically explained by secondary (or higher-order) endosymbiosis, the evolutionary history of nucleus-encoded plastid-targeted genes is often markedly mosaic and difficult to reconcile with simple single-donor models. In many complex algal lineages, plastid-targeted proteins do not derive exclusively from the lineage that gave rise to the plastid itself, suggesting that additional processes have contributed to plastid acquisition. Among the hypotheses proposed to explain this pattern, kleptoplasty is particularly attractive, as it combines repeated exposure to diverse algal prey with direct selective pressure to maintain stolen plastids, thereby providing a plausible route toward plastid pre-adaptation and subsequent integration. This thesis investigates whether plastid-related gene mosaicism can be studied quantitatively at large scale, and whether donor diversity patterns and associated functional signals can help distinguish between different modes of plastid acquisition.

A first major objective of this work was methodological. Because plastid-related transfer inference is highly sensitive to incomplete taxonomic sampling, paralogy, and contamination, I developed and consolidated a controlled phylogenomic framework for large-scale analysis. This work included the formal benchmarking of Forty-Two, an orthology-enrichment tool designed to recover orthologous sequences from incomplete genomic and transcriptomic datasets while minimizing incorrect assignments due to paralogy or contamination. It also involved the construction of a broadly sampled plastid-related orthogroup dataset and the curation, decontamination, pooling, and dereplication of transcriptomic resources oriented toward photosynthetic lineages, particularly from the MMETSP (Marine Microbial Eukaryote Transcriptome Sequencing Project). These preparatory steps were essential, as contamination can easily mimic transfer events and thus distort biological interpretations of plastid-related gene histories. Together, these three chapters provide the technical backbone required for the comparative analyses developed in the second half of the thesis.

On this basis, I implemented an automated phylogenetic pipeline to infer plastid-related transfer events from single-gene trees in a taxonomy-aware and discovery-driven manner. Rather than relying primarily on best-hit approaches or simple sister-group relationships, the method detects transfer-compatible topologies based on paraphyletic donor assemblages surrounding monophyletic receiver clades. Individual donor–receiver events are then summarized at the species and lineage levels using donor α -diversity indices, with weighted Shannon–Wiener diversity as the main metric. This framework transforms gene-wise transfer inference into a single comparative descriptor of plastid-related donor diversity and enables direct comparison among complex algal lineages.

The global analyses revealed that nuclear-encoded plastid-targeted gene repertoires in complex algae are indeed strongly mosaic, but that this mosaicism is structured rather than random. A control based on plastid-encoded plastid-targeted genes recovered the expected donor signals, confirming that low-diversity, single-origin histories can be detected when present. Independent expert analyses of hundreds of trees further indicated that the pipeline performs as expected across many topologies. Across nucleus-encoded plastid-targeted genes, low donor α -diversity is associated with strong single-donor dominance, intermediate α -diversity with more composite but still structured histories, and high α -diversity with lineages compatible with transitional or kleptoplast-associated regimes. Although this does not establish a strict one-to-one correspondence between high donor diversity and kleptoplasty, it demonstrates that donor α -diversity captures biologically meaningful differences among lineages. Robustness analyses further show that this signal is broadly stable across alternative diversity indices, taxonomic samplings, phylogenetic models, and key event inference parameters, even though absolute values vary with analytical settings.

Several case studies illustrate the biological relevance of this framework. In euglenids, the inclusion of the kleptoplastidic species *Rapaza viridis* allowed the identification of a plastid-related core gene set shared with photosynthetic euglenids, consistent with a pre-adaptive phase preceding permanent plastid integration. Functional comparisons suggest a distinction between an early regime associated with plastid exploitation/maintenance and a later regime associated with long-term plastid deep metabolic integration beyond photosynthesis. In peridinin-containing dinoflagellates, donor diversity and functional partitioning point to a more complex plastid history than expected under a strictly canonical endosymbiotic scenario and reveal parallels with the euglenid system. Dinophysiales, although still limited by sparse and contamination-prone data, emerge as a promising model for transient plastid exploitation and maintenance. At a broader scale, donor-receiver flux analyses identify Ochrophyta as a major donor lineage across multiple recipients, highlighting the importance of reticulate gene exchange in plastid-related evolution.

Taken together, this thesis does not argue for a single universal mechanism underlying plastid evolution in complex algae. Rather, it shows that plastid-related gene mosaicism can be investigated quantitatively, reproducibly, and across multiple lineages, and that donor α -diversity provides a useful framework for interpreting plastid integration histories. Biologically, the results support a pluralistic view in which stable endosymbiosis, plastid replacement, and kleptoplasty-compatible pre-adaptive phases have all contributed, to varying extents, to the emergence and diversification of present-day complex algal lineages. Methodologically, this work provides a reusable phylogenomic framework for studying plastid-associated gene evolution while explicitly controlling for the data-quality issues that most strongly compromise transfer inference.

Table of Contents

<i>Acknowledgements</i>	v
<i>Abstract</i>	vii
<i>Table of Contents</i>	xi
I. Chapter 1 - General introduction	1
1. <i>Photosynthetic eukaryotes and plastid origins</i>	1
1.1. Eukaryotic diversity and evolution	1
2. <i>Anatomy of a complex plastid</i>	3
2.1. Endosymbiosis, nucleomorphs and EGT	3
2.2. Protein targeting to complex plastids	5
2.3. Membrane architecture of complex plastids and implications	7
3. <i>Hypotheses for the origin of complex red plastids</i>	10
3.1. The chromalveolate hypothesis	10
3.2. Chromalveolate-like hypotheses	11
3.3. Rhodoplex-like models: serial and nested endosymbioses	12
4. <i>Mosaic origin of nucleus-encoded and plastid-targeted genes</i>	15
5. <i>Pre-adaptation models and selective pressure</i>	16
5.1. Models explaining gene mosaicism	16
5.2. Pre-adaptation and selective pressure: endosymbiosis VS kleptoplasty	19
6. <i>Kleptoplasty across eukaryotes</i>	23
6.1. Diversity of prey and donor sources	23
7. <i>Diversity of complex algae</i>	24
8. <i>Working hypotheses</i>	30
9. <i>Outline of the thesis</i>	32
9.1. Objectives and strategies	32
9.2. Predictions	33
II. Chapter 2 - Forty-Two: a next-generation tool for enriching alignments of orthologous genes and gene families.	35
1. <i>Abstract</i>	36
2. <i>Orthology enrichment</i>	37
2.1. Introduction	37
2.2. Objectives and strategies	40

2.3.	Materials and Methods-----	41
2.3.1.	Implementation-----	41
2.3.2.	Orthology assessment -----	41
2.3.3.	Paralog-aware enrichment (PARA mode) -----	42
2.3.4.	Orthologue merging (COMPRESS mode)-----	43
2.3.5.	Homologue trimming and orthologue splicing -----	43
2.3.6.	Experimental design: pruning/enrichment protocol -----	43
2.3.7.	F1-score -----	44
2.3.8.	Comparison with Orthograph: experimental setup-----	45
2.4.	Results -----	45
2.4.1.	OG reconstruction tests -----	45
2.4.1.1.	Proving heuristics - Complete proteome data-----	46
2.4.1.2.	Assessing performance - Unannotated genomic data-----	49
2.4.1.3.	Comparison versus Orthograph-----	49
2.5.	Conclusions -----	51
3.	<i>Contamination detection: principles, implementation and applications</i> -----	51
3.1.	Introduction-----	51
3.2.	Materials and Methods-----	52
3.2.1.	Contamination detection implementation-----	52
3.2.2.	Metagenomic mode and debriefing -----	53
3.3.	Results -----	54
3.3.1.	Consensus assessment of cyanobacterial genome contamination -----	54
3.3.2.	Decontamination of transcriptomic data in metazoan phylogenomics -----	54
3.3.3.	Benchmarking on simulated chimeric genomes-----	55
3.3.4.	Integration in genome quality assessment pipelines -----	56
3.4.	Conclusions -----	56
4.	<i>Transition: the backbone</i> -----	57
III.	Chapter 3 - Broadly sampled orthologous groups of eukaryotic proteins for the phylogenetic study of plastid-bearing lineages -----	59
1.	<i>Abstract</i> -----	60
2.	<i>Objective</i> -----	60
3.	<i>Data description</i> -----	61
4.	<i>References</i> -----	65
5.	<i>Supplementary material</i> -----	67
5.1.	Supplementary file 1 - Methods-----	67
5.1.1.	Proteome pre-processing -----	67

5.1.2.	Organelle complementation -----	67
5.1.3.	Proteome dereplication -----	67
5.1.4.	Sequence naming -----	67
5.1.5.	Localization tagging -----	68
5.1.6.	Completeness analysis - BUSCO -----	68
5.1.7.	Orthology inference -----	68
5.1.8.	Orthogroup classification -----	68
5.1.9.	Photosynthetic clade detection and pruning -----	69
5.1.10.	Annotation with eggNOG-mapper -----	69
5.2.	Supplementary figures -----	70
5.3.	Supplementary tables -----	73
6.	<i>Transition: towards a broader complex algal phylogenetic sampling</i> -----	76
IV.	Chapter 4 - Decontamination, pooling and dereplication of the 678 samples of the Marine Microbial Eukaryote Transcriptome Sequencing Project -----	77
1.	<i>Abstract</i> -----	78
2.	<i>Objective</i> -----	78
3.	<i>Data description</i> -----	79
4.	<i>References</i> -----	82
5.	<i>Supplementary material</i> -----	85
5.1.	Data file 1 - Methods -----	85
5.1.1.	Sample selection and combination -----	85
5.1.1.1.	MMETSP re-assembly pre-processing -----	85
5.1.1.2.	Contamination detection - Forty-Two -----	85
5.1.1.3.	Sample pooling -----	86
5.1.2.	Transcriptome dereplication and quality survey -----	87
5.1.2.1.	Cross-contamination detection - Sobek -----	87
5.1.2.2.	Transcriptome dereplication -----	87
5.1.2.3.	Contamination detection - Forty-Two -----	87
5.1.2.4.	Completeness analysis - BUSCO -----	88
5.2.	Supplementary figures -----	89
5.3.	Supplementary tables -----	97
6.	<i>Transition: towards a large-scale complex algae analysis</i> -----	113
V.	Chapter 5 - A Quantitative Phylogenomic Approach -----	115
1.	<i>Introduction</i> -----	116
1.2.	Assembling a high-quality dataset -----	117

1.2.1.	Orthoclan selection -----	119
1.2.2.	Event classification-----	119
1.3.	Gene transfer inference -----	120
1.3.1.	Paraphyly-Aware Gene Transfer Inference (PAGTI) -----	120
1.3.2.	Illustrative topologies -----	121
1.3.3.	Inferred transfer events-----	122
1.3.4.	Objective and strategies-----	123
2.	<i>Results</i> -----	125
3.1.	Global analysis-----	125
3.1.1.	Large-scale gene fluxes-----	125
3.1.2.	Interpretation of the α -diversity metric -----	128
3.1.2.1.	Per species distribution-----	128
3.1.2.2.	Per lineage distribution-----	132
3.1.2.3.	Plastidial control: cpcpt# genes recover the expected donor signal -----	134
3.2.	Case studies -----	136
3.2.1.	Euglenids: a case study for a nucpt# core gene set associated with kleptoplast pre-adaptation? -----	136
3.2.1.1.	Identification of a shared euglenid pre-adaptation nucpt# gene set-----	136
3.2.1.2.	Looking for an α -diversity-driven signature-----	137
3.2.1.3.	Looking for a donor-driven differential signature-----	138
3.2.1.4.	Looking for a function-driven differential signature -----	140
3.2.2.	Peridinin dinoflagellates: a kleptoplastid-acquisition signature?-----	143
3.2.2.1.	Identification of an ancestral core peridinin plastid event set -----	144
3.2.2.2.	Functional characterization-----	146
3.2.2.3.	Donor lineages-----	148
3.2.3.	Dinophysiales: non-permanent kleptoplasts-----	148
3.2.3.1.	Enrichment with new Dinophysis species -----	149
3.2.3.2.	A perfect example of contaminated data -----	149
3.2.3.3.	Orthoclan analysis -----	151
3.2.4.	Ochrophyta: a major donor lineage -----	152
3.2.4.1.	Assessing contributions of ochrophyte subclades-----	154
3.2.4.2.	Functional characterisation-----	157
3.2.4.3.	Ochrophyta and Haptophyta: hints to unresolved plastids origin? -----	159
4.	<i>Methodological results</i> -----	162
4.1.	How robust is the signal? -----	162
4.1.1.	Pipeline robustness -----	162
4.1.1.1.	Variations with limited effect on the signal -----	164
4.1.1.2.	Variations with stronger effects-----	165
4.1.2.	Does the signal depend on the choice of the α -diversity index? -----	170

4.1.3.	Expert validation of transfer inferences -----	175
4.2.	What does donor α -diversity actually measure?-----	176
4.2.1.	Confounding factors -----	176
5.	<i>Methods</i> -----	180
5.1.	Orthoclan enrichment-----	180
5.2.	Phylogenetic inference-----	181
5.3.	Gene transfer inference-----	181
5.3.1.	Core principles-----	181
5.3.2.	Complex topologies -----	182
5.3.3.	Early vs late transfers: transfer compatibility rules -----	183
5.4.	Event classification -----	185
5.4.1.	Case 1: Tagged receiver (no ambiguity)-----	185
5.4.2.	Case 2: Tagless receiver, tagged donor-----	185
5.4.3.	Case 3: Neither receiver nor donor is tagged -----	185
5.4.4.	Case 4: Tagless receiver with a cpcpt# donor-----	185
5.5.	α -diversity indices and weights -----	186
5.5.1.	Shannon-Wiener (Shannon entropy)-----	186
5.5.2.	Simpson ("reverse Simpson")-----	186
5.5.3.	Chao1 (richness estimator)-----	186
5.5.4.	Minimum species threshold -----	187
5.5.5.	Lineage-specific weights -----	187
5.6.	KDE-based classification of α -diversity-----	189
5.7.	KEGG module assignation -----	189
5.8.	Updating the dataset: integrating new kleptoplastidic species -----	192
5.9.	Ribosomal phylogenomic tree-----	192
5.10.	Balanced sampling-----	192
6.	<i>Transition: towards an α-diversity framework for plastid integration</i> -----	193
VI.	Chapter 6 - Discussion -----	195
1.	<i>The α-diversity framework</i> -----	195
1.1.	Low α -diversity: single-donor dominance -----	195
1.2.	Intermediate α -diversity: complex histories-----	197
1.3.	High α -diversity: wider prey range-----	198
1.4.	Interpretative framework linking α -diversity to plastid integration stage -----	201
2.	<i>How does pre-adaptation shape plastid integration?</i> -----	202
2.1.	The euglenid system: pre-adaptation identified under kleptoplastic selective pressure ---	202
2.2.	Parallel with peridinin dinoflagellates: a convergent functional partitioning? -----	204
2.3.	Dinophysiales: the limits of large-scale functional inference-----	205

3.	<i>What shapes high α-diversity in receivers?</i>	207
3.1.	Predators, intermediaries, and donor exposure	208
3.2.	Ochrophytes as dominant donors in a reticulate transfer network	209
3.3.	Availability matters: you eat what is there	209
4.	<i>Toward a predictive framework for plastid integration</i>	210
4.1.	The gene reset as a natural perturbation	210
4.2.	Plastid integration as a genomic continuum	211
4.3.	Testable predictions	211
5.	<i>Limitations & perspectives</i>	212
6.	<i>Closing statement</i>	216
VII.	General conclusion	217
VIII.	Contributions	219
IX.	REFERENCES	222

I. Chapter 1 - General introduction

Photosynthetic eukaryotes are scattered across the tree of life, and the mechanisms by which their plastids were acquired remain among the most debated questions in evolutionary biology. Classical models invoke endosymbiosis as the primary driver of plastid spread, yet nucleus-encoded plastid-targeted genes often reveal mosaic phylogenetic origins that are difficult to reconcile with simple single-donor scenarios. Several hypotheses have been proposed to explain this pattern, including serial (possibly cryptic) endosymbiosis, the related Shopping Bag and Red Carpet models, and the Rhodoplex family of hypotheses, but the selective processes linking transient plastid associations to permanent integration remain incompletely understood. Kleptoplasty provides one possible solution to this problem, because it creates an immediate selective pressure for maintaining stolen plastids. This thesis explores the idea that kleptoplasty may therefore have played a broader role in plastid evolution than is usually assumed. The following sections introduce the biological background and conceptual framework underlying this hypothesis.

1. Photosynthetic eukaryotes and plastid origins

1.1. Eukaryotic diversity and evolution

The origin of eukaryotic cells remains a major question in evolutionary biology. Because the events underlying eukaryogenesis occurred deep in time, the traces they left in extant genomes are often difficult to disentangle and interpret (Donoghue et al., 2023). Nevertheless, comparative genomics and phylogenomics have clarified several key points. Current models no longer treat eukaryotes as a lineage that simply diverged alongside Bacteria and Archaea; rather, eukaryotes are generally inferred to have emerged from within the archaeal domain, in association with the acquisition of an alphaproteobacterial endosymbiont that gave rise to the mitochondrion (Vosseberg et al., 2024). Although the endosymbiotic origin of mitochondria is now strongly supported (Roger et al., 2017), important questions remain regarding the nature of the host cell, the identity of the mitochondrial ancestor, and the sequence of events that led from symbiosis to full organelle integration.

Plastids represent a second landmark example of organellogenesis. In the ancestor of Archaeplastida, a heterotrophic eukaryote established a primary endosymbiosis with a cyanobacterium (Reyes-Prieto et al., 2007; Ponce-Toledo et al., 2017), giving rise to the primary plastids. These primary algae, i.e., glaucophytes, rhodophytes, and Viridiplantae (i.e., green algae and land plants), are supposedly monophyletic (Rodriguez-Ezpeleta et al., 2006; Irisarri et al., 2021) but laid the foundation for the subsequent diversification of the so-called complex algae, a polyphyletic assemblage of photosynthetic

eukaryote lineages (Sibbald and Archibald, 2020). That diversification took its root into eukaryote-to-eukaryote endosymbioses forming a network of photosynthetic organisms spread across the eukaryotic tree (Figure 1). Among those, CASH lineages (Cryptophytes, Alveolates, Stramenopiles or Ochrophytes, Haptophytes) harbour plastids related to those of red algae while euglenids and Chlorarachniophyceae have plastids related to those of green algae. Whereas the distinct origins of green-derived plastid were relatively quickly settled (Rogers et al., 2007), emergence of complex red plastids led to much more controversy, from the single-origin chromalveolate hypothesis (Cavalier-Smith, 1999) to its phylogenomic falsification (Baurain et al., 2010), and the mechanisms by which plastids were acquired, be it red or green, remain unclear even to this day (Ponce-Toledo et al., 2025).

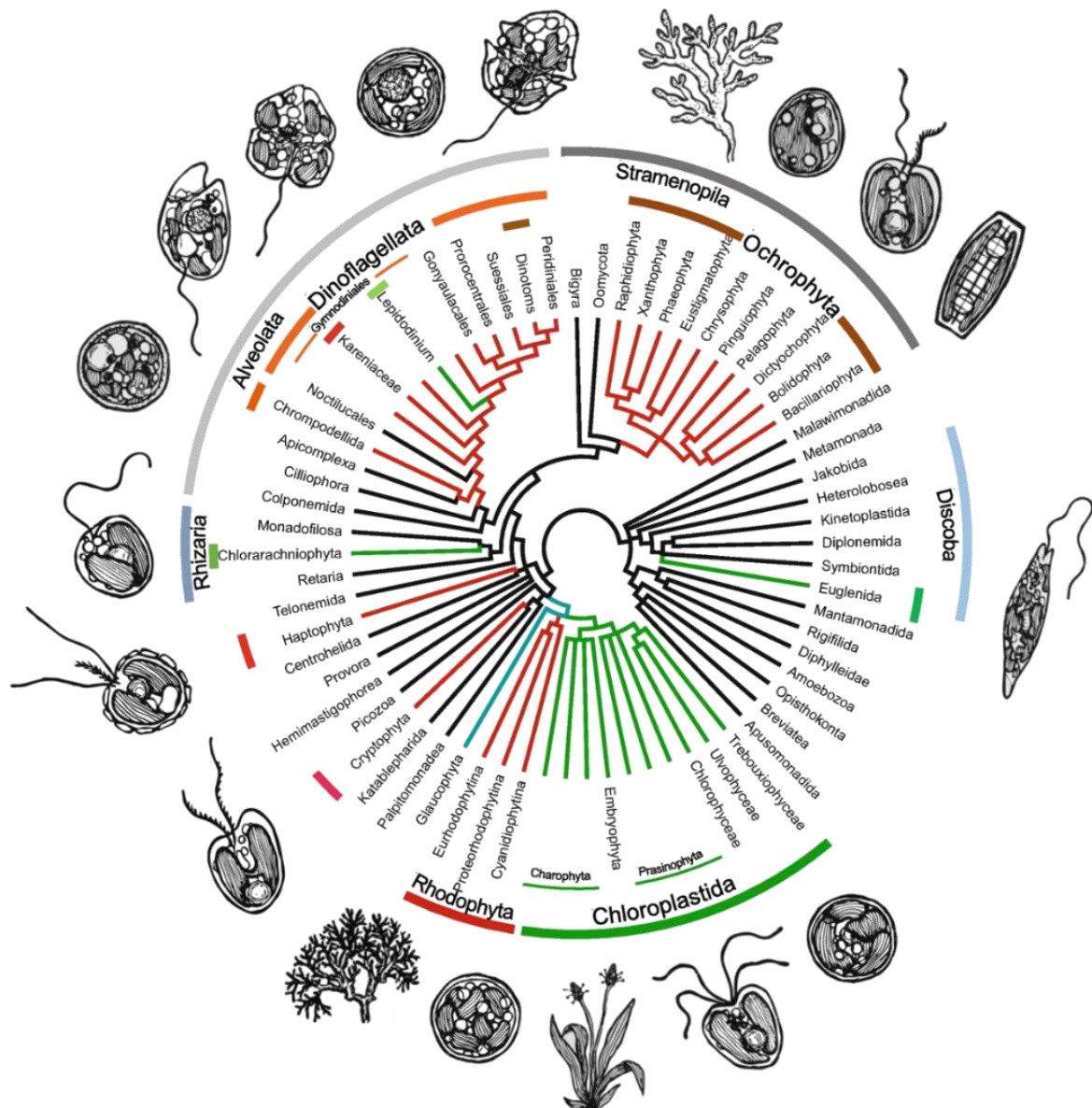


Figure 1: Phylogenetic tree representation of eukaryotic diversity illustrating the patchy distribution of photosynthetic lineages (from Novák Vanclová and Dorrell, 2024). Branch colours indicate photosynthetic lineages and plastid ancestry: green branches mark green-plastid-bearing lineages, including Chloroplastida (= Viridiplantae), chlorarachniophytes, euglenids, and Lepidodinium; teal branches indicate Glaucophyta; red branches indicate red-plastid-bearing lineages, including Rhodophyta, Cryptophyta, and red-derived CASH lineages. Outer coloured arcs indicate major host groups: Discoba in slate blue, Rhizaria in dark slate blue, Stramenopiles in dark grey, ochrophyte-derived plastids in brown (including

dinotoms), Alveolata in light grey, and dinoflagellates in orange. The discontinuous distribution of coloured branches highlights the repeated acquisition, replacement, and loss of plastids across eukaryotic evolution.

2. Anatomy of a complex plastid

2.1. Endosymbiosis, nucleomorphs and EGT

Several feeding mechanisms in protists differ in whether prey are internalized whole (phagocytosis), partially extracted (e.g., myzocytosis), or digested externally (e.g., pallium feeding) (Schnepf and Elbrächter, 1992) (Figure 2). In phagocytosis, the whole prey cell is engulfed by the host membrane system and internalized into a food vacuole, a process on which relies endosymbiosis (Reyes-Prieto et al., 2007). An important implication of endosymbiosis is that the prey is still alive and becomes an (endo)symbiont. In contrast, myzocytosis is a specialized variant of phagocytosis, in which the predator extracts the internal contents of a prey cell thanks to a feeding tube (peduncle) that pierces the prey, that dies in the process, and extracts its content (excluding its plasma membrane) into a food vacuole with potential organelle retention (Schnepf and Deichgräber, 1984). In the case of plastid retention, the phenomenon is called “kleptoplasty” and the plastid is referred to a “kleptoplast” (Gast et al., 2007; Hehenberger et al., 2019), while nucleus retention is called “karyoklepty” and the retained nucleus a “kleptokaryon” (Johnson et al., 2007; Kim et al., 2017). Finally, the pallium feeding mechanism involves the enclosure of the prey within a pseudopodal structure (the pallium or feeding veil), where digestion occurs extracellularly before nutrient uptake (Jacobson and Anderson, 1986), thereby precluding the retention of intact organelles (Figure 2).



Figure 2: Illustration of the main feeding mechanisms in dinoflagellates (prey in yellow): phagocytosis (complete prey engulfment), myzocytosis via a feeding tube (“peduncle feeding”), and pallium feeding (extracellular digestion in a “feeding veil”) (adapted from García-Oliva and Wirtz, 2022). Phagocytosis and myzocytosis, which involve total or partial prey ingestion with potential preservation of intact organelles, are also observed across diverse protist lineages and are compatible with plastid (kleptoplasty, kleptoplasts) and nucleus (karyoklepty; kleptokaryon) retention. In contrast, pallium feeding relies on extracellular digestion and is generally considered incompatible with the retention of functional organelles.

Although different mechanisms exist, the main mechanism invoked in eukaryotic plastid spread is endosymbiosis (which involves phagocytosis) (Keeling, 2010), a phenomenon where a heterotrophic eukaryote predator engulfs its prey and, instead of digesting it, establishes a stable symbiotic association (Bhattacharya et al., 2007; Nowack and Melkonian, 2010). When the prey (or symbiont) is a primary

(eukaryotic) alga, the endosymbiosis is termed secondary (McFadden, 2001), giving rise to a secondary alga; when the prey is a secondary alga, the endosymbiosis is termed tertiary and so on (Baurain et al., 2010; Bodył and Moszczyński, 2006). From secondary endosymbiosis and onward, the resulting plastid-bearing eukaryotes are more generally referred to as complex algae, or more precisely as lineages with complex plastids, because their plastids derive from a eukaryotic algal endosymbiont rather than directly from a cyanobacterium and are typically surrounded by more than two membranes (Archibald, 2015; Reyes-Prieto et al., 2007).

Among complex algae, some of the clearest evidence for secondary endosymbiosis comes from nucleomorph-bearing lineages, the plastids of which are surrounded by four membranes (Figure 3). Nucleomorphs are highly reduced relic nuclei of the engulfed eukaryotic algal endosymbionts and are known in two independent lineages: cryptophytes, whose plastids derive from a red algal endosymbiont, and chlorarachniophytes, whose plastids derive from a green algal endosymbiont (Archibald, 2007; Douglas et al., 2001). These structures persist in the periplastidial compartment, located between the second and third plastid membranes (Figure 3), and they retain highly compact genomes whose gene content is phylogenetically traceable to the corresponding algal endosymbiont lineage (Curtis et al., 2012). Their existence therefore provides direct evidence that plastids in these groups originated from a eukaryotic endosymbiont rather than from a cyanobacterium, and very likely by endosymbiosis rather than any another mechanism.

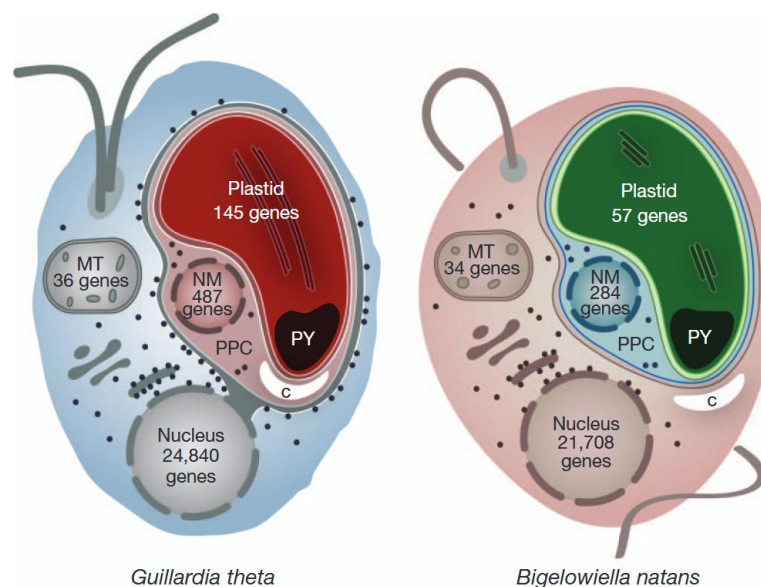


Figure 3: Illustration of the cryptophyte *Guillardia theta* and the chlorarachniophyte *Bigelowiella natans*, two independently evolved complex algae that retain a nucleomorph (NM), the highly reduced relic nucleus of the engulfed eukaryotic algal endosymbiont. In *G. theta*, the nucleomorph and plastid are derived from a red algal endosymbiont, whereas in *B. natans* they derive from a green algal endosymbiont. Both organisms therefore contain four genetically distinct compartments: the host nucleus, mitochondrion (MT), plastid, and nucleomorph, with the latter located in the periplastidial compartment (PPC) between the inner and outer plastid membrane pairs. Reproduced from Curtis et al., 2012.

At the onset of the endosymbiotic association, genes are massively and progressively transferred from the endosymbiont to the host nucleus, a process known as EGT (Endosymbiotic Gene Transfer) (Timmis et al., 2004). In parallel, endosymbiont genes that are rendered obsolete by intracellular life or redundant with pre-existing host functions undergo mutational decay and loss (Timmis et al., 2004). EGT of the remaining essential genes, coupled with the progressive genome erosion, gradually renders the endosymbiont dependent on host-encoded proteins for its continued operation, driving its transition from autonomous organism to permanent organelle (Timmis et al., 2004). Each round of endosymbiosis is expected to contribute additional waves of EGT to the host nucleus, alongside genes acquired by horizontal transfer (HGT) from other sources than the endosymbiont (Teich et al., 2007), resulting in chimeric nuclear genomes whose composition reflects their complex evolutionary history (Elias and Archibald, 2009) (Figure 4). However, for this dependency to take effect, the genes acquired by EGT need to be targeted back to the plastid, implying that the host has to evolve a protein-targeting machinery at an early stage of plastid integration in order to make proteins cross plastid membranes, since the onset of nucleus-to-plastid protein trafficking is a central requirement for organelle establishment (Timmis et al., 2004).

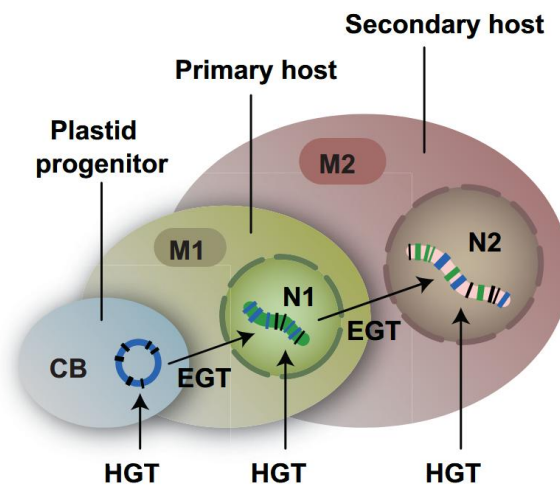


Figure 4: **Schematic representation of endosymbiotic and horizontal gene flow during primary and secondary endosymbiosis.** During primary endosymbiosis, genes are transferred from the cyanobacterial plastid progenitor (CB) to the nucleus of the primary host (N1) through EGT. During secondary endosymbiosis, additional EGT occur from the primary algal endosymbiont to the nucleus of the secondary host (N2). In parallel, HGT from other sources can contribute additional genes to each genome, making it difficult to distinguish endosymbiotic from non-endosymbiotic gene acquisitions. M1/M2, mitochondria. Reproduced from Elias and Archibald, 2009.

2.2. Protein targeting to complex plastids

Primary plastids are bounded by two membranes derived from the cyanobacterial endosymbiont, implying the loss of the host phagosomal membrane during early stages of organelle integration (McFadden, 1999). In contrast, secondary endosymbiosis classically results in four membrane-bound plastids: from inner to outer membranes, the primary plastid envelope, i.e., two membranes, the primary alga cytoplasmic membrane and the host phagosomal membrane, like in Cryptophyceae (McFadden, 2001). The multiplication of membranes surrounding complex plastids must have created a major

challenge for protein targeting: nuclear-encoded and plastid-targeted (termed nucpt# in the present manuscript) proteins has to cross four lipid bilayers (or three in euglenids and peridinin dinoflagellates, and up to five in dinotoms, see below) to reach the plastid stroma (Grosche et al., 2014) (Figure 5). In primary plastids (two membranes), protein import relies on the TOC/TIC translocon system (Translocon at the Outer/Inner envelope membrane of Chloroplasts), which handles transport across the two cyanobacterium-derived membranes (Soll and Schleiff, 2004) (Figure 6). The TOC/TIC system has been retained in the inner membranes of complex plastids, where it still mediates crossing of the two innermost membranes, corresponding to the original primary plastid envelope (Figure 5).

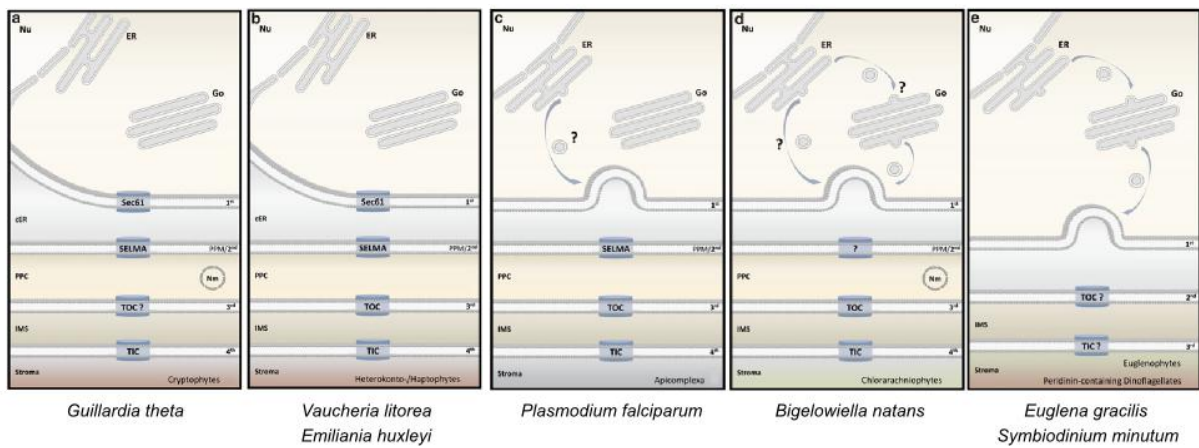


Figure 5: Schematic overview of representative complex plastids assumed to derive from a red algal endosymbiont (*Guillardia theta*, *Vaucheria litorea*, *Emiliania huxleyi*, *Plasmodium falciparum*, and peridinin-containing dinoflagellates) or a green algal endosymbiont (*Bigelowiella natans* and *Euglena gracilis*). Most lineages shown harbour plastids surrounded by four membranes, whereas euglenophytes and peridinin-containing dinoflagellates have plastids surrounded by three membranes. The outermost membrane may be continuous with the endoplasmic reticulum (chloroplast ER), as in cryptophytes, heterokonts and haptophytes, or disconnected from it, as in apicomplexans and chlorarachniophytes, implying different routes for plastid preprotein trafficking. In four-membrane plastids, protein import involves successive translocation into the host endomembrane system (via the Sec61 translocon), into the periplastidial compartment (via the SELMA machinery), and into the plastid stroma (via the canonical TOC/TIC translocons), whereas three-membrane plastids rely on vesicle-mediated transport through the ER/Golgi system followed by translocation across the remaining plastid membranes. Question marks indicate unresolved steps in preprotein import. Abbreviations: Nu, nucleus; Nm, nucleomorph; ER, endoplasmic reticulum; Go, Golgi apparatus; cER, chloroplast ER; PPC, periplastidial compartment; PPM, periplastidial membrane; IMS, intermembrane space; Sec61, ER Sec61 translocation complex; SELMA, Symbiont-specific ERAD-like machinery; TOC/TIC, translocons of the outer/inner chloroplast membranes. Adapted from Grosche et al., 2014.

For the additional outer membranes acquired through subsequent plastid acquisition, distinct mechanisms have evolved. In most CASH lineages (notably cryptophytes, haptophytes and ochrophytes but not alveolates), the outermost membrane is continuous with the host endoplasmic reticulum (the chloroplastic endoplasmic reticulum, or cER), and proteins targeted to the plastid carry a bipartite signal peptide (Patron and Waller, 2007) whose N-terminal portion directs cotranslational import across the cER via the Sec61 translocon (Bolte et al., 2009) (Figure 6). Integrated in the periplastidial membrane (third layer), the SELMA translocon (Symbiont-specific ERAD-Like MACHinery) mediates preprotein transport across the second outermost membrane (Hempel et al., 2009; Sommer et al., 2007) (Figure 6). The SELMA system is derived from the endoplasmic-reticulum-associated degradation (ERAD) pathway of the engulfed alga, repurposed for protein import rather than protein degradation. Proteins

targeted to complex plastids thus carry bipartite targeting peptides composed of an N-terminal signal peptide (Patron and Waller, 2007) (for ER entry) followed by a transit peptide (Hof and Kruijff, 1995) (for TOC/TIC-mediated import into the plastid stroma).

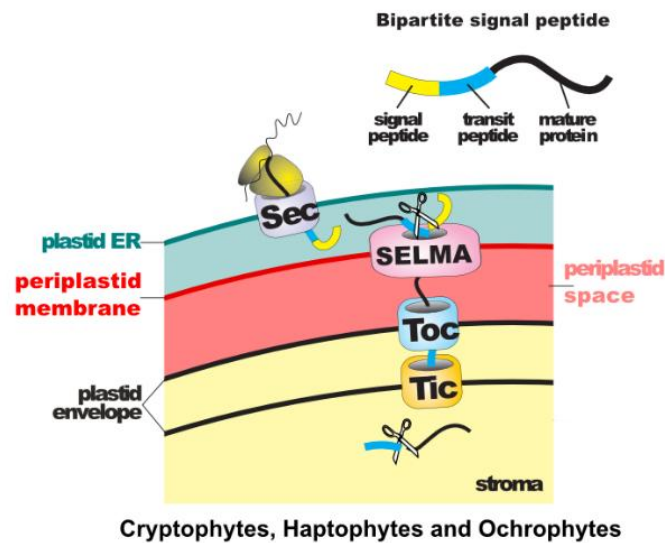


Figure 6: Schematic representation of the protein targeting route across the four membranes of complex plastids in cryptophytes, haptophytes, and ochrophytes (Adapted from Bodýl et al., 2009a). Nucleus-encoded plastid-targeted proteins carry a bipartite targeting presequence consisting of a signal peptide (yellow), which directs cotranslational import across the outermost membrane (plastid ER) via the Sec61 translocon, followed by a transit peptide (blue) that mediates downstream import into the plastid stroma. After cleavage of the signal peptide in the ER lumen, the transit peptide is exposed, and the protein is translocated across the periplastid membrane by the SELMA translocon. Finally, the transit peptide directs the protein across the two innermost membranes (plastid envelope) via the Toc and Tic complexes, where it is cleaved to release the mature protein into the stroma.

Not all complex algae use the same targeting route, however. Euglenids and peridinin dinoflagellates, whose plastids are bounded by three membranes and lack a cER, rely on vesicular transport through the Golgi apparatus to deliver plastid-targeted proteins to the outermost plastid membrane (Nassoury et al., 2003; Sulli et al., 1999) (Figure 5). Meanwhile, phylogenetic analyses of the SELMA system indicate that it was independently recruited from the ERAD machinery in different red complex lineages rather than inherited from a single endosymbiotic event (Felsner et al., 2011). The independent emergence of distinct solutions for the same transport challenge across separately derived complex plastids underscores both the constraints imposed by multi-membrane architectures and the evolutionary flexibility of eukaryotic protein trafficking systems (Grosche et al., 2014; Novák Vanclová and Dorrell, 2024).

2.3. Membrane architecture of complex plastids and implications

Since the maximum number of plastid membranes observed in complex algae is most often four (Figure 5) (five in dinotoms; Hehenberger et al., 2014), membranes stopped stacking at some point, probably for reasons of protein transport to the plastid (Gould et al., 2015). Indeed, the development of a transport system to cross the plastid membranes after tertiary or quaternary endosymbiosis would be immensely

complex if membranes kept accumulating (Petersen et al., 2014). This constraint may be particularly acute for ER-derived membranes, where protein import via the Sec61 translocon is cotranslational (Bolte et al., 2009) and the ribosome binds directly to the translocon during translation. As translation can occur only once, crossing ER-derived membrane can only occur once as well (R. Dorrell, personal communication). This may select for the loss or fusion of supernumerary ER-derived membranes during higher-order endosymbioses, concurrent with EGT to the host nucleus, hence the drive to develop a mechanism to reduce or maintain a low membrane count in the process of plastid acquisition. Besides, among complex algae, euglenids and peridinin dinoflagellates stand out with three membrane-bound plastids (Figure 5) (interpreted as the primary plastid envelope and the host phagosomal membrane) (Moog and Maier, 2018), which again suggest that some mechanism had to be developed to get rid of extra membranes. One strategy is to capture organelles only (kleptoplasty and/or karyoklepty) instead of engulfing the whole prey, with mechanisms such as myzocytosis (Figure 7). Myzocytosis bypasses at least the prey plasma membrane, so that plastids bound by two membranes in the donor would theoretically yield a three-membrane-bound kleptoplast instead of four in the case of whole-cell endosymbiosis by phagocytosis (Figure 7).

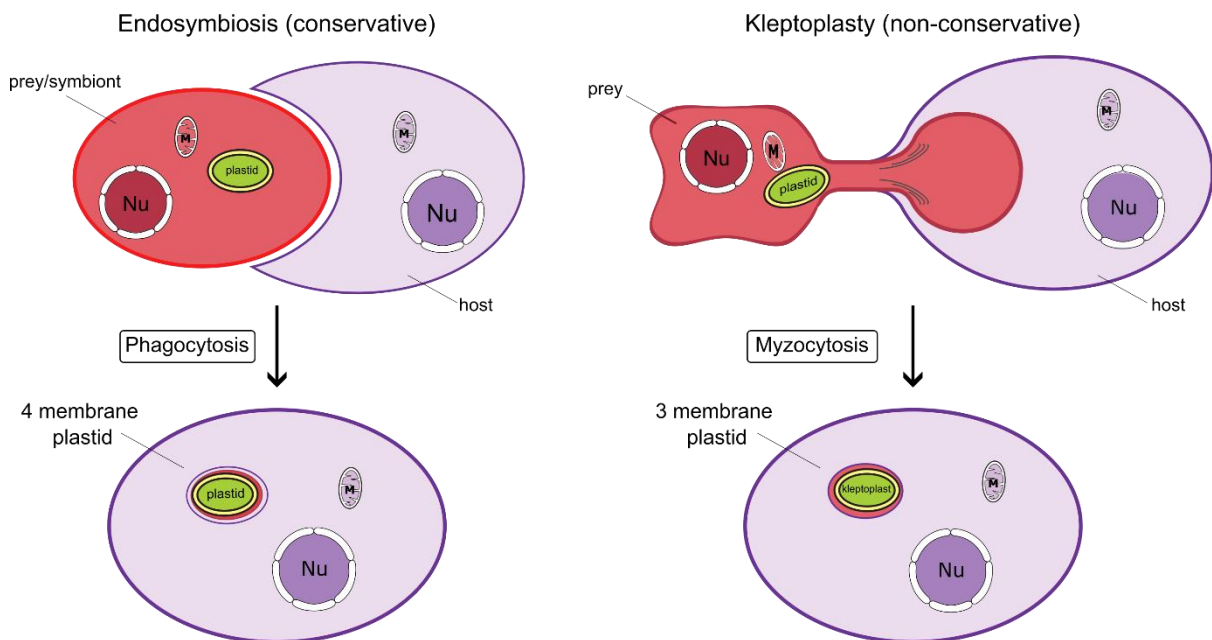


Figure 7: Simplified illustration of membrane accumulation around plastids through endosymbiosis versus kleptoplasty. Only the final outcome is depicted; intermediate steps and the fate of other organelles, including the symbiont/prey nucleus, are not illustrated. Left panels illustrate endosymbiosis, in which the host (purple) engulfs the symbiont (red) by phagocytosis, retaining the plastid (green) surrounded by additional membranes derived from the symbiont and host endomembrane system (conservative with respect to membranes). Right panels illustrate kleptoplasty, in which the host extracts prey cell contents through myzocytosis, retaining only the plastid (kleptoplast) with fewer surrounding membranes, as the prey's plasma membrane is not conserved during this process (non-conservative). Organelles are labelled as: Nu, nucleus; M, mitochondrion. Colours indicate cellular origin (purple, host; red, symbiont/prey).

In practice, however, membrane numbers are much more variable. For instance, the Ross Sea Dinoflagellate (RSD; *Karenia* sp. Ross Sea; member of the *Karenia* family) (Gast et al., 2006;

Sellers et al., 2014a) acquires haptophyte-derived kleptoplasts directly from its prey via myzocytosis, yet retains only three membranes (Sellers et al., 2014b). *Dinophysis acuminata* (Dinophysiales) kleptoplasts have only two membranes (Wisecaver and Hackett, 2010), despite also being acquired by myzocytosis. In this case, however, the plastids are not obtained directly from the donor alga: *Dinophysis* feeds on the kleptoplastidic ciliate *Mesodinium rubrum* (Alveolates), which has itself already sequestered plastids from its cryptophyte prey *Teleaulax amphioxeia* (Anschütz et al., 2022) (Figure 8). Thus, even when the same cellular mechanism is involved, the number of retained membranes depends on both the donor plastid architecture and the feeding mechanism through which it is acquired.

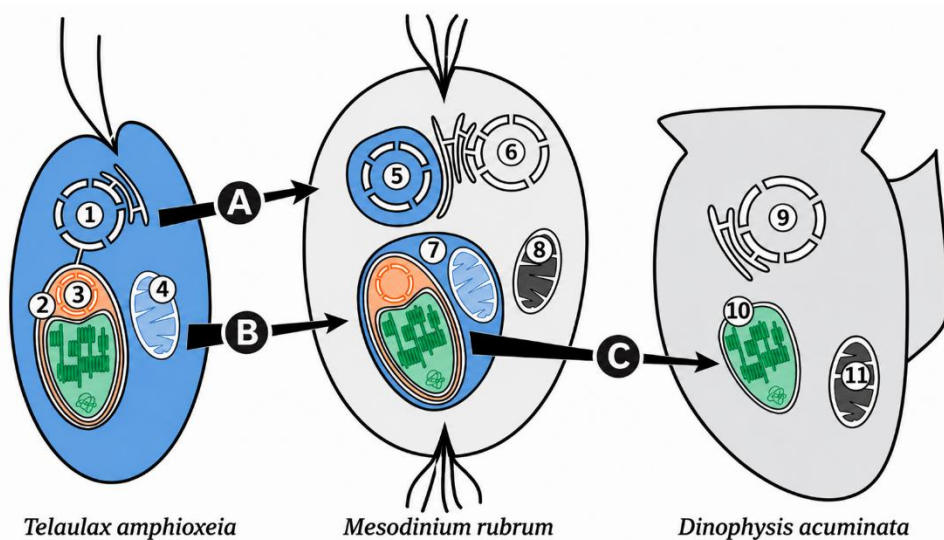


Figure 8: Schematic representation of the indirect acquisition of cryptophyte-derived plastids by *Dinophysis acuminata*. In this system, the cryptophyte *Teleaulax amphioxeia* acts as the primary plastid donor and is consumed by the ciliate *Mesodinium rubrum*, which retains cryptophyte-derived kleptoplast and kleptokaryon in different compartments. *Dinophysis acuminata* then feeds on the plastid-bearing ciliate and retains the kleptoplast while stripping the two outermost membranes and nucleomorph. Numbers indicate: 1, cryptophyte nucleus; 2, plastid; 3, nucleomorph; 4, cryptophyte mitochondrion; 5, retained cryptophyte nucleus i.e., kleptokaryon; 6, ciliate nucleus; 7, kleptoplast-mitochondrial complex; 8, ciliate mitochondrion; 9, dinoflagellate nucleus; 10, kleptoplast; 11, dinoflagellate mitochondrion. Adapted from Wisecaver and Hackett, 2010 and Anschütz et al., 2022.

However, kleptoplasty also occurs via whole-cell engulfment by phagocytosis: the euglenophyte *Rapaza viridis* ingests the entire prey, *Tetraselmis* sp. (green algae; Chlorophyta) (Yamaguchi et al., 2012), then segregates the plastid from the other cytoplasmic components within the phagosome (Karnkowska et al., 2023). The isolated plastid becomes tightly packed by the phagosomal membrane, which together with the double membrane of the plastidial envelope, forms the triple-membrane kleptoplastic envelope of *R. viridis* (Karnkowska et al., 2023) (Figure 9). In this respect, it is noteworthy that the three membrane-bound plastids of peridinin dinoflagellates have also been hypothesized to be

acquired through kleptoplasty (Bodyl, 2018; Moog and Maier, 2018), suggesting that membrane reduction through kleptoplastic acquisition may be a recurrent theme in plastid evolution.

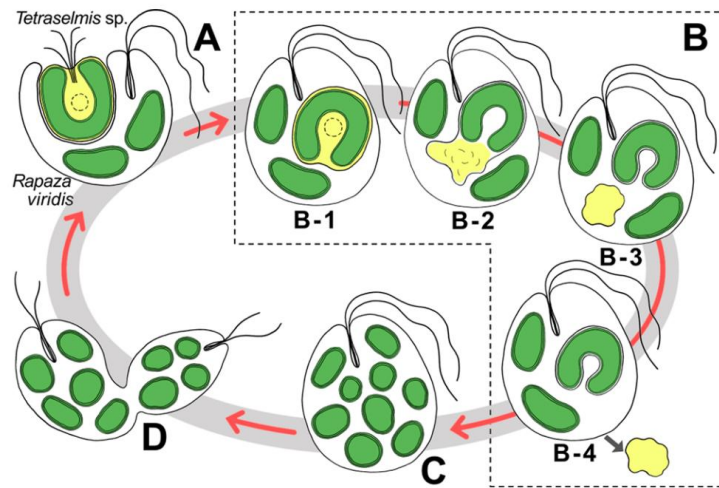


Figure 9: **Kleptoplast acquisition and inheritance in *Rapaza viridis*.** (A) *R. viridis* ingests an entire *Tetraselmis* cell by phagocytosis. (B) Within the phagosome, the prey plastid is segregated from the other cytoplasmic components and becomes tightly enclosed by the phagosomal membrane, forming a three membrane-bound kleptoplast together with the two plastid envelope membranes (B-1 to B-3). The remaining prey cytoplasm is rapidly excreted (B-4). (C) The acquired kleptoplast is subdivided within the *R. viridis* cytoplasm. (D) Kleptoplast fragments are distributed to daughter cells during cell division, but because they remain only temporarily functional, new *Tetraselmis* prey must be acquired regularly. Adapted from Karnkowska et al., 2023.

3. Hypotheses for the origin of complex red plastids

3.1. The chromalveolate hypothesis

While the distinct origins of plastids in green complex algae, Chlorarachniophyceae (Rhizaria) and Euglenophyceae (Discoba), two distantly related lineages, have been settled relatively early (Rogers et al., 2007), the question of how many times red algal plastids were acquired by eukaryotic hosts of the “CASH” (Cryptophyta, Alveolates, Stramenopiles, Haptophyta) lineages (Baurain et al., 2010) has been debated for over two decades, and several models have been proposed (Figure 10). First, three complex red algal lineages, cryptophytes and “chromophytes” (haptophytes and stramenopiles) (Cavalier-Smith, 1981) were grouped based on ultrastructural and biochemical similarities, i.e., four-membrane-bound plastids, the chloroplastic endoplasmic reticulum (cER), and chlorophyll c-type pigments into the “kingdom” Chromista, whereas future alveolate lineages (i.e., ciliates, apicomplexans, dinoflagellates) remained independent in the kingdom “Protozoa” (Cavalier-Smith, 1981). Building on these potential synapomorphies and additional features shared with plastid-bearing alveolates (i.e., dinoflagellates and apicomplexans ; see Archibald and Keeling, 2002), T. Cavalier-Smith later hypothesized that all four CASH lineages had a single common origin (Cavalier-Smith, 1999; Cavalier-Smith, 2003). Around the same time an alternative view, allowing for multiple independent origins, was also considered (Delwiche, 1999). The latter proposed that each major CASH lineage acquired its plastid through an

independent secondary endosymbiosis with a red alga, implying up to four separate events (Delwiche, 1999). This model requires that the structural and biochemical similarities of CASH lineages evolved independently, which might be considered unlikely. In this regard, the model of T. Cavalier-Smith, formalized as the “Chromalveolate hypothesis” (Cavalier-Smith, 1999), was much more parsimonious: all eukaryotes bearing “secondary” red plastids (the so-called “chromalveolates” = CASH lineages) originated from a single secondary endosymbiotic event with a red alga, and all the cellular adaptations took place in a common ancestor of them all (e.g., most of EGT, including the emergence of the protein import system and the acquisition of a bipartite targeting signal by hundreds of genes; Cavalier-Smith, 1999). Importantly, this hypothesis has three consequences: (i) all “chromalveolates” should be monophyletic, (ii) all non-photosynthetic lineages scattered among them (e.g., ciliates, oomycetes) must have experienced independent plastid loss, and (iii) many algal genes should be found in heterotrophic “chromalveolates”, reflecting their photosynthetic past (Keeling, 2009), key elements that soon came to be falsified. First, dedicated multigene phylogenies failed to recover the monophyly of “chromalveolates” as a whole (Baurain et al., 2010): using a phylogenomic approach comparing signal strength across nuclear, plastid, and mitochondrial compartments, Baurain et al., 2010, provided strong evidence for separate plastid acquisitions in “chromists” (cryptophytes, haptophytes, and ochrophytes). Second, the host phylogenies of individual CASH lineages placed them within supergroups that also contain non-photosynthetic relatives with no evidence of plastid ancestry, making the numerous plastid losses required by the hypothesis increasingly implausible (Stiller et al., 2014a). Furthermore, detailed analyses of plastid-targeted proteomes revealed a degree of mosaicism difficult to reconcile with a single endosymbiotic origin (Dorrell et al., 2017). As a consequence, the chromalveolate hypothesis in its strict form is now considered falsified by the majority of the community (Burki et al., 2016; Keeling, 2013), leaving the stage for hypotheses of a new genre.

3.2. Chromalveolate-like hypotheses

Before the definitive falsification of the chromalveolate hypothesis (Baurain et al., 2010), the discovery of unexpectedly abundant green algal genes in diatom (Ochrophyta) and other chromalveolate genomes (Moustafa et al., 2009) prompted a modified version of the chromalveolate hypothesis (Cavalier-Smith, 1999): a cryptic secondary green algal endosymbiosis would have preceded the red one in the common ancestor of all “chromalveolates” (Moustafa et al., 2009). In this scenario, the ancestral green plastid was subsequently replaced by a red algal plastid, leaving behind a residual green signal in the nuclear genome (Dorrell and Smith, 2011). However, Deschamps and Moreira (2012) substantially tempered these quantitative claims by reanalysing diatom genomes with additional red algal genomic data and stricter analytical criteria, and more recently, Morozov and Galachyants (2019) reported that the relative contributions of red and green algal EGTs in diatom nuclear genomes are approximately equal and proposed additional mechanisms to explain this chimeric signal.

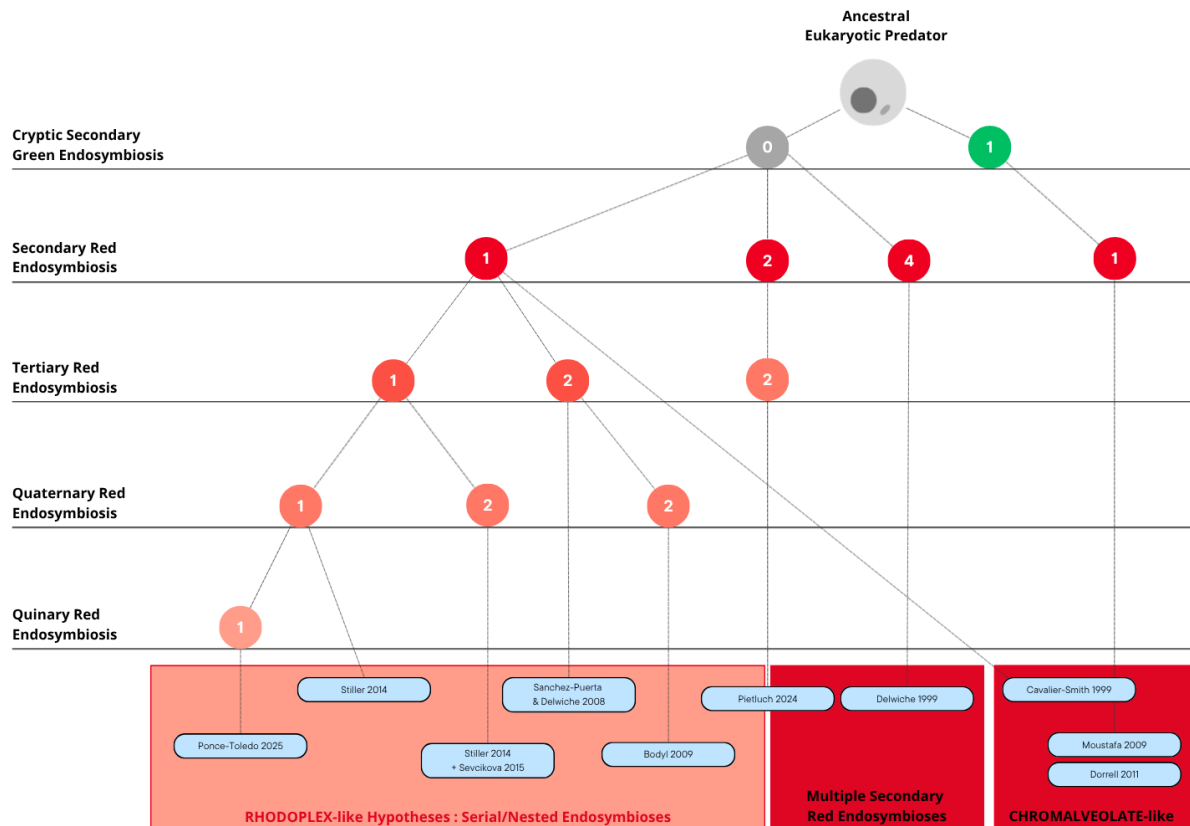


Figure 10: Graphical representation of the main published models (non-exhaustive) for the origin of red plastid-bearing complex algae, organized by the number and order of endosymbiotic events that each model invokes. Rows represent successive levels of plastid acquisition, from an optional cryptic green endosymbiosis (top) through secondary, tertiary, quaternary, and quinary red endosymbioses (bottom). Numbers within circles indicate how many independent events of that order each model requires. Circle colour reflects the order of endosymbiosis: green for the cryptic green endosymbiosis (Dorrell and Smith, 2011; Moustafa et al., 2009), and a dark-to-light red gradient for increasing orders of red endosymbiosis (dark red, secondary; lighter shades for tertiary through quinary); grey circles indicate the absence of a given level in that model. Lines connect each model to the combination of events it invokes. Models are grouped into three families (coloured boxes at bottom): chromalveolate-like models invoking a single common ancestor (Cavalier-Smith, 1999; Dorrell and Smith, 2011; Moustafa et al., 2009), models proposing multiple independent secondary endosymbioses with different host lineages (and potentially different red algae) (Delwiche, 1999; Pietluch et al., 2024), and Rhodoplex-like models involving serial (nested) higher-order endosymbioses, allowing for a single plastid origin but multiple host lineages (Sanchez-Puerta and Delwiche, 2008; Stiller et al., 2014a; Bodyl et al., 2009a; Ševčíková et al., 2015; Pietluch et al., 2024).

3.3. Rhodoplex-like models: serial and nested endosymbioses

Following the chromalveolate falsification (Baurain et al., 2010), a family of models has emerged proposing that CASH plastids spread horizontally through multiple eukaryote-to-eukaryote endosymbiotic events of increasing order. These scenarios were collectively formalized under the “Rhodoplex hypothesis” (Petersen et al., 2014), although some authors preceded that formalization. The objective of these models was to reconcile the monophyletic nature of complex red plastids, as supported by multiple lines of evidence, including plastid phylogenies (Keeling, 2009) with the apparent polyphyly of the corresponding host lineages. The earliest formulations came from Sanchez-Puerta et

al. (2007) and Sanchez-Puerta and Delwiche (2008), who proposed a single secondary endosymbiosis with a red alga in the ancestor of cryptophytes and haptophytes, complemented by one or two tertiary endosymbioses during the evolution of ochrophytes and myzozoans (i.e., dinoflagellates + chromerids + apicomplexans). In addition to the single secondary endosymbiosis of a red alga by cryptophytes, Bodyl et al. (2009b) proposed a model invoking two tertiary (cryptophyte-to-haptophyte and cryptophyte-to-ochrophyte) and two quaternary endosymbioses (ochrophyte-to-chromerid and haptophyte-to-dinoflagellate) (Figure 11, panel A). Chromerids (*Chromera velia* and *Vitrella brassicaformis*) were a newly discovered “lineage” of photosynthetic alveolates (within myzozoans), closely related to apicomplexans (Moore et al., 2008), and often referred to as “colpodellids” to emphasize their possible paraphyly (Oborník and Lukeš, 2015). Along the same lines, Stiller et al. 2014a proposed a serial chain of events (Figure 11, panel B): a cryptophyte plastid was acquired by the ancestor of ochrophytes (tertiary) and an ochrophyte by the ancestor of haptophytes (quaternary); dinoflagellates were not included in their analysis. These scenarios were further supported by phylogenomic evidence placing the chromerid plastids within the ochrophyte radiation (Ševčíková et al., 2015), who further suggested that this ochrophyte origin could extend to myzozoan plastids as a whole, consistent with the common ancestry of all myzozoan plastids proposed by Janouškovec et al., (2015a). However, Petersen et al., (2014) cast doubt on this common ancestry based on phylogenies of nucleus-encoded plastid-targeted proteins. Nevertheless, this point is retained in the model illustration as a plastid acquired by the ancestor of all Myzozoa (Figure 11).

More recently, Ponce-Toledo et al., (2025) proposed a model extending the serial chain to quinary endosymbiosis (Figure 11, panel C): a red alga was first acquired by cryptophytes (secondary), then a cryptophyte by the ancestor of haptophytes (tertiary), a haptophyte by the ancestor of ochrophytes (quaternary), and finally an ochrophyte by the ancestor of myzozoans (quinary). This scenario is supported mainly by SELMA phylogenies in which Ponce-Toledo et al., 2025 identified two SELMA components in ASH lineages, Der1 and Cdc48-1, as monophyletic endosymbiotic gene transfers related to cryptophyte nucleomorph genes (Nm EGT, Figure 11), which argue against scenarios in which cryptophyte and ASH plastids originated from independent secondary endosymbioses (Ponce-Toledo et al., 2025). Another marker that has historically complicated these scenarios is the plastid-encoded *rpl36* gene: haptophyte and cryptophyte plastids share a bacterial-derived c-type *rpl36*, distinct from the plastidial p-type *rpl36* found in Archaeplastida and stramenopiles, a pattern originally interpreted as evidence that haptophyte and cryptophyte plastids are sister lineages (Rice and Palmer, 2006) and compatible with early Hacrobia-type interpretations grouping haptophytes and cryptophytes together (with two additional lineages, i.e., telonemids and katablepharids) (Okamoto et al., 2009). However, large-scale host phylogenomic analyses later failed to recover their monophyly (Baurain et al., 2010; Burki et al., 2012).

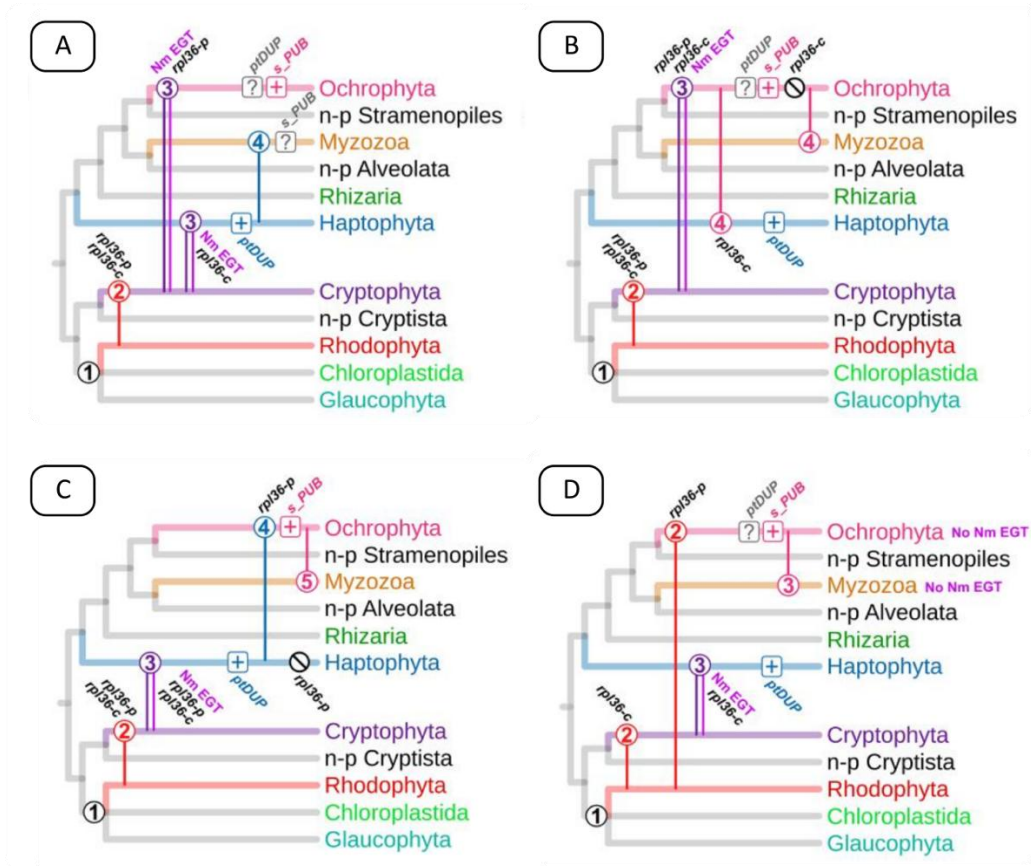


Figure 11: Competing scenarios for the spread of red-lineage plastids across CASH lineages (adapted from Ponce-Toledo et al., (2025). Each panel depicts the host phylogeny (grey tree) onto which plastid acquisition events are mapped (coloured lines), numbered in order of endosymbiotic transfer. Non-photosynthetic lineages are prefixed "n-p". Key markers used to distinguish or support alternative scenarios are indicated directly on the figure, including the bacterial-derived c-type rpl36 shared by haptophyte and cryptophyte plastids, putative nucleomorph-derived EGTs (Nm EGT) among SELMA components, and inferred neofunctionalizations or uncertain origins of SELMA-associated proteins such as ptDUP and s_PUB. (A) Model of Stiller et al., (2014a), complemented by Ševčíková et al., (2015). (B) Model of Bodyl et al., (2009b). (C) Model of Ponce-Toledo et al., (2025). (D) Hybrid model of Pietluch et al., (2024).

In an intriguing study, Pietluch et al. (2024) introduced a hybrid model that challenges both the chromalveolate and the Rhodoplex hypothesis (Figure 11, panel D). Based on their plastid phylogenies, they propose two independent, secondary endosymbioses with distinct red algal donors, one giving rise to the cryptophyte plastid and one to the ochrophyte plastid, complemented by two tertiary transfers (cryptophyte-to-haptophyte and ochrophyte-to-chromerid). The aim of this model is to explain their failure to recover monophyletic CASH plastids within red algae, a very unusual result that might be due to an uncontrolled phylogenetic artefact (D. Baurain and H. Brinkmann, personal communications).

Despite variations in the specific number and order of plastid acquisition events proposed, all Rhodoplex-like models share a common feature: they mainly invoke a single (or at most two) secondary red endosymbiosis, followed by several higher-order eukaryote-to-eukaryote transfers that spread plastids horizontally across host lineages. However, the specific succession of donors and receivers

differs between models, and the mechanisms underlying these higher-order transfers remain unspecified. Moreover, none of these models explicitly accounts for the mosaic composition of nucpt# gene observed in CASH lineages (Dorrell et al., 2017), in which genes trace back not to a single donor but to multiple algal sources. This mosaicism is the subject of the following section.

4. Mosaic origin of nucleus-encoded and plastid-targeted genes

In the classical endosymbiotic model for plastid acquisition, EGT progressively relocates genes from the endosymbiont genome to the host nucleus (and successfully targets them back to the plastid); the resulting nucpt# genes should predominantly trace back to a single ancestral endosymbiont (Timmis et al., 2004), just like the remaining genes still encoded in the endosymbiont/plastid genome (cpcpt# genes, for chloroplast-encoded chloroplast-targeted genes). However, most studies reveal a different picture: a strong genetic mosaicism, where plastid-targeted genes often derive from multiple, sometimes phylogenetically distant, algal lineages (Dorrell et al., 2017; Nosenko et al., 2006; Ponce-Toledo et al., 2019). Those nucpt# genes that are not derived from the plastid ancestor are generally interpreted as the result of horizontal gene transfer (HGT), possibly from distinct prey but frequently occurring in an endosymbiotic context. This process, referred to as endosymbiosis-related horizontal gene transfer (ERGT), notably for primary plastids (Ball et al., 2025), contributes to the chimeric nature of plastid proteomes (Dorrell et al., 2017) (Figure 12).

As mentioned in the previous section, diatoms were initially reported to contain more nucpt# genes of green algal origin (16%) than of red algal origin (10%) (Moustafa et al., 2009). However, a reanalysis of these diatom genomes using additional red algal genomic data and a stricter set of analytical criteria found that only 13% of nucpt# could be confidently traced from cyanobacteria to green algae to diatoms, with 66% clearly traceable to red algae (Deschamps and Moreira, 2012). Yet, in Ochrophyta as a whole, up to 25% of plastid-targeted proteins in the predicted LCA (Lowest Common Ancestor) were inferred to be green in origin (Dorrell et al., 2017). Similarly, cryptophytes (red) and chlorarachniophytes (green), although they are the most fitting lineages for their EGTs to trace back to their respective endosymbiont (because of their remnant nucleomorph), both harbour a sizable part of algal genes of the opposite “colour” (Archibald, 2009; Curtis et al., 2012). Moreover, euglenids, which possess green plastids, have a significant fraction of nucpt# genes from red algal origin (Ponce-Toledo et al., 2018). These findings challenge the simplicity of a scenario where all EGTs trace back to a single origin and suggest that additional and more complex processes have shaped plastid evolution.

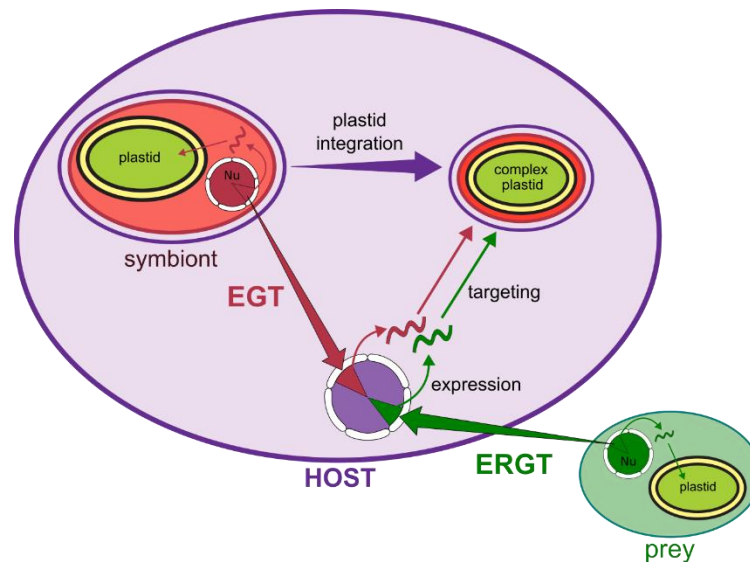


Figure 12: Simplified illustration of EGT and ERGT. Endosymbiotic gene transfer (EGT) and endosymbiosis-related horizontal gene transfer (ERGT) during plastid integration. The host cell (purple) harbours an algal endosymbiont (red; left) retaining its own nucleus (Nu) and plastid (green). During plastid integration, the endosymbiont is progressively reduced to a complex plastid surrounded by multiple membranes (right), while genes are transferred from the endosymbiont nucleus to the host nucleus (EGT, red arrow). These nucleus-encoded, plastid-targeted (nucpt#) gene products are expressed and imported back into the complex plastid (targeting, red arrows). In parallel, genes of plastid-related function can also be acquired from prey organisms that are not the endosymbiont itself (ERGT, green arrow from prey cell), contributing additional nucpt# genes to the host nuclear genome. Over successive rounds of feeding on diverse photosynthetic prey, ERGT can generate a mosaic of donor signatures in the nucpt# gene repertoire, distinct from the signal expected under EGT alone (Ball et al., 2025b).

This gene mosaicism, where hosts carry nucpt# genes from a different origin than their current plastid donor, is a central observation that any model of plastid evolution must account for. It resonates with the idea, articulated early on by Doolittle (1998), that eukaryotes may acquire genes from the organisms they eat (“you are what you eat”), progressively building a chimeric genome through trophic interactions, and that the idea of plastids being acquired from single ancestors at a single point in time is probably an “oversimplification” (Larkum et al., 2007).

5. Pre-adaptation models and selective pressure

5.1. Models explaining gene mosaicism

To explain this gene mosaicism, several models have been considered. One of them is, the “Shopping Bag” model that was originally formulated in the context of the debate over plastid origins, as a challenge to the implicit view that plastids were acquired from a single ancestor at a single discrete point in time (Howe et al., 2008; Larkum et al., 2007). Rather than treating plastid acquisition as an instantaneous event, this model proposes that organelle establishment results from a gradual process in which repeated transient associations with photosynthetic prey allow the host lineage to accumulate nucpt# genes before the permanent integration of a “final” plastid; in summary, a pre-adaptation process. Under those assumptions, gene mosaicism is not an anomaly, but a possible trace of these

successive interactions (Howe et al., 2008a; Larkum et al., 2007). Indeed, major evolutionary innovations rarely occur as sudden “hopeful monsters” (Kutschera and Niklas, 2008).

The idea that plastid acquisition may be facilitated by a prior phase of genomic pre-adaptation is also central to the “Red Carpet” hypothesis (Ponce-Toledo et al., 2019). This hypothesis proposes that red algal genes in both chlorarachniophytes and euglenids were transferred to the host nucleus before and/or during the early steps of endosymbiosis, and that this breeding ground of genes acted as a facilitator for the subsequent acquisition of their green algal endosymbiont. In this sense, the Red Carpet hypothesis (Ponce-Toledo et al., 2019) can be viewed as a specific instance of the broader conceptual framework that is the Shopping Bag model (Howe et al., 2008; Larkum et al., 2007). Furthermore, the Shopping Bag model has also been invoked to explain the green “footprint” in red complex algae (the Shopping Bag Endosymbiosis model; Morozov and Galachyants, 2019), as an alternative to a cryptic serial endosymbiosis (Figure 13).

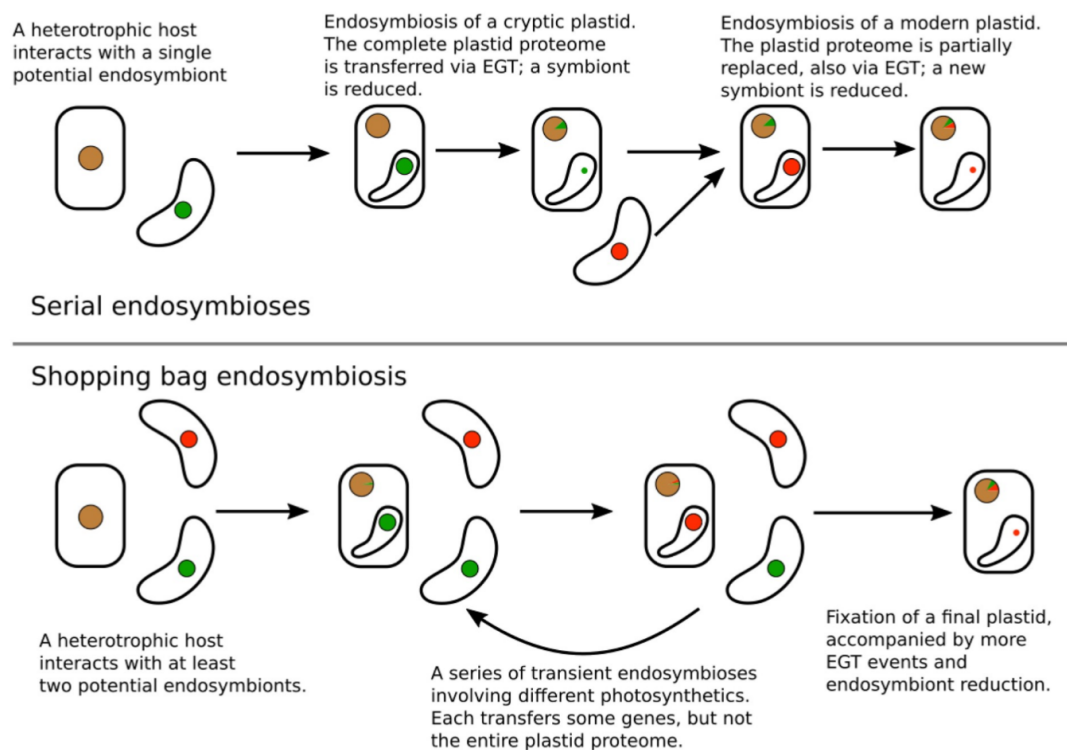


Figure 13: Schematic comparison of serial endosymbiosis and shopping-bag endosymbiosis models. In the serial endosymbiosis model, a heterotrophic host first acquires a plastid from a green endosymbiont, from which a substantial plastid-related gene set is transferred to the host nucleus through EGT. This (cryptic) plastid is later replaced by a modern plastid, whose integration involves additional EGT from a red algal symbiont. In the shopping-bag model, the host repeatedly interacts with different photosynthetic endosymbionts, each transient association contributing only part of the EGT. A final plastid is then stabilized in a host whose nucleus already contains genes of multiple algal origins. Pie charts represent the inferred origin of genes in the host genome; brown indicates host-derived genes, whereas green and red indicate genes derived from distinct algal endosymbionts. Small circles represent reduced organellar genomes (reproduced from Morozov and Galachyants, 2019).

Furthermore, these host-centred pre-adaptation models can be viewed alongside the “Portable Plastid” hypothesis (Grzebyk et al., 2003), which shifts part of the explanatory burden from the host to the plastid itself. In the Shopping Bag Endosymbiosis model (Morozov and Galachyants, 2019) and Red Carpet hypothesis (Ponce-Toledo et al., 2019), pre-adaptation resides primarily in the host, through the gradual accumulation of genes facilitating later plastid integration. By contrast, the Portable Plastid hypothesis (Grzebyk et al., 2003) raises the possibility that some plastids may themselves be “easily” transferable between eukaryotic hosts because they retain properties that make them more readily integrable, more “pre-adapted”. This especially applies to red plastids, whose genome retain $\sim 1.5\times$ more protein-coding genes than their green counterparts (Qiu et al., 2017), implying that the host needs to provide fewer nuclear proteins compared to hosts of green plastids. Recent work on serial plastid replacement in dinoflagellates is consistent with this idea, suggesting that at least some complex red plastids may carry with them part of the machinery required for successful transfer and integration (Lewis et al., 2024). In this sense, plastid evolution may have been shaped not only by host pre-adaptation, but also by differential plastid portability. This asymmetry may explain why red-derived plastids dominate the complex algal landscape, and conversely, why the independent acquisition of green plastids in euglenids and chlorarachniophytes (Rogers et al., 2007) remains exceptional.

While the Shopping Bag Endosymbiosis hypothesis (Morozov and Galachyants, 2019) and the Red Carpet model (Ponce-Toledo et al., 2019) can account for gene mosaicism, they mainly invoke repeated endosymbiotic or endosymbiosis-like interactions as sources of EGT/ERGT (Morozov and Galachyants, 2019; Keeling, 2024). Yet, increased exposure to foreign DNA through repeated transient interactions does not by itself explain the fixation of functional transfers. The Limited Transfer Window hypothesis posits that the frequency of plastid-to-nucleus gene transfer depends on the number of plastids in a cell: more plastids mean more DNA release events, and thus more opportunities for transfer (Barbrook et al., 2006; Smith et al., 2011). However, more recent studies show that greater opportunity for gene acquisition through phagotrophy or endosymbiosis does not necessarily translate into a higher frequency of retained HGT events (Keeling, 2024). One possible explanation is that phagocytosis increased the chance but lacked “necessity”: organisms that can continuously acquire plastids by feeding have less selective pressure to retain transferred genes permanently. Still, *bona fide* EGT (where the gene origin matches that of the plastid) does occur, but the mechanisms behind the selective pressure driving these transfers remain poorly understood.

5.2. Pre-adaptation and selective pressure: endosymbiosis VS kleptoplasty

The retention of photosynthetic prey has often been interpreted primarily as a means for heterotrophic hosts to access photosynthetically derived carbon or other photosynthetic products (Johnson, 2011; Stoecker et al., 2009). However, a recent experimental study challenged this carbon-centred view by suggesting that, at least in an experimentally induced early photosymbiosis model, oxygen supply rather than carbon transfer may provide the initial selective advantage for maintaining photosynthetic cells in protist host (Quevarec et al., 2026). Regardless of whether carbon, oxygen, or a combination of both represent the primary ecological benefit of early photosymbiosis, the selective pressures that specifically favour EGT/ERGT remain much less clearly formalized, particularly in the context of endosymbioses between eukaryotes.

In the classical view of endosymbiotic organelle acquisition, the endosymbiont is first taken up and retained by the host, and genetic integration follows as a consequence of this association (Dyall et al., 2004; Timmis et al., 2004). Under this scenario, because the endosymbiont is fixed in the host from the onset, this model predicts *nucpt#* genes should virtually all be derived from the endosymbiont itself by EGT. This model also implies that EGT to the host is possible before protein products can be targeted back to the organelle. This prediction has been used to suggest that the nucleus should contain a large number of genes that are derived from the endosymbiont but not targeted to the endosymbiont (i.e., endosymbiont-derived repurposed proteins) (Timmis et al., 2004). However, the hypothesis that early phases of endosymbiosis were marked by massive EGT has been considerably toned down: few verified cases of EGT have been confirmed (Moreira and Deschamps, 2014), and many *nucpt#* turn out to be repurposed genuine host proteins rather than genes of endosymbiont origin (EGT) (Bernabeu and Gabaldón, 2026; Nosenko et al., 2006; Patron et al., 2006).

An alternative model, formalized as the “Targeting-ratchet” hypothesis (Keeling, 2013), reverses the order of events: genetic integration precedes the fixation of the endosymbiont in the cell. In this framework, an ordinary cycle of uptake and digestion is progressively prolonged by delayed digestion or transient retention, and selection on the host to extract more benefit from the association leads it to target proteins, such as nutrient transporters, to the retained cell. Once this targeting capacity is established, genes acquired from previous transient endosymbionts can also be targeted to subsequent ones, creating a ratchet that drives incremental integration (Keeling, 2013, 2024). This cycle can proceed indefinitely, and eventually culminate in permanent endosymbiont retention if digestion is completely halted (Keeling, 2013, 2024).

These models describe how plastid integration can proceed incrementally, which is certainly compatible with the Shopping Bag model (Larkum et al., 2007), but they do not fully resolve which selective

pressures initially favour the reduction of endosymbiont-derived genetic compartments, especially in the context of secondary endosymbiosis where the engulfed alga initially retains a functional nucleus and a self-maintained plastid, having been swallowed intact. Several, non-mutually exclusive hypotheses have been proposed, mostly from studies of mitochondria and primary plastids, to explain why genes may be lost from organellar genomes or transferred to the host nucleus. First, endosymbiotic gene transfer can contribute to organelle integration by relocating genes to the host nucleus (EGT), where their expression can become embedded within host regulatory systems, provided that the encoded proteins are retargeted to the organelle (Martin and Herrmann, 1998; Timmis et al., 2004). In this view, the progressive establishment of host-controlled protein import and sorting systems is central to the transition from endosymbiont to organelle (Gross and Bhattacharya, 2009, 2011). Second, the photosynthetic machinery represents a source of reactive oxygen species (ROS), and it has been argued that genes encoding ROS-sensitive proteins would benefit from nuclear location, where they are less exposed to oxidative stress than in the organelle (Allen and Raven, 1996). Third, from an energetic perspective, maintaining multiple copies of the organellar genome per cell is costly, and encoding a gene in the nucleus may be favoured when this cost outweighs the cost of importing the protein back to the organelle (Kelly, 2021), particularly given the high copy number of organellar genomes relative to the nuclear genome (Cole, 2016; Kelly, 2021). Finally, selection for smaller and more rapidly replicated organellar genomes may further favour gene loss or transfer to the nucleus (Selosse et al., 2001). Yet, all these hypotheses address the selective advantages of gene transfer in the context of an established endosymbiosis, where the endosymbiont genome is, or is already being, reduced. They do not explain what initiates this reduction when the endosymbiont is self-sustaining. The selective pressure they describe remains indirect: there is an advantage to transferring a gene, but no immediate penalty for failing to do so as long as the endosymbiont nucleus continues to express it.

Conversely, the question of why certain genes are retained in the organelle despite this pressure to transfer has been addressed by complementary hypotheses. The “CoRR” hypothesis (CO-location for Redox Regulation) posits that core regulatory plastid genes benefit from remaining in the organelle because their expression must be directly coupled to the local redox state (Allen, 2017). The “CES” principle (Control by Epistasy of Synthesis) proposes that synthesis of organelle-encoded subunits is regulated by unassembled components of the same complex in a negative feedback loop (Wietrzynski et al., 2021). More recently, the “COCOA” principle (CO-location for COntrol of Assembly) was introduced (Howe and Barbrook, 2024), arguing that organelles preferentially retain genes for proteins that initiate assembly of electron transfer chain complexes. COCOA adds to, but does not replace, CoRR and CES, and together these hypotheses define a minimal set of genes that resist transfer regardless of the selective pressures acting on the rest of the organellar genome.

In contrast, kleptoplasty, inherently generates a more intense and direct selective pressure for ERGT compared to endosymbiosis. Unlike endosymbionts, which can self-maintain within the host, kleptoplasts are short-lived, persisting typically from days to months, depending on the host species (Nowack and Melkonian, 2010), thereby putting direct selective pressure on the retention of any gene that contributes to the extended duration of kleptoplast retention and function, regardless of phylogenetic source (Cordoba et al., 2021). This constitutes the “necessity” that Keeling, (2024) identifies as the missing driver in opportunity-based models (Figure 14). Moreover, retaining a functional kleptoplast imposes an immediate cost on the host: the photosynthetic machinery generates reactive oxygen species, and the host must cope with this phototoxicity to maintain plastid functionality (Uzuka et al., 2019). This creates a concrete selective pressure for the acquisition of genes involved in plastid maintenance and photoprotection, a pressure that is renewed with each kleptoplastic cycle. A striking illustration of this principle is provided by the comparison between *Mesodinium* and *Dinophysis* (Wisecaver and Hackett, 2010). *Mesodinium rubrum* retains the cryptophyte prey nucleus (kleptokaryon), which remains transcriptionally active and supplies the genes needed for kleptoplast maintenance (Johnson et al., 2007; Lasek-Nesselquist et al., 2015). Accordingly, no photosynthesis-related (nucpt#) genes have been found to be transcribed from the ciliate genome (Altenburger et al., 2020). In *Dinophysis*, however, which acquires plastids of cryptophyte origin indirectly through predation on *Mesodinium* but does not retain the prey nucleus, nucpt# genes of diverse phylogenetic origins (haptophytes, dinoflagellates, chlorarachniophytes, cyanobacteria, and cryptophytes) are actively transcribed for plastid maintenance and pigment biosynthesis (Garrić et al., 2024; Hongo et al., 2019).

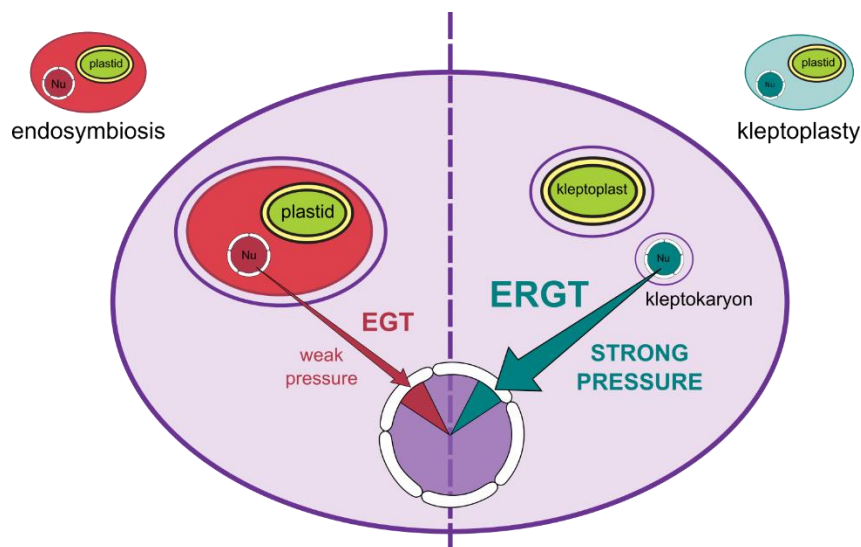


Figure 14: Simplified illustration depicting the contrasting selective pressures on gene transfer during endosymbiosis and kleptoplasty. The host cell (purple) is depicted with two concurrent sources of plastid-related genes. Left of the dashed line: a stable algal endosymbiont (red) retains its own nucleus (Nu), which continues to service its plastid. Thus, EGT to the host nucleus (red arrow) occurs but is under comparatively weak selective pressure, as the endosymbiont nucleus already provides the gene products needed for plastid maintenance. Right of the dashed line: plastids stolen from prey (kleptoplasts) and, in some cases, temporarily retained prey nuclei (kleptokaryon) lack long-term genetic autonomy. Consequently, ERGT from the kleptokaryon to the host nucleus (teal arrow) is under strong selective pressure, because only host-encoded copies of plastid-maintenance genes can ensure the continued function of kleptoplasts once the stolen nucleus degrades.

As noted by Mansour and Anestis (2021), the inability to retain the prey nucleus may impose even stronger selective pressure for ERGT, because plastid maintenance can no longer rely on continuous nuclear support from the prey. *Dinophysis* provides a particularly relevant example: unlike *Mesodinium*, it retains cryptophyte-derived plastids but not the kleptokaryon, meaning that any genetic contribution from prey is likely restricted to transient exposure during feeding and digestion. This narrow window of genetic exposure would further intensify the selective pressure for ERGT, as attested by the phylogenetically diverse nucleotide repertoire of *D. acuminata* (Hongo et al., 2019; Wisecaver and Hackett, 2010). The kleptoplastidic Ross Sea Dinoflagellate (RSD) further supports this view: despite retaining only temporary haptophyte-derived kleptoplasts from *Phaeocystis antarctica* (Gast et al., 2007; Stamatakis et al., 2017), and despite the continued presence of its ancestral peridinin plastid, this lineage encodes kleptoplast-targeted proteins acquired by HGT from multiple sources unrelated to the current kleptoplast donor (i.e., ERGT). The presence of many of these genes in closely related kareniaceans with fully integrated haptophyte-derived plastids indicates that nucleotide genes can accumulate before organelle fixation (Hehenberger et al., 2019), consistent with a pre-adaptation scenario mediated by repeated transient associations with photosynthetic prey. Furthermore, Bodyl (2018) argued that the peridinin plastid of dinoflagellates, may have evolved through kleptoplasty rather than endosymbiosis. This hypothesis is based mainly on the membrane topology of the peridinin plastid (three membranes), the feeding biology and recurrent plastid replacements in dinoflagellates, and the existence of modern kleptoplastidic systems showing partial host control over transient plastids. For these reasons, the role of kleptoplasty in the chain of events leading to plastid acquisition in eukaryotes may have been underestimated.

As recently emphasized by Sprecher and Johnson, 2025, kleptoplasty has long been regarded as an evolutionary dead-end, in contrast to endosymbiosis, which is usually framed as the main route toward permanent plastid acquisition. Yet extant systems that satisfy strict criteria for ongoing plastid organogenesis, namely detectable EGT, reduction or reconfiguration of the endosymbiont nucleus, and stable transmission of the symbiont to host progeny, are comparatively rare. The clearest case is *Paulinella*, which involves a cyanobacterium-derived chromatophore rather than a eukaryotic algal endosymbiont (Gabr et al., 2020), hence an ongoing instance of primary endosymbiosis. The only other described cases, the dinoflagellates MGD and TGD, occur in a serial secondary endosymbiotic context (replacement of the ancestral peridinin plastid by a primary green algal endosymbiont): they retain nucleomorphs and show evidence of ongoing EGT, suggesting an intermediate stage of secondary plastid organogenesis (Matsuo et al., 2022; Sarai et al., 2020).

6. Kleptoplasty across eukaryotes

Kleptoplasty has evolved independently across a broad phylogenetic range of eukaryotes, although it is especially common in microbial lineages. Among dinoflagellates, kleptoplasty is well documented. In the *Dinophysis acuminata* complex, plastids are typically of cryptophyte origin, but are not acquired directly from cryptophytes, *Dinophysis* exploits the ciliate *Mesodinium rubrum* as an intermediary, a trophic relay (Wisecaver and Hackett, 2010; Minnhagen et al., 2011) whereas, in *Dinophysis mitra*, the retained plastids are instead of haptophyte type (Koike et al., 2005), indicating that even within a single dinoflagellate genus, kleptoplast sources can differ substantially. The RSD retains plastids derived from the haptophyte *Phaeocystis antarctica* and can maintain them for up to several months in the absence of prey, longer than any other known kleptoplastic system in protists (Gawryluk et al., 2019; Stamatakis et al., 2017). In Kryptoperidiniaceae, *Durinskia capensis* temporarily retains plastids from diatom prey, suggesting that dinotoms may span intermediate states between transient plastid retention and more stable plastid integration given the permanent state of its sister species, namely *K. foliaceum* and *K. kwazulunatalensis* (Yamada et al., 2019, 2023).

Ciliates (which are also alveolates) provide another major reservoir of kleptoplastidic diversity. Classical oligotrichs such as *Laboea strobila*, retain functional kleptoplasts (Stoecker et al., 1988), and later literature reviews made clear that kleptoplasty is widespread among plastid-retaining ciliates (Johnson, 2011). More recent work on *Strombidium* has reinforced the idea that plastid retention in ciliates is a recurrent trophic solution (Maselli et al., 2021; McManus et al., 2018). Particularly relevant here is *Mesodinium chamaeleon*, a species capable of retaining kleptoplasts from at least six distinct genera of cryptophyte algae (Paight et al., 2023). Additional examples broaden this picture: kleptoplasty is well established in benthic foraminifera (Jesus et al., 2022) and in the marine centrohelid *Meringosphaera*, which harbours kleptoplasts derived from dictyochophytes (Ochrophyta) (Sørensen et al., 2023). Foraminifera belong to Rhizaria while centrohelids are classified with Haptista, with haptophytes (Burki et al., 2020). Among metazoans, however, long-term functional kleptoplasty remains largely restricted to sacoglossan sea slugs, including species of *Elysia*, *Plakobranthus*, and *Costasiella* (Cruz and Cartaxana, 2022).

6.1. Diversity of prey and donor sources

Beyond its phylogenetic diversity across host lineages, kleptoplasty can also involve a remarkably broad prey spectrum within individual species. In this respect, *Dinophysis mitra* stands out, since it feeds mainly on haptophytes, but plastids from other microalgal groups, including Bolidophyceae, Bacillariophyceae, Dictyochophyceae, Pelagophyceae (all ochrophytes), and prasinophytes (Viridiplantae), have also been detected (Nishitani et al., 2012). A similarly broad prey spectrum has

been described in the dinoflagellate *Shimiella gracilentia*, which feeds on species from at least three major algal lineages, including the prymnesiophytes (Haptophyta) *Chrysochromulina* sp. and *Phaeocystis antarctica*, the cryptophytes *Teleaulax amphioxeia*, *Storeatula major*, and *Rhodomonas salina*, and the prasinophyte *Pyramimonas* sp. (Ok et al., 2021).

This diversity of prey sources provides a concrete basis for the genetic mosaicism observed in kleptoplastic lineages. As already discussed, *Dinophysis* and the RSD both harbour nucleotide genes of phylogenetically mixed origins, which do not match their current kleptoplast donor (Hehenberger et al., 2019; Hongo et al., 2019). This diversity of prey offers an ecological explanation for such patterns: repeated interactions with photosynthetic organisms from different lineages multiply the opportunities for ERGT from distinct donors, a process whose selective logic was developed in section 5.2.

7. Diversity of complex algae

The preceding sections have introduced all major lineages of complex algae (i.e., Chlorarachniophyceae, Euglenophyceae, Cryptophyceae, Haptophyta, peridinin Dinophyceae, Chromerida (Colpodellida), Ochrophyta), whose distinguishing traits are compiled in Table 1. Among them, two independent lineages harbour plastids of green algal origin, each acquired through a separate event : Euglenophyceae, whose plastid derives from a donor related to *Pyramimonas* (Chlorophyta; Pyramimonadophyceae), and Chlorarachniophyceae, whose plastid originates from a core chlorophyte related to Bryopsidales (Chlorophyta; core chlorophytes; Ulvophyceae) (Rogers et al., 2007).

The red plastid-bearing lineages are more numerous and their evolutionary relationships more contentious. There is broad agreement that the cryptophyte plastid is of secondary red algal origin (Baurain et al., 2010; Burki et al., 2012) and, at the other end of the spectrum, that the chromerid and apicomplexan plastids share a common ancestry within the Myzozoa (at least quaternary endosymbiosis), with phylogenomic evidence placing chromerid plastids within the ochrophyte radiation (Janouškovec et al., 2015a; Ševčíková et al., 2015). The plastids of ochrophytes and haptophytes (chromists) (Cavalier-Smith, 1981) share several structural and biochemical features, including four bounding membranes, a chloroplastic ER, chlorophyll c, and the SELMA protein import machinery (Hempel et al., 2009; Sommer et al., 2007). These similarities were initially interpreted as evidence for a single common plastid acquisition (Teich et al., 2007). However, extending any common ancestry to peridinin dinoflagellate plastids was particularly problematic: they possess only three bounding membranes, lack both SELMA and a chloroplastic ER, and carry a highly reduced plastid genome organized as minicircles (Barbrook et al., 2019; Nassoury et al., 2003). Although the plastid origins of haptophytes, ochrophytes and peridinin dinoflagellates remains highly debated (Archibald,

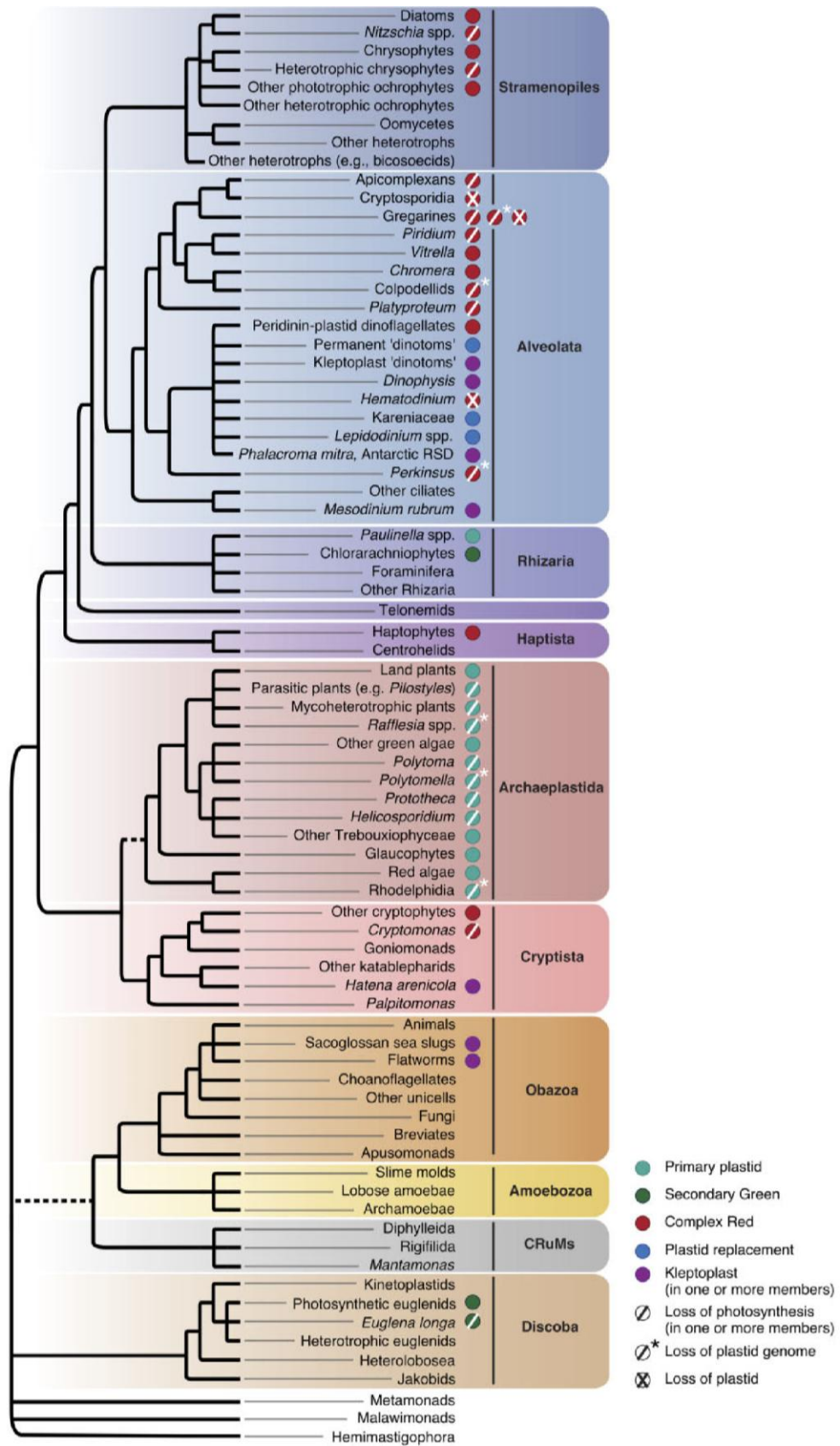
2015; Burki et al., 2020; Dorrell et al., 2017; Pietluch et al., 2024; Ponce-Toledo et al., 2025; Stiller et al., 2014a; Strassert et al., 2021), current evidence increasingly favours independent acquisitions through serial higher-order endosymbioses, sometimes implying that the same donor lineage may have been involved in separate events (reviewed in Ponce-Toledo et al. 2025).

This overview is not exhaustive. Within several of these major groups, particular cases further illustrate the remarkable plasticity of plastid evolution. Among complex green algae, the kleptoplastic *Rapaza viridis* (Euglenophyceae) sheds light on a potential plastid acquisition by kleptoplasty in euglenids (Karnkowska et al., 2023; Yamaguchi et al., 2012). Dinoflagellates, which are discussed extensively in this thesis, display a uniquely high propensity for plastid turnover. Kleptoplastic lineages such as *Dinophysis* (Wisecaver and Hackett, 2010) and the RSD (Hehenberger et al., 2019), as well as cases of plastid replacement in Kareniaceae (haptophyte-derived plastid; Tengs et al., 2000) and Kryptoperidiniaceae (diatom-derived plastid; Hehenberger et al., 2014; Imanian et al., 2010), have been introduced in the preceding sections. Beyond these, dinoflagellates have independently replaced their ancestral peridinin plastid with green plastids of pedinophyte origin (Chlorophyta) in three distantly related species by serial secondary endosymbiosis: in *Lepidodinium* spp. (Kamikawa et al., 2015; Watanabe et al., 1990), and in two as yet undescribed strains, MGD and TGD, both of which retain a nucleomorph-like structure in the periplastidial compartment (Matsuo et al., 2022; Sarai et al., 2020). The fact that a complex red plastid has been replaced by a primary green one three times independently is in itself a striking illustration of the evolutionary lability of the peridinin dinoflagellate plastid. More recently, the boundary between kleptoplasty and permanent plastid retention has been blurred further within these very lineages: Novák Vanclová et al. (2024) have revealed a diversity of plastid origins within the Kareniaceae itself, while Yamada et al. (2025) have shown that within the Kryptoperidiniaceae, kleptoplastic and permanently integrated diatom plastids coexist across closely related species, forming a functional continuum.

Finally, complex plastid evolution also encompasses trajectories of loss. Several parasitic or heterotrophic lineages retain a non-photosynthetic plastid for essential metabolic functions, as exemplified by the apicoplast of Apicomplexa (Janouškovec et al., 2015a; Ralph et al., 2004). Rarer instances of apparently complete plastid loss have been reported in several lineages, notably in *Cryptosporidium* (Apicomplexa; Zhu et al., 2000), *Hematodinium* sp. (dinoflagellates, Gornik et al., 2015), *Actinophrys sol* (Ochrophyta; Azuma et al., 2022) and *Picophagus flagellatus* (Ochrophyta; Terpis et al., 2024). Cases of plastid genome loss in lineages that nonetheless retain a plastid organelle and plastid-derived metabolic pathways have also been documented, such as *Polytomella* (Chlorophyta), whose plastid has completely lost its genome yet remains functional (Smith and Lee, 2014), and the parasitic angiosperm *Rafflesia lagascae* (Molina et al., 2014). Together with plastid replacements and kleptoplasty, these trajectories of loss underscore that plastid evolution is not a one-

way process of progressive integration but a dynamic landscape in which acquisition, replacement, and loss can all occur, sometimes within closely related lineages (Figure 15; Sibbald and Archibald, 2020).

Figure 15: Figure on the next page. Simplified phylogenetic overview of major eukaryotic lineages showing the discontinuous distribution of plastid-bearing lineages and plastid-related traits across the tree. Coloured circles indicate the type of plastid present in each lineage: primary plastids, complex green plastids, complex red plastids, plastid replacements, and kleptoplasts. Additional symbols mark lineages that have secondarily lost photosynthesis, lost the plastid genome, or lost the plastid entirely. The figure highlights both the broad phylogenetic spread of plastid-bearing organisms and the remarkable diversity of plastid evolutionary states among extant eukaryotes (reproduced from Sibbald and Archibald, 2020).



<i>Feature</i>	<i>Euglenophyceae</i>	<i>Chlorarachniophyceae</i>	<i>Cryptophyceae</i>	<i>Haptophyta</i>	<i>peridinin Dinophyceae</i>	<i>Chromerida (Colpodellida)</i>	<i>Ochrophyta</i>
<i>Importance (species)</i>	2,037	at least 6 species	245	517	2,956	8	21,052
<i>Pigments</i>	chlorophylls a and b; β-carotene, neo- and diadinoxanthin [echinenone, diato- and zeaxanthin]	chlorophylls a and b; β- carotene, lutein, neo-, viola- and lorenzoanthin dodecanoate	chlorophylls a and c; phycoerythrin (or phycocyanin); α- carotene, allo-, croco-, zea-, and monadoxanthin	chlorophylls a and c; β-carotene, diadino- and fucoxanthin	chlorophylls a and c; β-carotene and peridinin [diadino-, diato- and dinoxanthin]	chlorophyll a (no chlorophyll c); β- carotene, violaxanthin; pseudo-isofucoxanthin or vaucherixanthin	chlorophylls a and c; β-carotene, fucoxanthine, violaxanthine [diato- and diadinoxanthin]
<i>Storage compounds</i>	paramylon (β1-3 branched β1-6), insoluble cytoplasmic	pseudo-paramylon (β1-3 branched β1-6), soluble vesicular cytoplasmic	rhodamylon (α1-4 branched α1-6), insoluble in periplastidial compartment	laminarin (β1-3 branched β1-6), soluble vacuolar cytoplasmic	starch (α1-4 branched α1-6), insoluble cytoplasmic	amylopectin (α1-4 branched α1-6)?, insoluble cytoplasmic	laminarin (β1-3 branched β1-6), soluble vacuolar cytoplasmic
<i>Plastid type</i>	complex green plastid	complex green plastid	complex red plastid	complex red plastid	complex red plastid	complex red plastid	complex red plastid
<i>Plastid membranes</i>	3 membranes	4 membranes + nucleomorph	4 membranes + cER + nucleomorph	4 membranes + cER	3 membranes	4 membranes	4 membranes + cER
<i>Thylakoids</i>	in triplets (no girdle lamella)	in triplets (no girdle lamella)	in pairs	in triplets (no girdle lamella)	in triplets (no girdle lamella)	in triplets (no girdle lamella)	in triplets (with girdle lamella)
<i>Mitochondria</i>	discoidal cristae (with basal constriction)	tubular cristae (with basal constriction)	flattened cristae	tubular cristae	tubular cristae	tubular cristae	tubular cristae
<i>Flagella</i>	2 (or 4) anisokont; in anterior invagination; often only one emergent	1; coiled around cell body in zoospores	2 iso- or anisokont; anterior invagination or median position	2 iso- or anisokont; + 1 haptonema	2 anisokont; perpendicular in specialized grooves	2 anisokont	generally 2 heterokont

<i>Stigma</i>	extraplastidial	absent	absent [intraplastidial]	absent or intraplastidial	absent [intra- or extraplastidial, ocellus]	absent / present (extraplastidial ?)	absent [intra- or extraplastidial]
<i>Morphology</i>	monadoid [non- flagellated colonies]	monadoid, coccoid, amoeboid (plasmodia) [palmelloid]	monadoid [palmelloid]	monadoid [palmelloid, coccoid, amoeboid, filamentous]	monadoid [palmelloid, coccoid, amoeboid, filamentous]	monadoid (coccoid stages)	n.a.
<i>Habitats</i>	marine, freshwater or brackish; soils, muds, or parasitic	benthic marine (and [pico-]planktonic)	marine (50%) and freshwater (50%) or symbiotic	marine planktonic [freshwater or brackish]; symbiotic (zooxanthellae)	marine planktonic (90%) or freshwater; endozoic, symbiotic or parasitic	marine symbiotic (zooxanthellae)	n.a.

Table 1: Data compiled from (de Reviere, 2002) for all lineages except Chromerida (Oborník et al., 2012; Weatherby and Carter, 2013). Numbers of species were updated (Guiry, 2024) except Chlorarachniophyceae. cER: chloroplastic endoplasmic reticulum; n.a.: too variable to be tabulated.

8. Working hypotheses

The preceding sections introduced several non-exclusive hypotheses to explain the mosaic phylogenetic origins of nucleus-encoded plastid-targeted genes (nucpt#) in complex algae. Among them, the Shopping Bag Endosymbiosis model invokes repeated transient symbioses preceding plastid fixation, with genes from earlier partners facilitating the integration of a later plastid. The central question of this thesis is whether such pre-adaptive histories can be detected quantitatively across complex algae. Kleptoplasty provides a concrete mechanism for pre-adaptation, because it combines repeated exposure to potentially diverse photosynthetic prey with direct selective pressure to retain functions that sustain stolen plastid activity, and a demonstrated ability to prevent membrane accumulation. Under these assumptions, in lineages whose plastid was acquired through a kleptoplastic route, nucpt# genes should be of more diverse origins than in lineages whose plastid was acquired through more canonical endosymbiotic routes, where the endosymbiont is long-lived and exerts weaker, indirect selective pressure on genes favouring plastid integration. However, this prediction is not a strict diagnostic criterion. As outlined conceptually in Figure 16 (on the next page), several alternative scenarios can break a one-to-one correspondence between plastid-acquisition mode and donor diversity. Indeed, a lineage with a stable endosymbiotic plastid may still accumulate a diverse nucpt# repertoire although repeated HGT/ERGT from unrelated prey or endosymbionts. Conversely, a kleptoplastic lineage may display low donor diversity if it repeatedly exploits a narrow or highly specialized prey spectrum. Thus, mosaicism itself may be more widespread among complex algae, regardless of the exact mechanism behind plastid acquisition; the key question addressed here is not whether nucpt# gene repertoires are mosaic, but how diverse their donor composition is across lineages. In summary, fundamentally different plastid-acquisition routes can lead to identical terminal signals.

We also suggest that kleptoplasty may impose a functional progression in the types of genes favoured during plastid integration. First, the immediate benefit of retaining a stolen plastid is to extract fixed carbon (and possibly nitrogen), energy, or photosynthetically-derived metabolites (Johnson, 2011; Stoecker et al., 2009). Whether this metabolite import into the host occurs through dedicated transporters, passive leakage or a mix of both (as discussed in Ball et al., 2025b in the context of primary endosymbioses) is currently unknown, even in intensively studied model systems such as sacoglossan sea slugs (Cruz and Cartaxana, 2022), albeit numerous putative kleptoplast transporters have been reported in *Rapaza viridis* (Karnkowska et al., 2023). In any case, selection must be strong to keep the kleptoplasts photosynthetically active for as long as possible. Hence, any gene transfer whose product can act or help to build “spare parts” for the maintenance of the aging kleptoplast is very likely to constitute an advantage to the host. Such ratchet would progressively pre-adapt the host to exploit the photosynthates of freshly captured kleptoplasts and maintain them for ever longer periods of time

through, e.g., photoprotection, photosynthetic complex repair and redox stress management (Wisecaver and Hackett, 2010; Hehenberger et al., 2019; Hongo et al., 2019; Karnkowska et al., 2023; Kashiya et al., 2026). Second, if the interaction persists over evolutionary time, the plastid may become increasingly embedded in host genetic and metabolic networks, involving biosynthetic functions beyond photosynthesis such as fatty acid biosynthesis (Ralph et al., 2004). This integration would eventually make the host dependent on its plastid for metabolic pathways not directly related to photosynthesis, as exemplified by the essential retention of the apicoplast in non-photosynthetic Apicomplexa (Ralph et al., 2004). Thus, in our view, kleptoplasty may drive a gradual shift from immediate kleptoplast exploitation and maintenance to an eventually broader metabolic integration of a now permanent organelle.

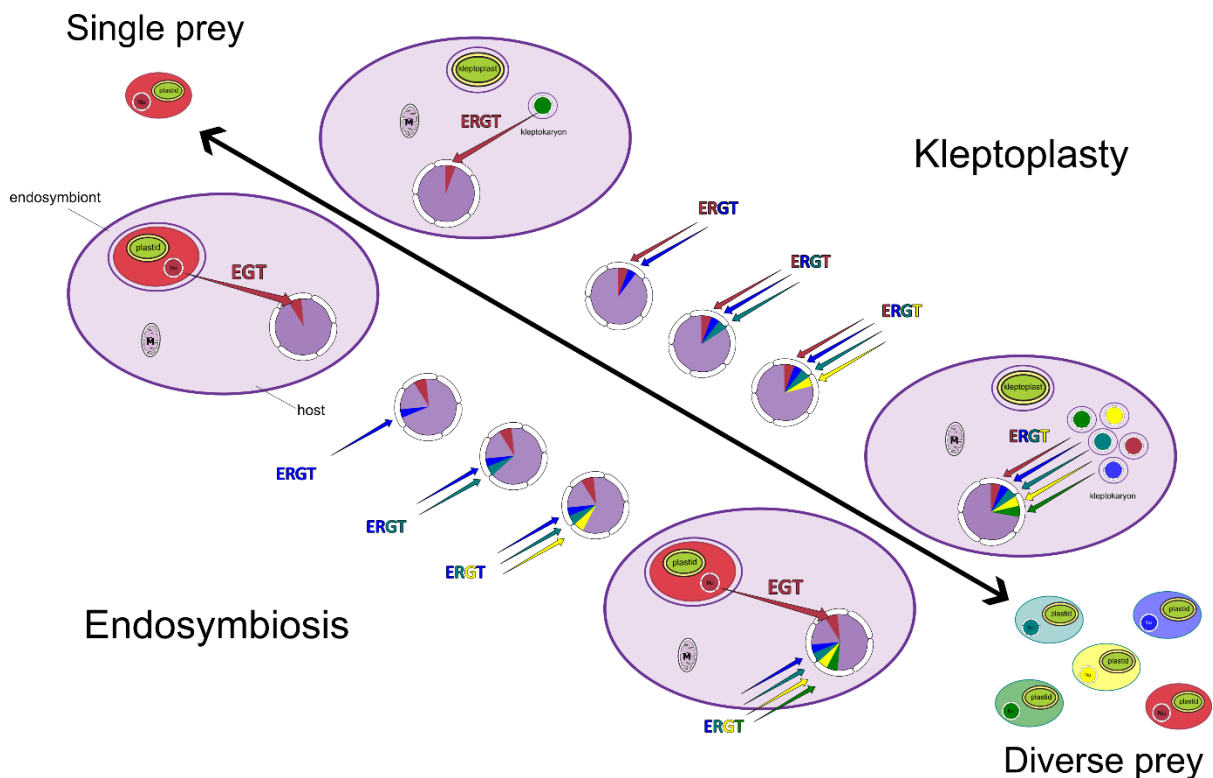


Figure 16: Schematic representation of the diversity of *nucpt#* gene mosaicism under scenarios combining the mode of plastid acquisition (endosymbiosis vs kleptoplasty) and the diversity of prey (or endosymbiont) sources (single vs diverse). Pie charts on transferred genes in the nucleus of the host represent the phylogenetic origin of donors. In a single-prey kleptoplasty scenario (top left), ERGT originates from a restricted donor pool, producing a low α -diversity indistinguishable from that expected under endosymbiosis with EGT from a single endosymbiont (bottom left). Conversely, an endosymbiotic host exposed to diverse prey through continued phagotrophy (bottom right) can accumulate ERGT from multiple sources, generating a mosaic signal comparable to that of a kleptoplasty-mediated acquisition marked by feeding on diverse prey (top right).

9. Outline of the thesis

9.1. Objectives and strategies

The primary goal of this thesis was to measure the diversity of donor origins among nucleus-encoded plastid-targeted (nucpt#) genes across complex algae, using an automated, phylogeny-based framework. This required a large collection of high-quality single-gene trees, which in turn depended on broadly sampled sequence data with well-controlled orthology and low contamination, a challenge for organisms with phagotrophic ancestors and extant relatives. We therefore organized the thesis as a series of methodological steps, each addressed in a dedicated chapter, before proceeding to the main evolutionary analysis.

The first step was to assemble a reliable reference dataset of proteomes (Chapter III). Proteomes served as seeds for orthology definition and were evaluated for completeness (BUSCO) and contamination. For contamination detection, we relied primarily on Forty-Two, a tool developed by my supervisor, to which I contributed taxonomic classification features relevant to contamination screening and of which I performed extensive benchmarking (Chapter II). We also developed a tool to assign encoding localization to each gene, allowing us to filter orthogroups for plastid-targeted candidates in downstream analyses.

However, proteomes alone are too sparsely sampled to capture the full diversity of complex algae. To broaden taxonomic coverage, we complemented the 73 selected proteomes with transcriptomes drawn from the MMETSP resource, one of the largest publicly available collections at the time (Chapter IV). MMETSP was known to be heavily contaminated, a problem our laboratory had previously documented using Forty-Two. We therefore developed a dedicated decontamination strategy. Because eliminating divergent sequences outright is problematic when studying horizontal gene transfer (as genuine transfers can resemble contamination), we instead focused on cross-contamination between sequencing libraries. We adapted CroCo, a tool originally designed for pairwise transcriptome comparison, and we parallelized it (the Sobek tool) to process all 678 MMETSP transcriptomes simultaneously, enabling the identification and removal of cross-contaminated samples. Closely related and highly redundant libraries were pooled, a choice justified by our focus on lineage-level rather than species-level histories. Individual contamination levels were further assessed with Forty-Two, and the most contaminated samples were excluded; however, individual contaminated sequences within retained samples were not removed, to avoid discarding genuine transfer candidates.

The cleaned proteome and transcriptome datasets were then combined (Chapter V). Transcriptomes were enriched into the proteome-seeded multiple sequence alignments using Forty-Two, for which I

developed a feature preventing the assignment of a single sequence to multiple alignments (to avoid conflating paralogs). The resulting MSAs were used to compute phylogenetic trees with IQ-TREE, yielding our final dataset of high-quality single-gene trees.

With this dataset in hand, we developed an automated phylogenetic pipeline (the PAGTI tool) to infer transfer events from plastid-related gene trees in a taxonomy-aware yet discovery-driven manner (Chapter V). Rather than relying on best-hit approaches or simple sister-group relationships, our method detects transfer-compatible topologies based on paraphyletic donor assemblages surrounding quasi-monophyletic receiver clades, intruding sequences being tolerated at different levels. Transfer events are then summarized as organism-level donor α -diversity profiles, providing a quantitative metric to compare the diversity of donor contributions across receiver lineages.

9.2. Predictions

This whole framework, including both a large dataset of phylogenetic trees of plastid-targeted genes and our transfer event inference pipeline, was then used to test four predictions (Chapter V). First, lineages with documented kleptoplastic behaviour should show higher donor α -diversity than lineages with well-established endosymbiotic plastids. Second, α -diversity should partition complex algae into distinct regimes corresponding to different modes of plastid acquisition. Third, if pre-adaptation precedes permanent plastid establishment, then pre-adaptive gene sets should be functionally biased toward immediate plastid exploitation and maintenance rather than long-term plastid integration (i.e., host metabolic integration beyond photosynthesis). Fourth, plastid replacement events should compress α -diversity toward single-donor dominance, providing a natural perturbation of the framework.

The main chapter (Chapter V) is thus organized accordingly: first, a global analysis of transfer fluxes and donor α -diversity across lineages; second, targeted case studies focused on euglenids, peridinin dinoflagellates, Dinophysiales, and ochrophytes.

II. Chapter 2 - Forty-Two: a next-generation tool for enriching alignments of orthologous genes and gene families.

Mick Van Vlierberghe¹, Arnaud Di Franco², Hervé Philippe², Denis Baurain¹

¹ InBioS–PhytoSYSTEMS - Eukaryotic Phylogenomics, ULiège, Liège, Belgium

² SETE, UMR CNRS 5321, Moulis, France

Contributions

D. Baurain developed the core codebase, while I contributed substantially to the taxonomic methods implemented in the Bio::MUST::Core modules, as well as to unit testing. I also wrote scripts for run debriefing, preventing multiple integration of identical transcripts across MSAs, and generating YAML configuration files.

H. Philippe provided manually curated eukaryotic ribosomal MSAs. I performed taxonomic standardization (i.e., file and sequence renaming) to enable their use in our contamination detection framework.

1. Abstract

Forty-Two was developed as a flexible tool for enriching predefined orthologous groups from heterogeneous sequence datasets, including proteomes, transcriptomes, and unannotated genomes. Its first originality lies in its multiple triangular best reciprocal hit strategy: candidate sequences are not validated against a single reference organism, but against a set of independent reference proteomes, using an orthology checklist built separately for each orthologous group. The `ref_org_mul` parameter further makes this procedure adaptive by retaining only the most informative reference proteomes for a given orthologous group, thereby balancing orthology stringency and recovery rate across gene families of different taxonomic richness and complexity.

A second major advance is Forty-Two's ability to enrich orthologous groups directly from unannotated genomic data. Through an Exonerate- (Slater and Birney, 2005) based splicing step followed by HSP sorting and trimming, Forty-Two can recover full-length coding sequences from long genomic fragments or scaffolds, including scaffolds containing multiple genes. This feature distinguishes it from transcriptome-oriented tools such as HaMStR (Ebersberger et al., 2009) and Orthograph (Petersen et al., 2017) and explains its much higher performance on unannotated genomes in the benchmark presented here.

Third, Forty-Two integrates taxonomic quality control into the enrichment process. After orthology assessment, optional taxonomic filters can classify candidate orthologs using an LCA-based procedure inspired by MEGAN (Huson et al., 2016). Depending on user settings, sequences whose inferred taxonomic affiliation does not match the expected source taxon can be discarded or explicitly tagged as putative contaminants. In addition, Forty-Two can be used as a stand-alone contamination detection tool when applied to curated marker-gene datasets, producing reports on completeness, purity, contamination level, and compile contamination sources with the companion script `debrief-42.pl`.

Together, these features make Forty-Two more than an orthogroup-enrichment program. It provides a modular toolkit for orthology-aware sequence recovery, genomic extraction from raw nucleotide data, contamination-aware filtering, run debriefing, and reproducible dataset construction. Its associated utilities and derived tools, including `yaml-generator-42.pl`, `one-on-one.pl`, `Leel` (Rodríguez et al., 2017), `MESYRES` (Cornet et al., 2021), and `Physeter` (Lupo et al., 2021), further illustrate how the original enrichment framework has been extended into a broader ecosystem for phylogenomic data curation.

This chapter is organized in two main parts: the first describes the orthology enrichment capabilities of Forty-Two, from its methodological foundations to its benchmarking against comparable existing tools,

while the second presents the contamination detection capabilities of Forty-Two and their application in published studies.

2. Orthology enrichment

2.1. Introduction

In phylogenetics and evolutionary genomics, distinguishing between orthologous and paralogous genes is fundamental. Orthologs are homologous genes that diverged following a speciation event, whereas paralogs arise from gene duplication events within a lineage. This distinction is a specific case of homology, which broadly refers to common ancestry (Fitch, 1970; Koonin, 2005; Altenhoff et al., 2019). Orthologs tend to retain similar biological functions across species, whereas paralogs are more likely to diverge functionally following duplication (Altenhoff et al., 2012; Gabaldón and Koonin, 2013). Because speciation events mirror species divergence, orthologous genes are therefore generally considered the appropriate units for reconstructing species trees, whereas paralogs can confound phylogenetic inference by reflecting gene-level rather than species-level histories (Koonin, 2005; Kristensen et al., 2011).

Accurate identification of orthologs in genomic datasets is non-trivial, and this is not only due to algorithmic issues. Indeed, gene duplication, gene loss, and horizontal gene transfer can obscure evolutionary relationships and genuinely complicate the inference of orthology. In particular, many gene families, some of which can be very large, exhibit complex evolutionary histories, including the presence of co-orthologs (resulting from lineage-specific duplications) and hidden paralogs that may be difficult to distinguish using sequence similarity alone (Koonin, 2005; Tekai, 2016; Natsidis et al., 2021). These complexities make orthology inference a central and still unresolved challenge in evolutionary genomics (Glover et al., 2019; Linard et al., 2021).

A widely used operational heuristic for orthology inference is the best reciprocal hit (BRH), also referred to as reciprocal or bidirectional best hit (RBH/BBH; Overbeek et al., 1999). Under this approach, two genes from different genomes are considered putative orthologs if they are each other's best match in pairwise sequence similarity searches (Kristensen et al., 2011; Wolf and Koonin, 2012; Altenhoff et al., 2019). BRH is a heuristic method rather than a phylogenetically rigorous definition, but it has been widely adopted due to its simplicity and computational efficiency, although it is not typically attributed to a single original publication, the approach was formalized through early comparative-genomic procedures such as the COG framework, which relied on the consistency of genome-specific best hits to delineate orthologous groups (Tatusov et al., 1997, 2000).

Nevertheless, BRH has several limitations. First, when applied strictly (i.e., enforcing a single best hit), it cannot accommodate co-orthologs arising from lineage-specific duplications (Dalquen and Dessimoz, 2013). Second, BRH relies on sequence similarity scores (e.g., BLAST scores), which are not direct measures of evolutionary distance and can be biased by factors such as alignment parameters or substitution rate variation among lineages (Wall et al., 2003). Third, substantial shifts in substitution rates following duplication or horizontal transfer events can lead to incorrect identification of best hits, thereby confounding orthology inference (Lafond et al., 2018).

Despite its limitations, BRH forms the conceptual basis of many orthology inference algorithms. Graph-based clustering approaches extend BRH relationships to identify orthologous groups (OGs), which partition proteins from multiple species into homologous sets inferred to derive from a common ancestor (Tatusov et al., 1997; Enright, 2002). Popular software packages implementing such approaches or related strategies include Proteinortho (Lechner et al., 2011), OrthoMCL (Li et al., 2003), OrthoFinder (Emms and Kelly, 2015) and OMA (Roth et al., 2008). These methods are typically applied to complete proteomes derived from the conceptual translation of coding sequences to generate genome-wide OGs. However, many of these OGs correspond to large multigene families that include both orthologs and paralogs. This has motivated the development of more refined frameworks, such as hierarchical orthologous groups (HOGs), which aim to reconcile gene family structure with the species tree under study and to provide a more fine-grained representation of orthology relationships (Koonin, 2005; Gabaldón and Koonin, 2013; Glover et al., 2019).

In addition to de novo orthology inference methods, several curated orthology databases have been developed to provide precomputed OGs across large taxonomic scales. Prominent examples include eggNOG (Jensen et al., 2008), OrthoDB (Kriventseva et al., 2008), and Ensembl Compara (Vilella et al., 2009), which integrate genome-wide comparisons to define OGs and hierarchical relationships across hundreds to thousands of species. These resources are widely used both as references for functional annotation and as standardized gene sets in evolutionary genomics (Altenhoff and Dessimoz, 2009; Glover et al., 2019). In particular, they form the basis of software packages designed to estimate the completeness and quality of genome assemblies, transcriptomes, or gene predictions, such as BUSCO (Simão et al., 2015), which relies on sets of near-universal single-copy orthologs expected to be present in a given lineage (Simão et al., 2015). These approaches therefore require that genomes are not only available but also accurately assembled and annotated, in particular with reliable protein-coding gene predictions.

Consequently, the effective use of such reference ortholog sets is often limited by the availability of high-quality genomic data and annotations. A major limitation in evolutionary genomics is that many

organisms of interest lack fully sequenced and well-annotated genomes. Instead, available data may consist of fragmented genomic assemblies or transcriptomic datasets such as expressed sequence tags (ESTs) or de novo assembled transcripts (Philippe and Telford, 2006; Yang et al., 2014). Although such datasets are inherently incomplete, they remain highly valuable for phylogenomic analyses. Highly expressed and conserved genes, such as ribosomal proteins and other housekeeping genes, are frequently recovered in transcriptomic data and have been widely used to construct phylogenomic supermatrices (Baptiste et al., 2002; Philippe et al., 2004; Baurain et al., 2007; Wickett et al., 2014). However, low-quality genomic and transcriptomic datasets are often affected by contamination from foreign sequences (i.e., coming from organisms other than the one from which the dataset is purported to originate; Cornet and Baurain, 2022), which can dramatically impair phylogenomic analyses and evolutionary conclusions if not properly filtered (Philippe et al., 2011; Laurin-Lemay et al., 2012).

A common strategy in phylogenomics consists of first inferring OGs from complete proteomes representing a taxonomic diversity appropriate to the question under consideration. These groups are then enriched with additional sequences derived from genomic or transcriptomic datasets (Irisarri et al., 2017; Simion et al., 2017). This two-step approach allows for increased taxonomic sampling while preserving the reliability of the initial orthology inference (Altenhoff et al., 2019; Glover et al., 2019). However, its success critically depends on ensuring that only true orthologous and non-contaminant sequences are incorporated during the enrichment phase (Philippe et al., 2011; Simion et al., 2017).

Several tools have been developed to enrich predefined OGs with additional sequences, including full pipelines that provide collections of curated, aligned OGs aimed at covering varying subsets of the Tree of Life, e.g., GToTree (Lee, 2019) for prokaryotes, PhyloFisher (Tice et al., 2021) and EukPhylo (Katz et al., 2025) for eukaryotes. Among the earlier tools specifically dedicated to the enrichment step from nucleotide data, two standalone solutions are particularly noteworthy: HaMStR (Ebersberger et al., 2009) and Orthograph (Petersen et al., 2017).

HaMStR implements a profile-based, reciprocal best hit strategy targeted at incomplete datasets. It starts from a curated set of “core orthologs” from reference taxa; for each orthologous group, profile hidden Markov models (HMMs; Eddy, 1998) are used to search candidate transcriptomic or low-coverage genomic sequences, and significant hits are then reciprocally searched against the reference proteome. A candidate is added to the orthologous group only if its best hit in the reference taxon belongs to the same group, enforcing a modified reciprocal similarity criterion tailored to incomplete and error-prone data (Ebersberger et al., 2009). Orthograph adopts a related but more global approach. It also relies on a reference set of OGs and a best reciprocal hit strategy using profile HMMs, but addresses a key limitation of HaMStR, namely that transcripts can be assigned redundantly to multiple OGs. Orthograph

instead maps each transcript to the globally best-matching OG, ensuring that none is assigned more than once, while still identifying additional homologous transcripts such as splice variants and inparalogs within each OG (Petersen et al., 2017).

These two software packages rely on reference ortholog sets and sequence similarity searches and thus inherit limitations related to sequence divergence, gene family complexity, and contamination. Moreover, they use profile HMMs for similarity searches, which certainly increases sensitivity (Eddy, 1998), potentially beyond what is desirable, as highly divergent sequences may complicate downstream phylogenetic analyses (Philippe et al., 2011). In this context, we have developed over the past two decades a suite of tools including “forty” and its successor “Forty-Two”, which, in contrast to profile-based approaches, still use similarity searches and BRH for recovering orthologs in various types of genomic datasets, but also come with a number of heuristics that allow them to better cope with multigene families, horizontal gene transfer, and contamination.

Although Forty-Two has been used in several high-impact studies, especially in the field of animal (Irisarri et al., 2017; Simion et al., 2017; Delsuc et al., 2018; Philippe et al., 2019) and eukaryotic (Di Franco et al., 2022) phylogenomics, it has not yet been formally described in the literature. The objective of this chapter is therefore to present the Forty-Two software package, describe its methodological foundations, and compare its performance to existing tools such as Orthograph across a range of biologically realistic scenarios. Particular attention will be given to its ability to accurately enrich OGs while minimizing the inclusion of paralogous or contaminant sequences.

2.2. Objectives and strategies

First, we validated Forty-Two's core heuristic, a multiple triangular Best Reciprocal Hit (multi-BRH) strategy against complete reference proteomes, under controlled conditions. Using a dataset of 480 aligned OGs (= multiple sequence alignments or MSAs) derived from 57 eukaryotic proteomes spanning eight taxonomic groups, we designed pruning/enrichment experiments in which sequences from a given taxonomic group were removed and subsequently recovered. We systematically explored the impact of the `ref_org_mul` parameter (ranging from 0.1 to 1.0), yielding 38,400 experiments in total. Second, we assessed Forty-Two's performance in a real-world scenario by replacing complete proteomes with unannotated nucleotide genomes as the mining set. This step requires an additional splicing and trimming step to recover full-length coding sequences from genomic scaffolds, a key capability distinguishing Forty-Two from profile-based tools such as HaMStR and Orthograph. Finally, we benchmarked Forty-Two against Orthograph using the latter's own reference dataset: 4,625 single-copy OGs from Hymenoptera, applied to both the proteomes and unannotated genomes of *Apis mellifera* and *Harpegnathos saltator*.

2.3. Materials and Methods

2.3.1. Implementation

The Forty-Two software package is available on CPAN and is divided into four main tools that handle (i) configuration setup (`yaml-generator-42.pl`), (ii) MSA enrichment by orthology assessment based on a multi-BRH approach (Forty-Two), (iii) a debriefing tool (`debrief-42.pl`) and (iv) an multi-MSA enrichment prevention tool (`one-on-one.pl`) while the user-friendly assistant configurator helps configuring and reproducing previous runs. Usually, Forty-Two's target for enrichment are OGs that are aligned (MSAs) but this is not a pre-requisite, as Forty-Two can also enrich non-aligned OGs, and more generally any collection of homologous yet preferentially orthologous sequences.

Forty-Two requires only one pre-processing step: formatting the genome/transcriptome databases used to identify candidate orthologs and the reference proteome databases used for orthology assessment. Formatting implies two steps: (i) abbreviating sequence identifiers to single words with a script such as `inst-abbr-ids.pl` from Bio::MUST::Core followed by (ii) turning files into BLAST databases (nucleotide or protein, depending on the data type) with the `makeblastdb` command. All that is left is to follow the instructions given by the wizard (`yaml-generator-42.pl --wizard`).

2.3.2. Orthology assessment

Forty-Two's heuristic (Figure 17) is based on a multiple best reciprocal hit (BRH) criterion; it can be divided into three main steps:

Best-hit checklist building:

Orthologous queries are extracted from the current MSA to enrich (based on a species selection, user defined) and used to perform BLAST searches against the reference proteomes. Each best hit(s) in each proteome for each query is collected, thus constituting a checklist of orthologous sequences in the reference proteome set. The principle behind this checklist is that any other sequence that is orthologous to the concerned gene should recover the same best hit(s) in every reference proteome. The maximum number (in proportion) of reference proteomes actually retained for each MSA is governed by the `ref_org_mul` parameter of the YAML configuration file.

The checklist is further consolidated by performing a BLAST search of each best hit against the other reference proteomes. If one best hit does not recover the best hit(s) collected in the other reference proteomes, it is flagged with a warning, and the corresponding reference organism may be preferentially

excluded from the list (always if it fails all BRH checks). This serves as a quality control step: MSAs in which many reference organisms fail this mutual BRH check are likely to suffer from underlying orthology issues, such as poorly delimited gene families or patchy paralog distributions.

Candidate orthologs mining:

The same orthologous queries are then used to mine the candidate genomes/transcriptomes for homologs (candidate orthologs) with BLAST (e.g., TBLASTN for genomes, BLASTP for proteomes).

Orthology test:

Each homolog is used to perform a BLAST (e.g., BLASTX for nucleotide queries, BLASTP for protein queries) versus each proteome of the reference proteome set and their best hit(s) are collected. If a candidate ortholog recovers the same best hit(s) as in the checklist for every reference proteome, it is then considered as an ortholog.

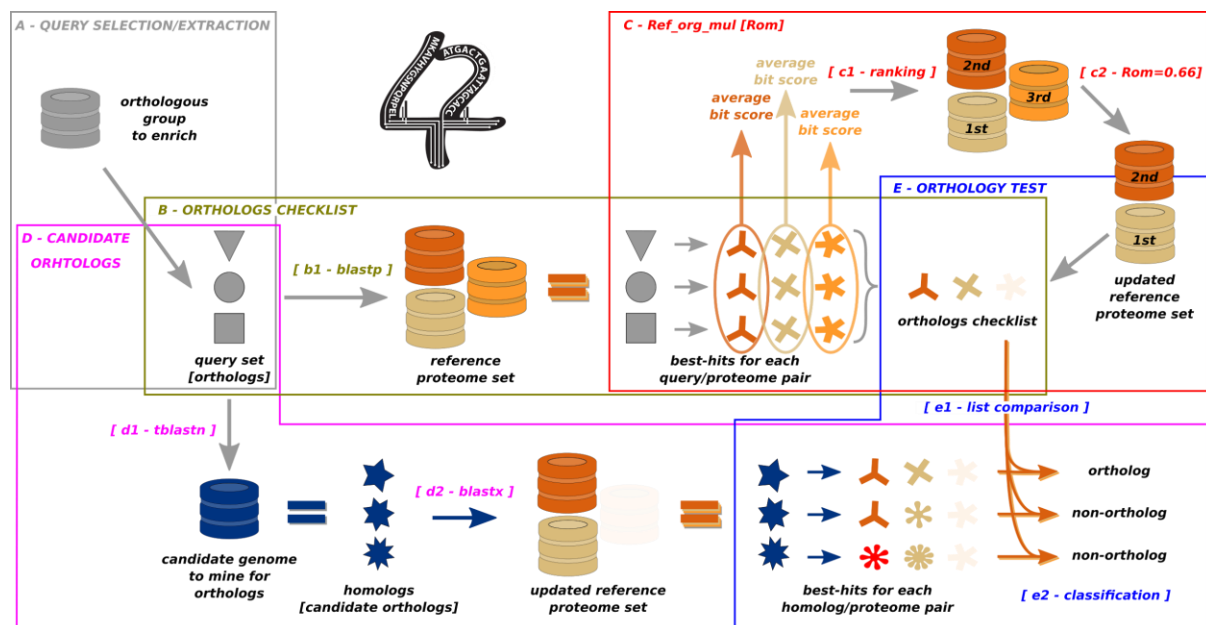


Figure 17: Forty-Two's heuristics illustrated. (A) Extraction of representative sequences from the orthologous group to enrich = “query set” used for subsequent BLAST searches. (B) Ortholog checklist construction with (b1) a BLASTP of the query set against a complete reference proteome set and storage of best hits for each query/reference proteome pairs in order to build an ortholog checklist. (C) Optional “ref_org_mul” parameter to filter out reference proteomes that might hinder orthology assessment by (c1) ranking them according to their average bit score and only keeping the top fraction. (D) Candidate ortholog search with (d1) a TBLASTN of the query set against the genome(s) to mine them for orthologs and (d2) a BLASTX of the resulting homologs (candidate orthologs) against the updated reference proteome set and storage of best-hit lists for each homolog/reference proteome pairs. (E) Orthology assessment for each homolog by (e1) comparing each homolog best-hit list to the previously built ortholog checklist, followed by (e2) classification as ortholog or not.

2.3.3. Paralog-aware enrichment (PARA mode)

Forty-Two is also able to enrich MSAs resulting from the split of complex multigenic families after phylogenetic analysis. For this, it requires files composed of representative sequences of unwanted out-

paralogs (PARA files). When Forty-Two identifies a potential ortholog that is more similar to the sequences in the PARA file than to the sequences in the target MSA, it is rejected for paralogy.

2.3.4. Orthologue merging (COMPRESS mode)

When using highly redundant transcriptome assemblies for MSA enrichment, some genes can be represented by a series of very similar transcripts, either partially overlapping or containing minor sequencing errors. To deal with these situations, Forty-Two provides the parameter `merge_orthologues` in the configuration file. When enabled, orthologues are first fed to CAP3 (Huang and Madan, 1999) in an attempt to merge some of them into contigs. Successfully merged orthologues are identified by a trailing “+N” tag where N is the number of orthologous sequences removed in the merging process. The contig itself is named after the longest orthologous sequence composing it.

2.3.5. Homologue trimming and orthologue splicing

When mining unannotated genomes, additional steps are required to ensure that full-length proteins are recovered; these are termed “homologue trimming” and “orthologue splicing” in Forty-Two. Homologue trimming is triggered at the initial mining step if distinct clusters of BLAST hits fall on the same nucleotide sequence (often genomic scaffolds or contigs containing many genes). Its goal is to ensure proper extraction of (possibly multiple) candidate orthologous sequences, even if in case of tandem duplication and presence of introns. If multiple sequences extracted from the same nucleotide sequence survive until incorporation into the MSA, they will be disambiguated by a trailing “.k” tag where k is the rank of the extracted orthologue ($k > 1$). Orthologue splicing takes place when orthologous sequences are incorporated into the MSA. It can be performed using either the intron-aware aligner Exonerate (Slater and Birney, 2005) or BLAST; in the latter case, exons are assumed to correspond to BLAST HSP (high-scoring segment pairs). Alignment occurs on “template” sequences (one for Exonerate, possibly more for BLAST) that are the most similar to the orthologue to incorporate. Splicing is always followed by another merging step in which all exons are combined into a single contiguous sequence per orthologue. When exons overlap, the merger gives precedence to the those corresponding to the highest-scoring templates and HSPs. Sophisticated homologue trimming and Exonerate-powered orthologous splicing are two key capabilities that distinguishes Forty-Two from simpler profile-based tools such as HaMStR and Orthograph.

2.3.6. Experimental design: pruning/enrichment protocol

To evaluate Forty-Two’s multi-BRH heuristic, we designed “pruning/enrichment” experiments using 480 OGs selected from 210,451 OGs generated by OrthoFinder (v0.4.0) on 57 eukaryotic proteomes spanning eight taxonomic groups (Figure 18). Each experiment depleted OGs of sequences from one taxonomic group (pruning, using `prune-ali.pl` from Bio::MUST::Core), then attempted to reconstruct

the original OG composition with Forty-Two. The mining set comprised the pruned species; the reference proteome set comprised all remaining species. We executed this protocol eight times (one per group) and tested a range of 10 ref_org_mul values (0.1-1.0, step 0.1), for 38,400 proteome-based experiments (480×8×10). In the genome-based scenario, we used ref_org_mul = 0.1 only, giving 3,840 experiments (480×8×1). For proteome mining, reconstruction success was simply assessed by protein identifier comparison, whereas for genome mining, it required phylogenetic tree reconstruction: removed sequences were reinstated, then genomic fragments were assigned to their expected protein counterparts by tree-based bipartition parsing.

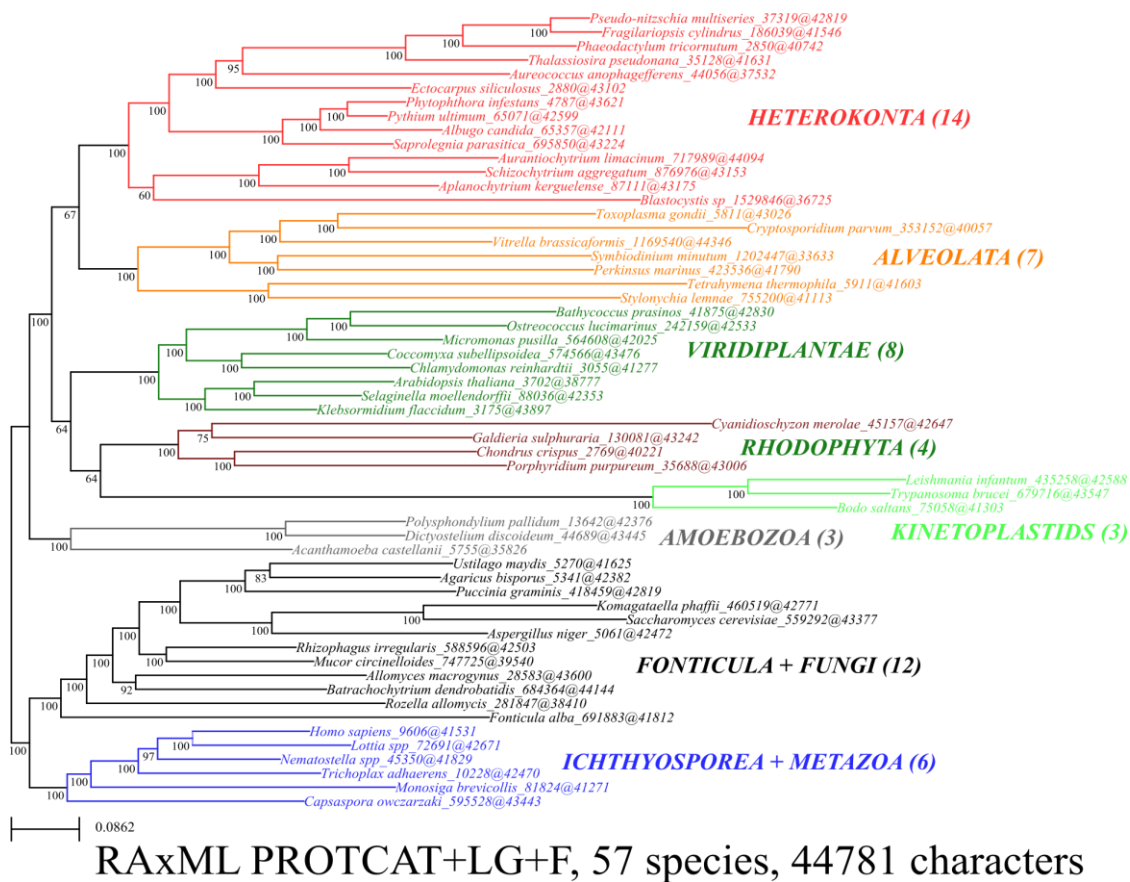


Figure 18: Supermatrix tree built with RAxML PROTCAT+LG+F model from 57 species resulting into 44,781 unambiguously aligned amino acid positions.

2.3.7. F1-score

To measure Forty-Two's performance, we used the F1-score, which is the harmonic mean of precision and recall, thus ignoring true negatives (TN). Precision corresponds to the proportion of predicted positives that are true positives, i.e. $TP/(TP + FP)$, whereas recall corresponds to the proportion of actual positives that are correctly recovered, i.e. $TP/(TP + FN)$. The F1-score therefore summarizes the balance between false positives and false negatives while ignoring true negatives (TN). This methodological choice is due to the complexity of defining what TN are in our experimental setup.

Indeed, they could be all the sequences constituting a genome, but this definition would tend to overestimate specificity $TN/(TN + FP)$ and accuracy $(TP + TN)/(TP + TN + FP + FN)$ because of the huge difference in numbers with TP, FP and FN.

2.3.8. Comparison with Orthograph: experimental setup

We compared Forty-Two with Orthograph (v0.6.3) (Petersen et al., 2017) using the same benchmark dataset previously employed by Petersen et al., (2017) to compare Orthograph (v0.6.3) with HaMStR (Ebersberger et al., 2009). This benchmark consists of 4,625 single-copy OGs from OrthoDB 7, derived from four Hymenoptera species (*Apis mellifera*, *Camponotus floridanus*, *Harpegnathos saltator*, *Nasonia vitripennis*) and the outgroup beetle *Tribolium castaneum*. We tested both packages on proteomes and unannotated genomes of *A. mellifera* (15,314 genes) and *H. saltator* (18,564 genes), then measured the sensitivity and specificity of sequence enrichment.

2.4. Results

2.4.1. OG reconstruction tests

Forty-Two relies on a multiple triangular best reciprocal hit strategy for orthology assessment (Figure 17). To evaluate this heuristic, we used a dataset generated by running OrthoFinder on the complete proteomes of 57 broadly sampled eukaryotes distributed across eight major taxonomic groups (Figure 18). This clustering yielded 210,451 orthologous groups (OGs), of which 177,197 singletons. We thus filtered the remaining 33,254 OGs according to two main criteria, each one divided into three levels. The first criterion was taxonomic richness, with OGs classified as universal when all 57 species were represented, widespread when approximately 66% of species in each taxonomic group were present, and sparse when only approximately 33% were represented. The second criterion was the average gene (i.e., protein) copy number per organism, using thresholds of 1.1, 1.2, and 2.0. This filtering procedure yielded a final benchmark dataset of 480 OGs, as shown in Table 2.

Taxonomic richness	Maximum copy number		
	1.1	1.2	2.0
Universal	17	33	36
Widespread	76	130	59
Sparse	32	60	37

Table 2: Characterisation of the 480 OGs after filtering our 33,254 non-singleton OGs on taxonomic richness (Universal, Widespread, Sparse) and maximum copy number (1.1, 1.2, 2.0).

The tests we designed were called “pruning/enrichment” experiments and consisted in depleting all OGs of targeted sequences according to a taxonomic target (“pruning”) across, followed by an attempt to restore their original taxonomic sampling by running Forty-Two. Forty-Two was run with a mining

set constituted of the species that were removed from the OGs, a complete reference proteome set with all species but the ones that were targeted by the pruning and a query set constituted of all the remaining species in the pruned OGs. The full experimental protocol is described in Materials and Methods.

2.4.1.1. Proving heuristics - Complete proteome data

As mentioned above, the main Forty-Two's heuristic is based on a multiple best reciprocal hit (multi-BRH) approach, termed 'triangular' because it involves the query, the candidate, and multiple independent reference proteomes (each considered in turn, but the condition must be satisfied for all). Hence, for two sequences to be orthologs, they must recover the same best hit(s) when BLASTing against those reference proteomes, which can be organisms that are different from the two other sequences of the triangle. For assessing orthology, Forty-Two allows the use of any number of reference organisms/proteomes and, owing to its "ref_org_mul" feature, it allows a dynamic and automatically tailored selection of the best reference proteomes/organisms for any OG to enrich. By default, orthology assessment requires a successful BRH with every reference proteome (100% match). Yet, the ref_org_mul parameter relaxes this by selecting only a fraction of the available reference organisms, since using a reference organism that lacks the ortholog to identify in the mined genome/transcriptome would obviously hinder orthology assessment. As this parameter was expected to strongly influence Forty-Two's performance, we extensively evaluated the impact of the 'ref_org_mul' feature, it a total of 38,400 pruning/enrichment experiments (see Materials and Methods for details).

Figure 19 shows three precision-recall curves for OGs belonging to each level of taxonomic richness where precision and recall values were computed for 10 different 'ref_org_mul' values. The first thing that stands out is that Forty-Two is precise whatever the taxonomic richness of the OG and 'ref_org_mul' value; only recall values show high variability. Indeed, the higher the 'ref_org_mul', the lower the recall. This can be explained by the stringent nature of the orthology test implemented in Forty-Two. Because a candidate sequence must recover the expected best-hit pattern across all retained reference proteomes, increasing the value of ref_org_mul also increases the probability that at least one reference comparison will fail, thereby causing the orthology test to fail. Such failures may occur for several reasons, including the absence of the ortholog in one or more reference proteomes, but also strong sequence divergence or lineage-specific rate acceleration in some taxa, which can alter best-hit relationships and prevent complete BRH recovery. For example, the group OG0000611-Vir-1 (pruned taxonomic group: Viridiplantae = 19 sequences, ref_org_mul: 1), all of the 19 candidate sequences for enrichment failed the orthology assessment because out of the 49 reference organisms, one lacked the orthologous gene (BRH count was 48 instead of 49). In contrast, in group OG0000611-Vir-0.1 (pruned taxonomic group: Viridiplantae = 19 sequences, ref_org_mul: 0.1) where only 10% (5) of the best (highest average bit scores) reference organisms were used (Table 3), all 19 candidate sequences were

successfully recovered. This reasoning explains the higher variability of recall values when the taxonomic richness decreases. Indeed, for an OG characterized as sparse, constituted of $\frac{1}{3}$ of species within each taxonomic group, it means that only $\frac{1}{3}$ of our 57 organisms (19) possess this orthologous gene. In addition, we used on average 50 reference organisms (different but largely overlapping sets, according to the pruned taxonomic groups); the results show that using high ‘ref_org_mul’ values in such OGs are more likely to include reference proteomes lacking the ortholog, thereby impairing orthology assessment.

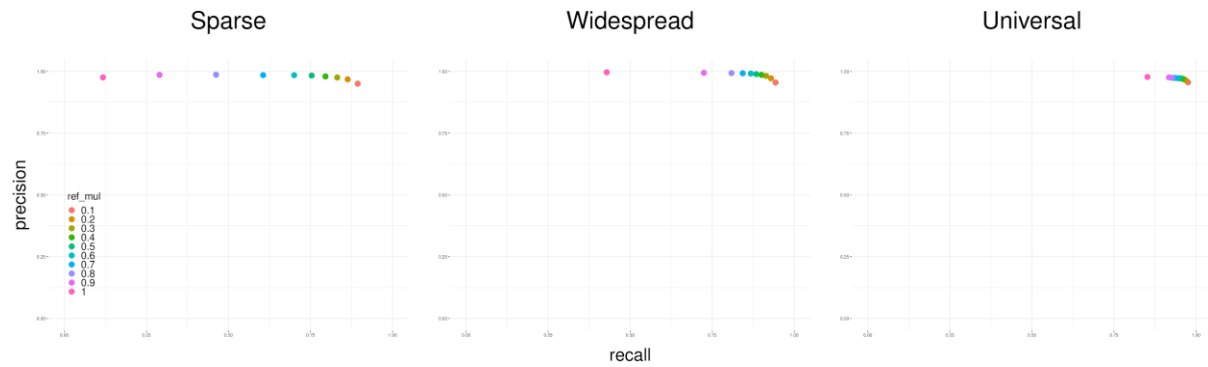


Figure 19: Precision-recall curves computed for the three levels of taxonomic richness (a. Sparse, b. Widespread and c. Universal) according to ‘ref_org_mul’ values.

OG	serie	pres	copy mean	ref_mul	Ps	TPs	FPs	FNs	precision	recall	seqs
OG0000611	Vir	widespread	2	0.1	19	19	0	0	1	1.0000000	104
OG0000611	Vir	widespread	2	0.2	19	18	0	1	1	0.9473684	104
OG0000611	Vir	widespread	2	0.3	19	18	0	1	1	0.9473684	104
OG0000611	Vir	widespread	2	0.4	19	18	0	1	1	0.9473684	104
OG0000611	Vir	widespread	2	0.5	19	18	0	1	1	0.9473684	104
OG0000611	Vir	widespread	2	0.6	19	18	0	1	1	0.9473684	104
OG0000611	Vir	widespread	2	0.7	19	18	0	1	1	0.9473684	104
OG0000611	Vir	widespread	2	0.8	19	18	0	1	1	0.9473684	104
OG0000611	Vir	widespread	2	0.9	19	18	0	1	1	0.9473684	104
OG0000611	Vir	widespread	2	1.0	19	0	0	19	NA	0.0000000	104

Table 3: Precision-recall values for orthologous group OG0000611 when pruning/enriching Viridiplantae across the whole range of ‘ref_org_mul’ values.

Then we computed an AUC (Area Under Curve) for the eight series (8 taxonomic groups) of 10 experiments (10 ‘ref_org_mul’ values) for each OG ($8 \times 480 = 3840$ AUCs). Figure 20a shows that Forty-Two has a good overall ability to reconstruct the original OGs with a lower quartile equal to 0.94. While it is slightly influenced by the taxonomic group (Figure 20b) and gene copy number (Figure 20c) reconstruction is more impacted by the taxonomic richness (Figure 20c). In fact, this is more an artifact

introduced by experimental design than an actual caveat of the software. Indeed, since we are dealing with AUCs, we take into account all ‘ref_org_mul’ values (1 AUC = experiments of 1 OG for 10 ‘ref_org_mul’ values) and we already know that high ‘ref_org_mul’ values tend to hinder orthology assessment for low taxonomic sampling whereas this effect gradually fades away as richness increases. This trend is further captured in Figure 21, which reports F1 scores (harmonic mean of precision and recall) providing a single balanced performance measure that accounts for both missed orthologs and spurious assignments.

This first batch of results shows the importance of the ‘ref_org_mul’ feature of Forty-Two. This parameter allows dealing with large numbers of OGs that can be quite different according to taxonomic and paralogy criteria, as depicted previously, thereby ensuring a robust and adaptive orthology assessment.

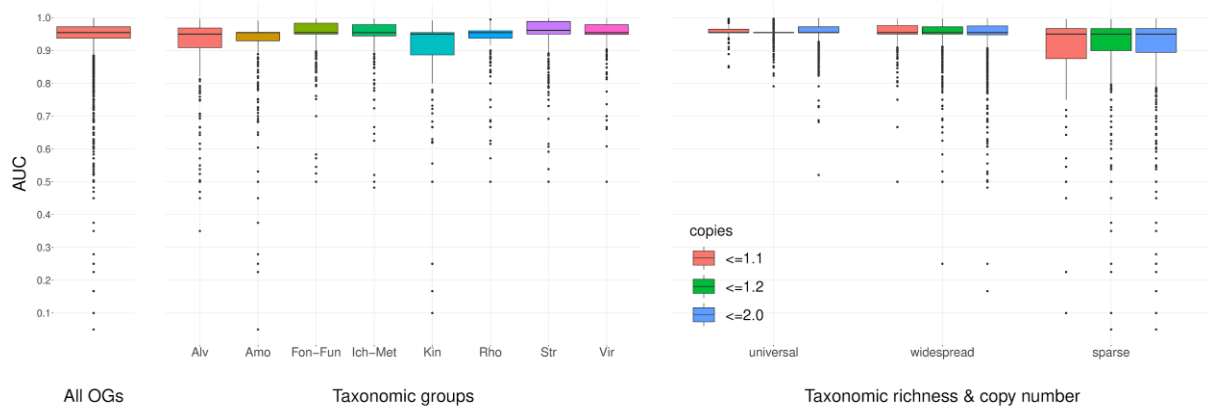


Figure 20: Boxplots showing AUC distributions computed (a) for all OGs, for each OG and for each ‘ref_org_mul’ values according to (b) pruned taxonomic group and (c) OG complexity (i.e., taxonomic richness and average copy number) and for all ‘ref_org_mul’ values.

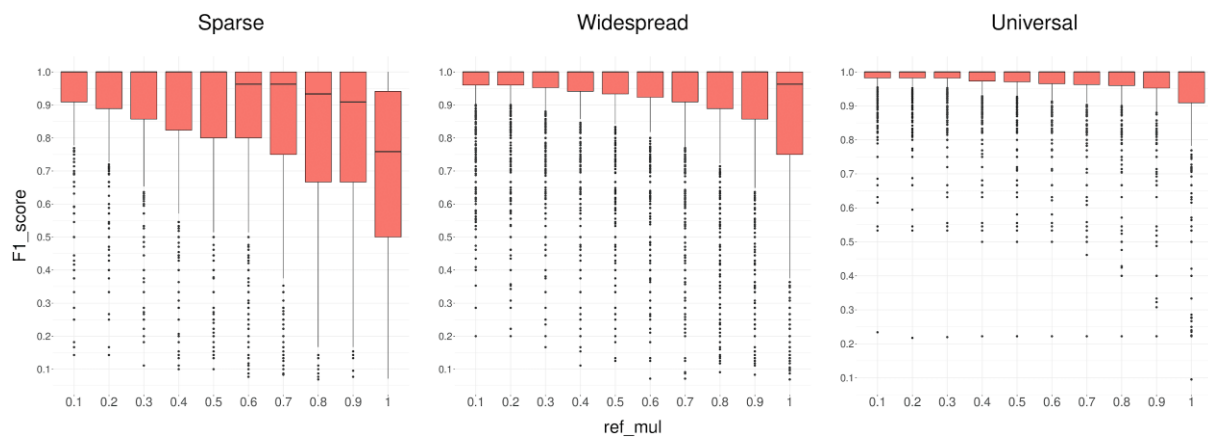


Figure 21: Boxplots showing F1-score distributions computed for the three levels of taxonomic richness (a. Sparse, b. Widespread and c. Universal) according to ‘ref_org_mul’ values.

2.4.1.2. Assessing performance - Unannotated genomic data

To test Forty-Two’s heuristics in a real-world case scenario we did the same OG reconstruction experiments, exactly as described above, except that we used unannotated genomes (“real-world”) instead of complete proteomes (“ideal”) for the mining set and then compared the results for only one ‘ref_org_mul’ value (0.1) to spare computation time and maximize recall, making 3,840 pruning/enrichment experiments ($480 \text{ OGs} \times 8 \text{ taxonomic groups} \times 1 \text{ 'ref_org_mul' value}$).

As shown in Figure 22, it is harder for Forty-Two to find orthologous sequences from genomes. Indeed, compared to previous experiments, this one is more challenging considering that during the enrichment process, an additional splicing step is performed using Exonerate to refine BLAST HSP boundaries, followed by a sorting/combining step for BLAST’s high-scoring segment pairs (HSPs) called ‘trimming’ in order to ensure that the full length protein is considered. This trimming step has also a major role in handling long DNA fragments, such as scaffolds, but this will be further discussed when comparing Forty-Two with another software package, Orthograph. Nonetheless, when considering “universal” OGs, we observe a high reconstruction performance, regardless of mean copy numbers.

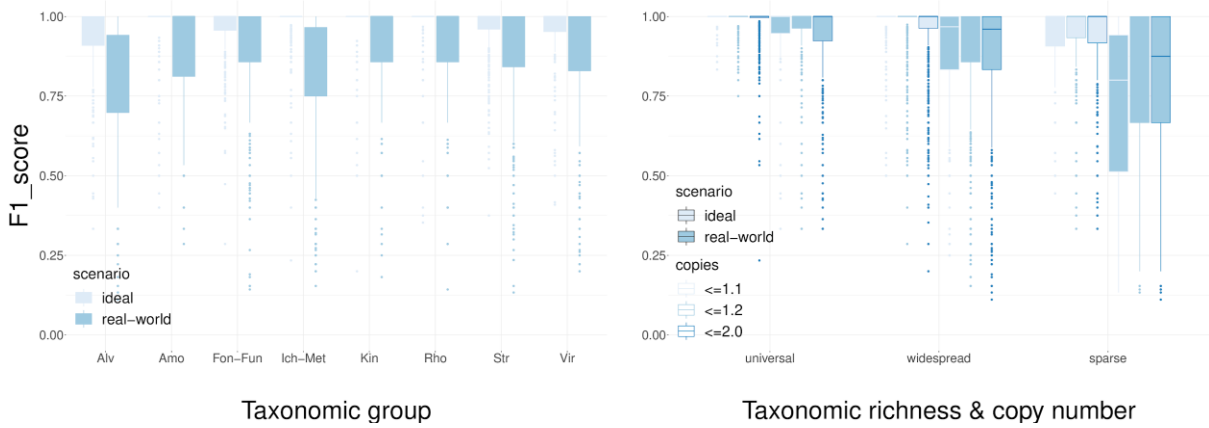


Figure 22: Comparison of OG reconstruction starting from complete proteomes (‘ideal’) and unannotated genomes (‘real-world’). Boxplots showing F1 score distribution computed for each OG according to (a) pruned taxonomic group and (b) OG complexity (i.e., taxonomic richness and average copy number) and for a ‘ref_org_mul’ of 0.1.

2.4.1.3. Comparison versus Orthograph

We chose to compare Forty-Two performance with Orthograph, considered as the upgrade of HaMStR, first by enriching OGs starting from transcriptomic data and then from genomic data. We anticipated that both software packages would be on par when considering transcriptomic data, since algorithms are alike even though Forty-Two does not rely on a profile-based approach. On the contrary, we expected our software to perform better when considering genomic data and long DNA fragments thanks to its “homologue trimming” and (optionally Exonerate-powered) “orthologue splicing” features maximizing the odds of reconstructing full-length proteins. As in the original benchmark comparison between Orthograph/HaMStR and because true negatives (TN) were defined, we were able to use

accuracy and sensitivity metrics instead of precision and recall (F1-score) used for the complete proteome data benchmarking (section 2.4.1.1). First, we used a dataset of 4,625 single-copy (i.e., present in exactly one copy per species) OGs from OrthoDB, built from five species where Forty-Two managed to correctly identify 4611 genes (99.70%) in the proteome of *A. mellifera* and 4604 genes (99.55%) in the proteome of *H. saltator*, which is slightly more than Orthograph with 4582 (99.07%) and 4590 (99.24%), respectively (Table 4). Both software packages showed very high accuracy (0.999 vs 0.997 for *A. mellifera* and 0.998 vs 0.998 for *H. saltator*) of both software packages. Forty-Two found two more false positives than Orthograph for *A. mellifera* and 7 more for *H. saltator*, which is explained by its slightly higher sensitivity (0.996 vs 0.990 for *A. mellifera* and 0.993 vs 0.992 for *H. saltator*).

Software	Test	Genes	Species	OGS	Found	TP	FP	FN	Sensitivity	Accuracy
Forty-Two	Single-copy	4625	<i>A. mellifera</i>	15314	4611	4604	7	21	0.996	0.999
Orthograph	Single-copy	4625	<i>A. mellifera</i>	15314	4582	4577	5	48	0.990	0.997
Forty-Two	Single-copy	4625	<i>H. saltator</i>	18564	4604	4594	10	31	0.993	0.998
Orthograph	Single-copy	4625	<i>H. saltator</i>	18564	4590	4587	3	38	0.992	0.998

Table 4: Performance comparison when using complete proteomes to recover the missing genes in the OGs by measuring sensitivity and accuracy. TP: true positives, FP: false positives, TN: true negatives and FN: false negatives. Sensitivity $TP/(TP+FN)$. Accuracy $(TP+TN)/(TP+FN+TN)$.

Although performances are alike despite the algorithmic specificities of each package, the search for orthologs in unannotated genomes shows higher variation (Table 5), which is expected since Forty-Two was designed to handle long genomic fragments such as scaffolds, in contrast to Orthograph that was exclusively designed to handle transcriptomic data. This gap in performance is mainly due to scaffolds holding more than one gene. On the one hand, Forty-Two found 4442 genes in 243 different scaffolds of *A. mellifera*, with 7 scaffolds holding more than half of those genes. With this fragmented distribution of genes over scaffolds, Orthograph struggles to identify genes (only 1567), which translates into a huge drop in sensitivity (0.313) and accuracy (0.400), compared to Forty-Two (0.981 and 0.927 respectively). On the other hand, Forty-Two found 4355 genes in 666 different scaffolds of *H. saltator*, with 95 scaffolds hosting half of those genes. With the genes more evenly distributed across scaffolds, Orthograph performed a lot better while still being beneath Forty-Two in both sensitivity (0.784 vs 0.964) and accuracy (0.879 vs 0.944).

Software	Test	OGS	Species	Scaffolds	Found	TP	FP	FN	Sensitivity	Accuracy
Forty-Two	Single-copy	4625	<i>A. mellifera</i>	5321	4442	4238	204	84	0.981	0.927
Orthograph	Single-copy	4625	<i>A. mellifera</i>	5321	1567	1430	137	3144	0.313	0.400
Forty-Two	Single-copy	4625	<i>H. saltator</i>	8893	4355	4125	230	154	0.964	0.944
Orthograph	Single-copy	4625	<i>H. saltator</i>	8893	3793	3550	243	980	0.784	0.879

Table 5: Performance comparison when using unannotated genomes to find the missing genes in the OGs, by measuring sensitivity and accuracy. TP: true positives, FP: false positives, TN: true negatives and FN: false negatives. Sensitivity $TP/(TP+FN)$. Accuracy $(TP+TN)/(TP+FN+TN)$.

2.5. Conclusions

In this chapter, we have presented Forty-Two, a bioinformatics tool designed to enrich multiple sequence alignments (MSAs) of orthologous genes (OGs) using a multiple best reciprocal hit (multi-BRH) heuristic. Our pruning/enrichment experiments on complete proteome data and unannotated genomic data have demonstrated that Forty-Two accurately identifies orthologs with high precision and recall, as measured by F1-scores. Furthermore, our comparison with Orthograph showed that while both tools perform similarly on complete proteomes, Forty-Two vastly outperforms Orthograph on unannotated genomic data, achieving up to 98% sensitivity where Orthograph reaches only 31-78%. Combined with its modular architecture and configurable parameters such as `ref_org_mul`, Forty-Two thus represents a significant advance over existing tools. Finally, Forty-Two has spawned two other tools with more specialized purposes. The first is Leel (1331) (Rodríguez et al., 2017), which aims at reconstructing nucleotide MSAs precisely corresponding to protein MSAs from the same (or equivalent) nucleotide datasets as those used to build the original MSAs. It was first used in a phylogenomic study about true salamanders, and it is distributed in the same package as Forty-Two, as a large part of the code base is shard across both programs. The second is MESYRES (Cornet et al., 2021), which is a post-processor that further considers synteny over contigs to ensure orthology when dealing with metagenomic datasets. It was developed in the context of cyanolichens, and it is available on our Bitbucket repository [<https://bitbucket.org/phylogeno/mesyres/>].

3. Contamination detection: principles, implementation and applications

3.1. Introduction

Genomic contamination, defined as the accidental inclusion of sequences from organisms other than the expected one, is a pervasive issue in genomics and phylogenomics (Cornet and Baurain, 2022b). Its sources are multiple: biological contamination of cultures, accidental introduction of foreign DNA during extraction or sequencing on shared platforms, and *in silico* contamination arising from chimeric assemblies or erroneous metagenomic binning (Cornet and Baurain, 2022b). Overlooked contamination can lead to severe analytical artifacts, as illustrated by the erroneous report of extensive horizontal gene transfer in the tardigrade genome (Boothby et al., 2015), which was later shown to be due to undetected bacterial contamination (Koutsovoulos et al., 2016), or by incorrect conclusions in high-profile phylogenomic studies of animal (Philippe et al., 2011) and plant (Laurin-Lemay et al., 2012) evolution.

Over the past decade, numerous contamination detection tools have been developed, based on diverse algorithmic principles (Cornet and Baurain, 2022b). These can be broadly divided into database-free

methods relying on DNA composition (e.g., BlobTools, Anvi'o), gene marker-based estimators that exploit the expected single-copy nature of conserved genes (e.g., CheckM, BUSCO), and genome-wide approaches that classify sequences by comparison against reference databases (e.g., Kraken2, Physeter, GUNC). A systematic benchmarking of six widely used tools on simulated contamination at six taxonomic ranks revealed that no single tool accurately estimates contamination levels across all scenarios and that the use of multiple complementary methods is necessary for reliable detection (Cornet et al., 2024). In this context, Forty-Two occupies a distinctive position as the only gene marker-based tool capable of performing inter-domain contamination detection, while simultaneously fulfilling its primary function of orthologous sequence enrichment (Cornet and Baurain, 2022b).

3.2. Materials and Methods

3.2.1. Contamination detection implementation

Beyond its orthology assessment capabilities (described in Section 2 in this chapter), Forty-Two integrates a contamination detection feature that operates as an optional post-processing step of the standard “phylogenomic” mode aimed at enriching MSAs. The principle is as follows: when a MSA is enriched with sequences from a genome or transcriptome, by default, Forty-Two adds all sequences that pass the multi-BRH orthology test. If the source data are contaminated, both genuine orthologs from the target organism and contaminant sequences (originating from a different organism but nonetheless truly orthologous) will be incorporated. When enabled, the contamination detection module determines, for each newly added sequence, which existing sequences in the MSA are the most closely related. If a putative ortholog turns out to be closest to a taxonomic group distinct from the expected source taxon, it is flagged as a likely contaminant. Two main approaches are available for this taxonomic verification, implemented as exclusive options.

The first approach, termed “positive filtering”, relies on taxonomic filters (`tax_filter`) that are activated after the orthology assessment step. Forty-Two performs a BLAST search of each putative ortholog against all sequences in the targeted MSA and computes its taxonomic affiliation from the closest sequence(s). Two predefined modes are available for this computation: a “best-hit” mode, in which only the single most closely related sequence is used, and a “megan-like” mode, in which multiple close sequences are considered, and a Lowest Common Ancestor (LCA) is inferred from their taxonomic affiliations. The LCA algorithm takes inspiration from the approach implemented in the MEGAN software (Huson et al., 2016) and relies on two user-configurable parameters: a minimum bit-score threshold (default: 80) to consider the first hit and a percentage threshold (default: 95%) of the first hit's bit-score to accumulate sub-best hits used for LCA computation. If the inferred taxonomy satisfies the taxonomic filter, the ortholog is added to the MSA; otherwise, it is tagged as a contaminant ('c#') and

the name of the organism at the origin of the taxonomic affiliation is appended to its identifier. The "megan-like" mode often yields more reliable taxonomic affiliations at the cost of reduced precision (e.g., affiliation at the genus or family level instead of species level).

The second approach uses decoy proteomes sampled across the diversity of the Tree of Life (TOL). Under this scheme, an ortholog must obtain a higher BLAST bit-score when compared to the sequences present in the MSA than to those of the decoy TOL proteomes. To avoid rejecting genuinely uncontaminated orthologs that happen to be closely related to a decoy organism, Forty-Two skips all decoy hits that satisfy the user-defined taxonomic filter for the organism being added.

3.2.2. Metagenomic mode and debriefing

Forty-Two can also be used as a stand-alone contamination detection tool, without performing any MSA enrichment, through its "metagenomic" mode. In this mode, Forty-Two recovers ribosomal protein orthologs from a genome or transcriptome and labels each one by computing the LCA of their closest relatives in curated reference MSAs, using the "megan-like" algorithm described above. Two sets of reference MSAs are available for this purpose: a set of 78 eukaryotic ribosomal protein MSAs, manually curated and continuously enriched in H. Philippe's laboratory, and a set of 90 prokaryotic ribosomal protein MSAs fetched from RiboDB (Jauffrit et al., 2016). An inter-domain dataset combining both sets has also been assembled. Ribosomal proteins are among the most highly expressed genes in any transcriptome (Philippe and Telford, 2006), and the same applies to potential contaminant organisms, making them ideal references for reliable taxonomic affiliation. For the metagenomic mode, a more stringent parameterization of the "megan-like" algorithm is recommended, with a `tax_score_mul` of 0.99 instead of the default 0.95 and a `tax_min_ident` of 50 instead of 0, to reduce false positives during LCA computation.

The follow-up analysis is performed by `debrief-42.pl`, a companion script that parses the taxonomic reports produced by Forty-Two. For each identified ortholog, `debrief-42.pl` compares the taxonomic label (LCA) computed by Forty-Two with the source organism lineage (according to NCBI Taxonomy) and classifies the sequence as a contaminant if the two differ at a predefined taxonomic rank, based on a user-defined list of taxa. Contaminations are further classified to determine the main sources of contaminants, based on a second user-defined list of taxa. In this regard, two types of contaminating sequences are distinguished: classified contaminations, for which the taxonomic label is informative enough to point to a specific lineage, and unclassified contaminations, which bear an uninformative label too broad to indicate a precise origin (e.g., `Sar`). Sequences that can only be affiliated at the highest taxonomic levels (e.g., "cellular organisms", "Eukaryota", "Bacteria", "Archaea") are classified as unknown sequences. This framework allows `debrief-42.pl` to compute several quality metrics for each

genome or transcriptome, including completeness (proportion of expected ribosomal proteins recovered), purity (proportion of sequences affiliated to the expected taxon), and contamination level (proportion of sequences affiliated to a foreign taxon), as well as the identification of contamination sources.

3.3. Results

3.3.1. Consensus assessment of cyanobacterial genome contamination

The contamination detection capabilities of Forty-Two were first deployed at large scale in a study assessing the contamination level of 440 publicly available cyanobacterial genome assemblies (Cornet et al., 2018). In this study, Forty-Two was used to mine the genome assemblies against curated ribosomal protein alignments (from RiboDB), controlling for orthology relationships. This analysis revealed that 21 assemblies contained at least one ribosomal protein gene of foreign (i.e., non-cyanobacterial) origin: 8 assemblies harboured between 1 and 8 foreign proteins, 9 assemblies between 16 and 41, and 4 assemblies between 56 and 80 (Cornet et al., 2018). Almost all assemblies featuring at least one foreign SSU rRNA (16S) gene also showed at least one foreign ribosomal protein gene (14 out of 16 assemblies), confirming the concordance between the two ribosomal gene methods. The ribosomal protein-based analysis by Forty-Two was then used, together with the SSU rRNA (16S) analysis, as a gold standard to calibrate the three genome-wide methods (Kraken, DIAMOND blastx, CONCOCT) also employed in the study. Notably, four assemblies that were flagged by Forty-Two as contaminated at more than 10% in terms of ribosomal proteins were only tagged as lowly contaminated by CheckM (3.41%, 1.96%, 0.88% and 0.84%, respectively), highlighting the complementarity of the two approaches (Cornet et al., 2018). Overall, the consensus of all six methods indicated that more than 5% of publicly available cyanobacterial genome assemblies were highly contaminated by foreign DNA.

3.3.2. Decontamination of transcriptomic data in metazoan phylogenomics

Forty-Two also played a central role in the construction of a large-scale metazoan phylogenomic dataset aimed at resolving the debated branching order among early-diverging animal lineages (Simion et al., 2017). In this study, transcriptomic data from 126 non-bilaterian species were incorporated into 4,002 core orthologous clusters using Forty-Two. Prior to integration, transcriptomes were screened for cross-contamination between datasets obtained in the same laboratory or sequencing project (with Sobek), revealing contamination levels ranging from 0.16% to 70.64% of the reads depending on the species. A three-step decontamination procedure was then applied at different taxonomic scales: removal of non-holozoan contaminants (DC1), of bilaterian contaminants (DC2), and of reciprocal contamination between sponge and cnidarian transcriptomes (DC3). After integration by Forty-Two, an additional manual BLAST-based filtering step removed 4,885 sequences whose best hit belonged to a clade other

than their expected source, further eliminating residual contaminants and recent paralogs. The resulting dataset of 1,719 genes for 90 species constituted the largest and most internally consistent metazoan supermatrix assembled at the time, and provided strong support for sponges as the sister group to all other animals under the site-heterogeneous CAT model (Simion et al., 2017). The whole pipeline described here, combining Forty-Two and a manual BLAST-based filtering step, has now been fully automated in Forty-Two itself (i.e., the second approach described in 3.2.1 using decoy proteomes).

3.3.3. Benchmarking on simulated chimeric genomes

In a review of contamination detection methods in genomic data, six widely used tools were benchmarked on controlled simulated data (Cornet and Baurain, 2022b). To this end, two chimeric genomes were constructed: a bacterial chimera combining a Firmicutes master genome (GCF_000003645.1) with Gammaproteobacteria CDS (GCF_000007145.1) at a simulated contamination level of 5% redundant and 5% non-redundant (based on single-copy core gene exchanges), resulting in 88.1% Firmicutes and 11.9% Gammaproteobacteria; and an inter-domain chimera concatenating this bacterial chimera with an uncontaminated *Ustilago maydis* genome, yielding 80.1% Opisthokonta, 16.1% Firmicutes and 3.8% Gammaproteobacteria (Cornet and Baurain, 2022b). The following six tools were benchmarked on these chimeras: CheckM (Parks et al., 2015), EukCC (Saary et al., 2020), Forty-Two, GUNC (Orakov et al., 2021), Physeter (Lupo et al., 2021) and Kraken2 (Wood et al., 2019).

Using the inter-domain ribosomal protein dataset (98 MSAs) in metagenomic mode, Forty-Two correctly identified both reference bacterial genomes as clean (0 contaminant sequences) and detected massive contamination in the bacterial chimera (40 contaminated alignments out of 48 recovered). On the inter-domain chimera, Forty-Two recovered 84 ribosomal protein alignments, of which 48 were flagged as contaminated (87 contaminant sequences in total), and correctly identified the contamination sources as Firmicutes (*Bacillus*) and Gammaproteobacteria (*Xanthomonas*), with mitochondrial ribosomal proteins appearing as Alphaproteobacteria. In comparison, CheckM (Parks et al., 2015) detected only 1.65% redundant contamination in the bacterial chimera and misclassified the eukaryotic genome as bacterial, while GUNC (Orakov et al., 2021) accurately estimated 10% contamination on the bacterial chimera but was unsuitable for the eukaryotic genome. This benchmark confirmed the unique inter-domain detection capability of Forty-Two and highlighted the complementarity of gene marker-based and genome-wide approaches (Cornet and Baurain, 2022b).

3.3.4. Integration in genome quality assessment pipelines

The contamination detection capabilities of Forty-Two has also been integrated into ToRQuEMaDA (TQMD), a tool for retrieving, storing and dereplicating prokaryotic genomes on high-performance computing clusters (Léonard et al., 2021). TQMD uses Forty-Two in its metagenomic mode to assess both the completeness and the contamination level of each genome based on ribosomal protein recovery. These metrics are used alongside other quality criteria (assembly statistics, annotation richness) to rank genomes during the dereplication process, ensuring that representative genome sets are composed of high-quality, minimally contaminated assemblies (Léonard et al., 2021). This integration demonstrates the utility of Forty-Two's contamination detection capabilities beyond phylogenomics, as a general-purpose genome quality estimator applicable to large-scale comparative genomics projects.

3.4. Conclusions

Taken together, these applications illustrate the versatility of Forty-Two's contamination detection framework. As a gene marker-based tool relying on ribosomal proteins, Forty-Two occupies a distinctive niche among contamination detection methods: it does not aim at exhaustive genome-wide classification but instead leverages the universal presence and high expression of ribosomal proteins to provide a rapid and reliable diagnostic of contamination levels and sources. Its unique inter-domain detection capability, confirmed by benchmarking on simulated chimeric genomes (Cornet and Baurain, 2022b), sets it apart from tools such as CheckM (Parks et al., 2015) or GUNC (Orakov et al., 2021) that are restricted to a single domain of life. The LCA-based taxonomic affiliation algorithm, originally implemented in Forty-Two and later adopted by Physeter for genome-wide analyses (Lupo et al., 2021), has proven robust across a wide range of biological contexts, from large-scale surveys of public genome databases to the stringent decontamination of phylogenomic datasets (Simion et al., 2017). The integration of Forty-Two into automated pipelines such as TQMD (Léonard et al., 2021) further confirms its suitability as a lightweight yet informative quality estimator. These results underscore the recommendation that no single tool is sufficient for reliable contamination detection, and that a consensus approach combining complementary methods remains necessary, as formally demonstrated by a systematic benchmarking of six tools across six taxonomic ranks on simulated contamination data, further mimicking horizontal gene transfer (Cornet et al., 2024).

4. Transition: the backbone

Forty-Two will be used extensively in the remainder of this thesis and literally is the backbone of the work presented in this thesis. Forty-Two's ability to identify the closest taxonomic relatives of newly added sequences within a MSA provides a straightforward mechanism for detecting contamination, an extremely important step when studying horizontal gene transfers, and will be thoroughly used in Chapter III, IV and V. Additionally, Forty-two's ability to perform ortholog enrichment will be use in Chapter IV and V in a crucial step of enriching alignments with newly sequenced transcriptomic data, thereby expanding taxonomic sampling for our phylogenomic analyses.

III. Chapter 3 - Broadly sampled orthologous groups of eukaryotic proteins for the phylogenetic study of plastid-bearing lineages

Mick Van Vlierberghe (mvanvlierberghe@doct.uliege.be)¹, Hervé Philippe (herve.philippe@sete.cnrs.fr)^{2,3} and Denis Baurain (denis.baurain@uliege.be)^{1,*}

¹InBioS - PhytoSYSTEMS, Eukaryotic Phylogenomics, University of Liège, Liège, Belgium

²Station d'Ecologie Théorique et Expérimentale de Moulis, UMR CNRS 5321, Moulis, France

³Département de Biochimie, Centre Robert-Cedergren, Université de Montréal, Montréal, Québec, Canada

* corresponding author

This chapter has been published in *BMC Research Notes*; as a consequence, it has its own “References” section based on a numbered list.

1. Abstract

Objectives:

Identifying orthology relationships among sequences is essential to understand evolution, diversity of life and ancestry among organisms. To build alignments of orthologous sequences, phylogenomic pipelines often start with all-vs-all similarity searches, followed by a clustering step. For the protein clusters (orthogroups) to be as accurate as possible, proteomes of good quality are needed. Here, our objective is to assemble a data set especially suited for the phylogenomic study of algae and formerly photosynthetic eukaryotes, which implies the proper integration of organellar data, to enable distinguishing between several copies of one gene (paralogs), taking into account their cellular compartment, if necessary.

Data description:

We submitted 73 top-quality and taxonomically diverse proteomes to OrthoFinder. We obtained 47,266 orthogroups and identified 11,775 orthogroups with at least two algae. Whenever possible, sequences were functionally annotated with eggNOG and tagged after their genomic and target compartment(s). Then we aligned and computed phylogenetic trees for the orthogroups with IQ-TREE. Finally, these trees were further processed by identifying and pruning the subtrees exclusively composed of plastid-bearing organisms to yield a set of 39,784 clans suitable for studying photosynthetic organism genome evolution.

Keywords:

Orthology, Phylogenomics, Algae, CASH, Proteomes, Eukaryotic evolution, Contamination, Organelles, Endosymbiotic gene transfer (EGT), Horizontal or lateral gene transfer (HGT/LGT), Kleptoplasty.

2. Objective

Our main objective is to analyse the phylogenetic origin of plastid-targeted genes in complex algae [1, 2, 3] in a fully automated fashion. To do so, we designed and developed a series of strategies and tools around a large-scale single-gene tree analysis pipeline. The first step was to build alignments of orthologous sequences with OrthoFinder, a high accuracy orthogroup inference algorithm [4]. We focused on top-quality proteomes, especially with high completeness, which is essential to obtain the most complete and balanced OGs possible [5,6]. In order to maximize completeness and to facilitate the phylogenetic analysis, we complemented beforehand the proteomes having no or only incomplete

plastid and/or nucleomorph sequences. Then we processed the resulting OGs, first by isolating the OGs containing photosynthetic organisms, and second by sorting out gene copies shared by plastid-bearing algae from their paralogs. To this end, we built trees using IQ-TREE [7] and used our own tool (tree-clan-splitter.pl) to detect and prune the subtree(s) of interest.

3. Data description

We collected 73 top-quality eukaryotic proteomes (i.e., conceptually translated genomes; [Data file 1](#), [Data set 1](#), [Data set 2](#)) with high completeness (assessed by using BUSCO [5,6]) ([Data file 2](#)) [5,6] and low contamination levels (assessed by using Forty-Two) ([Data set 3](#)) [8,9]. Those were selected to be taxonomically diverse, covering all photosynthetic phyla [10,11], along with some non-photosynthetic organisms to be used as beacons by our clan-identifying algorithm (see [Supplementary file 1](#) for detailed Methods). Those proteomes were complemented with organellar (i.e., plastid and nucleomorph) proteins if they were partly or fully missing in the original source. Hence, 16 were complemented with plastid proteomes whereas two were complemented with nucleomorph proteomes. All proteomes (complemented or not) were dereplicated with CD-HIT [12]. In addition, we used tag-loc-ids.pl, a custom tool designed to tag sequences identifiers (see [Supplementary file 1](#)) according to their encoding genome and cellular localization, such as nuclear-encoded-and-plastid-targeted (nucpt#), nuclear-encoded-periplastid-compartment-targeted (nuppct#), plastid-encoded-plastid-targeted (cpcpt#), nucleomorph-encoded (nm#), and mitochondrion-encoded (mt#), to facilitate subsequent phylogenetic analyses. Then, we used OrthoFinder [4] for orthology inference, which resulted in 47,266 OGs ([Data file 3](#), [Data set 4](#)), composed of two or more sequences belonging to eleven main taxonomic groups (according to NCBI Taxonomy [13]), either classified as “primary algae” (Glaucocystophyceae, Rhodophyta, Viridiplantae) or “complex algae” (Apicomplexa, Colpodellida, Dinophyceae, Cryptophyceae, Euglenozoa, Ochrophyta (including Pelagophyceae), Haptophyta, and Chlorarachniophyceae). Hence, OGs were tabulated into three different categories: “two-algae” (at least one complex alga from two different groups or at least one complex alga and one primary alga, n=11,775), “one-alga” (at least one alga, n=18,844) and “zero-algae” (no algae, n=16,647) using the script classify-mcl-out.pl (see [Supplementary file 1](#)). In order to address the issue of multiple-copy genes (paralogs), we developed a strategy to isolate subtrees (“clans”) of interest, i.e., including only plastid-bearing organisms. Briefly, we computed trees for the 11,775 “two-algae” OGs when possible (i.e., ≥ 3 sequences, n=11,499) with IQ-TREE [7] and developed a tool for identifying and pruning subtrees fulfilling user-specified taxonomic filters (tree-clan-splitter.pl, see [Supplementary file 1](#)). This way, we obtained 31,784 “photosynthetic” clans ([Data set 5](#)) only composed of plastid-bearing organisms (including species with a non-photosynthetic plastid, such as *Plasmodium falciparum*). Additionally, we provide detailed annotation reports obtained with eggNOG [14].

Table 1: Overview of data files/data sets.

Label	Name of data file/data set	File types (file extension)	Data repository and identifier (DOI or accession number)
Supplementary file 1	Methods	PDF file (.pdf)	Figshare https://doi.org/10.6084/m9.figshare.13604102.v3 [18]
Data file 1	Taxonomic sampling	Image file (.png)	Figshare https://doi.org/10.6084/m9.figshare.13603511.v1 [19]
Data set 1	Proteome set description	Text files (.csv, .html)	Figshare https://doi.org/10.6084/m9.figshare.13113893.v1 [20]
Data set 2	Proteome files	FASTA files (.tar.gz)	Figshare https://doi.org/10.6084/m9.figshare.13573424.v2 [21]
Data file 2	BUSCO report	Text file (.csv)	Figshare https://doi.org/10.6084/m9.figshare.13235045.v1 [22]
Data set 3	Forty-Two reports and configuration files	Text files (.tsv, .csv, .yaml)	Figshare https://doi.org/10.6084/m9.figshare.13235063.v3 [23]
Data file 3	Orthogroup properties	Image file (.pdf)	Figshare https://doi.org/10.6084/m9.figshare.13312622.v1 [24]
Data set 4	Orthogroups	FASTA files, YAML configuration file (.tar.gz)	Figshare https://doi.org/10.6084/m9.figshare.13573658.v3 [25]
Data set 5	Clans	FASTA files (.tar.gz)	Figshare https://doi.org/10.6084/m9.figshare.13573415.v1 [26]
Data file 4	Organelle database	Text file (.tsv)	Figshare https://doi.org/10.6084/m9.figshare.13246841.v1 [27]
Data file 5	Plastid-targeted proteins	Spreadsheet (.xlsx)	Figshare https://doi.org/10.6084/m9.figshare.13246784.v1 [28]
Data file 6	eggNOG OG annotations	Text file (.tsv)	Figshare https://doi.org/10.6084/m9.figshare.13415048.v1 [29]
Data file 7	eggNOG clan annotations	Text file (.tsv)	Figshare https://doi.org/10.6084/m9.figshare.13415060.v1 [30]

Limitations

- Occasionally, organellar genome sequences are from a different strain than the nucleus data; it could be an issue if we were trying to resolve relationships between close relatives of the same lineage. Nonetheless, it is not the case here, since the major endosymbiotic-like events we are tracking occurred most certainly between distinct lineages.
- The way we handle the tagging overwrites the information about potential NUMTs, NUNMs and NUPTs; this means that if a gene existed in both genomic compartments (nucleus and organelle) we always retained the organellar counterpart.
- Only a few of the nucleus-encoded-and-plastid-targeted proteins (nucpt#) were identified by proteomics (e.g., in *P. falciparum*) [17]; the remaining are the results of *in silico* predictions [15,16], which are less reliable than proteomic experiments.

Abbreviations

- OGs: orthologous groups or orthogroups
- NUMTs: nuclear mitochondrial DNAs
- NUNMs: nucleomorph-derived DNAs
- NUPTs: nuclear plastid DNAs

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

All data generated or analysed during this study are publicly available in the figshare repository (<https://doi.org/10.6084/m9.figshare.13604102.v3>, <https://doi.org/10.6084/m9.figshare.13603511.v1>, <https://doi.org/10.6084/m9.figshare.13113893.v1>, <https://doi.org/10.6084/m9.figshare.13573424.v2>, <https://doi.org/10.6084/m9.figshare.13235045.v1>, <https://doi.org/10.6084/m9.figshare.13235063.v3>, <https://doi.org/10.6084/m9.figshare.13312622.v1>, <https://doi.org/10.6084/m9.figshare.13573658.v3>, <https://doi.org/10.6084/m9.figshare.13573415.v1>, <https://doi.org/10.6084/m9.figshare.13246841.v1>, <https://doi.org/10.6084/m9.figshare.13246784.v1>, <https://doi.org/10.6084/m9.figshare.13415048.v1>, <https://doi.org/10.6084/m9.figshare.13415060.v1>). Please see [Table 1](#) and reference list [18 - 30] for details and links to the data.

Competing interests

The authors declare that they have no competing interests.

Funding

Mick Van Vlierberghe was a FRiA fellow of the FRS-FNRS (National Fund for Scientific Research of Belgium). Computational resources were provided through two grants to Denis Baurain (University of Liège “Crédit de démarrage 2012” SFRD-12/04; FRS-FNRS “Crédit de recherche 2014” CDR J.0080.15).

Authors’ contributions

Mick Van Vlierberghe and Denis Baurain designed the experiments, Mick Van Vlierberghe performed all the computational analyses and drew the figures, Mick Van Vlierberghe and Denis Baurain wrote the manuscript. Hervé Philippe substantively revised the work.

Acknowledgements

We are grateful to Ugo Cenci for sharing the protein sequences of *Goniomonas avonlea* (Cenci U, Sibbald SJ, Curtis BA, Kamikawa R, Eme L, Moog D, et al. Nuclear genome sequence of the plastid-lacking cryptomonad *Goniomonas avonlea* provides insights into the evolution of secondary plastids. doi:10.1186/s12915-018-0593-5.)

4. References

1. Petersen J, Teich R, Brinkmann H, Cerff R. A “green” phosphoribulokinase in complex algae with red plastids: evidence for a single secondary endosymbiosis leading to haptophytes, cryptophytes, heterokonts, and dinoflagellates. *J Mol Evol.* 2006;62:143-57.
2. Teich R, Zauner S, Baurain D, Brinkmann H, Petersen J. Origin and distribution of Calvin cycle fructose and sedoheptulose biphosphatases in plantae and complex algae: a single secondary origin of complex red plastids and subsequent propagation via tertiary endosymbioses. *Protist.* 2007;158:263-76. doi:10.1016/j.protis.2006.12.004.
3. Sibbald SJ, Archibald JM. Genomic Insights into Plastid Evolution. *Genome Biol Evol.* 2020;12:978-90.
4. Emms DM, Kelly S. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol.* 2015;16:157. doi:10.1186/s13059-015-0721-2.
5. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva E V., Zdobnov EM. BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics.* 2015;31:3210-2.
6. Waterhouse RM, Seppey M, Simao FA, Manni M, Ioannidis P, Klioutchnikov G, et al. BUSCO applications from quality assessments to gene prediction and phylogenomics. *Mol Biol Evol.* 2018;35:543-8.
7. Nguyen LT, Schmidt HA, Von Haeseler A, Minh BQ. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol.* 2015;32:268-74.
8. Simion P, Philippe H, Baurain D, Jager M, Richter DJ, Di Franco A, et al. A Large and Consistent Phylogenomic Dataset Supports Sponges as the Sister Group to All Other Animals. *Curr Biol.* 2017;27:958-67. doi:10.1016/J.CUB.2017.02.031.
9. Irisarri I, Baurain D, Brinkmann H, Delsuc F, Sire JY, Kupfer A, et al. Phylotranscriptomic consolidation of the jawed vertebrate timetree. *Nat Ecol Evol.* 2017;1:1370-8.
10. Blaby-Haas CE, Merchant SS. Comparative and Functional Algal Genomics. *Annu Rev Plant Biol.* 2019;70:605-38. doi:10.1146/annurev-arplant-050718-095841.
11. Hanschen ER, Starkenburg SR. The state of algal genome quality and diversity. *Algal Res.* 2020;50:101968. doi:10.1016/j.algal.2020.101968.
12. Li, W. and A. Godzik (2006). "Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences." *Bioinformatics* 22(13): 1658-1659.
13. Schoch CL, Ciufo S, Domrachev M, Hottel CL, Kannan S, Khovanskaya R, et al. NCBI Taxonomy: a comprehensive update on curation, resources and tools. *Database (Oxford).* 2020;2020:1-21.

14. Huerta-Cepas J, Forslund K, Coelho LP, Szklarczyk D, Jensen LJ, Von Mering C, et al. Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper. *Mol Biol Evol.* 2017;34:2115-22.
15. Dorrell RG, Gile G, McCallum G, Méheust R, Baptiste EP, Klinger CM, et al. Chimeric origins of ochrophytes and haptophytes revealed through an ancient plastid proteome. *Elife.* 2017;6:1-45.
16. Novák AMG, Orcid V, Füßy Z, Ebenezer TE, Dobáková EL, Eliáš M. Metabolic quirks and the colourful history of the *Euglena gracilis* secondary plastid. 2019;44:0-2.
17. Boucher MJ, Ghosh S, Zhang L, Lal A, Jang SW, Ju A, et al. Integrative proteomics and bioinformatic prediction enable a high-confidence apicoplast proteome in malaria parasites. *PLOS Biol.* 2018;16:e2005895. doi:10.1371/journal.pbio.2005895.
18. Van Vlierberghe M, Philippe H, Baurain D. Supplementary file 1 - Methods. Figshare 2021. <https://doi.org/10.6084/m9.figshare.13604102.v3>.
19. Van Vlierberghe M, Philippe H, Baurain D. Data file 1 - Taxonomic sampling. Figshare 2021. <https://doi.org/10.6084/m9.figshare.13603511.v1>.
20. Van Vlierberghe M, Philippe H, Baurain D. Data set 1 - Proteome set description. Figshare 2021. <https://doi.org/10.6084/m9.figshare.13113893.v1>.
21. Van Vlierberghe M, Philippe H, Baurain D. Data set 2 - Proteome set. Figshare 2021. <https://doi.org/10.6084/m9.figshare.13573424.v2>.
22. Van Vlierberghe M, Philippe H, Baurain D. Data file 2 - BUSCO report. Figshare 2021. <https://doi.org/10.6084/m9.figshare.13235045.v1>.
23. Van Vlierberghe M, Philippe H, Baurain D. Data set 3 - Forty-two reports and configuration files. Figshare 2021. <https://doi.org/10.6084/m9.figshare.13235063.v3>.
24. Van Vlierberghe M, Philippe H, Baurain D. Data file 3 - Orthogroup properties. Figshare 2021. <https://doi.org/10.6084/m9.figshare.13312622.v1>.
25. Van Vlierberghe M, Philippe H, Baurain D. Data set 4 - Orthogroups. Figshare 2021. <https://doi.org/10.6084/m9.figshare.13573658.v3>.
26. Van Vlierberghe M, Philippe H, Baurain D. Data set 5 - Clans. Figshare 2021. <https://doi.org/10.6084/m9.figshare.13573415.v1>.
27. Van Vlierberghe M, Philippe H, Baurain D. Data file 4 - Organelle database. Figshare 2021. <https://doi.org/10.6084/m9.figshare.13246841.v1>.
28. Van Vlierberghe M, Philippe H, Baurain D. Data file 5 - Plastid-targeted proteins. Figshare 2021. <https://doi.org/10.6084/m9.figshare.13246784.v1>.
29. Van Vlierberghe M, Philippe H, Baurain D. Data file 6 - eggNOG OG annotations. Figshare 2021. <https://doi.org/10.6084/m9.figshare.13415048.v1>.
30. Van Vlierberghe M, Philippe H, Baurain D. Data file 7 - eggNOG clan annotations. Figshare 2021. <https://doi.org/10.6084/m9.figshare.13415060.v1>.

5. Supplementary material

5.1. Supplementary file 1 - Methods

5.1.1. Proteome pre-processing

All 73 proteomes were downloaded as FASTA files (Data set 2) and renamed after their scientific binomial name and associated NCBI Taxonomy [13] taxon identifier, in the form of `genus_species_taxid.fasta` thanks to our taxonomic tools (Bio::MUST::Core suite, available at <https://metacpan.org/release/Bio-MUST-Core>). Similarly, sequence identifiers were reformatted in the form of `>taxid|protid` to ensure universal software compatibility (e.g., BLAST, CD-HIT). To carry out this step, we used **abbr-ids-fas.pl**, a script able to craft controlled identifiers out of (possibly unstructured) complex sequence identifiers based on a regular expression argument (`--id-regex`), which can be a predefined or a custom regex. In addition, we used the `--prefix` argument to prepend the organism taxon id to the sequence identifiers, followed by a “|”. The same procedure was applied to all the following assembled databases. New script (in bold) availability details are provided in Table S1.

5.1.2. Organelle complementation

We built a plastid-encoded and nucleomorph-encoded protein database (Data file 4) in order to screen the 73 proteomes with BLAST (`blastp -max_target_seqs 1 -outfmt 6`). When those proteins were missing in part or totally, we chose to complement proteomes with the adequate organellar proteome by concatenating both files (for details see Data set 1). Indeed, being able to recognize those proteins and where they are encoded is essential for the envisioned applications because they provide valuable insight for phylogenetic interpretations.

5.1.3. Proteome dereplication

To handle protein redundancy, especially after completing nuclear data with organellar data, we used CD-HIT (v4.6) [12] to dereplicate both complemented and uncomplemented proteomes with the accurate but slow mode (`-g 1`) and a sequence identity threshold of 99% (`-c 0.99`).

5.1.4. Sequence naming

Sequence identifiers were changed from their compact (`abbr`) form (see above) to a full length (`long`) form with full taxonomic information called “`must_id`” (`>genus_species_taxid@protid`). This step was performed using **change-ids-ali.pl**, which switches identifiers (from “`taxid`” to “`genus_species_taxid@`”) based on an organism mapper file (the first column is a “`must`” `base_id`, i.e., a “`must_id`” without the “`protid`”, and the second is the corresponding current taxon id).

5.1.5. Localization tagging

We used BLAST (blastp -max_target_seqs 1 -outfmt 6) to identify plastid and nucleomorph-encoded proteins within proteomes by querying a plastid and a nucleomorph database (Data file 4). In addition, we used the deeply rich HPPG (Homologous Plastid Protein Group) database [15], a database of proteins possessing plastid-targeting sequences that are homologous to one another. We further completed this database with *Euglena gracilis* [16] and *Plasmodium falciparum* [17] nuclear-encoded-and-plastid-targeted proteins (Data file 5). Again we queried the database to identify the nuclear-encoded-and-plastid-targeted sequences in our data. Then, we developed **tag-loc-ids.pl**, a tool to parse BLAST reports (-outfmt 6) and generate a sequence id mapper (.idm) file to prefix the sequence identifiers with the appropriate tag, when a protein has been accurately identified (99% identity and identical length). There are five different tags, each characterizing a distinct combination of encoding genome and targeted compartment (if known): cpcpt# (plastid-encoded-plastid-targeted), nucpt# (nucleus-encoded-plastid-targeted), nuppt# (nucleus-encoded-periplastidial-compartment-targeted), nm# (nucleomorph-encoded), and mt# (mitochondrion-encoded). The renaming step takes place using **rest-ids-ali.pl**, which also uses an .idm file (like change-ids-ali.pl) but, contrary to the latter, performs a full sequence identifier switch (including the “protid”).

5.1.6. Completeness analysis - BUSCO

Proteome quality was assessed with BUSCO v3.0.1 [5,6] mainly for completeness, using 303 single-copy orthologous eukaryotic markers (evalue = 1e-03, lineage data = “eukaryota_odb9”, mode = protein).

5.1.7. Orthology inference

We used OrthoFinder v2.3.3 with default parameters to perform the orthology inference on the 73 proteomes we collected and processed (Data file 1, Data set 1 and 2).

5.1.8. Orthogroup classification

We used **classify-mcl-out.pl** to bin OGs (Data set 4) into different categories based on a YAML configuration file where all categories are described. The configuration file defines different categories. The order of definition is relevant. Hence, if an OG matches more than one category, it is classified according to the first category defined. Each category has a label that is used to create the corresponding subdirectory for writing OG files. A category is characterized by one or more criteria. To match a category, an OG must satisfy all these criteria. Criteria are thus linked by logical AND operators (and their order of definition is irrelevant). Each criterion has a tax_filter describing its taxonomic requirements. Wanted taxa are to be prefixed by a “+” symbol, whereas unwanted taxa are to be prefixed

by a “-” symbol. Wanted and unwanted taxa are linked by logical ORs. Criteria may also have a `min_seq_count` argument and/or a `max_seq_count`. These respectively specify the minimum and maximum number of sequences that must pass the `tax_filter` for the OG to match the criterion. Minimum defaults to 1, while there is no upper (maximum) bound by default. Other conditions are available: `min_org_count` and `max_org_count` correspond to distinct organisms instead of sequences, whereas `min_copy_mean` and `max_copy_mean` allow bounding the mean number of gene copies per organism. All of these default to no bound.

5.1.9. Photosynthetic clade detection and pruning

OGs of interest (“two-algae”, $n=11,499$) were written to files with **extract-mcl-out.pl** and aligned with MAFFT (`mafft-linsi`), then the corresponding trees were computed with IQ-TREE (`-m LG4X`). When possible (i.e., ≥ 4 sequences, $n=11,076$) bootstraps were computed (`-bb 1000`) otherwise they were not as only one unrooted tree is possible (i.e., $3 \geq$ sequences, $n=423$). We used **tree-clan-splitter.pl** to identify photosynthetic clans based on an .otu (Operational Taxonomic Unit) configuration file where the user defines the target clan(s) by describing its/their taxonomic requirements. Wanted taxa are to be prefixed by a “+” symbol (e.g., target: +Viridiplantae, +Rhodophyta). Then, each species is labeled thanks to a local copy of the NCBI Taxonomy [13] and the largest clans containing only target species are identified and written to a file such as OGXXXXXXXX-N.fasta with XXXXXXXX being the OG number and “N” the n^{th} target clan identified. Following the same naming pattern, for each target clan, its complement is written to a file (all other sequences in the OG, i.e., potential paralogs) such as OGXXXXXXXX-N.para. Such .para files are useful to enforce orthology when enriching multiple sequence alignments corresponding to subfamilies of larger protein families using our dedicated software package Forty-Two [9] available at <https://metacpan.org/release/Bio-MUST-Apps-FortyTwo>.

5.1.10. Annotation with eggNOG-mapper

OGs and clans were annotated with eggNOG-mapper (`emapper.py -m diamond`). 11,769/47,266 OGs and 773/39,784 clans did not yield any match in the eggNOG database (Data file 6 and 7).

5.2. Supplementary figures

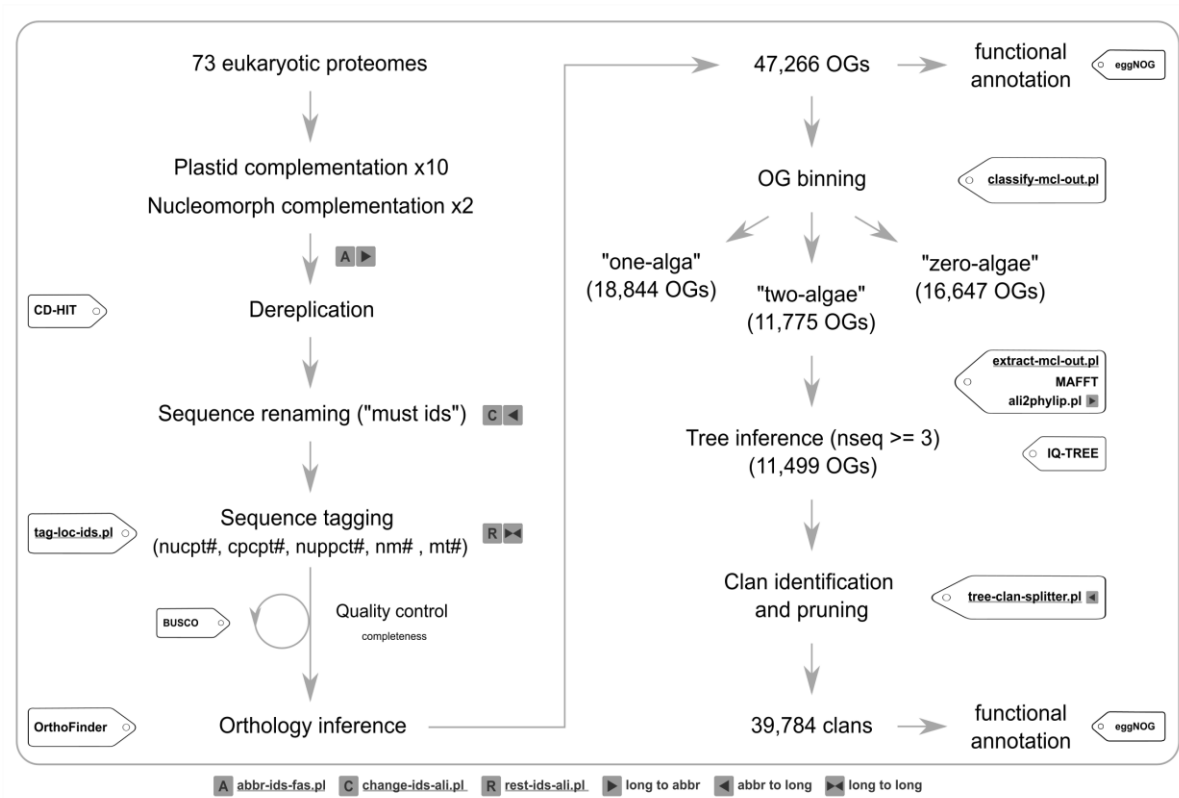
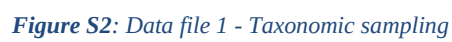


Figure S1: Graphical summary of the pipeline described in this study (see text for details). “long to abbr” stands for abbreviating sequence identifiers mainly for CD-HIT that may truncates longer ids, “abbr to long” for restoring full length sequence identifiers with full taxonomic information, and “long to long” for renaming purposes, such as adding a prefix tag (e.g., nucpt#) to a full length sequence identifier.

Script	Purpose	Distribution
change-ids-ali.pl	Abbreviate or restore the organism component of full seq ids	BMC
rest-ids-ali.pl	Change (restore) full seq ids	BMC
abbr-ids-fas.pl	Abbreviate (standardize) full seq ids	BMT-Mcl
tag-loc-ids.pl	Compute seq id organelle tags based on sequence identity	BMT-Mcl
classify-mcl-out.pl	Classify MCL clusters based on taxonomic filters	BMT-Mcl
extract-mcl-out.pl	Extract OGs (FASTA files) from MCL clusters	BMT-Mcl
tree-clan-splitter.pl	Extract clans (FASTA files) from trees based on taxonomic filters	BMT-TreeParsing

Table S1: New script availability details. Other scripts are from Bio::MUST::Core (BMC), which is a dependency of both Bio::MUST::Tools::Mcl (BMT-Mcl) and Bio::MUST::Tools::TreeParsing (BMT-TreeParsing). All Bio::MUST-like distributions can be installed from CPAN (i.e., assuming a properly configured Perl environment, with \$ cpanm Bio::MUST::Core).



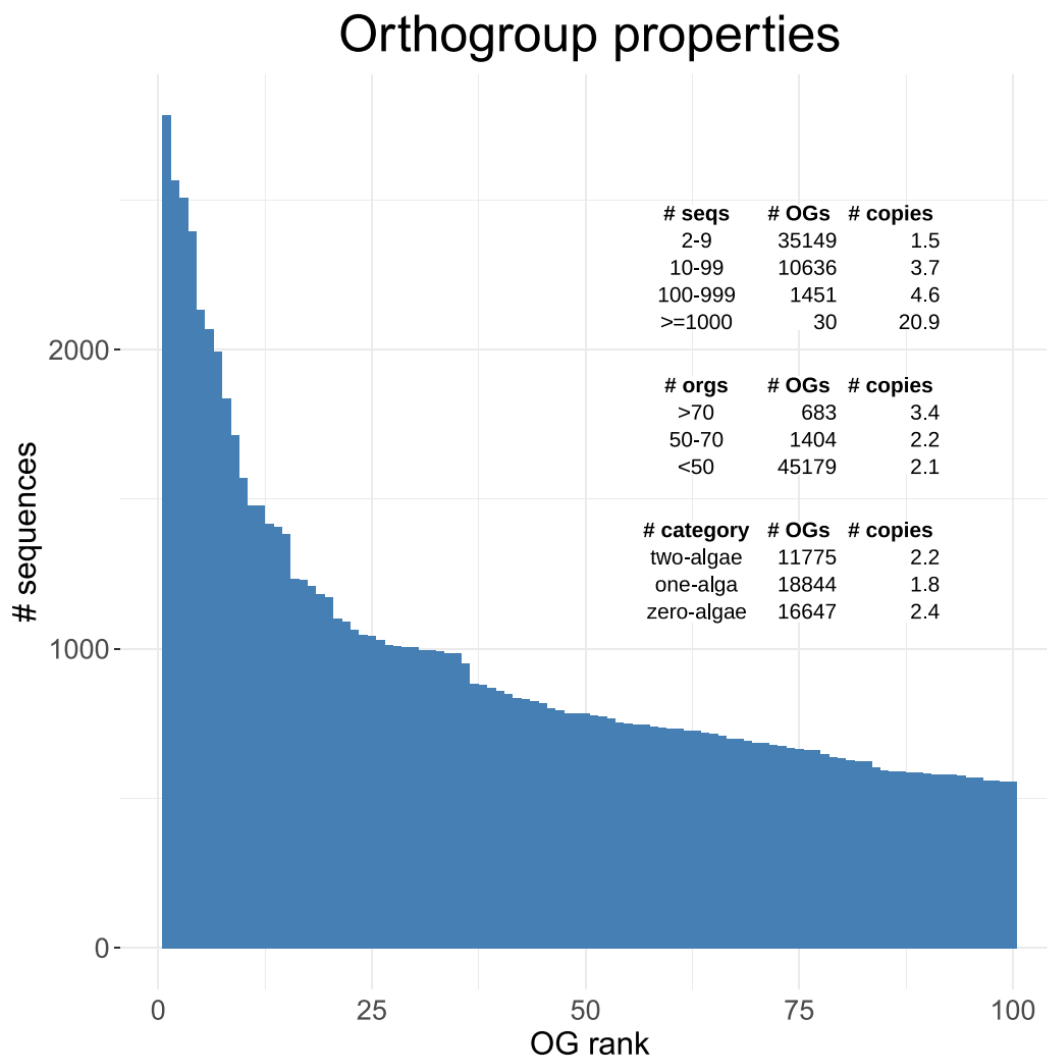


Figure S3 : Data file 3 - Orthogroup bar chart. Distribution of the number of sequences across the 100 largest orthogroups. The **upper inset table** summarizes the number of sequences per orthogroup range (**# seqs**), the number of orthogroups in each range (**# OGs**), and the mean copy number per organism (**copies**). The **middle inset table** summarizes the number of distinct organisms represented in each orthogroup (**# orgs**), the number of orthogroups in the corresponding range (**# OGs**), and the mean copy number per organism (**copies**). The **lower inset table** summarizes the orthogroup categories defined in the note (**category**), the number of orthogroups in each category (**# OGs**), and the mean copy number per organism (**copies**).

5.3. Supplementary tables

taxid	organism	nucleus.url	plastid.entrez	nm_plastid.url
1257118	Acanthamoeba castellanii_1257118	http://AmoebaDB.org/common/downloads/release-26/AcastellaniiNeff/fasta/data/amoebaDB-26_Acasti		
507362	Agaricus bisporus_507362	https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/300/555/GCF_000300555.1_Agabi_varbun_1/GCF_000		
65357	Albugo candida_65357	https://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/001/678/535/GCA_001678535.1_ASM107853v1/GCA_001		
578402	Alomyces macrosporus_578402	https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/151/295/GCF_000151295.1_ASM74375v1/GCF_00007		
87111	Alphacystidium virgatum_87111	https://genome.jgi.doe.gov/pal1/download/pal1_GeneCatalog_proteins_20121220_aa.fasta.gz		
3702	Avallbopsis thaliana_3702	ftp://ftp.psb.ugent.be/pub/plaza/plaza_pico_02/Fasta/protome.proteome.aan.tfa.gz		
380704	Aspergillus niger_380704	https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/002/855/GCF_00002855.1_ASM285v2/GCF_000002		
717989	Aurantiochytrium limacium_717989	http://genome.jgi.doe.gov/Aur11/download/Aur11_GeneCatalog_proteins_20120618_aa.fasta.gz		
44056	Aureococcus anophagefferens_44056	ftp://ftp.psb.ugent.be/pub/plaza/plaza_pico_02/Fasta/protome.proteome.aan.tfa.gz	("Bathycoccus prasinos"[Organism]) AND (refseq[filter] AND (chloroplast[filter] OR plastid[filter]))	
41875	Bathycoccus prasinos_41875	ftp://ftp.psb.ugent.be/pub/plaza/plaza_pico_02/Fasta/protome.proteome.bprcc1105.tfa.gz		
684364	Batrachochytrium dendrobatidis_684364	https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/203/795/GCF_000203795.1_v1.0/GCF_000203795.1		
753081	Bigelelliella natans_753081	https://genome.jgi.doe.gov/portal/Big1a1/download/Big1a1_filtered_proteins.fasta.gz		ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/002/455/GCF_000002455.1_ASM245v1/GCF_00000245
1529846	Blastocystis sp._1529846	https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/743/755/GCF_000743755.1_ASM74375v1/GCF_00007		
75058	Bodo saltans_75058	https://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/001/146/835/GCA_001146835.1_BSA/GCA_001146835.1		
596528	Capsaspora encarsae_596528	https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/151/015/GCF_000151015.2_C_ewczaraki_V2/GCF_0		
3055	Chlamydomonas reinhardtii_3055	ftp://ftp.psb.ugent.be/pub/plaza/plaza_pico_02/Fasta/protome.proteome.cre.tfa.gz		
2769	Chondrus crispus_2769	https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/350/225/GCF_000350225.1_ASM350225v2/GCF_0003		
1169474	Chromera velia_1169474	https://cryptodb.org/common/downloads/release-45/CveliaCCMP2878/fasta/data/CryptODB-45_CveliaC	("Chromera velia"[Organism]) AND (refseq[filter] AND (chloroplast[filter] OR plastid[filter]))	
1460289	Chrysochromulina sp_1460289	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/001/275/005/GCA_001275005.1_Clobinv2/GCA_001275005	("Chrysochromulina sp"[Organism]) AND (refseq[filter] AND (chloroplast[filter] OR plastid[filter]))	
574566	Coccomyxa subellipsoidea_574566	ftp://ftp.psb.ugent.be/pub/plaza/plaza_pico_02/Fasta/protome.proteome.cvu.tfa.gz		
353152	Cryptosporidium parvum_353152	http://CryptODB.org/common/downloads/release-26/Cparvum10wall/fasta/data/CryptODB-26_Cparvum1		
45157	Cyanidioschyzon merolae_45157	ftp://ftp.psb.ugent.be/pub/plaza/plaza_pico_02/Fasta/protome.proteome.cme.tfa.gz		
2762	Cyanophora paradoxa_2762	http://cyanophora.rutgers.edu/cyanophora_v2013/Cyanophora_MAY2017-augustus_predictions.fasta		
352472	Dicystodiscus discoidium_352472	https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/004/095/GCF_00004095.1_dich_2.7/6CF_0000409		
2880	Ectocarpus siliculosus_2880	ftp://ftp.psb.ugent.be/pub/plaza/plaza_pico_02/Fasta/protome.proteome.esi.tfa.gz		
280403	Emiliania huxleyi_280403	https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/372/725/GCF_000372725.1_Emiliana_huxleyi_COMP1	("Emiliania huxleyi"[Organism]) AND (refseq[filter] AND (chloroplast[filter] OR plastid[filter]))	
3039	Euglena gracilis_3039	ftp://ftp.pride.ebi.ac.uk/pride/data/archive/2019/01/PXD009998/Euglena_translated_transcriptome.fast	("Euglena gracilis"[Organism]) AND (refseq[filter] AND (chloroplast[filter] OR plastid[filter]))	
691883	Fonticula alba_691883	ftp://ftp.ensemblgenomes.org/pub/protists/release-30/fasta/protists_nuclear/daalafonticulaagroup1_c		
186039	Fragilariopsis cylindrus_186039	ftp://ftp.psb.ugent.be/pub/plaza/plaza_pico_02/Fasta/protome.proteome.fcy.tfa.gz		
1389228	Galdieria phlegrea_1389228	http://cyanophora.rutgers.edu/gphlegrea/G_phlegrea.Predicted_nuclear_ProteinSeq_aa	("Galdieria phlegrea"[Organism]) AND (chloroplast[filter] OR plastid[filter]))	
130081	Galdieria sulphuraria_130081	https://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/341/285/GCF_000341285.1_ASM341285v1/GCF_0003		
1255295	Gonimomonas ivontiae_1255295	U. Cenci (personal communication)		
905079	Gullardia thea_905079	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/315/825/GCF_000315825.1_Gullm1/GCF_000315825.1		ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/002/875/GCF_000002875.1_ASM287v1/GCF_00000287
9606	Homo sapiens_9606	ftp://ftp.ncbi.nlm.nih.gov/genomes/Homo_sapiens/protein/protein.fa.gz		
3175	Klebsoridium flaccidum_3175	https://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/708/835/GCA_000708835.1_ASM70883v1/GCA_0007		
644223	Komagataella phaffii_644223	https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/027/005/GCF_000027005.1_ASM2700v1/GCF_00002		
435258	Leishmania infantum_435258	http://TrITrypDB.org/common/downloads/release-26/InfantumPCMS/fasta/data/TrITrypDB-26_Infantu		
225164	Lotia gigantea_225164	https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/327/385/GCF_000327385.1_Helro1/GCF_000327385		
564608	Micromonas pusilla_564608	ftp://ftp.psb.ugent.be/pub/plaza/plaza_pico_02/Fasta/protome.proteome.mpu.tfa.gz		
431895	Monosiga brevicollis_431895	https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/002/865/GCF_000002865.3_v1.0/GCF_000002865.3		
747725	Mucor circinelloides_747725	http://genome.jgi.doe.gov/Mucc2/download/Mucc2_circinelloides_v2_filtered_proteins.fasta.gz		
72520	Nannochloropsis gaditana_72520	https://plasmodb.org/common/downloads/Current_Release/Placiparum307/fasta/data/PlasmODB-48_f		
435351	Nematostella vectensis_435351	https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/209/225/GCF_000209225.1_ASM209225v1/GCF_0002		
70448	Ostreococcus tauri_70448	https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/214/015/GCF_000214015.3_version_14/06/GCF_00		
423536	Perkinsus marinus_423536	https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/006/405/GCF_000006405.1_JCVI_PMG_1.0/GCF_000		
2850	Phaeodactylum tricornutum_2850	ftp://ftp.psb.ugent.be/pub/plaza/plaza_pico_02/Fasta/protome.proteome.ptri.tfa.gz		
403677	Phytophthora infestans_403677	https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/142/945/GCF_000142945.1_ASM14294v1/GCF_0001		
37360	Plasmodiophora brassicae_37360	https://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/001/049/375/GCA_001049375.1_pbe3.h15/GCA_0010493		
36329	Plasmodium falciiparum_36329	https://plasmodb.org/common/downloads/Current_Release/Placiparum307/fasta/data/PlasmODB-48_f		
670386	Polysphondylium pallidum_670386	https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/004/825/GCF_000004825.1_PolPa_Dec2009/GCF_00		
2786	Porphyra umbilicalis_2786	https://phytozome.jgi.doe.gov/p2/portal.html#info?alias=Org_Pumbilicalis_e		
55680	Porphyridium purpureum_55680	http://cyanophora.rutgers.edu/porphyridiumPorphyridium_genomodels_UPDATED.fasta		
635002	Pseudo-nitzschia multiseries_635002	http://genome.jgi.doe.gov/Psmu1/download/Psmu1_GeneCatalog_proteins_20111011_aa.fasta.gz		
418459	Puccinia graminis_418459	https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/149/925/GCF_000149925.1_ASM14992v1/GCF_0001		
2788	Pyropia yezoensis_2788	http://nirfs.fra.affrc.go.jp/cgi-bin/time_download/time.cgi?norl_FASTA_AminoAcid		
65071	Pythium ultimum_65071	ftp://ftp.ensemblgenomes.org/pub/protists/release-30/fasta/pythium_ultimum/pep/Pythium_ultimum.pu		
46433	Reticulomyxa flosa_46433	https://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/512/085/GCA_000512085.1_Retl_assembly1.0/GCA_0		
747089	Rhizophagus irregularis_747089	https://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/439/145/GCA_000439145.2_Gloin1/GCA_000439145		
988480	Rozella alomyces_988480	https://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/442/015/GCA_000442015.1_Rozella_k41_1100/GCA_0		
559292	Saccharomyces cerevisiae_559292	https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/146/045/GCF_000146045.2_R64/GCF_000146045.2_1		
665850	Saprotegia parafatica_665850	https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/151/545/GCF_000151545.1_ASM15154v2/GCF_0001		
878978	Schizochytrium aggregatum_878978	http://genome.jgi.doe.gov/chag1/download/chag1_GeneCatalog_proteins_20121220_aa.fasta.gz		
88036	Selaginella moellendorffii_88036	https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/143/415/GCF_000143415.3_v1.0/GCF_000143415.3_1		
755200	Stylochytrium lemnae_755200	https://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/000/751/175/GCA_000751175.1_Stylochytrium_lemnae_asn		
2951	Symbiodinium microadriaticum_2951	https://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/001/939/145/GCA_001939145.1_ASM193914v1/GCA_001		
1202447	Symbiodinium minutum_1202447	http://marinegenomics.oist.jp/symb/download/symb8.v1.2.augustus.pro.tfa.gz		
312017	Tetrahymena thermophila_312017	https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/189/635/GCF_000189635.1_JCVI-TTA1-2.2/GCF_0001		
35128	Thalassiosira pseudonana_35128	ftp://ftp.psb.ugent.be/pub/plaza/plaza_pico_02/Fasta/protome.proteome.tps.tfa.gz		
1321669	Tisochrysis lutea_1321669	https://www.seaonee.org/data/0036147171/data/471137.fasta		
506771	Toxoplasma gondii_506771	ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/006/565/GCF_000006565.2_TGaa/GCF_000006565.2_1		
679716	Trypanosoma brucei_679716	http://TrITrypDB.org/common/downloads/release-28/bruceiGenomeD4/972/fasta/data/TrITrypDB-26		
237631	Ustilago maydis_237631	https://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/328/475/GCF_000328475.1_Umaydis521.2.0/GCF_00		
1169540	Vitrella brassicaformis_1169540	http://CryptODB.org/common/downloads/release-26/VbrassicaformisCCMP3155/fasta/data/CryptODB-2		

Table S1 : Data set 1 - Proteomes set description including urls for nuclear, plastid and nucleomorph genomes

organism	completeness	single.copy	duplicate	fragmented	missing	total.genes
Acanthamoeba_castellanii_5755	8,18	7,85	0,33	0,59	1,23	303
Agaricus_bisporus_5341	9,5	9,37	0,13	0,17	0,33	303
Albugo_candida_65357	9,04	8,35	0,69	0,33	0,63	303
Allomyces_macrogyne_28583	9,67	3,2	6,47	0,07	0,26	303
Aplanochytrium_kerguelense_87111	8,88	8,65	0,23	0,46	0,66	303
Arabidopsis_thaliana_3702	9,9	7,89	2,01	0,07	0,03	303
Aspergillus_niger_5061	9,81	9,64	0,17	0,03	0,16	303
Aurantiochytrium_limacinum_717989	8,92	8,75	0,17	0,56	0,52	303
Aureococcus_anophagefferens_44056	6,77	6,37	0,4	0,89	2,34	303
Bathycoccus_prasinus_41875	8,41	8,18	0,23	0,43	1,16	303
Batrachomyxium_dendrobatidis_684364	9,67	9,41	0,26	0,17	0,16	303
Bigelowella_natans_753081	8,75	8,32	0,43	0,5	0,75	303
Blastocystis_sp_1529846	6,93	6,6	0,33	1,19	1,88	303
Bodo_saltans_75058	6,57	6,11	0,46	0,96	2,47	303
Capsaspora_owczarzakii_595528	9,47	9,47	0	0,1	0,43	303
Chlamydomonas_reinhardtii_3055	9,01	8,81	0,2	0,59	0,4	303
Chondrus_crispus_2769	7,65	7,39	0,26	0,79	1,56	303
Chromera_velia_1169474	8,09	7,69	0,4	0,89	1,02	303
Chrysochromulina_sp_1460289	6,34	6,14	0,2	0,96	2,7	303
Coccomyxa_subellipsoidea_574566	8,98	8,91	0,07	0,5	0,52	303
Cryptosporidium_parvum_353152	6,86	6,63	0,23	0,63	2,51	303
Cyanidioschyzon_merolae_45157	9,34	9,17	0,17	0,33	0,33	303
Cyanophora paradoxa_2762	8,98	8,65	0,33	0,33	0,69	303
Dictyostelium_discoideum_44689	9,6	9,47	0,13	0,07	0,33	303
Ectocarpus_siliculosus_2880	9,07	8,81	0,26	0,56	0,37	303
Emiliania_huxleyi_280463	7,5	6,11	1,39	1,42	1,08	303
Euglena_gracilis_3039	8,45	7,99	0,46	0,76	0,79	303
Fonticula_alba_691883	8,05	7,72	0,33	0,99	0,96	303
Fragilariopsis_cylindrus_186039	8,12	7,69	0,43	0,86	1,02	303
Galdieria_phlegrea_1389228	6,84	6,67	0,17	1,98	1,18	303
Galdieria_sulphuraria_130081	9,2	8,28	0,92	0,23	0,57	303
Goniomonas_avonlea_1255295	7,82	7,62	0,2	1,52	0,66	303
Guillardia_theta_905079	8,42	7,79	0,63	0,73	0,85	303
Homo_sapiens_9606	10	6,2	3,8	0	0	303
Klebsormidium_flaccidum_3175	9,84	9,74	0,1	0,1	0,06	303
Komagataella_phaffii_460519	9,57	9,47	0,1	0,03	0,4	303
Leishmania_infantum_435258	7,13	6,8	0,33	0,83	2,04	303
Lottia_spp_72691	9,73	9,7	0,03	0,03	0,24	303
Micromonas_pusilla_564608	8,58	8,35	0,23	0,43	0,99	303
Monosiga_brevicollis_81824	8,26	8,09	0,17	0,76	0,98	303
Mucor_circinelloides_747725	9,34	8,48	0,86	0,46	0,2	303
Nannochloropsis_gaditana_72520	7,96	7,56	0,4	0,99	1,05	303
Nematostella_spp_45350	9,47	9,27	0,2	0,26	0,27	303
Ostreococcus_tauri_70448	8,82	8,65	0,17	0,3	0,88	303
Perkinsus_marinus_423536	7,42	4,42	3	0,89	1,69	303
Phaeodactylum_tricornutum_2850	8,32	8,02	0,3	0,66	1,02	303
Phytophthora_infestans_4787	9,34	8,88	0,46	0,13	0,53	303
Plasmodiophora_brassicae_37360	9,44	9,27	0,17	0,26	0,3	303
Plasmodium_falciparum_36329	7,13	6,93	0,2	0,26	2,61	303
Polysphondylium_pallidum_13642	9,27	9,27	0	0,36	0,37	303
Porphyra_umbilicalis_2786	6,24	6,04	0,2	1,22	2,54	303
Porphyridium_purpureum_35688	9,01	8,78	0,23	0,33	0,66	303
Pseudo-nitzschia_multiseries_37319	8,48	8,22	0,26	0,79	0,73	303
Puccinia_graminis_418459	9,64	9,14	0,5	0,13	0,23	303
Pyropia_zeoensis_2788	4,49	4,29	0,2	2,31	3,2	303
Pythium_ultimum_65071	9,27	8,91	0,36	0,33	0,4	303
Reticulomyxa_filosa_46433	5,71	4,98	0,73	2,18	2,11	303
Rhizophagus_irregularis_588596	8,95	8,88	0,07	0,83	0,22	303
Rozella_ellomycis_281847	8,25	8,15	0,1	0,53	1,22	303
Saccharomyces_cerevisiae_559292	9,5	9,24	0,26	0,03	0,47	303
Saprolegnia_parasitica_695850	9,17	8,61	0,56	0,26	0,57	303
Schizochytrium_aggregatum_876976	8,38	8,12	0,26	0,66	0,96	303
Selaginella_moellendorffii_88036	9,6	5,61	3,99	0,26	0,14	303
Stylonychia_lemnae_755200	8,25	7,49	0,76	0,59	1,16	303
Symbiodinium_microadriaticum_2951	6,07	5,64	0,43	1,25	2,68	303
Symbiodinium_minutum_1202447	6,3	5,31	0,99	1,62	2,08	303
Tetrahymena_thermophila_5911	8,38	7,92	0,46	0,13	1,49	303
Thalassiosira_pseudonana_35128	7,69	7,39	0,3	0,79	1,52	303
Tisochrysis_lutea_1321669	6,99	6,63	0,36	0,89	2,12	303
Toxoplasma_gondii_508771	7,92	7,62	0,3	0,59	1,49	303
Trypanosoma_brucei_679716	7,23	6,93	0,3	0,59	2,18	303
Ustilago_maydis_5270	9,9	9,9	0	0,03	0,07	303
Vitrella_brassicaformis_1169540	8,75	8,35	0,4	0,3	0,95	303

Table S2: Data file 2 - Sample quality statistics inferred with BUSCO v3.0.1 based on 303 single-copy orthologous markers from the eukaryotic dataset.

organelle	query.entrez
Apicoplast	apicoplast[All Fields] AND (refseq[filter] AND plastid[filter])
Plastid	plastid[All Fields] AND (refseq[filter] AND plastid[filter])
Nucleomorph	nucleomorph[All Fields] AND (protists[filter] AND refseq[filter])

Table S3: Data file 4 - NCBI queries to download and build the organellar protein database used with tag-loc-ids.pl.

Nuclear encoded and plastid-targeted proteins			
Organism(s)	Source	Method	Additional information
HPPGs (Homologous plastid protein group)	Dorrell RG, Gile G, McCallum G, Méheust R, Baptiste EP, Klinger CM, et al. Chimeric origins of ochrophytes and haptophytes revealed through an ancient plastid proteome. <i>Elife</i> . 2017;6:1–45. doi:10.7554/eLife.23717.	Predictions	Supplementary materials
<i>Euglena gracilis</i>	Novák Vanclová AMG, Zoltner M, Kelly S, Soukal P, Záhonová K, Füßy Z, et al. Metabolic quirks and the colourful history of the <i>Euglena gracilis</i> secondary plastid. <i>New Phytol</i> . 2020;225:1578–92. doi:10.1111/nph.16237.	Predictions	Supplementary materials
<i>Plasmodium falciparum</i>	Boucher MJ, Ghosh S, Zhang L, Lal A, Jang SW, Ju A, et al. Integrative proteomics and bioinformatic prediction enable a high-confidence apicoplast proteome in malaria parasites. <i>PLOS Biol</i> . 2018;16:e2005895. doi:10.1371/journal.pbio.2005895.	Proteomics	Table S8
Nuclear encoded and periplastidial-compartment-targeted proteins			
Organism(s)	Source	Method	Additional information
<i>Bigelowiella natans</i> , <i>Guillardia theta</i>	Curtis BA, Tanifuji G, Burki F, Gruber A, Irimia M, Maruyama S, et al. Algal genomes reveal evolutionary mosaicism and the fate of nucleomorphs. <i>Nature</i> . 2012;492:59–65. doi:10.1038/nature11681.	Predictions	Tables S1.10.1 and S1.10.2

Table S5: Data file 5 - Links to the different sources used to build the plastid-targeted protein database.

6. Transition: towards a broader complex algal phylogenetic sampling

The orthologous clans established in this chapter, from curated reference proteomes provided a robust phylogenetic backbone, but their taxonomic diversity remained necessarily limited by the restricted availability of high-quality algal proteomes. To broaden the sampling of complex algae, transcriptomic resources therefore became essential. In particular, the Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP; Keeling et al. 2014) constituted one of the richest publicly available sources of transcriptomic data for complex algae, making it a natural target for expanding our taxonomic coverage. Yet this gain in diversity comes at a cost: transcriptomes are intrinsically more heterogeneous than reference proteomes, often incomplete, frequently redundant, and particularly prone to contamination. In the context of plastid-related gene evolution, this issue is especially critical, because genuinely transferred genes can be difficult to distinguish from foreign (i.e., contaminant) sequences introduced during sampling, culturing, or sequencing. Expanding the dataset therefore required a dedicated pipeline for selecting, consolidating, and filtering transcriptomic resources before they could be incorporated into downstream phylogenomic analyses. This pipeline is the object of the next chapter, published in *BMC Research Notes*.

IV. Chapter 4 - Decontamination, pooling and dereplication of the 678 samples of the Marine Microbial Eukaryote Transcriptome Sequencing Project

Mick Van Vlierberghe (mvanvlierberghe@doct.uliege.be)¹, Arnaud Di Franco (arnaud.difranco@gmail.com)², Hervé Philippe (herve.philippe@sete.cnrs.fr)² and Denis Baurain (denis.baurain@uliege.be)^{1,*}

¹InBioS - PhytoSYSTEMS, Eukaryotic Phylogenomics, University of Liège, Liège, Belgium

²Station d'Ecologie Théorique et Expérimentale de Moulis, UMR CNRS 5321, Moulis, France

* corresponding author

This chapter has been published in *BMC Research Notes*; as a consequence, it has its own “References” section based on a numbered list.

1. Abstract

Objectives:

Complex algae are photosynthetic organisms resulting from eukaryote-to-eukaryote endosymbiotic-like interactions. Yet the specific lineages and mechanisms are still under debate. That is why large scale phylogenomic studies are needed. Whereas available proteomes provide a limited diversity of complex algae, MMETSP (Marine Microbial Eukaryote Transcriptome Sequencing Project) transcriptomes represent a valuable resource for phylogenomic analyses, owing to their broad and rich taxonomic sampling, especially of photosynthetic species. Unfortunately, this sampling is unbalanced and sometimes highly redundant. Moreover, we observed contaminated sequences in some samples. In such a context, tree inference and readability are impaired. Consequently, the aim of the data processing reported here is to release a unique set of clean and non-redundant transcriptomes produced through an original protocol featuring decontamination, pooling and dereplication steps.

Data description:

We submitted 678 MMETSP re-assembly samples to our parallel consolidation pipeline. Hence, we combined 423 samples into 110 consolidated transcriptomes, after the systematic removal of the most contaminated samples (186). This approach resulted in a total of 224 high-quality transcriptomes, easy to use and suitable to compute less contaminated, less redundant and more balanced phylogenies.

Keywords:

Bioinformatics, Decontamination, Transcriptomes, Gene phylogenies, Phylogenomics, MMETSP, Algae, Eukaryotic evolution, Endosymbiotic gene transfer (EGT), Kleptoplastidy

2. Objective

Plastid-bearing organisms are scattered across the eukaryotic tree. Among those, CASH lineages have plastids related to red algae, but the mechanisms by which they were acquired and the ancestors involved remain unclear [1,2,3]. To overcome the inconsistencies of purely endosymbiotic models [4], we propose kleptoplastidy [5] as an additional mechanism for explaining plastid spread among CASH lineages. In line with the shopping bag model [6], our hypothesis posits multiple transient interactions with prey of diverse origins but also proposes a rationale for the selective force driving the progressive accumulation of plastid-targeted genes. In such a scenario, also recently proposed by Bodyl [7], the phylogenetic diversity of plastid-targeted genes would be higher than predicted with endosymbiotic models, where genes originate mostly from a single source [8]. Obviously, testing this hypothesis requires a scrupulous phylogenetic study and at the largest scale possible. In addition, the inherent

nature of those genes, which are transferred from one lineage to another, makes them hardly distinguishable from actual contaminations, prompting the use of data as clean as possible to avoid false positives. To overcome the limited diversity of complex algal proteomes and reduce the high redundancy of the richly sampled MMETSP transcriptomes [9,10], we designed a consolidation pipeline. However, we chose to remove the most contaminated samples instead of trying to decontaminate them. This somewhat radical method has the advantage of allowing a more reliable interpretation of downstream phylogenetic analyses at the expense of a moderate loss of taxonomic breadth and richness. We used two methods to estimate contamination levels, one targeting cross-contaminations [11] (between MMETSP samples, using raw reads) and the other targeting more regular contaminations [12,13] (from any eukaryotic source, using ribosomal markers). Finally, we decided to release the resulting transcriptomes in line with the FAIR data principles [14].

3. Data description

MMETSP transcriptomic samples [9] cover a wide range of primary and complex algae and were recently re-assembled [10]. Despite showing better annotation overall, we observe high redundancy and contaminations, which complicates the study of plastid spread across eukaryotic lineages. To improve this dataset, we developed a pipeline to eliminate taxonomic redundancy while maximizing data purity and completeness. In detail, out of the 678 samples in MMETSP re-assemblies, 16 were discarded from the start (14 with less than 5000 sequences, 1 with a corrupted FASTQ file and another one because the organism was unknown). The remaining 662 were screened for contaminations with Forty-Two [12,13] using 78 ribosomal protein markers, and 89 samples were further discarded because they were too contaminated (Data file 1, Data set 1), leaving 573 samples. To minimize taxonomic redundancy, we aimed to combine closely related samples. However, 124 samples represented unique genera and were not combined. As phylogenetic diversity can be very different across protist genera, we determined whether multiple samples within a given genus should be all combined or if some should only be combined at the species/strain level. To this end, we used CroCo [11] and combined the samples that shared a “cross-contamination” value $\geq 10\%$ (interpreted here as close relatedness [11]) (Data file 2). Otherwise, they were left uncombined, even if from the same genus. This dual strategy yielded 110 combined samples (out of 423 candidate samples) and 26 uncombined samples (additional singletons), thereby aggregating 313 redundant samples (Data file 1, Data file 3). From a taxonomic point of view, this combination had the effect of reducing the number of samples by more than 50% for the majority of main phyla (Data file 4), resulting in a total of 260 transcriptomes (combined and singletons, Data set 2). Furthermore, the remaining most contaminated transcriptomes were removed using two strategies, i) by targeting cross-contaminations with Sobek, a new parallel implementation of CroCo [11] (Data file 5), ii) after removing intra-sample redundancy with CD-HIT-EST [15,16] (sequence

identity threshold of 95%), by targeting regular contaminations with Forty-Two (Data set 3) [12,13]. Both detection methods combined pinpointed 36 transcriptomes, corresponding to 97 original samples (Data file 6, Data file 7, Data file 1). Finally, sample completeness was assessed with BUSCO [17,18] (Data file 8). The resulting data set of 224 transcriptomes (Data set 2) spans a wide range of organisms, especially among photosynthetic species (172/224), with a majority of complex algae (145/224) (Data set 4). Its usefulness is illustrated with the GAPDH phylogeny (Data set 5). Briefly, two different trees were built after enriching the same starting set of sequences from high-quality proteomes [19], one using all 570 samples from plastid-bearing organisms and the other using 172 transcriptomes resulting from our pipeline. The comparison of these two trees shows that the enrichment performed with the data set described here is less contaminated, less redundant and more balanced from a taxonomic point of view, thereby allowing more robust evolutionary interpretations.

Table 1: Overview of data files/data sets.

Label	Name of data file/data set	File types (file extension)	Data repository and identifier (DOI or accession number)
<i>Data file 1</i>	<i>Methods</i>	<i>PDF file (.pdf)</i>	Figshare https://doi.org/10.6084/m9.figshare.14079866.v5 [20]
<i>Data set 1</i>	<i>Forty-Two reports and configuration files (662 individual samples)</i>	<i>Text files (.tsv, .csv, .yaml)</i>	Figshare https://doi.org/10.6084/m9.figshare.12362699.v1 [21]
<i>Data file 2</i>	<i>Consolidation table</i>	<i>Spreadsheet (.xlsx)</i>	Figshare https://doi.org/10.6084/m9.figshare.14727411.v3 [22]
<i>Data file 3</i>	<i>Sample consolidation report</i>	<i>Image file (.pdf)</i>	Figshare https://doi.org/10.6084/m9.figshare.12154824.v3 [23]
<i>Data file 4</i>	<i>Redundancy drop analysis</i>	<i>Spreadsheet (.xlsx)</i>	Figshare https://doi.org/10.6084/m9.figshare.12213731.v3 [24]
<i>Data set 2</i>	<i>Transcriptomes</i>	<i>FASTA files (.tar.gz)</i>	Figshare https://doi.org/10.6084/m9.figshare.13634840.v1 [25]
<i>Data file 5</i>	<i>Sobek analysis summary</i>	<i>Text file (.csv)</i>	Figshare https://doi.org/10.6084/m9.figshare.12410522.v3 [26]
<i>Data set 3</i>	<i>Forty-Two reports and configuration files (260 transcriptomes)</i>	<i>Text files (.tsv, .csv, .yaml)</i>	Figshare https://doi.org/10.6084/m9.figshare.13006622.v1 [27]
<i>Data file 6</i>	<i>Consolidated sample purity (cross-contaminations)</i>	<i>Image file (.pdf)</i>	Figshare https://doi.org/10.6084/m9.figshare.12173235.v3 [28]
<i>Data file 7</i>	<i>Consolidated sample purity (contaminations)</i>	<i>Image file (.pdf)</i>	Figshare https://doi.org/10.6084/m9.figshare.12998726.v3 [29]
<i>Data file 8</i>	<i>Completeness analysis</i>	<i>Text file (.csv)</i>	Figshare https://doi.org/10.6084/m9.figshare.12154833.v3 [30]

<i>Data set 4</i>	<i>Taxonomic samplings</i>	<i>Image files (.png, .html,)</i>	Figshare https://doi.org/10.6084/m9.figshare.12401639.v1 [31]
<i>Data set 5</i>	<i>GAPDH phylogenies</i>	<i>Image files, text file (.pdf)</i>	Figshare https://doi.org/10.6084/m9.figshare.13096208.v2 [32]

Limitations

- The way some samples were combined at the genus level might lead to a loss of resolution at the species/strain level within a given genus.
- Even though combined samples were dereplicated, sequence redundancy remains substantial, especially among dinoflagellates, which are known for bearing multiple copies of their genes.
- Similarly, even if we provide a set of cleaned transcriptomes, contaminations remain. First, in some cases, such as ciliates, it is a “feature” of the phylum that all its members are highly contaminated. Therefore, we used a phylum-based outlier approach to determine whether a sample was too contaminated, because setting a global threshold for all phyla at once was impracticable. Second, in other cases, contamination by foreign sequences is a “necessary evil” for the reason that we plan to study transferred genes in a plastid acquisition context, and those would be undetectable in a completely “uncontaminated” transcriptome.

Abbreviations

- CASH: Cryptophytes, Alveolates, Stramenopiles, Haptophytes
- MMETSP: Marine Microbial Eukaryote Transcriptome Sequencing Project

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

The MMETSP re-assemblies analysed during this study are available in the Zenodo repository (<https://zenodo.org/record/1212585/#.YA1WXdZ7lX0>, 10.5281/zenodo.1212585, [9,10]). All the remaining data generated or analysed during this study are publicly available in the figshare repository (<https://doi.org/10.6084/m9.figshare.14079866.v5>, <https://doi.org/10.6084/m9.figshare.12362699.v1>, <https://doi.org/10.6084/m9.figshare.14727411.v3>, <https://doi.org/10.6084/m9.figshare.12154824.v3>,

<https://doi.org/10.6084/m9.figshare.12213731.v3>, <https://doi.org/10.6084/m9.figshare.13634840.v1>,
<https://doi.org/10.6084/m9.figshare.12410522.v3>, <https://doi.org/10.6084/m9.figshare.13006622.v1>,
<https://doi.org/10.6084/m9.figshare.12173235.v3>, <https://doi.org/10.6084/m9.figshare.12998726.v3>,
<https://doi.org/10.6084/m9.figshare.12154833.v3>, <https://doi.org/10.6084/m9.figshare.12401639.v1>,
<https://doi.org/10.6084/m9.figshare.13096208.v2>). Please see Table 1 and references [20-32] for details and links to the data.

Competing interests

The authors declare that they have no competing interests.

Funding

Mick Van Vlierberghe was a FRIA fellow of the F.R.S.-FNRS (National Fund for Scientific Research of Belgium). Computational resources were provided through two grants to Denis Baurain (University of Liège “Crédit de démarrage 2012” SFRD-12/04; FRS-FNRS “Crédit de recherche 2014” CDR J.0080.15).

Authors' contributions

Mick Van Vlierberghe and Denis Baurain designed the experiments, Mick Van Vlierberghe performed all the computational analyses and drew the figures, Mick Van Vlierberghe and Denis Baurain wrote the manuscript. Arnaud Di Franco and Hervé Philippe prototyped contamination analyses based on ribosomal markers and Hervé Philippe further substantively revised the work.

Acknowledgements

Not applicable.

4. References

- [1] Zimorski, V., et al. (2014). "Endosymbiotic theory for organelle origins." *Curr Opin Microbiol* 22: 38-48.
- [2] Ponce-Toledo, R. I., et al. (2019). "Horizontal and endosymbiotic gene transfer in early plastid evolution." *New Phytol* 224(2): 618-624.
- [3] Sibbald SJ, Archibald JM. Genomic Insights into Plastid Evolution. *Genome Biol Evol.* 2020;12:978-90.
- [4] Keeling, P. J. (2013). "The number, speed, and impact of plastid endosymbioses in eukaryotic evolution." *Annu Rev Plant Biol* 64: 583-607.

- [5] Nowack, E. C. and M. Melkonian (2010). "Endosymbiotic associations within protists." *Philos Trans R Soc Lond B Biol Sci* 365(1541): 699-712.
- [6] Larkum, A. W., et al. (2007). "Shopping for plastids." *Trends Plant Sci* 12(5): 189-195.
- [7] Bodyl, A. (2018). "Did some red alga-derived plastids evolve via kleptoplastidy? A hypothesis." *Biol Rev Camb Philos Soc* 93(1): 201-222.
- [8] Archibald, J. M. (2015). "Genomic perspectives on the birth and spread of plastids." *Proc Natl Acad Sci U S A* 112(33): 10147-10153.
- [9] Keeling, P. J., et al. (2014). "The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP): illuminating the functional diversity of eukaryotic life in the oceans through transcriptome sequencing." *PLoS Biol* 12(6): e1001889.
- [10] Johnson, L. K., et al. (2019). "Re-assembly, quality evaluation, and annotation of 678 microbial eukaryotic reference transcriptomes." *Gigascience* 8(4).
- [11] Simion, P., et al. (2018). "A software tool 'CroCo' detects pervasive cross-species contamination in next generation sequencing data." *BMC Biol* 16(1): 28.
- [12] Simion, P., et al. (2017). "A Large and Consistent Phylogenomic Dataset Supports Sponges as the Sister Group to All Other Animals." *Curr Biol* 27(7): 958-967.
- [13] Irisarri, I., et al. (2017). "Phylotranscriptomic consolidation of the jawed vertebrate timetree." *Nat Ecol Evol* 1(9): 1370-1378.
- [14] <https://www.force11.org/group/fairgroup/fairprinciples>
- [15] Li, W., et al. (2001). "Clustering of highly homologous sequences to reduce the size of large protein databases." *Bioinformatics* 17(3): 282-283.
- [16] Li, W. and A. Godzik (2006). "Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences." *Bioinformatics* 22(13): 1658-1659.
- [17] Simao, F. A., et al. (2015). "BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs." *Bioinformatics* 31(19): 3210-3212.
- [18] Waterhouse, R. M., et al. (2018). "BUSCO Applications from Quality Assessments to Gene Prediction and Phylogenomics." *Mol Biol Evol* 35(3): 543-548.
- [19] Van Vlierberghe M, Philippe H, Baurain D. Broadly sampled orthologous groups of eukaryotic proteins for the phylogenetic study of plastid-bearing lineages. *BMC Res Notes*. 2021;14:21-4.
- [20] Van Vlierberghe M, Di Franco A, Philippe H, Baurain D. Data file 1 - Methods. Figshare 2021. <https://doi.org/10.6084/m9.figshare.14079866.v5>
- [21] Van Vlierberghe M, Di Franco A, Philippe H, Baurain D. Data set 1 - Forty-Two reports and configuration files (662 individual samples). Figshare 2021. <https://doi.org/10.6084/m9.figshare.12362699.v1>
- [22] Van Vlierberghe M, Di Franco A, Philippe H, Baurain D. Data file 2 - Consolidation table. Figshare 2021. <https://doi.org/10.6084/m9.figshare.14727411.v3>,

- [23] Van Vlierberghe M, Di Franco A, Philippe H, Baurain D. Data file 3 - Sample consolidation report. Figshare 2021. <https://doi.org/10.6084/m9.figshare.12154824.v3>
- [24] Van Vlierberghe M, Di Franco A, Philippe H, Baurain D. Data file 4 - Redundancy drop analysis. Figshare 2021. <https://doi.org/10.6084/m9.figshare.12213731.v3>
- [25] Van Vlierberghe M, Di Franco A, Philippe H, Baurain D. Data set 2 - Transcriptomes. Figshare 2021. <https://doi.org/10.6084/m9.figshare.13634840.v1>
- [26] Van Vlierberghe M, Di Franco A, Philippe H, Baurain D. Data file 5 - Sobek analysis summary. Figshare 2021. <https://doi.org/10.6084/m9.figshare.12410522.v3>
- [27] Van Vlierberghe M, Di Franco A, Philippe H, Baurain D. Data set 3 - Forty-Two reports and configuration files (260 transcriptomes). Figshare 2021. <https://doi.org/10.6084/m9.figshare.13006622.v1>
- [28] Van Vlierberghe M, Di Franco A, Philippe H, Baurain D. Data file 6 - Consolidated sample purity (cross-contaminations). Figshare 2021. <https://doi.org/10.6084/m9.figshare.12173235.v3>
- [29] Van Vlierberghe M, Di Franco A, Philippe H, Baurain D. Data file 7 - Consolidated sample purity (contaminations). Figshare 2021. , <https://doi.org/10.6084/m9.figshare.12998726.v3>
- [30] Van Vlierberghe M, Di Franco A, Philippe H, Baurain D. Data file 8 - Completeness analysis. Figshare 2021. <https://doi.org/10.6084/m9.figshare.12154833.v3>
- [31] Van Vlierberghe M, Di Franco A, Philippe H, Baurain D. Data set 4 - Taxonomic samplings. Figshare 2021. <https://doi.org/10.6084/m9.figshare.12401639.v1>
- [32] Van Vlierberghe M, Di Franco A, Philippe H, Baurain D. Data set 5 - GAPDH phylogenies. Figshare 2021. <https://doi.org/10.6084/m9.figshare.13096208.v2>
- [33] Huson, D. H., Auch, A. F., Qi, J., & Schuster, S. C. (2007). MEGAN analysis of metagenomic data. *Genome research*, 17(3), 377-386.
- [34] Cornet, L., Meunier, L., Van Vlierberghe, M., Leonard, R. R., Durieu, B., Lara, Y., ... & Wilmotte, A. (2018). Consensus assessment of the contamination level of publicly available cyanobacterial genomes. *PloS one*, 13(7).
- [35] Srivastava, A., Sarkar, H., Gupta, N., & Patro, R. (2016). RapMap: a rapid, sensitive and accurate tool for mapping RNA-seq reads to transcriptomes. *Bioinformatics*, 32(12), i192-i200.

5. Supplementary material

5.1. Data file 1 - Methods

The workflow (see Figures S1 and S2) described below was designed to reduce taxonomic redundancy among MMETSP samples while enhancing completeness and controlling the contamination level of the consolidated (i.e., combined and dereplicated) samples.

5.1.1. Sample selection and combination

5.1.1.1. MMETSP re-assembly pre-processing

All 678 transcriptome samples from the MMETSP re-assembly [9,10] were downloaded as FASTA files, along with the associated sequencing read (FASTQ) files. Transcriptomes with <5000 contigs were discarded from the analysis (MMETSP0044, MMETSP0186, MMETSP0223, MMETSP0225, MMETSP0229, MMETSP0252, MMETSP0329, MMETSP0372, MMETSP0439, MMETSP0451, MMETSP0452, MMETSP0716, MMETSP1147 and MMETSP1148), as well as two other samples (MMETSP1115: FASTQ file corrupted, MMETSP1475: unknown organism), yielding a total of 662 samples to enter our consolidation pipeline and quality survey. We renamed all the samples after their scientific binomial name and associated NCBI Taxonomy taxon identifier, in the form of `genus_species_MMETSPXXXX_taxid.fasta` thanks to our taxonomic tools (Bio::MUST::Core suite, available at <https://metacpan.org/release/Bio-MUST-Core>). Similarly, sequence identifiers were reformatted in the form of `>MMETSPXXXX_contig_id` to ensure universal software compatibility (e.g., BLAST, CD-HIT). The corresponding FASTQ files were also renamed accordingly.

5.1.1.2. Contamination detection - Forty-Two

We performed a contamination detection with Forty-Two (v0.202160) [12,13], a software whose original aim is to add (and optionally align) sequences to thousands of preexisting multiple sequence alignments (MSAs) while controlling for orthology relationships and potentially contaminating sequences. Forty-Two uses advanced heuristics based on a multiple Best Reciprocal (BLAST) Hit (multi-BRH) strategy against reference proteomes to distinguish orthologous from paralogous sequences among homologues. Nowadays, Forty-Two is routinely used in our labs as a stand-alone contamination detection tool to spot foreign sequences in transcriptomic or genomic data. To this end, Forty-Two comes with a set of 78 eukaryotic ribosomal protein MSAs, manually curated and continuously enriched with new species in H. Philippe's lab. These MSAs are available at <https://bitbucket.org/phylogeno/42-ribo-msas/>. For the full configuration, see Data set 1. In short, for each transcriptome, Forty-Two recovers ribosomal protein orthologs and then labels each one by computing the lowest common ancestor (LCA) of their closest relatives (best BLAST hits) in the

corresponding MSA (excluding self-matches), provided they have a bit-score ≥ 80 and are within 99% of the bit-score of the first hit (MEGAN-like algorithm [33,34]). Then, we used **debrief-42.pl** (available at <https://metacpan.org/release/Bio-MUST-Apps-FortyTwo>) to compute contamination percentages; the LCA computed by Forty-Two is compared with the source organism lineage (NCBI Taxonomy) and labeled as a contaminant if it differs at a predefined taxonomic rank. Contamination percentages resulting from this analysis were pooled according to the organisms phylum for box-plot visualization: Amoebozoa, Archaeplastida, Ciliophora, Colpodellida, Cryptophyceae, Dinophyceae, Euglenozoa, Haptophyta, Opisthokonta, Rhizaria, Stramenopiles and “other Eukaryotes”, with the latter group being an aggregate of Heterolobosea, *Palpitomonas* and Perkinsozoa. For each phylum, we considered outliers (i.e., those points further than 1.5x interquartile range from the upper box hinge) as transcriptomes too contaminated to be retained, thereby excluding 89 samples from subsequent consolidation.

5.1.1.3. Sample pooling

In order to reduce sample redundancy (taxonomically speaking) and maximize completeness, we used CroCo [11], a cross-contamination detection tool in a diverted fashion. Its optimal performance is obtained by working on relatively distant organisms (sequence divergence $>4\%$ [11]). Here, we used CroCo on closely related samples, i.e., at the genus level or even with identical strains. This way, we could consolidate (combine and dereplicate files) samples that “overlapped” too much in terms of genetic similarity. To determine which of the remaining 573 samples had to undergo combination, we grouped them by genus, resulting in 85 candidate clusters collectively composed of 449 samples, the other 124 samples belonging to unique genera and being called “singletons” in the following. We submitted each cluster to a cross-contamination analysis with CroCo [11] using default settings (fold-threshold = 2, minimum-coverage = 0.2, overexp = 300, suspect-id = 95, suspect-len = 40) to classify transcripts into “suspected” and “unsuspected” with respect to potential “cross-contamination”. Then, we used RapMap (v0.6.0) [35] for mapping the reads to the suspected transcripts. We chose to combine several samples of the same cluster (i.e., genus) together when their “cross-contamination” level was $\geq 10\%$. Otherwise, samples were combined at the species level, i.e., all strains from a given species were pooled together. This step resulted in 110 pooled samples (8 at genus level according to CroCo and 102 at species/strain level) and 26 uncombined samples (also considered as singletons from now on). In detail, 71 clusters yielded only one combination while the remaining 14 clusters yielded several combinations: 2 combinations (8 clusters), 3 combinations (4), 4 combinations (1) and 7 combinations (1). This way, out of the 449 candidates, we combined 423 samples into 110 while 26 were left uncombined (singletons), reducing taxonomic redundancy based on genuine sequence diversity, which along with the initial 124 singletons, yielded a total of 260 representative samples (see Data file 3).

5.1.2. Transcriptome dereplication and quality survey

To reduce transcript redundancy, especially after pooling multiple samples, we used CD-HIT-EST to derePLICATE the 260 combined samples. Then we characterized the consolidated data set through three quality control analyses: (i) a full-scale cross-contamination detection using Sobek, (ii) a second round of contamination detection using Forty-Two and (iii) a completeness analysis using BUSCO. We used the same criterion as before to discard the most contaminated samples, i.e., the removal of upper outliers. Beware that Sobek was run on transcriptomes prior to dereplication. Indeed, Croco's engine, used by Sobek, compares expression levels between transcriptomes in order to estimate cross-contamination levels [11]. Therefore, dereplicating transcriptomes and their associated sequence read (FASTQ) files would have crippled the functioning of Croco/Sobek.

5.1.2.1. Cross-contamination detection - Sobek

We submitted the 260 samples to a global purity analysis, i.e., aimed at identifying potential cross-contaminations across any of the combined MMETSP transcriptomes. As the original CroCo was designed to perform on a low number of samples, this required the development of a divide-and-conquer approach, notably due to a memory usage limitation affecting the read mapping step by RapMap [35]. Our new implementation of CroCo is called Sobek (see Figure S3 below). It benefits from parallel optimizations at multiple levels and relies on the splitting of the RapMap index (quasi-index) into 10 sub-indexes to solve the memory issue. This large-scale implementation allows us to perform a multiple-source cross-contamination detection, thereby pinpointing the lowest quality samples. Sobek was run with default settings, corresponding to the ones of CroCo, and the resulting cross-contamination values allowed us to pinpoint 31 over-contaminated combined samples (17) and singletons (14) that we suggest to discard. The Sobek parallel pipeline for SGE is available at <https://bitbucket.org/phylogeno/sobek/>.

5.1.2.2. Transcriptome dereplication

The transcriptomes corresponding to the 260 combined samples (including the singletons for consistency) were dereuplicated using CD-HIT-EST (v4.6) [15,16] with a sequence identity threshold of 95% (-c 0.95) and the accurate but slow mode (-g 1). This identity threshold was selected to stay in line with both CroCo and Sobek sensitivity settings.

5.1.2.3. Contamination detection - Forty-Two

We performed an additional contamination detection with Forty-Two, exactly as described at Section 1.2. According to our criterion (see above), it allowed us to identify 6 over-contaminated samples, either consolidated (3) or singletons (3). For the full configuration, see Data set 2. Based on the results of both

detection methods, we suggest discarding a total of 36 transcriptomes (1 sample being pointed out by both methods), hence leaving a total of 224 “clean” transcriptomes.

5.1.2.4. Completeness analysis - BUSCO

Third, transcriptome quality was assessed with BUSCO v4.0.5 [17,18] for completeness, using 255 single-copy orthologous eukaryotic markers (evaluate = 1e-03, lineage data = “eukaryota_odb10”, mode = transcriptome). The first goal of this test was to measure the impact of the consolidation on the samples and the second was to provide quality metrics for choosing the best transcriptomes for subsequent studies (Data file 7). Note that, for the contextual study described here, we chose to keep all transcriptomes as low completeness values does not impair the quality of our analyses.

5.2. Supplementary figures

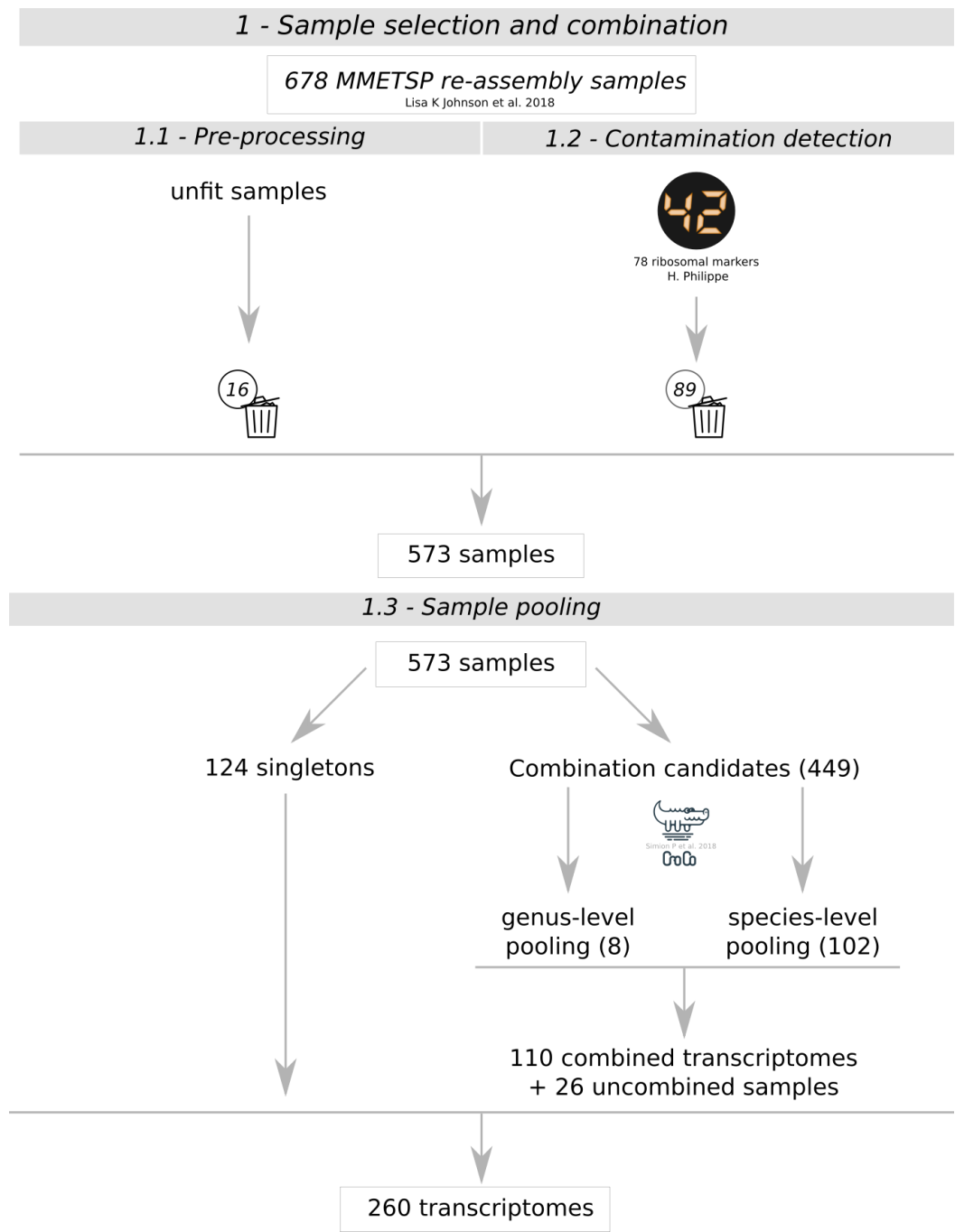


Figure S1: Graphical summary of the sample pooling strategy designed for this study.

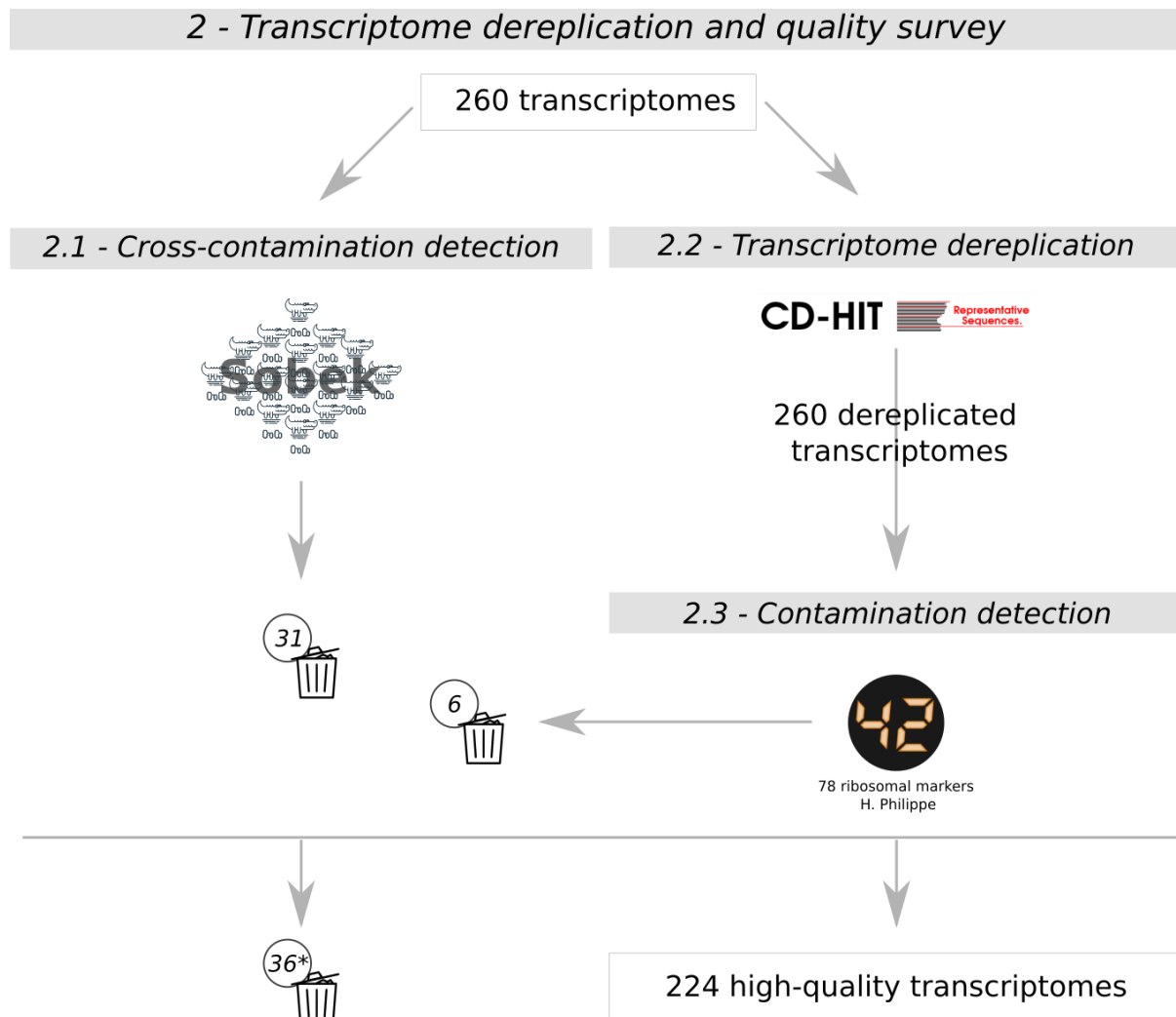


Figure S2: Graphical summary of the dereplication and contamination detection pipeline designed for this study. All 260 consolidated samples, including the $31+6 = 36^*$ (due to one sample flagged by both methods) that we suggest removing in our study, are made available in the data files described in this Data Note.

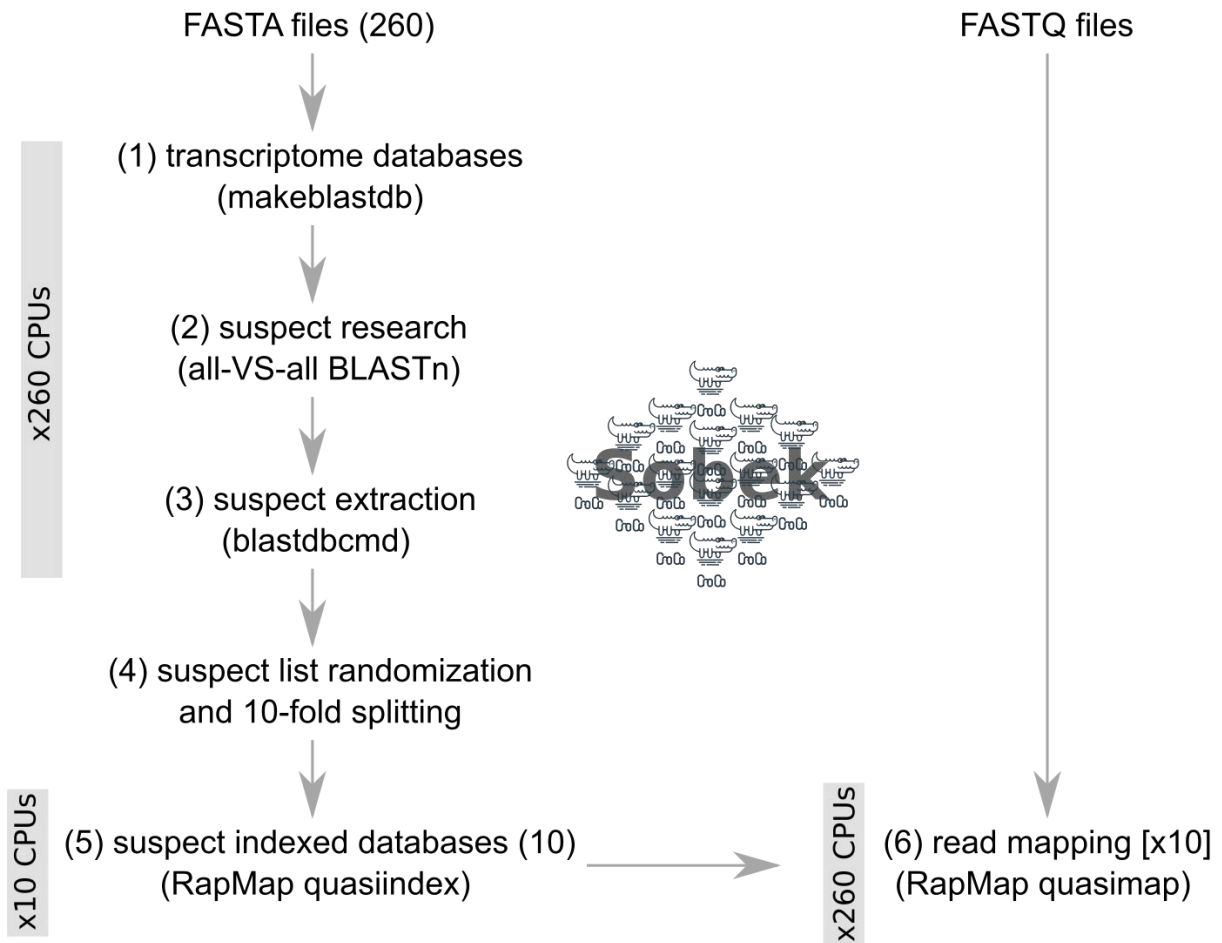


Figure S3: Overview of the divide-and-conquer strategy of Sobek. Running CroCo with 260 samples exceeded the memory capacity of our computing cluster, during the step caching the transcript database used for the read-to-transcript mapping by RapMap (rapmap quasimap). To solve this issue, we implemented a divide-and-conquer version of CroCo, termed Sobek. First, to directly overcome the memory shortage, we decided to split the index into 10 chunks, i.e., to partition the transcript list used to build the database list in 10 sub-lists (4). This approach required us to run CroCo 10 times for the set of 260 samples and ultimately, to consolidate the 10 mapping reports obtained for each transcriptome. Preliminary tests showed that random partitioning was the least biased yet reasonably reproducible strategy. Second, we parallelized several steps of a genuine CroCo run, i.e., (1) making BLAST (v2.7.1+) databases out of transcriptomes, (2) running an all-versus-all BLASTn with the transcriptomes, (3) extracting suspected transcripts from BLAST databases (blastdbcmd), (5) mapping each transcriptome to each sub-index (rapmap quasimap). As all four steps were embarrassingly parallel, we simply parallelized them based on the extraction of the corresponding commands from the original bash script. Mere wrapping into a Sun Grid Engine (SGE)-friendly interface allowed each job to run independently with a maximum of concurrent jobs equal to the number of samples at play, here 260. This pipeline has been fully automated and requires no additional step compared to the original CroCo software.

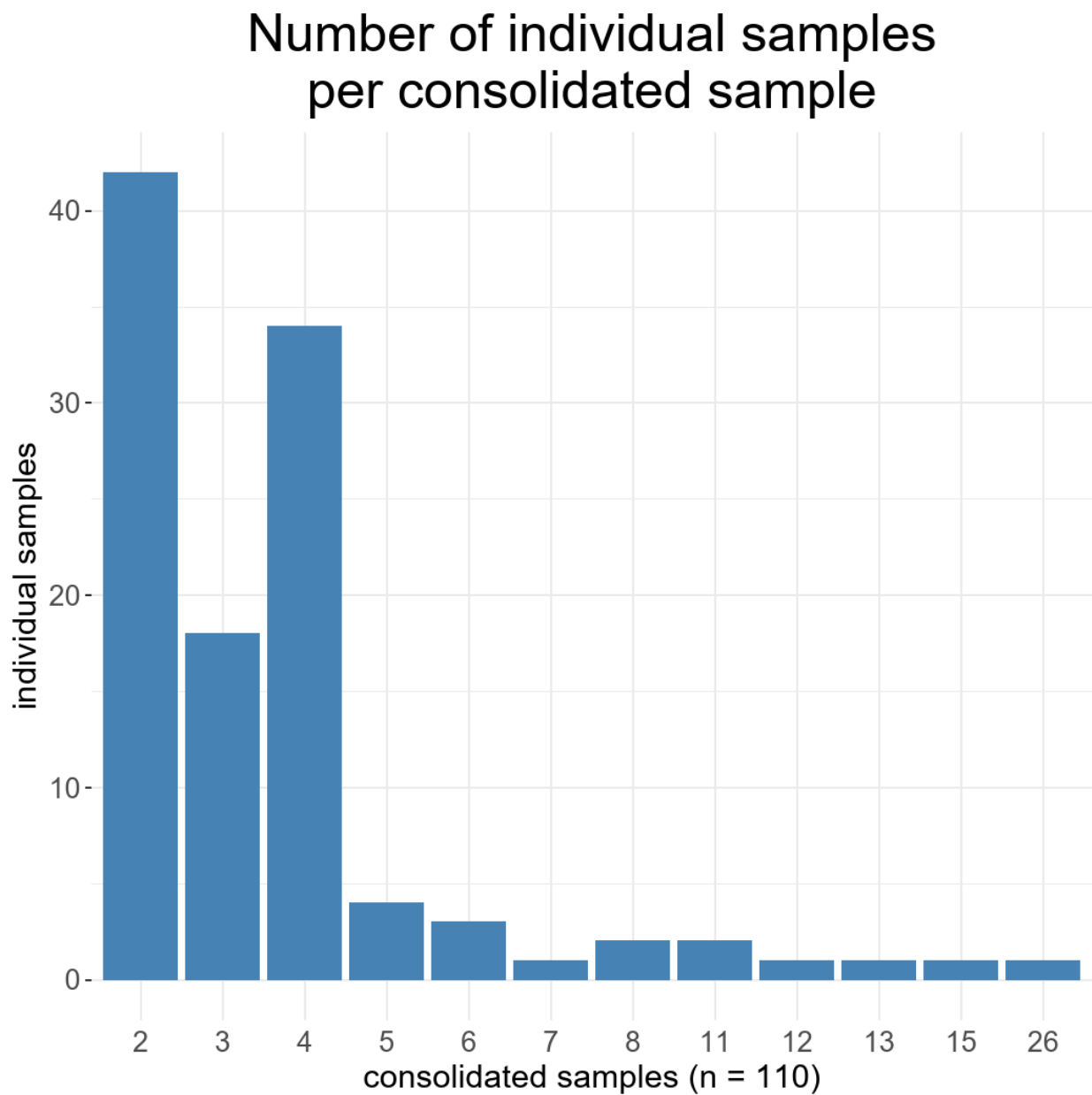


Figure S4: The consolidation pipeline we applied resulted in 110 consolidated samples. Most of them are made by the aggregation of 2 samples (42), 3 samples (18) and 4 samples (34). The remaining are constituted of 5 samples and more.

Sample purity (Sobek, n = 260)

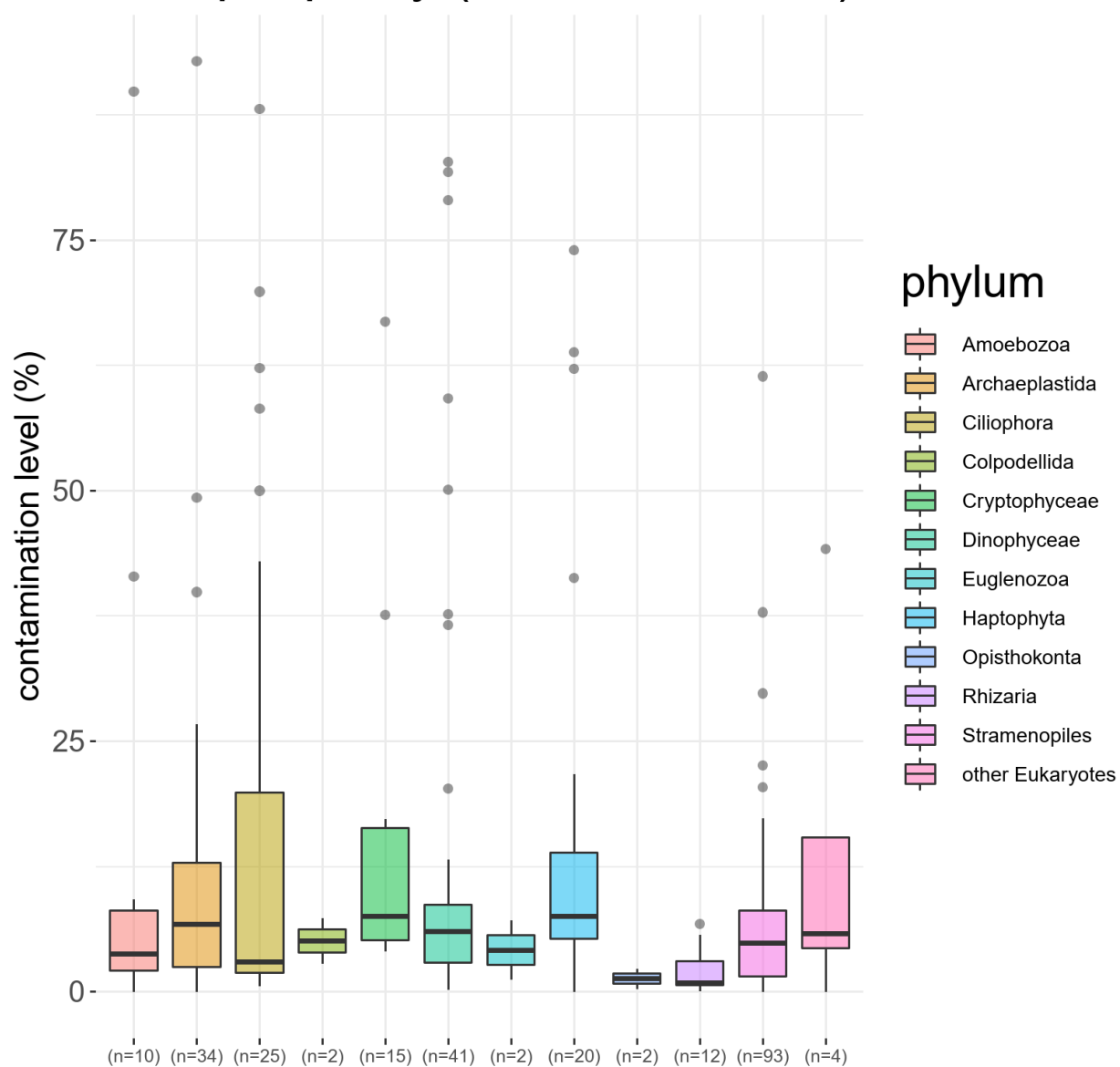


Figure S5: Data file 6

Sample purity (Forty-Two, n = 260)

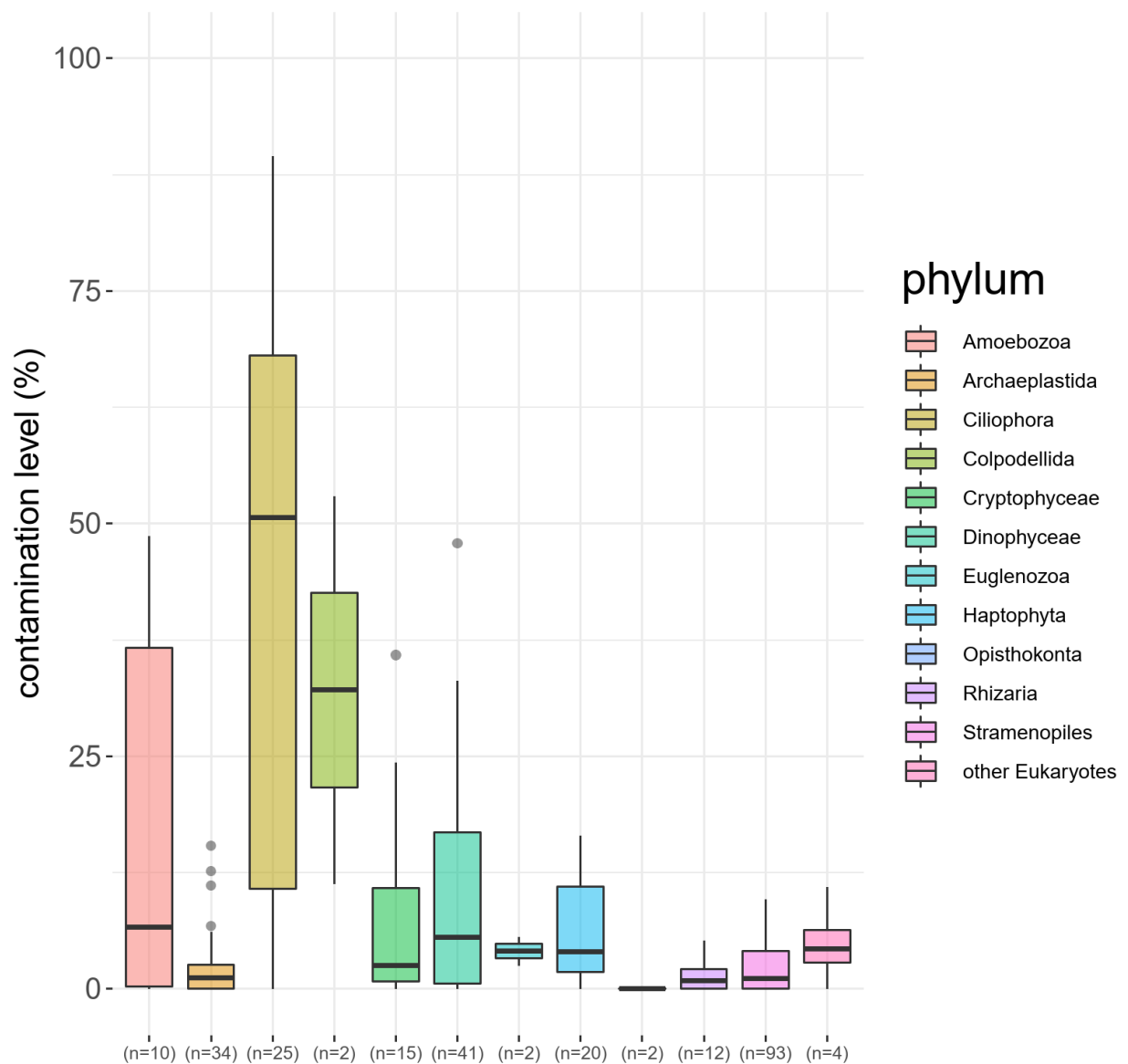


Figure S6 : Data file 7

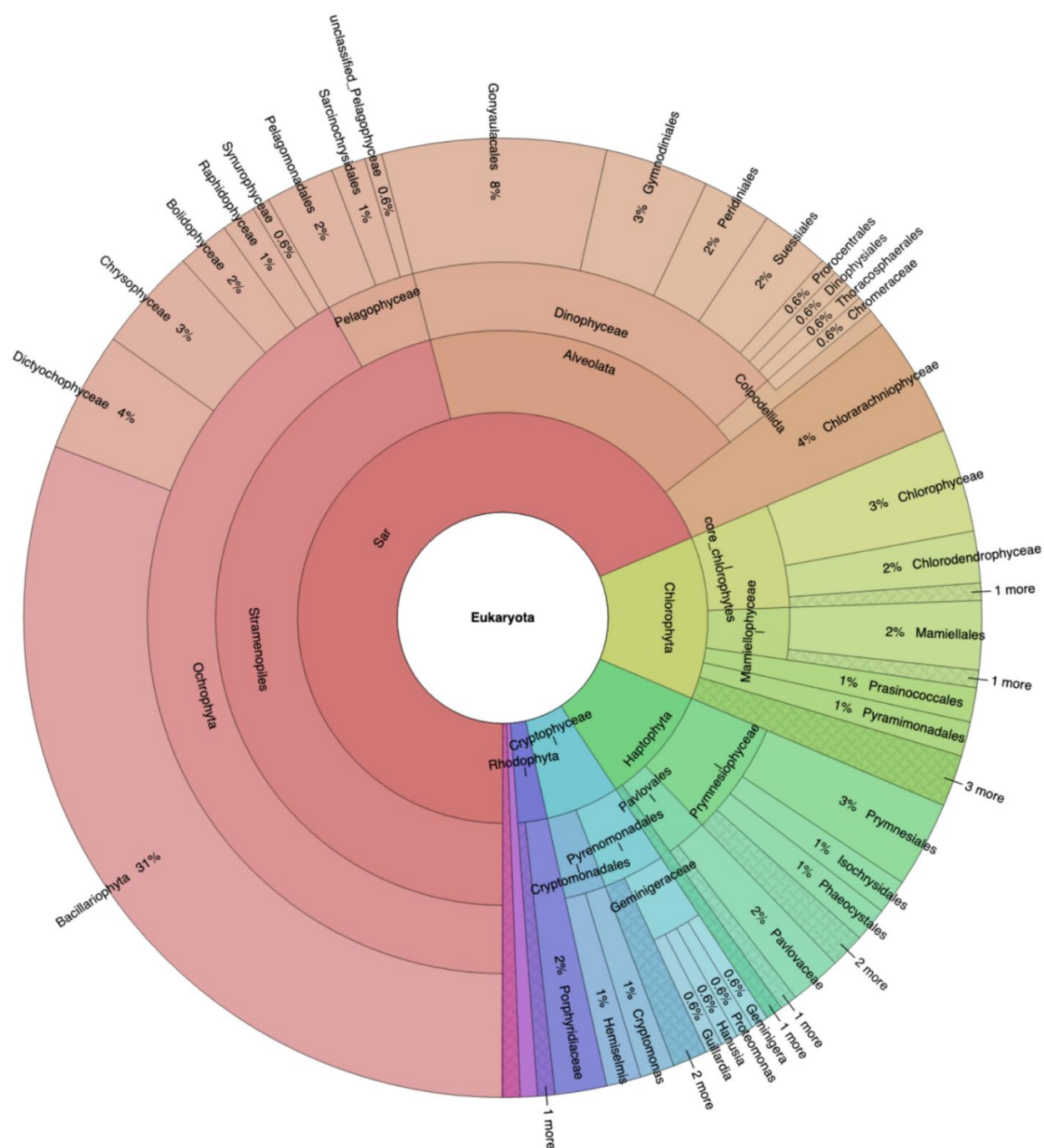


Figure S7: Data set 4.1 - Taxonomic sampling for the 224 "clean" samples. Interactive taxonomic visualization generated using Krona [Ondov BD, Bergman NH, Phillippy AM. Interactive metagenomic visualization in a Web browser. BMC Bioinformatics. 2011;12:385. doi:10.1186/1471-2105-12-385.]

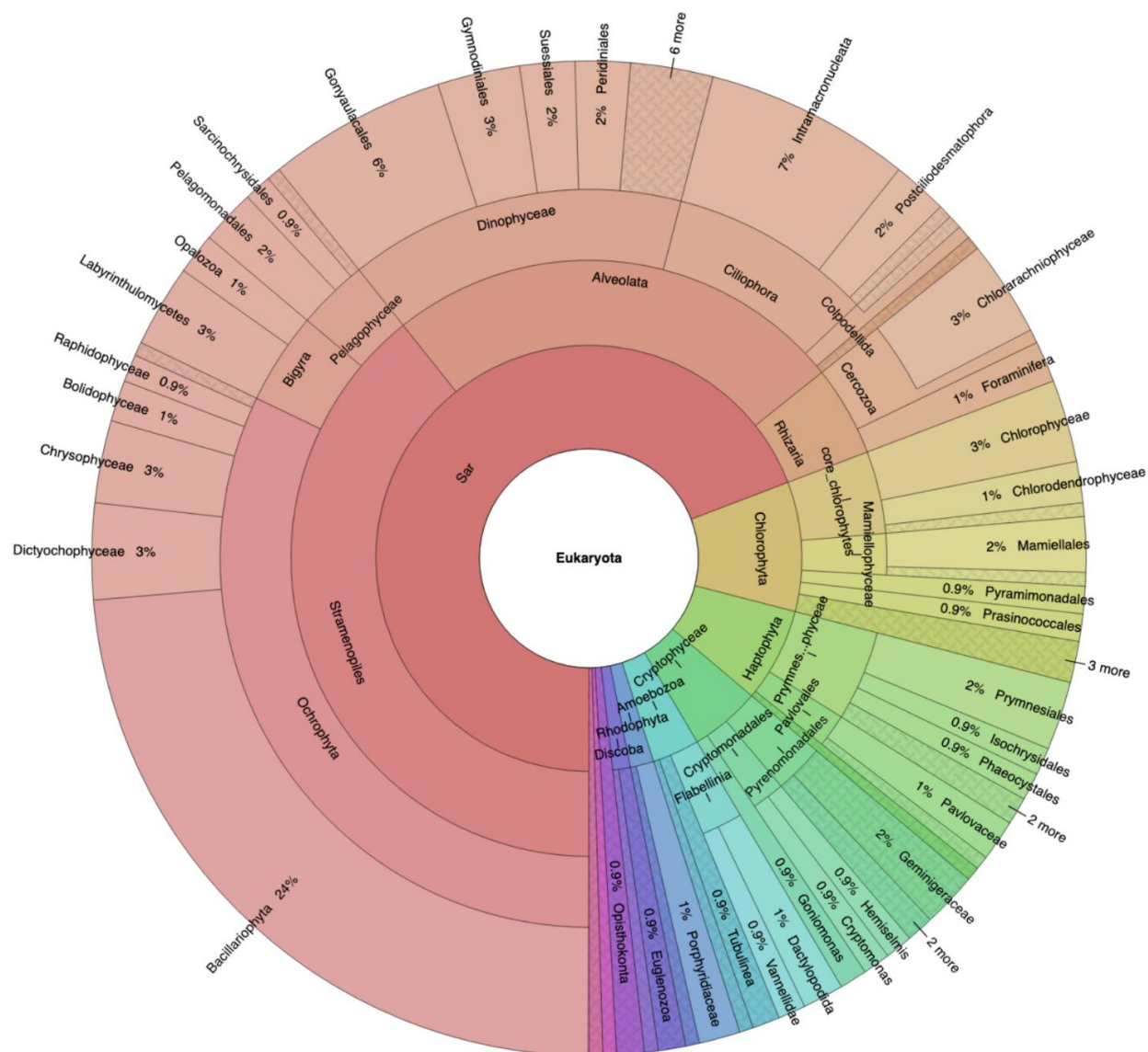


Figure S8: Data set 4.2 - Taxonomic sampling for the 172 photosynthetic "clean" samples. Interactive taxonomic visualization generated using Krona [Ondov BD, Bergman NH, Phillippy AM. Interactive metagenomic visualization in a Web browser. BMC Bioinformatics. 2011;12:385. doi:10.1186/1471-2105-12-385.]

5.3. Supplementary tables

cluster	cluster content	#input seqs	output transcriptome	#output seqs	#output seqs (cd-hit-est)	#derep percentage
cluster-1	Chromera velia_MMETSP0290_505693	56191	Chromera velia_MMETSP0290_505693	56191	53387	4,99
	Chromerida_sp_MMETSP0288_348535	45732				
	Vitrella_brassicaformis_MMETSP1451_1169540	54204	Chromerida_sp_877183	99936	58757	41,21
cluster-2	Anophryoides_haemophila_MMETSP1018_46462	17679				
	Anophryoides_haemophila_MMETSP1019_46462	21468	Anophryoides_haemophila_46462	39147	26423	32,50
cluster-3	Schmidingerella_taraikaensis_MMETSP0434_626734	48003				
	Schmidingerella_taraikaensis_MMETSP0436_626734	33913	Schmidingerella_taraikaensis_626734	81916	50627	38,20
cluster-4	Euplotes_focardii_MMETSP0205_36767	33076				
	Euplotes_focardii_MMETSP0206_36767	25681	Euplotes_focardii_36767	58757	41686	29,05
cluster-5	Azadinium_spinosum_MMETSP1036_632150	134290				
	Azadinium_spinosum_MMETSP1037_632150	137944				
	Azadinium_spinosum_MMETSP1038_632150	136736	Azadinium_spinosum_632150	408970	210879	48,44
cluster-6	Neoceratium_fusum_MMETSP1074_2916	137033				
	Neoceratium_fusum_MMETSP1075_2916	139430	Neoceratium_fusum_2916	276463	172376	37,65
cluster-7	Crypthecodinium_cohnii_MMETSP0323_2866	42298				
	Crypthecodinium_cohnii_MMETSP0324_2866	99510				
	Crypthecodinium_cohnii_MMETSP0325_2866	97130				
	Crypthecodinium_cohnii_MMETSP0326_2866	44293	Crypthecodinium_cohnii_2866	283231	127167	55,10
cluster-8	Lingulodinium_polyedrum_MMETSP1032_160621	123210				
	Lingulodinium_polyedrum_MMETSP1033_160621	120654				
	Lingulodinium_polyedrum_MMETSP1034_160621	121881				
	Lingulodinium_polyedrum_MMETSP1035_160621	119621	Lingulodinium_polyedrum_160621	485366	180234	62,87
cluster-9	Amphidinium_carterae_MMETSP0258_2961	58916				
	Amphidinium_carterae_MMETSP0259_2961	61314				
	Amphidinium_carterae_MMETSP0398_2961	9161				
	Amphidinium_carterae_MMETSP0399_2961	8890	Amphidinium_carterae_2961	138281	72203	47,79
	Amphidinium_massartii_MMETSP0689_160604	76432	Amphidinium_massartii_MMETSP0689_160604	76432	74099	3,05
cluster-10	Karenia_brevis_MMETSP0027_156230	162399				
	Karenia_brevis_MMETSP0029_156230	160017				
	Karenia_brevis_MMETSP0030_156230	163695				
	Karenia_brevis_MMETSP0031_156230	151377				
	Karenia_brevis_MMETSP0201_156230	169586				
	Karenia_brevis_MMETSP0202_156230	137008				
	Karenia_brevis_MMETSP0527_156230	157218				
	Karenia_brevis_MMETSP0528_156230	158863				
	Karenia_brevis_MMETSP0573_156230	164762				
	Karenia_brevis_MMETSP0574_156230	169234				
	Karenia_brevis_MMETSP0648_156230	155870				
	Karenia_brevis_MMETSP0649_156230	161176	Karenia_brevis_156230	1911205	378359	80,20
cluster-11	Karlodinium_veneticum_MMETSP1015_342587	125515				
	Karlodinium_veneticum_MMETSP1016_342587	127015				
	Karlodinium_veneticum_MMETSP1017_342587	124838	Karlodinium_veneticum_342587	377368	188037	50,17
cluster-12	Oxyrrhis_marina_MMETSP0471_2969	60232				
	Oxyrrhis_marina_MMETSP1424_2969	43252	Oxyrrhis_marina_2969	103484	100141	3,23
cluster-13	Durinskia_baltica_MMETSP0116_400756	128413				
	Durinskia_baltica_MMETSP0117_400756	129520	Durinskia_baltica_400756	257933	174216	32,46
cluster-14	Kryptoperidinium_foliaceum_MMETSP0118_160619	130753				
	Kryptoperidinium_foliaceum_MMETSP0119_160619	123721				
	Kryptoperidinium_foliaceum_MMETSP0120_160619	134492				
	Kryptoperidinium_foliaceum_MMETSP0121_160619	53496	Kryptoperidinium_foliaceum_160619	442462	313543	29,14
cluster-15	Prorocentrum_micans_MMETSP0251_2945	18649	Prorocentrum_micans_MMETSP0251_2945	18649	10046	46,13
	Prorocentrum_minimum_MMETSP0053_39449	119667				
	Prorocentrum_minimum_MMETSP0055_39449	109610				
	Prorocentrum_minimum_MMETSP0056_39449	129657				
	Prorocentrum_minimum_MMETSP0057_39449	132248				
	Prorocentrum_minimum_MMETSP0267_39449	111341				
	Prorocentrum_minimum_MMETSP0268_39449	68103				
	Prorocentrum_minimum_MMETSP0269_39449	115754	Prorocentrum_minimum_39449	786380	227338	71,09
cluster-16	Polarella_glacialis_MMETSP0227_89957	103802				
	Polarella_glacialis_MMETSP1440_89957	77363	Polarella_glacialis_89957	181165	119445	34,07
cluster-17	Peridinium_aciculiferum_MMETSP0370_268820	105955				
	Peridinium_aciculiferum_MMETSP0371_268820	107698	Peridinium_aciculiferum_268820	213653	134262	37,16
cluster-18	Perkinsus_chesapeakei_MMETSP0924_321610	31378				
	Perkinsus_chesapeakei_MMETSP0925_321610	29341	Perkinsus_chesapeakei_321610	60719	41361	31,88
	Perkinsus_marinus_MMETSP0923_1032511	26157	Perkinsus_marinus_MMETSP0923_1032511	26157	20565	21,38
cluster-19	Vannella_sp_MMETSP0166_95228	26006				
	Vannella_sp_MMETSP0168_95228	22740	Vannella_sp_95228	48746	43829	41,49
cluster-20	Goniomonas_sp_MMETSP0114_31324	36363	Goniomonas_sp_MMETSP0114_31324	36363	34955	3,87
	Goniomonas_pacifica_MMETSP0107_195067	66173				
	Goniomonas_pacifica_MMETSP0108_195067	76291	Goniomonas_pacifica_195067	142464	89982	36,84
cluster-21	Hemiselmis_andersenii_MMETSP0043_464988	63330				
	Hemiselmis_andersenii_MMETSP1041_464988	43207				
	Hemiselmis_andersenii_MMETSP1042_464988	50835				
	Hemiselmis_andersenii_MMETSP1043_464988	43557	Hemiselmis_andersenii_464988	200929	82386	59,00
	Hemiselmis_tepida_MMETSP1355_464990	38957	Hemiselmis_tepida_MMETSP1355_464990	38957	37611	3,46
cluster-22	Cryptophyceae_sp_MMETSP0986_697907	69530				
	Cryptophyceae_sp_MMETSP0987_697907	70994				
	Cryptophyceae_sp_MMETSP0988_697907	72388				
	Cryptophyceae_sp_MMETSP0989_697907	70975	Cryptophyceae_sp_697907	283887	117347	58,66
cluster-23	Eutreptiella_gymnastica_MMETSP0039_73025	154322				
	Eutreptiella_gymnastica_MMETSP0809_73025	70116				
	Eutreptiella_gymnastica_MMETSP0810_73025	89989				
	Eutreptiella_gymnastica_MMETSP0811_73025	110265	Eutreptiella_gymnastica_73025	424672	296196	30,25
cluster-24	Pleurochrysis_carterae_MMETSP1136_13221	60021				
	Pleurochrysis_carterae_MMETSP1137_13221	35988				
	Pleurochrysis_carterae_MMETSP1138_13221	45302	Pleurochrysis_carterae_13221	141311	72681	48,57

cluster-25	Emiliana_huxleyi_MMETSP0994_2903	41202			
	Emiliana_huxleyi_MMETSP0995_2903	40537			
	Emiliana_huxleyi_MMETSP0996_2903	40306			
	Emiliana_huxleyi_MMETSP0997_2903	41156			
	Emiliana_huxleyi_MMETSP1006_2903	29659			
	Emiliana_huxleyi_MMETSP1007_2903	29591			
	Emiliana_huxleyi_MMETSP1008_2903	35540			
	Emiliana_huxleyi_MMETSP1009_2903	27243			
	Emiliana_huxleyi_MMETSP1151_2903	50362			
	Emiliana_huxleyi_MMETSP1152_2903	52809			
	Emiliana_huxleyi_MMETSP1153_2903	51612			
	Emiliana_huxleyi_MMETSP1154_2903	52243			
	Emiliana_huxleyi_MMETSP1155_2903	52407			
	Emiliana_huxleyi_MMETSP1156_2903	51388			
	Emiliana_huxleyi_MMETSP1157_2903	52313	Emiliana_huxleyi_2903	648368	132312 79,59
cluster-26	Gephyrocapsa_oceanica_MMETSP1363_38817	49697			
	Gephyrocapsa_oceanica_MMETSP1364_38817	47723			
	Gephyrocapsa_oceanica_MMETSP1365_38817	51079			
	Gephyrocapsa_oceanica_MMETSP1366_38817	52229	Gephyrocapsa_oceanica_38817	200728	70163 65,05
cluster-27	Percolomonas_cosmopolitus_MMETSP0758_63605	14333			
	Percolomonas_cosmopolitus_MMETSP0759_63605	11692	Percolomonas_cosmopolitus_63605	26025	24420 6,17
cluster-28	Debaryomyces_hansenii_MMETSP0233_284592	9596			
	Debaryomyces_hansenii_MMETSP0234_284592	9454			
	Debaryomyces_hansenii_MMETSP0235_284592	9594	Debaryomyces_hansenii_284592	28986	12189 57,95
cluster-29	Aurantiochytrium_limacinum_MMETSP0958_717989	23524			
	Aurantiochytrium_limacinum_MMETSP0959_717989	26768			
	Aurantiochytrium_limacinum_MMETSP0960_717989	23557			
	Aurantiochytrium_limacinum_MMETSP0961_717989	22821	Aurantiochytrium_limacinum_717989	96670	31670 67,24
cluster-30	Bigelowiella_longifolia_MMETSP1359_552665	35304	Bigelowiella_longifolia_MMETSP1359_552665	35304	32064 9,18
	Bigelowiella_natans_MMETSP1052_227086	48484			
	Bigelowiella_natans_MMETSP1054_227086	43299			
	Bigelowiella_natans_MMETSP1055_227086	47299			
	Bigelowiella_natans_MMETSP1358_227086	36004			
	Bigelowiella_natans_MMETSP0045_753081	17857	Bigelowiella_natans_227086	192943	71921 62,72
cluster-31	Lotharella_globosa_MMETSP0041_91324	41011			
	Lotharella_globosa_MMETSP0111_91324	39306			
	Lotharella_globosa_MMETSP0112_91324	23258	Lotharella_globosa_91324	103575	57591 44,40
	Lotharella_oceanica_MMETSP0040_641309	37737	Lotharella_oceanica_MMETSP0040_641309	37737	35631 5,58
cluster-32	Rhodella_violacea_MMETSP0167_38298	29515			
	Rhodella_violacea_MMETSP0314_38298	25394	Rhodella_violacea_38298	54909	32428 40,94
cluster-33	Rhodorus_marinus_MMETSP0011_101924	45875			
	Rhodorus_marinus_MMETSP0315_101924	25364	Rhodorus_marinus_101924	71239	55659 21,87
cluster-34	Fragilariopsis_kerguelensis_MMETSP0733_186038	64661			
	Fragilariopsis_kerguelensis_MMETSP0734_186038	65906			
	Fragilariopsis_kerguelensis_MMETSP0735_186038	64090			
	Fragilariopsis_kerguelensis_MMETSP0736_186038	69704			
	Fragilariopsis_kerguelensis_MMETSP0906_186038	59819			
	Fragilariopsis_kerguelensis_MMETSP0907_186038	57054			
	Fragilariopsis_kerguelensis_MMETSP0908_186038	54452			
	Fragilariopsis_kerguelensis_MMETSP0909_186038	52264	Fragilariopsis_kerguelensis_186038	487950	205005 57,99
cluster-35	Tryblionella_compressa_MMETSP0744_913974	30737			
	Tryblionella_compressa_MMETSP0745_913974	28275			
	Tryblionella_compressa_MMETSP0746_913974	33793			
	Tryblionella_compressa_MMETSP0747_913974	28743	Tryblionella_compressa_913974	121548	43614 64,12
cluster-36	Amphiprora_sp_MMETSP0724_265536	30127			
	Amphiprora_sp_MMETSP0725_265536	32234			
	Amphiprora_sp_MMETSP0726_265536	23679			
	Amphiprora_sp_MMETSP0727_265536	29922	Amphiprora_sp_265536	115962	37136 67,98
	Amphiprora_paludosa_MMETSP1065_265537	41105	Amphiprora_paludosa_MMETSP1065_265537	41105	34999 14,85
cluster-37	Amphora_coffeiformis_MMETSP0316_265554	19655			
	Amphora_coffeiformis_MMETSP0317_265554	18385	Amphora_coffeiformis_265554	38040	23314 38,71
cluster-38	Leptocylindrus_danicus_MMETSP0321_163516	34549			
	Leptocylindrus_danicus_MMETSP0322_163516	22930			
	Leptocylindrus_danicus_MMETSP1362_163516	29356	Leptocylindrus_danicus_163516	86835	73189 15,71
cluster-39	Corethron_hystrix_MMETSP0010_216773	43375	Corethron_hystrix_MMETSP0010_216773	43375	35789 17,49
	Corethron_pennatum_MMETSP0169_218684	33597			
	Corethron_pennatum_MMETSP0171_218684	50958	Corethron_pennatum_218684	84555	76197 9,88
cluster-40	Proboscia_alata_MMETSP0174_216777	64653			
	Proboscia_alata_MMETSP0176_216777	65912	Proboscia_sp_216776	130565	78150 40,14
cluster-41	Skeletonema_grethae_MMETSP0578_267565	29138	Skeletonema_grethae_MMETSP0578_267565	29138	26002 10,76
	Skeletonema_costatum_MMETSP0013_2843	38251			
	Skeletonema_japonicum_MMETSP0503_267983	31357			
	Skeletonema_dohrnii_MMETSP0562_267566	39369			
	Skeletonema_dohrnii_MMETSP0563_267566	34864			
	Skeletonema_marinoi_MMETSP0319_267567	33129			
	Skeletonema_marinoi_MMETSP0320_267567	34068			
	Skeletonema_marinoi_MMETSP0918_267567	32428			
	Skeletonema_marinoi_MMETSP0920_267567	33308			
	Skeletonema_marinoi_MMETSP1039_267567	30079			
	Skeletonema_marinoi_MMETSP1040_267567	26989			
	Skeletonema_marinoi_MMETSP1428_267567	33141	Skeletonema_sp_2842	366983	134800 63,27
cluster-42	Asterionellopsis_glacialis_MMETSP0705_33640	24748			
	Asterionellopsis_glacialis_MMETSP0706_33640	30995			
	Asterionellopsis_glacialis_MMETSP0707_33640	31972			
	Asterionellopsis_glacialis_MMETSP0708_33640	24762			
	Asterionellopsis_glacialis_MMETSP0713_33640	29474			
	Asterionellopsis_glacialis_MMETSP1394_33640	29587	Asterionellopsis_glacialis_33640	171538	103604 39,60
cluster-43	Thalassionema_frauenfeldii_MMETSP0786_186022	33504	Thalassionema_frauenfeldii_MMETSP0786_186022	33504	31236 6,77
	Thalassionema_nitzschoides_MMETSP0156_33649	32947			
	Thalassionema_nitzschoides_MMETSP0158_33649	32424			
	Thalassionema_nitzschoides_MMETSP0693_33649	9760	Thalassionema_nitzschoides_33649	75131	36381 51,58

cluster-44	Thalassiothrix_antarctica_MMETSP0152_1049557	39510			
	Thalassiothrix_antarctica_MMETSP0154_1049557	44023	Thalassiothrix_antarctica_1049557	83533	51713 38,09
cluster-45	Extubocellulus_spinifer_MMETSP0696_265572	38082			
	Extubocellulus_spinifer_MMETSP0697_265572	39558			
	Extubocellulus_spinifer_MMETSP0698_265572	38813			
	Extubocellulus_spinifer_MMETSP0699_265572	40222	Extubocellulus_spinifer_265572	156675	53854 65,63
cluster-46	Minutocellus_polymorphus_MMETSP1070_265543	34797			
	Minutocellus_polymorphus_MMETSP1322_265543	30920			
	Minutocellus_polymorphus_MMETSP1434_265543	20091	Minutocellus_polymorphus_265543	85808	65401 23,78
cluster-47	Ditylum_brightwellii_MMETSP0998_49249	39071			
	Ditylum_brightwellii_MMETSP1001_49249	33934			
	Ditylum_brightwellii_MMETSP1002_49249	45855			
	Ditylum_brightwellii_MMETSP1010_49249	37871			
	Ditylum_brightwellii_MMETSP1062_49249	42787			
	Ditylum_brightwellii_MMETSP1063_49249	40929	Ditylum_brightwellii_49249	240447	75186 68,73
cluster-48	Bolidomonas_sp_MMETSP1320_91990	26882			
	Bolidomonas_sp_MMETSP1321_91990	29766	Bolidomonas_sp_91990	56648	53441 5,66
cluster-49	Ochromonas_sp_MMETSP0004_420556	33559			
	Ochromonas_sp_MMETSP0005_420556	32824	Ochromonas_sp_420556	66383	38732 41,65
cluster-50	Dinobryon_sp_MMETSP0019_98059	34469			
	Dinobryon_sp_MMETSP0020_98059	34369			
	Dinobryon_sp_MMETSP0812_98059	32473	Dinobryon_sp_98059	101311	42970 57,59
cluster-51	Paraphysomonas_sp_MMETSP1103_36769	27001	Paraphysomonas_sp_MMETSP1103_36769	27001	25548 5,38
	Paraphysomonas_imperforata_MMETSP0103_37642	38037			
	Paraphysomonas_imperforata_MMETSP0104_37642	29752	Paraphysomonas_imperforata_37642	67789	42904 36,71
	Paraphysomonas_vestita_MMETSP1107_36770	25172	Paraphysomonas_vestita_MMETSP1107_36770	25172	24464 2,81
cluster-52	Pseudopedinella_elastica_MMETSP1068_35684	71329			
	Pseudopedinella_elastica_MMETSP1097_35684	19296	Pseudopedinella_elastica_35684	90625	71784 20,79
cluster-53	Pedinellales_sp_MMETSP0992_483369	75372			
	Pedinellales_sp_MMETSP0993_483369	89404	Pedinellales_sp_483369	164776	100601 38,95
cluster-54	Schizochytrium_aggregatum_MMETSP0962_4773	10585			
	Schizochytrium_aggregatum_MMETSP0963_4773	21116			
	Schizochytrium_aggregatum_MMETSP0964_4773	27026			
	Schizochytrium_aggregatum_MMETSP0965_4773	26520	Schizochytrium_aggregatum_4773	85247	33227 61,02
cluster-55	Thraustochytrium_sp_MMETSP0198_1112570	24096			
	Thraustochytrium_sp_MMETSP0199_1112570	24103	Thraustochytrium_sp_1112570	48199	26790 44,42
cluster-56	Aureococcus_anophagefferens_MMETSP0914_44056	37411			
	Aureococcus_anophagefferens_MMETSP0915_44056	33212			
	Aureococcus_anophagefferens_MMETSP0916_44056	32727			
	Aureococcus_anophagefferens_MMETSP0917_44056	32246	Aureococcus_anophagefferens_44056	135596	51230 62,22
cluster-57	Aureombra_lagunensis_MMETSP0890_44058	24449			
	Aureombra_lagunensis_MMETSP0892_44058	25448			
	Aureombra_lagunensis_MMETSP0893_44058	26579	Aureombra_lagunensis_44058	76476	33309 56,45
cluster-58	Pelagococcus_subviridis_MMETSP0882_35679	37257			
	Pelagococcus_subviridis_MMETSP0883_35679	35078			
	Pelagococcus_subviridis_MMETSP0884_35679	35880			
	Pelagococcus_subviridis_MMETSP0885_35679	36878	Pelagococcus_subviridis_35679	145093	55619 61,67
cluster-59	Pelagomonas_calceolata_MMETSP0886_35677	24257			
	Pelagomonas_calceolata_MMETSP0887_35677	26269			
	Pelagomonas_calceolata_MMETSP0888_35677	27756			
	Pelagomonas_calceolata_MMETSP0889_35677	24178			
	Pelagomonas_calceolata_MMETSP1328_35677	27845	Pelagomonas_calceolata_35677	130305	57973 55,51
cluster-60	Pelagophyceae_sp_MMETSP0974_483370	34446			
	Pelagophyceae_sp_MMETSP0975_483370	32915			
	Pelagophyceae_sp_MMETSP0977_483370	32790	Pelagophyceae_sp_483370	100151	43162 56,90
cluster-61	Chattonella_subsalsala_MMETSP0947_44440	41406			
	Chattonella_subsalsala_MMETSP0948_44440	39613			
	Chattonella_subsalsala_MMETSP0949_44440	39051			
	Chattonella_subsalsala_MMETSP0950_44440	41979	Chattonella_subsalsala_44440	162049	63423 60,86
cluster-62	Heterosigma_akashiwo_MMETSP0894_536046	30847			
	Heterosigma_akashiwo_MMETSP0895_536046	27533			
	Heterosigma_akashiwo_MMETSP0896_536046	22791			
	Heterosigma_akashiwo_MMETSP0897_536046	25548			
	Heterosigma_akashiwo_MMETSP0414_536047	45862			
	Heterosigma_akashiwo_MMETSP0415_536047	49680			
	Heterosigma_akashiwo_MMETSP0416_536047	59619			
	Heterosigma_akashiwo_MMETSP0292_536047	44889			
	Heterosigma_akashiwo_MMETSP0294_536047	52840			
	Heterosigma_akashiwo_MMETSP0295_536047	45335			
	Heterosigma_akashiwo_MMETSP0296_536047	46390	Heterosigma_akashiwo_2829	451334	126315 72,01
cluster-63	Tetraselmis_chuii_MMETSP0491_63592	43997	Tetraselmis_chuii_MMETSP0491_63592	43997	41067 6,66
	Tetraselmis_striata_MMETSP0817_3165	45641			
	Tetraselmis_striata_MMETSP0818_3165	47874			
	Tetraselmis_striata_MMETSP0819_3165	48474			
	Tetraselmis_striata_MMETSP0820_3165	48764	Tetraselmis_striata_3165	190753	73138 61,66
	Tetraselmis_sp_MMETSP0419_582737	43453	Tetraselmis_sp_MMETSP0419_582737	43453	39702 8,63
cluster-64	Dunaliella_tertiolecta_MMETSP1126_3047	29944			
	Dunaliella_tertiolecta_MMETSP1128_3047	29099	Dunaliella_tertiolecta_3047	59043	33087 43,96
cluster-65	Bathycoccus_prasinus_MMETSP1399_41875	8931			
	Bathycoccus_prasinus_MMETSP1460_41875	11309	Bathycoccus_prasinus_41875	20240	18460 8,79
cluster-66	Ostreococcus_sp_MMETSP0929_548131	9005			
	Ostreococcus_sp_MMETSP0930_548131	9482			
	Ostreococcus_sp_MMETSP0932_548131	4990			
	Ostreococcus_sp_MMETSP0933_548131	9267			
	Ostreococcus_sp_MMETSP0936_548131	9258			
	Ostreococcus_sp_MMETSP0937_548131	9289			
	Ostreococcus_sp_MMETSP0938_548131	9193			

	Ostreococcus_sp_MMETSP0939_548131	8659	Ostreococcus_sp_548131	69143	27856	59.71
cluster-67	Mantoniella_antarctica_MMETSP1106_81844	39052				
	Mantoniella_sp_MMETSP1468_1461545	47261	Mantoniella_sp_13607	86313	62034	28.13
cluster-68	Micromonas_pusilla_MMETSP1080_38833	16277				
	Micromonas_pusilla_MMETSP1082_38833	12990				
	Micromonas_pusilla_MMETSP1084_38833	14898				
	Micromonas_pusilla_MMETSP1327_38833	15978				
	Micromonas_pusilla_MMETSP1366_38833	12696				
	Micromonas_pusilla_MMETSP1387_38833	13796				
	Micromonas_pusilla_MMETSP1390_38833	14410				
	Micromonas_pusilla_MMETSP1393_38833	16326				
	Micromonas_pusilla_MMETSP1400_38833	14968				
	Micromonas_pusilla_MMETSP1401_38833	15284				
	Micromonas_pusilla_MMETSP1402_38833	15692				
	Micromonas_pusilla_MMETSP1403_38833	15140				
	Micromonas_pusilla_MMETSP1404_38833	16116	Micromonas_pusilla_38833	194571	94557	51.40
cluster-69	Picocystis_salinarum_MMETSP0807_88271	11627				
	Picocystis_salinarum_MMETSP1159_88271	10098	Picocystis_salinarum_88271	21725	10134	53.35
cluster-70	prasinophyte_sp_MMETSP1309_464262	15611				
	prasinophyte_sp_MMETSP1311_464262	14515				
	prasinophyte_sp_MMETSP1312_464262	10749	Chloropicon_sp_2302914	53094	39047	4.47
cluster-71	Pycnococcus_sp_MMETSP1085_41879	10979	Pycnococcus_sp_MMETSP1085_41879	10979	9255	15.70
	Pycnococcus_provasoli_MMETSP1316_41880	17265				
	Pycnococcus_provasoli_MMETSP1459_41880	15204				
	Pycnococcus_provasoli_MMETSP1471_41880	14502	Pycnococcus_provasoli_41880	46971	24716	47.38
cluster-72	Alexandrium_andersonii_MMETSP1436_327968	104672	Alexandrium_andersonii_MMETSP1436_327968	104672	102034	2.52
	Alexandrium_catenella_MMETSP0790_2925	118407	Alexandrium_catenella_MMETSP0790_2925	118407	113628	4.04
	Alexandrium_fundyense_MMETSP0198_2932	3125				
	Alexandrium_fundyense_MMETSP0197_2932	21740				
	Alexandrium_fundyense_MMETSP0347_2932	28633	Alexandrium_fundyense_2932	53498	36493	31.79
	Alexandrium_margalefi_MMETSP0661_109239	160604	Alexandrium_margalefi_MMETSP0661_109239	160604	140398	12.58
	Alexandrium_minutum_MMETSP0328_39455	67858	Alexandrium_minutum_MMETSP0328_39455	67858	41337	39.08
	Alexandrium_monilatum_MMETSP0093_311494	110519				
	Alexandrium_monilatum_MMETSP0095_311494	112256				
	Alexandrium_monilatum_MMETSP0096_311494	109825				
	Alexandrium_monilatum_MMETSP0097_311494	116587	Alexandrium_monilatum_311494	449187	165271	63.21
	Alexandrium_tamarense_MMETSP0378_2926	42220				
	Alexandrium_tamarense_MMETSP0380_2926	38153				
	Alexandrium_tamarense_MMETSP0382_2926	126886	Alexandrium_tamarense_2926	207259	140957	31.99
cluster-73	Symbiodinium_kawaguti_MMETSP0132_104179	16419				
	Symbiodinium_kawaguti_MMETSP0133_104179	15572				
	Symbiodinium_kawaguti_MMETSP0134_104179	17666				
	Symbiodinium_kawaguti_MMETSP0135_104179	25355	Symbiodinium_kawaguti_104179	75012	40659	45.80
	Symbiodinium_sp_MMETSP1116_631057	68167				
	Symbiodinium_sp_MMETSP1117_631057	64980				
	Symbiodinium_sp_MMETSP1110_174948	138369	Symbiodinium_sp_2949	271516	206162	24.07
	Symbiodinium_sp_MMETSP1122_230996	73822				
	Symbiodinium_sp_MMETSP1123_230996	64145				
	Symbiodinium_sp_MMETSP1124_230996	63088				
	Symbiodinium_sp_MMETSP1125_230996	68385	Symbiodinium_sp_230996	269238	95210	64.64
cluster-74	Scrippsiella_cf_MMETSP0367_447022	106068				
	Scrippsiella_cf_MMETSP0368_447022	107051				
	Scrippsiella_cf_MMETSP0369_447022	115322				
	Scrippsiella_hangoi_MMETSP0359_268821	178850				
	Scrippsiella_hangoi_MMETSP0360_268821	108509				
	Scrippsiella_hangoi_MMETSP0361_268821	184980	Scrippsiella_sp_54902	800780	259426	67.60
	Scrippsiella_trochoidea_MMETSP0270_71861	149312				
	Scrippsiella_trochoidea_MMETSP0271_71861	153198				
	Scrippsiella_trochoidea_MMETSP0272_71861	149749	Scrippsiella_trochoidea_71861	452259	215574	52.33
cluster-75	Rhodomonas_lens_MMETSP0484_354590	64603				
	Rhodomonas_salina_MMETSP1047_52970	63132	Rhodomonas_sp_37096	127735	87596	31.42
	Rhodomonas_sp_MMETSP1091_302021	51198				
	Rhodomonas_sp_MMETSP1389_302021	54048	Rhodomonas_sp_302021	105246	70278	33.23
cluster-76	Isochrysis_sp_MMETSP1090_37098	39455				
	Isochrysis_sp_MMETSP1388_37098	51823	Isochrysis_sp_37098	91278	54201	40.62
	Isochrysis_galbana_MMETSP0595_37099	32131				
	Isochrysis_galbana_MMETSP0943_37099	38099	Isochrysis_galbana_37099	70221	66992	4.60
	Isochrysis_sp_MMETSP1129_40639	28337				
	Isochrysis_sp_MMETSP1130_40639	16994				
	Isochrysis_sp_MMETSP1131_40639	15320				
	Isochrysis_sp_MMETSP1132_40639	31184	Isochrysis_sp_40639	91835	38429	58.15
cluster-77	Pavlova_gyrans_MMETSP1466_44452	29304	Pavlova_gyrans_MMETSP1466_44452	29304	27572	5.91
	Pavlova_sp_MMETSP1139_298111	25070				
	Pavlova_sp_MMETSP1140_298111	30485				
	Pavlova_sp_MMETSP1381_298111	21667	Pavlova_sp_298111	77231	36942	52.17
	Pavlovaes_sp_MMETSP0982_483367	51740				
	Pavlovaes_sp_MMETSP0983_483367	50372				
	Pavlovaes_sp_MMETSP0984_483367	45204				
	Pavlovaes_sp_MMETSP0985_483367	44622	Pavlovaes_sp_483367	191938	73028	61.95
cluster-78	Phaeocystis_sp_MMETSP1162_33656	39133				
	Phaeocystis_sp_MMETSP1178_33656	25477	Phaeocystis_sp_33656	64610	58030	10.18
	Phaeocystis_antarctica_MMETSP1100_33657	83790				
	Phaeocystis_antarctica_MMETSP1444_33657	39194	Phaeocystis_antarctica_33657	122984	94151	23.44
cluster-79	Prymnesium_parvum_MMETSP0006_97485	47463				
	Prymnesium_parvum_MMETSP0007_97485	30822				
	Prymnesium_parvum_MMETSP0814_97485	32869				
	Prymnesium_parvum_MMETSP1083_97485	44417	Prymnesium_parvum_97485	155671	64341	58.67
	Prymnesium_polylepis_MMETSP0143_72548	95064				
	Prymnesium_polylepis_MMETSP0145_72548	68584				
	Prymnesium_polylepis_MMETSP0146_72548	76269				
	Prymnesium_polylepis_MMETSP0147_72548	76587				
	Prymnesium_polylepis_MMETSP0286_72548	43188	Prymnesium_polylepis_72548	359692	149189	58.52
cluster-80	Pseudo-nitzschia_australis_MMETSP0139_44445	28966				
	Pseudo-nitzschia_australis_MMETSP0140_44445	29277				
	Pseudo-nitzschia_australis_MMETSP0141_44445	28457				
	Pseudo-nitzschia_australis_MMETSP0142_44445	28040	Pseudo-nitzschia_australis_44445	114740	40335	64.85

	Pseudo-nitzschia_delicatissima_MMETSP1432_44447	21996	Pseudo-nitzschia_delicatissima_MMETSP1432_44447	21996	20568	6,49
	Pseudo-nitzschia_fraudulenta_MMETSP0850_183588	27679				
	Pseudo-nitzschia_fraudulenta_MMETSP0852_183588	30776	Pseudo-nitzschia_fraudulenta_183588	58455	34234	57,45
	Pseudo-nitzschia_heimi_MMETSP1423_476441	28036	Pseudo-nitzschia_heimi_MMETSP1423_476441	28036	26377	5,92
	Pseudo-nitzschia_pungens_MMETSP1060_37318	29381				
	Pseudo-nitzschia_pungens_MMETSP1061_37318	34725	Pseudo-nitzschia_pungens_37318	64106	41323	35,54
cluster-81	Chaetoceros_affinis_MMETSP0088_426623	7970				
	Chaetoceros_affinis_MMETSP0090_426623	27381				
	Chaetoceros_affinis_MMETSP0091_426623	25981				
	Chaetoceros_affinis_MMETSP0092_426623	9548	Chaetoceros_affinis_426623	70880	31273	55,88
	Chaetoceros_curvisetus_MMETSP0717_230516	14536				
	Chaetoceros_curvisetus_MMETSP0718_230516	9747				
	Chaetoceros_curvisetus_MMETSP0719_230516	7884	Chaetoceros_curvisetus_230516	32167	21563	61,27
	Chaetoceros_sp_MMETSP1429_49237	23506				
	Chaetoceros_debilis_MMETSP0149_122233	27809				
	Chaetoceros_brevis_MMETSP1435_426625	8658				
	Chaetoceros_debilis_MMETSP0190_122233	30465	Chaetoceros_sp_49237	90438	53934	19,42
	Chaetoceros_dichaeta_MMETSP1447_420782	37747	Chaetoceros_dichaeta_MMETSP1447_420782	37747	34140	9,56
	Chaetoceros_neogracilis_MMETSP0751_240364	28421				
	Chaetoceros_neogracilis_MMETSP0752_240364	33789				
	Chaetoceros_neogracilis_MMETSP0753_240364	31834				
	Chaetoceros_neogracilis_MMETSP0754_240364	34999				
	Chaetoceros_neogracilis_MMETSP1336_240364	28868	Chaetoceros_neogracilis_240364	157911	69873	55,75
cluster-82	Thalassiosira_weissflogi_MMETSP0878_67004	31740				
	Thalassiosira_weissflogi_MMETSP0879_67004	31919				
	Thalassiosira_weissflogi_MMETSP0880_67004	31748				
	Thalassiosira_weissflogi_MMETSP0881_67004	31920				
	Thalassiosira_weissflogi_MMETSP0886_67004	34041				
	Thalassiosira_weissflogi_MMETSP0899_67004	33903				
	Thalassiosira_weissflogi_MMETSP0900_67004	33908				
	Thalassiosira_weissflogi_MMETSP0901_67004	31884				
	Thalassiosira_weissflogi_MMETSP1405_67004	34792				
	Thalassiosira_weissflogi_MMETSP1406_67004	32728				
	Thalassiosira_weissflogi_MMETSP1407_67004	34960				
	Thalassiosira_weissflogi_MMETSP1408_67004	32100				
	Thalassiosira_weissflogi_MMETSP1409_67004	32186				
	Thalassiosira_weissflogi_MMETSP1410_67004	33354				
	Thalassiosira_weissflogi_MMETSP1411_67004	33394				
	Thalassiosira_weissflogi_MMETSP1412_67004	32610				
	Thalassiosira_weissflogi_MMETSP1413_67004	32010				
	Thalassiosira_weissflogi_MMETSP1414_67004	32138				
	Thalassiosira_weissflogi_MMETSP1415_67004	33092				
	Thalassiosira_weissflogi_MMETSP1416_67004	32717				
	Thalassiosira_weissflogi_MMETSP1417_67004	31856				
	Thalassiosira_weissflogi_MMETSP1418_67004	32035				
	Thalassiosira_weissflogi_MMETSP1419_67004	32494				
	Thalassiosira_weissflogi_MMETSP1420_67004	32135				
	Thalassiosira_weissflogi_MMETSP1421_67004	32560				
	Thalassiosira_weissflogi_MMETSP1422_67004	32732	Thalassiosira_weissflogi_67004	850956	56032	93,42
	Thalassiosira_sp_MMETSP1059_35127	22519				
	Thalassiosira_sp_MMETSP1071_35127	40061	Thalassiosira_sp_35127	62580	56005	10,51
	Thalassiosira_aff_MMETSP0902_420261	43953				
	Thalassiosira_aff_MMETSP0903_420261	44530				
	Thalassiosira_aff_MMETSP0904_420261	41604				
	Thalassiosira_aff_MMETSP0905_420261	43618	Thalassiosira_aff_420261	173705	67042	61,40
	Thalassiosira_gravida_MMETSP0402_420259	27492				
	Thalassiosira_gravida_MMETSP0403_420259	35050				
	Thalassiosira_gravida_MMETSP0404_420259	34814	Thalassiosira_gravida_420259	97356	47630	51,08
	Thalassiosira_minuscula_MMETSP0737_385413	47167				
	Thalassiosira_minuscula_MMETSP0738_385413	34790				
	Thalassiosira_minuscula_MMETSP0739_385413	43965				
	Thalassiosira_minuscula_MMETSP0740_385413	36324	Thalassiosira_minuscula_385413	162246	72960	55,03
	Thalassiosira_oceanica_MMETSP0970_159749	37281				
	Thalassiosira_oceanica_MMETSP0971_159749	39377				
	Thalassiosira_oceanica_MMETSP0972_159749	37371				
	Thalassiosira_oceanica_MMETSP0973_159749	38954	Thalassiosira_oceanica_159749	152983	51013	66,65
	Thalassiosira_punctigera_MMETSP1067_265564	59660	Thalassiosira_punctigera_MMETSP1067_265564	59660	54823	8,11
	Thalassiosira_rotula_MMETSP0404_49265	42839				
	Thalassiosira_rotula_MMETSP0911_49265	28915				
	Thalassiosira_rotula_MMETSP0912_49265	30225				
	Thalassiosira_rotula_MMETSP0913_49265	25207	Thalassiosira_rotula_49265	127066	52121	58,99
cluster-83	Florenziella_sp_MMETSP1324_236786	42429				
	Florenziella_sp_MMETSP1325_236786	7956	Florenziella_sp_236786	50385	45206	10,28
	Florenziella_parvula_MMETSP1323_236787	30809				
	Florenziella_parvula_MMETSP1344_236787	110828	Florenziella_parvula_236787	141637	104886	25,95
cluster-84	Aplanochytrium_sp_MMETSP0954_627963	20518				
	Aplanochytrium_sp_MMETSP0955_627963	18613				
	Aplanochytrium_sp_MMETSP0956_627963	18498				
	Aplanochytrium_sp_MMETSP0957_627963	17377	Aplanochytrium_sp_627963	75006	26332	64,89
	Aplanochytrium_stochinoi_MMETSP1346_215587	30490				
	Aplanochytrium_stochinoi_MMETSP1347_215587	18946				
	Aplanochytrium_stochinoi_MMETSP1348_215587	22995				
	Aplanochytrium_stochinoi_MMETSP1349_215587	21247	Aplanochytrium_stochinoi_215587	93678	37681	59,78
cluster-85	Pyramimonas_sp_MMETSP1061_36882	40209				
	Pyramimonas_sp_MMETSP1169_36882	40009	Pyramimonas_sp_36882	80218	75437	5,96
	Pyramimonas_parkeae_MMETSP0058_36894	50667				
	Pyramimonas_parkeae_MMETSP0059_36894	51327	Pyramimonas_parkeae_36894	101994	64934	36,34

Table S1: Data file 2 - Summary of genus-level sample clustering and consolidation. Samples that were not combined are highlighted in alternating yellow and green, while samples grouped into consolidated entries are highlighted in alternating orange and pink to improve readability

Phylum	Before	After	Redundancy-drop
Amoebozoa	11	10	9.1%
Archaeplastida	90	34	62.2%
Cryptophyceae	27	15	44.4%
Euglenozoa	5	2	60.0%
Haptophyta	66	20	69.7%
Rhizaria	21	12	42.9%
Stramenopiles	285	93	67.4%
Ciliophora	28	25	10.7%
Colpodellida	3	2	33.3%
Dinophyceae	113	41	63.7%
Perkinsozoa	4	2	50.0%
other_Eukaryotes	9	4	55.6%

Tables S2: Data file 4 - Taxonomic redundancy reduction after consolidation.

Column1	total.contigs	never.suspected	nb.suspects	nb.clean	nb.lowcov	nb.overexp	nb.dubious	nb.contam	clean.percent	low.cov.percent	overexp.percent	dubious.percent	contam.percent
Acanthoea_sp_81532	37731	33558	4173	1981	0	5	141	114	9419	0	1	37	30
Alexandrium_andersonii_MMETSP1436_327968	104672	96713	7959	4426	0	8	2145	1380	9662	0	0	204	131
Alexandrium_catenella_MMETSP0790_2925	118407	107395	11012	9637	0	9	1041	325	9883	0	0	87	27
Alexandrium_fundyense_2932	53498	15055	38443	538	4272	47	6777	26797	2914	798	8	1266	5008
Alexandrium_margalefii_MMETSP0661_109239	160604	139693	20911	8897	2211	55	4371	5376	9251	137	3	272	334
Alexandrium_minutum_MMETSP0328_39455	67858	56825	11033	6237	302	7	1168	3319	9293	44	1	172	489
Alexandrium_monilatum_311494	449187	407257	41930	35003	0	21	4786	2057	9845	0	0	106	45
Alexandrium_tamarense_2926	207259	131314	75945	37640	487	21	5984	2911	8151	23	1	288	140
Ammonia_sp_MMETSP1384_29189	75739	73330	2409	2300	0	2	57	50	9985	0	0	7	6
Amoebophrya_sp_MMETSP0795_88552	29024	27805	1219	1030	0	1	50	138	9934	0	0	17	47
Amorphochlora_amoebiformis_MMETSP0042_168855	35570	34280	1290	984	0	13	58	235	9913	0	3	16	66
Amphidinium_carerae_2961	138281	130291	7990	2731	1523	26	1241	2463	9619	110	1	89	178
Amphidinium_massartii_MMETSP0689_160604	76432	72756	3676	3458	0	4	56	157	9971	0	0	7	20
Amphiprora_paludosa_MMETSP1065_265537	41105	39468	1637	1472	0	7	48	110	9959	0	1	11	26
Amphiprora_sp_265536	115962	107724	8238	7291	1	28	462	453	9918	0	2	39	39
Amphora_coffeiformis_265554	38040	34623	3417	1646	2	16	167	316	9534	0	4	43	83
Anophryoides_haemophila_46462	39147	37330	1817	824	28	15	114	836	9746	7	3	29	213
Aplanochytrium_sp_627963	75006	68099	6907	6338	22	27	82	436	9924	2	3	10	58
Aplanochytrium_stocchini_215587	93678	85446	8232	7744	0	5	104	377	9947	0	0	11	40
Aristerostoma_sp_MMETSP0125_485358	41818	40910	908	628	0	4	34	241	9933	0	0	8	57
Asterionellopsis_glacialis_33640	171538	159914	11624	6978	1650	35	846	2112	9729	96	2	49	123
Astrosyne_radiata_MMETSP0418_1158023	34509	32565	1944	1278	0	22	245	399	9807	0	6	70	115
Aulacoseira_subarctica_MMETSP1064_202472	29571	28639	932	779	0	8	21	124	9948	0	2	7	41
Aurantiochytrium_limacinum_717989	96670	90067	6603	4786	0	63	258	1486	9812	0	6	26	153
Aureococcus_anophagefferens_44056	135596	118176	17420	15633	0	23	1210	540	9868	0	1	89	39
Aureoumbra_lagunensis_44058	76476	74077	2399	1593	0	10	29	39	9894	0	1	3	5
Azadinium_spinosum_632150	408970	213110	195860	186844	131	25	5919	2533	9779	3	0	144	61
Bathycoccus_prasinus_41875	20240	10722	9518	0	0	0	0	0	5297	0	0	0	0
Bicosoecida_sp_MMETSP0115_33651	17988	15840	2148	2004	0	5	92	47	9919	0	2	51	26
Bigelowiella_longifila_MMETSP1359_552665	35304	32921	2383	1562	2	16	463	340	9767	0	4	131	96
Bigelowiella_natans_227086	192943	179558	13385	10785	34	53	1304	1200	9865	1	2	67	62
Blepharisma_japonicum_MMETSP1395_5961	33020	31738	1282	526	5	10	101	640	9771	1	3	30	193
Bolidomonas_pacifica_MMETSP0785_91992	24608	23184	1424	1254	0	12	59	99	9930	0	4	23	40
Bolidomonas_pacifica_MMETSP1319_172717	32559	9821	22738	9599	0	4	9021	4113	5964	0	1	2770	1263
Bolidomonas_sp_91990	56648	32787	23861	0	0	0	0	0	5787	0	0	0	0
Brandtodinium_nutricula_MMETSP1462_1333877	105264	97099	8165	7746	0	10	216	193	9960	0	0	20	18
Cafeteria_roenbergensis_MMETSP0942_33653	26963	22391	4572	2450	0	6	470	1646	9212	0	2	174	610
Cafeteria_sp_MMETSP1104_33652	31307	27486	3821	3716	0	0	69	36	9966	0	0	22	11
Calcidiscus_leptoporus_MMETSP1334_127549	60887	56793	4094	3721	0	3	214	156	9938	0	0	35	25
Cercozoa_sp_136419	51229	47203	4026	2068	0	1	482	347	9617	0	0	94	67
Chaetoceros_affinis_426623	70880	64473	6407	5216	2	24	544	621	9831	0	3	76	87
Chaetoceros_curvisetus_230516	32167	30431	1736	803	95	37	111	659	9709	29	11	34	204
Chaetoceros_dichaeta_MMETSP1447_420782	37747	35875	1872	1611	0	6	104	151	9930	0	1	27	40
Chaetoceros_neogracilis_240364	157911	147317	10594	9479	0	35	586	488	9929	0	2	37	30
Chaetoceros_sp_49237	90438	84351	6087	5184	0	15	488	400	9900	0	1	53	44
Chattonella_subsalsa_44440	162049	158882	3167	2725	0	5	109	326	9972	0	0	6	20
Chlamydomonas_chlamydogama_MMETSP1392_225041	40748	38404	2344	2120	0	12	124	88	9945	0	2	30	21
Chlamydomonas_leiostraca_MMETSP1391_1034604	38192	35360	2832	2711	0	5	88	28	9968	0	1	23	7
Chlamydomonas_sp_MMETSP0063_327068	45607	42469	3138	3057	0	8	44	29	9982	0	1	9	6
Chlamydomonas_sp_MMETSP1180_3052	22693	21824	869	724	0	5	38	102	9936	0	2	16	44
Chlorarachnion_reptans_MMETSP0109_29199	35061	32784	2277	1779	0	22	184	292	9857	0	6	52	83
Chlorarachniophyceae_sp_MMETSP0113_91329	29314	28674	640	583	0	10	25	22	9980	0	3	8	7
Chloropicon_sp_2302914	53094	-37300	90394	0	2908	25	18781	21177	-7025	547	4	3537	3988
Chromera_velia_MMETSP0290_505693	56191	54611	1580	0	0	8	0	1572	9718	0	1	0	279
Chromerida_sp_877183	99936	92395	7541	46	8	14	99	7372	9250	0	1	9	737

Column1	total.contigs	never.suspected	nb.suspects	nb.clean	nb.lowcov	nb.overexp	nb.dubious	nb.contam	clean.percent	low.cov.percent	overexp.percent	dubious.percent	contam.percent
Chromulina_nebulosa_MMETSP1095_96789	25346	24882	464	0	0	0	0	464	9816	0	0	0	183
Chroomonas_mesostigmatica_MMETSP0047_195065	68249	56438	11811	1	2	19	4	11784	8269	0	2	0	1726
Chrysochromulina_rotalis_MMETSP0287_412157	54406	51010	3396	0	0	13	2	3381	9375	0	2	0	621
Chrysoculter_rhomboides_MMETSP1335_259385	26549	24503	2046	1	0	6	5	2034	9229	0	2	1	766
Chrysoreinhardia_sp._1104430	27772	24004	3768	0	0	0	0	0	8643	0	0	0	0
Climacostomum_virens_MMETSP1397_49980	34800	31748	3052	1	39	48	0	2964	9123	11	13	0	851
Compsopogon_caeruleus_MMETSP0312_31354	18679	18241	438	0	0	3	0	434	9765	0	1	0	232
Condylostoma_magnum_MMETSP0210_40633	37783	37398	385	0	0	9	0	376	9898	0	2	0	99
Corethron_hystrix_MMETSP0010_216773	43375	40888	2487	5	4	23	1	2454	9427	0	5	0	565
Corethron_pennatum_218684	84555	81462	3093	4	0	12	19	3058	9634	0	1	2	361
Coscinodiscus_wailesii_MMETSP1066_671091	65714	63913	1801	0	0	5	0	1796	9725	0	0	0	273
Craspedostauros_australis_MMETSP1442_1486917	25015	23022	1993	0	0	9	0	1984	9203	0	3	0	793
Cryptocodinium_cohnii_2866	283231	265265	17966	4	1	14	19	17918	9365	0	0	0	632
Cryptomonas_curvata_MMETSP1050_233186	61034	57897	3137	0	0	16	1	3120	9486	0	2	0	511
Cryptomonas_paramecium_MMETSP0038_2898	90051	82989	7062	35	61	36	155	6775	9219	6	3	17	752
Cryptophyceae_sp._697907	283887	269225	14662	9	1	66	24	14541	9483	0	2	0	512
Cyanoptycha_sp._MMETSP1086_77921	27634	26848	786	0	0	4	0	782	9715	0	1	0	282
Cyclophora_tenuis_MMETSP0397_216820	25997	23915	2082	2	0	19	9	2052	9199	0	7	3	789
Cyclotella_meneghiniana_MMETSP1057_29205	32886	31620	1266	0	0	4	0	1262	9615	0	1	0	383
Cylindrotheca_closterium_MMETSP0017_2856	46484	43020	3464	1	2	25	10	3426	9255	0	5	2	737
Dactylosolen_fragilissimus_MMETSP0580_1118495	23345	21887	1458	0	0	13	0	1445	9375	0	5	0	618
Debaryomyces_hansenii_284592	28986	28313	673	0	2	0	0	671	9767	0	0	0	231
Detonula_confervacea_MMETSP1058_83371	33378	31008	2370	0	0	6	0	2364	9289	0	1	0	708
Dictyocha_speculum_MMETSP1174_35687	65184	63533	1651	0	0	2	0	1649	9746	0	0	0	252
Dinobryon_sp._98059	101311	96006	5305	0	0	22	11	5269	9476	0	2	1	520
Dinophysis_acuminata_MMETSP0797_47934	128844	118155	10689	0	0	8	0	10681	9170	0	0	0	828
Ditylum_brightwellii_49249	240447	219299	21148	0	220	115	0	11881	9120	9	4	0	494
Dolichomastix_tenuilepis_MMETSP0033_195969	31615	28136	3479	0	15	4	0	3460	8899	4	1	0	1094
Dunaliella_tertiolecta_3047	59043	52339	6704	0	0	19	4	4142	8864	0	3	0	701
Durinskia_baltica_400756	257933	242112	15821	0	0	22	2	15791	9386	0	0	0	612
Elphidium_margaritaceum_MMETSP1385_933848	60450	57016	3434	0	0	3	0	3431	9431	0	0	0	567
Emiliania_huxleyi_2903	648368	98062	550306	72	30250	84	216	479615	1513	466	1	3	7397
Entomoneis_sp._MMETSP1443_186043	22267	21139	1128	0	0	12	0	1116	9493	0	5	0	501
Erythrolobus_australicus_MMETSP1353_1077150	15676	14926	750	0	0	5	0	745	9521	0	3	0	475
Erythrolobus_madagascarensis_MMETSP1354_708628	17856	17059	797	3	0	1	2	791	9555	0	0	1	442
Eucampia_antarctica_MMETSP1437_49252	31321	29986	1335	0	0	11	1	1323	9573	0	3	0	422
Euplotes_focardii_36767	58757	56157	2600	0	1	51	6	2541	9557	0	8	1	432
Euplotes_harpa_MMETSP0213_151035	43586	42735	851	1	0	5	1	844	9804	0	1	0	193
Eutreptiella_gymnastica_73025	424672	419401	5271	0	0	37	0	5228	9875	0	0	0	123
Exanthemachrysis_gayraliae_MMETSP1464_119497	27945	25095	2850	0	0	7	0	2843	8980	0	2	0	1017
Extubocellulus_spinifer_265572	156675	143391	13284	0	0	41	5	13229	9152	0	2	0	844
Fabrea_salina_MMETSP1345_342563	16640	16027	613	0	0	15	0	598	9631	0	9	0	359
Favella_ehrenbergii_MMETSP0123_182087	38054	4354	33700	0	165	15	0	33517	1144	43	3	0	8807
Fibrocapsa_japonica_MMETSP1339_94617	24389	23315	1074	0	0	7	0	1067	9559	0	2	0	437
Filamoeba_nolandii_MMETSP0413_136452	30376	29771	605	0	0	3	0	602	9800	0	0	0	198
Florenciella_parvula_236787	141637	46679	94958	0	0	0	0	0	3295	0	0	0	0
Florenciella_sp._236786	50385	12587	37798	0	0	0	0	0	2498	0	0	0	0
Fragilariopsis_kerguelensis_186038	487950	439048	48902	2	358	32	11	48323	8997	7	0	0	990
Gambierdiscus_austroaleis_MMETSP0766_439317	99150	93505	5645	11	185	15	3	5431	9431	18	1	0	547
Geminigera_cryophila_MMETSP0799_46947	105756	34991	70765	0	10	18	0	70726	3308	0	1	0	6687
Geminigera_sp._MMETSP1102_46946	100847	25548	75299	5915	19	10	52604	16746	3119	1	0	5216	1660
Gephyrocapsa_oceanica_38817	200728	72631	128097	0	3070	62	126	124757	3618	152	3	6	6215
Goniomonas_pacifica_195067	142464	119533	22931	0	7	18	2	22897	8390	0	1	0	1607
Goniomonas_sp._MMETSP0114_31324	36363	34882	1481	15	0	4	3	1459	9596	0	1	0	401
Gonyaulax_spinifera_MMETSP1439_66791	80982	73914	7068	1	0	6	2	7059	9127	0	0	0	871

Column1	total.contigs	never.suspected	nb.suspects	nb.clean	nb.lowcov	nb.overexp	nb.dubious	nb.contam	clean.percent	low.cov.percent	overexp.percent	dubious.percent	contam.percent
Grammatophora_oceanica_MMETSP0009_210454	54275	52808	1467	0	0	7	0	1460	9729	0	1	0	269
Guillardia_theta_MMETSP0046_905079	52725	47595	5130	0	5	47	9	5069	9027	0	8	1	961
Gymnochloa_sp._MMETSP0110_629695	24631	24014	617	0	0	8	0	609	9749	0	3	0	247
Gymnodinium_catenatum_MMETSP0784_39447	123545	115591	7954	0	0	12	0	7942	9356	0	0	0	642
Hanusia_phi_MMETSP1048_3032	36932	34823	2109	0	0	12	0	2097	9428	0	3	0	567
Haptolina_brevifila_MMETSP1094_156173	87610	82997	4613	0	0	15	0	4598	9473	0	1	0	524
Haptolina_ericina_MMETSP1096_156174	74018	70289	3729	6	0	9	16	3698	9497	0	1	2	499
Helicotheca_sp._MMETSP1171_374046	20449	19500	949	0	0	14	1	934	9535	0	6	0	456
Hemiselmis_andersenii_464988	200929	180256	20673	0	113	68	2	20476	8971	5	3	0	1019
Hemiselmis_tepida_MMETSP1355_464990	38957	36292	2665	98	316	16	309	1926	9341	81	4	79	494
Heterocapsa_arctica_MMETSP1441_192219	81670	70882	10788	0	0	9	0	10779	8679	0	1	0	1319
Heterocapsa_rotundata_MMETSP0503_89963	75283	11172	64111	155	11	13	2331	61572	1504	1	1	309	8178
Heterocapsa_triquetra_MMETSP0448_66468	103549	65308	38241	5	0	2	323	37900	6307	0	0	31	3660
Heterosigma_akashiwo_2829	451334	-84457	535791	26	133800	38	95	277146	-1870	2964	0	2	6140
Isochrysis_galbana_37099	70221	38113	32108	0	258	15	1	29006	5427	36	2	0	4130
Isochrysis_sp._37098	91278	30909	60369	0	2098	17	1	58239	3386	229	1	0	6380
Isochrysis_sp._40639	91835	6832	85003	25852	773	52	38286	19971	3558	84	5	4168	2174
Karenia_brevis_156230	1911205	1855131	56074	0	127	83	0	55730	9706	0	0	0	291
Karlodinium_veneticum_342587	377368	370129	7239	0	0	34	0	7196	9808	0	0	0	190
Kryptoperidinium_foliaceum_160619	442462	405063	37399	2	4	67	14	37302	9154	0	1	0	843
labyrinthulid_quahog_96639	17555	8640	8915	0	5	2	0	1113	4921	2	1	0	634
Leptocylindrus_danicus_163516	86835	82585	4250	0	0	36	0	4214	9510	0	4	0	485
Licmophora_paradoxa_MMETSP1360_515479	42249	39953	2296	0	0	5	0	2291	9456	0	1	0	542
Lingulodinium_polyedrum_160621	485366	386633	98733	2	77	25	2	98427	7965	1	0	0	2027
Litonotus_pictus_MMETSP0209_665100	42255	41645	610	0	0	7	0	603	9855	0	1	0	142
Lotharella_globosa_91324	103575	96512	7063	0	1	35	0	7027	9318	0	3	0	678
Lotharella_oceanica_MMETSP0040_641309	37737	36040	1697	1	0	10	0	1686	9550	0	2	0	446
Mantoniella_sp._13607	86313	74445	11868	0	2	6	2	11858	8625	0	0	0	1373
Mayorella_sp._MMETSP0417_242541	19033	1914	17119	0	3	7	6	17098	1005	1	3	3	8983
Mesodinium_pulex_MMETSP0467_283647	117595	49089	68506	4	10	45	2	68408	4174	0	3	0	5817
Micromonas_pusilla_38833	194571	167996	26575	2	262	91	12	24876	8634	13	4	0	1278
Minutocellus_polymorphus_265543	85808	79077	6731	1	0	39	0	6691	9215	0	4	0	779
Moneuplotes_crassus_MMETSP1380_5936	40432	39783	649	0	0	6	0	643	9839	0	1	0	159
Myrionecta_rubra_MMETSP0798_283649	67184	33573	33611	0	1	17	0	33590	4997	0	2	0	4999
Neobodo_designis_MMETSP1114_312471	55089	51165	3924	0	0	10	0	3914	9287	0	1	0	710
Neoceratium_fusus_2916	276463	267758	8705	0	0	39	1	8661	9685	0	1	0	313
Nephroselmis_pyriiformis_MMETSP0034_156128	73008	63174	9834	0	71	21	8	9734	8653	9	2	1	1333
Nitzschia_sp._MMETSP0014_2857	63046	50148	12898	0	1	18	20	12859	7954	0	2	3	2039
Noctiluca_scintillans_MMETSP0253_2966	66402	63351	3051	0	0	11	1	3039	9540	0	1	0	457
Ochromonadaceae_sp._483371	32338	20001	12337	0	0	2	1	2793	6184	0	0	0	863
Ochromonas_sp._420556	66383	59791	6592	1	0	23	1	6566	9007	0	3	0	989
Odontella_sp._MMETSP0015_2838	63642	55928	7714	2	10	26	5	7671	8788	1	4	0	1205
Ostreococcus_sp._548131	69143	54855	14288	9	115	41	28	14092	7934	16	5	4	2038
Oxvrrhis_marina_2969	103484	97697	5787	1	572	29	5	5174	9440	55	2	0	499
Palpitomonas_bilix_MMETSP0780_652834	50856	47871	2985	0	31	20	0	2934	9413	6	3	0	576
Paramoeba_aestuarina_MMETSP0161_180227	40496	39194	1302	0	0	9	0	1293	9678	0	2	0	319
Paramoeba_sp._MMETSP0151_200890	45281	44113	1168	0	0	7	1	1160	9742	0	1	0	256
Paraphysomonas_imperforata_37642	67789	56459	11330	0	0	10	0	11320	8328	0	1	0	1669
Paraphysomonas_sp._MMETSP1103_36769	27001	26313	688	0	0	24	0	664	9745	0	8	0	245
Paraphysomonas_vestita_MMETSP1107_36770	25172	23787	1385	0	0	7	2	1376	9449	0	2	0	546
Pavlova_gyrans_MMETSP1466_44452	29304	26655	2649	0	0	5	2	2642	9096	0	1	0	901
Pavlova_sp._483367	191938	176433	15505	0	3	66	7	15411	9192	0	3	0	802
Pavlova_sp._298111	77231	71434	5797	0	8	27	2	5758	9249	1	3	0	745
Pedinellales_sp._483369	164776	138323	26453	0	3	22	7	13415	8394	0	1	0	814
Pelagococcus_subviridis_35679	145093	131862	13231	0	0	33	1	13193	9088	0	2	0	909

Column1	total.contigs	never.suspected	nb.suspects	nb.clean	nb.lowcov	nb.overexp	nb.dubious	nb.contam	clean.percent	low.cov.percent	overexp.percent	dubious.percent	contam.percent
Pelagodinium_bei MMETSP1338_43686	107171	103529	3642	0	0	9	1	3632	9660	0	0	0	338
Pelagomonas_calceolata_35677	130305	117331	12974	0	0	26	15	12926	9004	0	1	1	991
Pelagophyceae_sp._483370	100151	86121	14030	0	0	32	12	10239	8599	0	3	1	1022
Percolomonas_cosmopolitus_63605	26025	25052	973	0	0	0	0	0	9626	0	0	0	0
Peridinium_aciculiferum_268820	213653	44018	169635	10	786	5	31	168732	2060	36	0	1	7897
Perkinsus_chesapeakei_321610	60719	33836	26883	0	38	13	0	26829	5572	6	2	0	4418
Perkinsus_marinus_1032511	26157	23025	3132	0	72	15	4	1510	8802	27	5	1	577
Pessonnella_sp._MMETSP0420_498008	23290	22159	1131	0	0	4	0	1127	9514	0	1	0	483
Phaeocystis_antarctica_33657	122984	109083	13901	0	0	17	0	13884	8869	0	1	0	1128
Phaeocystis_sp._33656	64610	56264	8346	0	0	0	0	0	8708	0	0	0	0
Phaeomonas_parva_MMETSP1163_124430	46917	36188	10729	1	94	29	1	10604	7713	20	6	0	2260
Picochlorum_sp._MMETSP1330_464278	16181	8176	8005	0	9	19	0	7977	5052	5	11	0	4929
Picocystis_salinarum_88271	21725	17737	3988	3	934	36	4	3011	8165	429	16	1	1385
Platyphrya_macrostoma_MMETSP0127_938130	83965	82699	1266	0	0	11	0	1255	9849	0	1	0	149
Pleurochrysis_carterae_13221	141311	132859	8452	0	0	26	7	8414	9401	0	1	0	595
Polarella_glacialis_89957	181165	170246	10919	0	0	24	0	10894	9397	0	1	0	601
Polytomella_parva_MMETSP0052_51329	35064	32263	2801	0	0	9	0	2792	9201	0	2	0	796
Porphyridium_aeruginum_MMETSP0313_2792	22715	21461	1254	0	0	1	0	1253	9447	0	0	0	551
Prasinococcus_capsulatus_MMETSP0941_156131	14102	12896	1206	0	0	4	1	1201	9144	0	2	0	851
Prasinoderma_singulare_MMETSP1315_676789	23145	20650	2495	0	0	5	0	2489	8922	0	2	0	1075
Proboscia_sp._216776	130565	124978	5587	0	0	3	6	4030	9572	0	0	0	308
Prorocentrum_micans_MMETSP0251_2945	18649	2738	15911	3	452	16	2	15438	1469	242	8	1	8278
Prorocentrum_minimum_39449	786380	728908	57472	17	31	62	79	57159	9269	0	0	1	726
Proteomonas_sulcata_MMETSP1049_77928	49537	46279	3258	0	0	2	0	3256	9342	0	0	0	657
Protoceratium_reticulatum_MMETSP0228_133427	112970	102155	10815	0	0	7	2	10806	9042	0	0	0	956
Protocruzia_adherens_MMETSP0216_223996	65706	64811	895	0	0	6	0	889	9863	0	0	0	135
Prymnesium_parvum_97485	155671	141548	14123	0	9	36	4	8287	9092	0	2	0	532
Prymnesium_polyleps_72548	359692	334187	25505	5	10	70	28	25378	9291	0	1	0	705
Pseudokeronopsis_sp._MMETSP0211_311385	44971	43905	1066	0	0	10	0	1055	9762	0	2	0	234
Pseudokeronopsis_sp._MMETSP1396_1634380	49169	47685	1484	0	0	20	0	1463	9698	0	4	0	297
Pseudo-nitzschia_australis_44445	114740	96650	18090	3	0	40	5	18034	8423	0	3	0	1571
Pseudo-nitzschia_delicatissima_44447	21996	19186	2810	0	0	8	0	1048	8722	0	3	0	476
Pseudo-nitzschia_fraudulenta_183588	58455	-40537	98992	0	1	24	1	4427	-6934	0	4	0	757
Pseudo-nitzschia_heimii_MMETSP1423_476441	28036	25840	2196	0	0	2	0	2194	9216	0	0	0	782
Pseudo-nitzschia_pungens_37318	64106	53561	10545	8	0	15	58	10464	8356	0	2	9	1632
Pseudopedinella_elastica_35684	90625	82828	7797	0	0	28	3	7766	9139	0	3	0	856
Pteridomonas_danica_38822	35999	23504	12495	1	0	1	0	2203	6529	0	0	0	611
Pterosperma_sp._MMETSP1438_1461541	43875	42829	1046	0	0	5	0	1041	9761	0	1	0	237
Pycnococcus_provasolii_41880	46971	39414	7557	24	89	10	37	6043	8396	18	2	7	1286
Pycnococcus_sp._MMETSP1085_41879	10979	767	10212	0	1	5	9	10195	698	0	4	8	9285
Pyramimonas_parkeae_36894	101994	97180	4814	0	0	7	0	4806	9528	0	0	0	471
Pyramimonas_sp._36882	80218	76632	3586	0	0	0	0	0	9552	0	0	0	0
Pyrodinium_bahamense_MMETSP0796_73915	150835	140748	10087	0	0	11	0	10076	9331	0	0	0	668
Rhizochromulina_marina_MMETSP1173_1034831	33405	29004	4401	0	0	21	1	4376	8682	0	6	0	1309
Rhodella_violacea_38298	54909	52014	2895	0	0	14	3	2878	9472	0	2	0	524
Rhodomonas_sp._302021	105246	100560	4686	1	0	34	5	4646	9554	0	3	0	441
Rhodomonas_sp._37096	127735	79607	48128	1	12	41	14	48052	6232	0	3	1	3761
Rhodorus_marinus_101924	71239	64792	6447	0	0	0	0	0	9095	0	0	0	0
Rosalina_sp._MMETSP0190_37353	77773	75744	2029	0	0	31	0	1998	9739	0	3	0	256
Sarcinochrysis_sp._MMETSP1170_54406	11581	10479	1102	0	0	5	0	1097	9048	0	4	0	947
Schizochytrium_aggregatum_4773	85247	78967	6280	0	42	18	0	6218	9263	4	2	0	729
Schmidingerella_taraikaensis_626734	81916	24494	57422	3	145	36	2	57221	2990	17	4	0	6985
Scrippsiella_sp._54902	800780	309486	491294	46	16640	93	108	474148	3865	207	1	1	5921
Scrippsiella_trochoidea_71861	452259	422327	29932	2	2	21	9	29866	9338	0	0	0	660
Scyphosphaera_apsteinii_MMETSP1333_418940	76564	74621	1943	0	0	7	0	1936	9746	0	0	0	252

Column1	total.contigs	never.suspected	nb.suspects	nb.clean	nb.lowcov	nb.overexp	nb.dubious	nb.contam	clean.percent	low.cov.percent	overexp.percent	dubious.percent	contam.percent
Sexangularia_sp._MMETSP0437_1486928	16940	15363	1577	1	0	8	8	1559	9069	0	4	4	920
Skeletonema_grethae_MMETSP0578_267565	29138	24074	5064	0	0	14	0	5050	8262	0	4	0	1733
Skeletonema_sp._2842	366983	326707	40276	4	912	123	14	39219	8902	24	3	0	1068
Stephanopyxis_turris_MMETSP0794_515487	55772	53939	1833	0	0	12	0	1821	9671	0	2	0	326
Stichococcus_sp._MMETSP1473_1461547	21325	19760	1565	0	0	5	0	1560	9266	0	2	0	731
Striatella_unipunctata_MMETSP0800_210618	37562	35616	1946	0	0	5	0	1941	9481	0	1	0	516
Strombidinopsis_acuminata_MMETSP0126_141414	119699	44884	74815	3	253	32	8	74507	3749	21	2	0	6224
Strombidinopsis_sp._MMETSP0463_181622	43507	34813	8694	0	1	47	3	8641	8001	0	10	0	1986
Strombidium_inclinatum_MMETSP0208_197538	42764	41091	1673	0	0	16	0	1657	9608	0	3	0	387
Strombidium_rassoulzadegani_MMETSP0449_1082188	19501	16150	3351	0	0	19	0	3331	8281	0	9	0	1708
Stygamoeba_regulata_MMETSP0447_1042828	53552	31348	22204	0	1	7	0	22192	5853	0	1	0	4144
Symbiodinium_kawagutii_104179	75012	46664	28348	4	48	28	1	28257	6221	6	3	0	3766
Symbiodinium_sp._230996	269238	257454	11784	0	30	32	4	11702	9562	1	1	0	434
Symbiodinium_sp._2949	271516	246257	25259	1	62	10	8	13259	9069	2	0	0	488
Synedropsis_cf._MMETSP1176_265551	33319	31707	1612	0	0	10	0	1602	9516	0	3	0	480
Tetraselmis_chuii_MMETSP0491_63592	43997	40365	3632	0	0	8	0	3624	9174	0	1	0	823
Tetraselmis_sp._MMETSP0419_582737	43453	41410	2043	0	0	8	0	2035	9529	0	1	0	468
Tetraselmis_striata_3165	190753	178362	12391	0	0	33	0	12350	9350	0	1	0	647
Thalassionema_frauenfeldii_MMETSP0786_186022	33504	32292	1212	0	0	6	0	1206	9638	0	1	0	359
Thalassionema_nitzschoides_33649	75131	71818	3313	0	0	27	1	3285	9559	0	3	0	437
Thalassiosira_aff._420261	173705	164006	9699	0	21	38	1	9634	9441	1	2	0	554
Thalassiosira_gravida_420259	97356	60432	36924	8	1	25	32	36855	6208	0	2	3	3785
Thalassiosira_minuscula_385413	162246	154434	7812	6	51	65	13	7673	9518	3	4	0	472
Thalassiosira_oceanica_159749	152983	146726	6257	0	0	23	16	6215	9591	0	1	1	406
Thalassiosira_punctigera_MMETSP1067_265564	59660	56613	3047	1	0	8	8	3030	9489	0	1	1	507
Thalassiosira_rotula_49265	127086	65372	61714	7	17	44	75	37846	5144	1	3	5	2977
Thalassiosira_sp._35127	62580	56285	6295	0	0	0	0	0	8994	0	0	0	0
Thalassiosira_weisstlogii_67004	850956	763918	87038	27	478	162	328	85769	8977	5	1	3	1007
Thalassiothrix_antarctica_1049557	83533	80295	3238	0	0	10	1	3227	9612	0	1	0	386
Thraustochytrium_sp._1112570	48199	46368	1831	0	11	12	0	1808	9620	2	2	0	375
Tiarina_fusa_MMETSP0472_693140	135916	77494	58422	0	3	57	0	58352	5701	0	4	0	4293
Timspurckia_oligopyrenoides_MMETSP1172_708627	13901	10187	3714	0	0	2	0	3712	7328	0	1	0	2670
Togula_jolla_MMETSP0224_285029	74092	69352	4740	0	0	11	0	4729	9360	0	1	0	638
Triceratium_dubium_MMETSP1175_1003142	51565	48967	2598	0	0	5	0	2593	9496	0	0	0	502
Trichosphaerium_sp._MMETSP0405_498012	31060	30569	491	0	0	3	0	488	9841	0	0	0	157
Tryblionella_compressa_913974	121548	116183	5365	0	0	23	0	5341	9558	0	1	0	439
uncultured_ciliate_MMETSP1317_187299	28878	28263	615	1	28	3	1	582	9787	9	1	0	201
Uronema_marinum_MMETSP0018_35107	19640	19190	450	0	0	4	0	446	9770	0	2	0	227
Vannella_sp._95228	48746	42477	6269	0	0	0	0	0	8713	0	0	0	0
Vexillifera_sp._MMETSP0173_218659	22363	21373	990	3	0	3	13	971	9558	0	1	5	434

Tables S3: Data file 5 - Overview of Sobek-based cross-contamination assessment across transcriptomes. The table reports the total number of contigs (total.contigs), the number of transcripts never suspected of cross-contamination (never.suspected), the number of suspected transcripts (nb.suspects), and, among the latter, the number assigned to the focal sample (nb.clean), showing insufficient expression in all samples (nb.lowcov), showing unusually high expression in at least three samples (nb.overexp), remaining ambiguous because of similar expression between focal and alien samples (nb.dubious), or assigned to an alien sample from the same experiment (nb.contam).

organism	completeness	single.copy	duplicate	fragmented	missing	total.genes
Acanthoea_sp._81532	7,38	7,14	0,24	1,45	1,17	255
Alexandrium_andersonii_MMETSP1436_327968	2,59	2,43	0,16	2,35	5,06	255
Alexandrium_catenella_MMETSP0790_2925	5,41	5,25	0,16	1,92	2,67	255
Alexandrium_fundyense_2932	2,43	2,16	0,27	0,98	6,59	255
Alexandrium_margalefii_MMETSP0661_109239	5,92	5,37	0,55	1,8	2,28	255
Alexandrium_minutum_MMETSP0328_39455	2,2	2,08	0,12	1,29	6,51	255
Alexandrium_monilatum_311494	6,9	6,39	0,51	0,94	2,16	255
Alexandrium_tamarense_2926	6,67	5,92	0,75	1,14	2,19	255
Ammonia_sp._MMETSP1384_29189	3,96	3,92	0,04	2,75	3,29	255
Amoebophrya_sp._MMETSP0795_88552	1,1	1,06	0,04	0,55	8,35	255
Amorphochlora_amoebiformis_MMETSP0042_168855	6,91	6,67	0,24	1,14	1,95	255
Amphidinium_carterae_2961	6,51	6,2	0,31	0,78	2,71	255
Amphidinium_massartii_MMETSP0689_160604	6,75	6,51	0,24	0,78	2,47	255
Amphiprora_paludosa_MMETSP1065_265537	4,55	4,43	0,12	2,39	3,06	255
Amphiprora_sp._265536	6,79	6,12	0,67	1,1	2,11	255
Amphora_coffeiformis_265554	2,59	2,47	0,12	1,02	6,39	255
Anophryoides_haemophila_46462	1,76	1,76	0	1,76	6,48	255
Aplanochytrium_sp._627963	6,79	5,61	1,18	1,33	1,88	255
Aplanochytrium_stocchini_215587	6	4,59	1,41	2	2	255
Aristerostoma_sp._MMETSP0125_485358	3,37	3,33	0,04	1,8	4,83	255
Asterionellopsis_glacialis_33640	7,41	2,31	5,1	1,1	1,49	255
Astrosyne_radiata_MMETSP0418_1158023	3,33	2,86	0,47	3,18	3,49	255
Aulacoseira_subarctica_MMETSP1064_202472	6,59	6,43	0,16	1,33	2,08	255
Aurantiochytrium_limacinum_717989	8,67	7,96	0,71	0,35	0,98	255
Aureococcus_anophagefferens_44056	5,69	4,75	0,94	1,88	2,43	255
Aureoumbra_lagunensis_44058	6,08	5,49	0,59	1,33	2,59	255
Azadinium_spinosum_632150	6	4,94	1,06	1,37	2,63	255
Bathycoccus_prasinus_41875	7,13	2,86	4,27	0,75	2,12	255
Bicosoecia_sp._MMETSP0115_33651	3,06	3,06	0	2,24	4,7	255
Bigelowiella_longifila_MMETSP1359_552665	4,62	4,31	0,31	2,08	3,3	255
Bigelowiella_natans_227086	7,37	6,35	1,02	0,78	1,85	255
Blepharisma_japonicum_MMETSP1395_5961	5,65	4,98	0,67	1,25	3,1	255
Bolidomonas_pacifica_MMETSP0785_91992	5,84	5,8	0,04	1,8	2,36	255
Bolidomonas_pacifica_MMETSP1319_172717	5,14	4,55	0,59	2,08	2,78	255
Bolidomonas_sp._91990	5,96	3,37	2,59	1,53	2,51	255
Brandtodinium_nutricula_MMETSP1462_1333877	5,53	5,14	0,39	1,29	3,18	255
Cafeteria_roenbergensis_MMETSP0942_33653	4,39	4,27	0,12	1,88	3,73	255
Cafeteria_sp._MMETSP1104_33652	6,2	6,08	0,12	2,04	1,76	255
Calcidiscus_leptoporus_MMETSP1334_127549	6,04	6	0,04	1,57	2,39	255
Cercozoa_sp._136419	7,69	7,53	0,16	1,25	1,06	255
Chaetoceros_affinis_426623	6,39	6,04	0,35	1,45	2,16	255
Chaetoceros_curvisetus_230516	2,67	2,55	0,12	2,27	5,06	255
Chaetoceros_dichaeta_MMETSP1447_420782	3,77	3,65	0,12	2,94	3,29	255
Chaetoceros_neogracilis_240364	7,88	3,45	4,43	0,59	1,53	255
Chaetoceros_sp._49237	7,37	5,06	2,31	0,9	1,73	255
Chattonella_subsals_44440	7,41	7,1	0,31	1,41	1,18	255
Chlamydomonas_chlamydogama_MMETSP1392_225041	7,3	7,22	0,08	1,65	1,05	255
Chlamydomonas_leiostraca_MMETSP1391_1034604	6,71	6,67	0,04	1,57	1,72	255
Chlamydomonas_sp._MMETSP0063_327068	5,65	5,57	0,08	1,96	2,39	255
Chlamydomonas_sp._MMETSP1180_3052	5,76	5,76	0	1,49	2,75	255
Chlorarachnion_reptans_MMETSP0109_29199	7,13	6,86	0,27	1,02	1,85	255
Chlorarachniophyceae_sp._MMETSP0113_91329	3,97	3,73	0,24	1,84	4,19	255
Chloropicon_sp._2302914	8,54	2,78	5,76	0,63	0,83	255
Chromera_velia_MMETSP0290_505693	5,14	5,06	0,08	1,65	3,21	255
Chromerida_sp._877183	6,74	5,88	0,86	1,22	2,04	255
Chromulina_nebulosa_MMETSP1095_96789	4,08	4	0,08	2,75	3,17	255
Chroomonas_mesostigmatica_MMETSP0047_195065	7,21	6,94	0,27	1,76	1,03	255
Chrysochromulina_rotalis_MMETSP0287_412157	5,73	5,61	0,12	1,45	2,82	255
Chrysoculter_rhomboides_MMETSP1335_259385	4,11	3,84	0,27	1,49	4,4	255
Chrysoreinhardtia_sp._1104430	4,66	4,31	0,35	1,96	3,38	255
Climacostomum_virens_MMETSP1397_49980	7,33	6,27	1,06	0,55	2,12	255
Compsopogon_caeruleus_MMETSP0312_31354	6,67	5,96	0,71	1,02	2,31	255
Condyllostoma_magnum_MMETSP0210_40633	5,26	5,02	0,24	1,65	3,09	255
Corethron_hystrix_MMETSP0010_216773	6,19	4,82	1,37	1,65	2,16	255

organism	completeness	single.copy	duplicate	fragmented	missing	total.genes
Corethron_pennatum_218684	4,94	4,86	0,08	2,35	2,71	255
Coscinodiscus_wailesii_MMETSP1066_671091	4,9	4,82	0,08	2,59	2,51	255
Craspedostauros_australis_MMETSP1442_1486917	3,42	3,22	0,2	1,96	4,62	255
Cryptocodinium_cohnii_2866	6,55	6,04	0,51	0,94	2,51	255
Cryptomonas_curvata_MMETSP1050_233186	4,47	4,31	0,16	2,39	3,14	255
Cryptomonas_paramecium_MMETSP0038_2898	6	5,18	0,82	2,35	1,65	255
Cryptophyceae_sp._697907	7,76	6,94	0,82	1,18	1,06	255
Cyanoptycha_sp._MMETSP1086_77921	3,81	3,73	0,08	1,65	4,54	255
Cyclophora_tenuis_MMETSP0397_216820	3,25	3,25	0	2,82	3,93	255
Cyclotella_meneghiniana_MMETSP1057_29205	6,07	5,76	0,31	1,49	2,44	255
Cylindrotheca_closterium_MMETSP0017_2856	7,61	7,18	0,43	0,86	1,53	255
Dactyliosolen_fragilissimus_MMETSP0580_1118495	6,08	5,88	0,2	1,8	2,12	255
Debaryomyces_hansenii_284592	7,34	6,67	0,67	1,57	1,09	255
Detonula_conservacea_MMETSP1058_83371	7,1	6,9	0,2	1,18	1,72	255
Dictyocha_speculum_MMETSP1174_35687	6,16	5,92	0,24	1,92	1,92	255
Dinobryon_sp._98059	6,75	6,2	0,55	1,29	1,96	255
Dinophysis_acuminata_MMETSP0797_47934	5,97	5,73	0,24	1,45	2,58	255
Ditylum_brightwellii_49249	7,49	5,61	1,88	0,94	1,57	255
Dolichomastix_tenuilepis_MMETSP0033_195969	7,49	7,41	0,08	1,22	1,29	255
Dunaliella_tertiolecta_3047	7,96	7,65	0,31	0,98	1,06	255
Durinskia_baltica_400756	6,98	3,92	3,06	1,49	1,53	255
Elphidium_margaritaceum_MMETSP1385_933848	4,67	4,67	0	2,35	2,98	255
Emiliania_huxleyi_2903	6,82	4,47	2,35	1,33	1,85	255
Entomoneis_sp._MMETSP1443_186043	4,04	4,04	0	1,41	4,55	255
Erythrolobus_australicus_MMETSP1353_1077150	3,84	3,76	0,08	2,31	3,85	255
Erythrolobus_madagascarensis_MMETSP1354_708628	2,67	2,59	0,08	2,71	4,62	255
Eucampia_antarctica_MMETSP1437_49252	3,96	3,96	0	2,63	3,41	255
Euplotes_focardii_36767	4,71	4	0,71	1,57	3,72	255
Euplotes_harpa_MMETSP0213_151035	3,53	3,41	0,12	2,12	4,35	255
Eutreptiella_gymnastica_73025	6,9	3,53	3,37	1,61	1,49	255
Exanthemachrysis_gayraliae_MMETSP1464_119497	5,02	4,98	0,04	2	2,98	255
Extubocellulus_spiniifer_265572	7,57	5,57	2	0,86	1,57	255
Fabrea_salina_MMETSP1345_342563	7,1	6,75	0,35	0,9	2	255
Favella_ehrenbergii_MMETSP0123_182087	5,69	5,53	0,16	1,1	3,21	255
Fibrocapsa_japonica_MMETSP1339_94617	3,45	3,37	0,08	2,47	4,08	255
Filamoeba_nolandii_MMETSP0413_136452	7,34	7,22	0,12	1,57	1,09	255
Florenciella_parvula_236787	5,96	4,39	1,57	1,96	2,08	255
Florenciella_sp._236786	3,72	3,45	0,27	2,08	4,2	255
Fragilariopsis_kerguelensis_186038	7,49	5,45	2,04	0,51	2	255
Gambierdiscus_australes_MMETSP0766_439317	5,29	5,25	0,04	1,49	3,22	255
Geminigera_cryophila_MMETSP0799_46947	6,54	6,27	0,27	1,92	1,54	255
Geminigera_sp._MMETSP1102_46946	5,17	4,9	0,27	2,71	2,12	255
Gephyrocapsa_oceanica_38817	6,31	5,29	1,02	1,8	1,89	255
Goniomonas_pacifica_195067	6,86	5,96	0,9	2,08	1,06	255
Goniomonas_sp._MMETSP0114_31324	4,39	4,35	0,04	1,69	3,92	255
Gonyaulax_spiniifera_MMETSP1439_66791	4,15	3,88	0,27	1,61	4,24	255
Grammatophora_oceanica_MMETSP0009_210454	3,53	3,45	0,08	2,78	3,69	255
Guillardia_theta_MMETSP0046_905079	8,32	7,14	1,18	0,82	0,86	255
Gymnoclora_sp._MMETSP0110_629695	7,06	6,94	0,12	0,86	2,08	255
Gymnodinium_catenatum_MMETSP0784_39447	6,79	6,16	0,63	0,98	2,23	255
Hanusia_phi_MMETSP1048_3032	5,05	4,78	0,27	1,69	3,26	255
Haptolina_brevifila_MMETSP1094_156173	4,55	4,51	0,04	2,47	2,98	255
Haptolina_ericina_MMETSP1096_156174	4,98	4,67	0,31	1,8	3,22	255
Helicotheca_sp._MMETSP1171_374046	2,39	2,35	0,04	1,29	6,32	255
Hemiselmis_andersenii_464988	6,31	5,96	0,35	2,27	1,42	255
Hemiselmis_tepida_MMETSP1355_464990	3,53	3,41	0,12	1,84	4,63	255
Heterocapsa_arctica_MMETSP1441_192219	4,83	4,71	0,12	1,65	3,52	255
Heterocapsa_rotundata_MMETSP0503_89963	5,45	5,33	0,12	1,65	2,9	255
Heterocapsa_triquetra_MMETSP0448_66468	5,22	5,02	0,2	1,76	3,02	255
Heterosigma_akashiwo_2829	6,31	4,82	1,49	2,04	1,65	255
Isochrysis_galbana_37099	7,65	3,53	4,12	1,02	1,33	255
Isochrysis_sp._37098	5,92	5,45	0,47	1,84	2,24	255
Isochrysis_sp._40639	5,02	4,67	0,35	1,73	3,25	255
Karenia_brevis_156230	6,78	5,92	0,86	0,82	2,4	255

organism	completeness	single.copy	duplicate	fragmented	missing	total.genes
Karlodinium_veneficum_342587	5,88	4,9	0,98	1,22	2,9	255
Kryptoperidinium_foliaceum_160619	7,57	2,67	4,9	0,98	1,45	255
labyrinthulid_quahog_96639	5,57	5,41	0,16	1,88	2,55	255
Leptocylindrus_danicus_163516	7,96	2,51	5,45	0,55	1,49	255
Licmophora_paradoxa_MMETSP1360_515479	3,33	3,25	0,08	2,82	3,85	255
Lingulodinium_polyedrum_160621	7,06	6,35	0,71	0,75	2,19	255
Litonotus_pictus_MMETSP0209_665100	3,85	3,65	0,2	1,76	4,39	255
Lotharella_globosa_91324	7,21	6,39	0,82	0,9	1,89	255
Lotharella_oceanica_MMETSP0040_641309	5,53	4,98	0,55	1,33	3,14	255
Mantoniella_sp._13607	6,87	6,12	0,75	1,76	1,37	255
Mayorella_sp._MMETSP0417_242541	7,85	7,69	0,16	0,98	1,17	255
Mesodinium_pulex_MMETSP0467_283647	5,49	4,78	0,71	1,57	2,94	255
Micromonas_pusilla_38833	8,94	0,9	8,04	0,31	0,75	255
Minutocellus_polymorphus_265543	7,22	3,53	3,69	1,14	1,64	255
Moneuplotes_crassus_MMETSP1380_5936	3,61	3,45	0,16	2,31	4,08	255
Myrionecta_rubra_MMETSP0798_283649	4,04	3,8	0,24	1,73	4,23	255
Neobodo_designis_MMETSP1114_312471	4,2	4,08	0,12	1,61	4,19	255
Neoceratium_fusus_2916	5,85	5,18	0,67	1,29	2,86	255
Nephroselmis_pyriiformis_MMETSP0034_156128	7,18	5,69	1,49	2,31	0,51	255
Nitzschia_sp._MMETSP0014_2857	6,83	4,75	2,08	0,78	2,39	255
Noctiluca_scintillans_MMETSP0253_2966	6,63	6,12	0,51	0,9	2,47	255
Ochromonadaceae_sp._483371	6,51	6,43	0,08	1,76	1,73	255
Ochromonas_sp._420556	6,24	5,61	0,63	1,41	2,35	255
Odontella_sp._MMETSP0015_2838	6,94	6,12	0,82	1,29	1,77	255
Ostreococcus_sp._548131	8,63	1,73	6,9	0,16	1,21	255
Oxyrrhis_marina_2969	5,8	1,88	3,92	0,9	3,3	255
Palpitomonas_bilix_MMETSP0780_652834	7,37	5,49	1,88	0,78	1,85	255
Paramoeba_aestuarina_MMETSP0161_180227	6,62	6,31	0,31	1,57	1,81	255
Paramoeba_sp._MMETSP0151_200890	5,89	5,73	0,16	2	2,11	255
Paraphysomonas_imperforata_37642	6,23	5,96	0,27	1,65	2,12	255
Paraphysomonas_sp._MMETSP1103_36769	8,39	7,8	0,59	0,55	1,06	255
Paraphysomonas_vestita_MMETSP1107_36770	5,45	5,45	0	1,84	2,71	255
Pavlova_gyrans_MMETSP1466_44452	3,37	3,37	0	1,14	5,49	255
Pavlova_sp._483367	7,69	6,75	0,94	0,86	1,45	255
Pavlova_sp._298111	4,55	4,39	0,16	1,57	3,88	255
Pedinellales_sp._483369	5,96	5,45	0,51	2,04	2	255
Pelagococcus_subviridis_35679	5,34	4,59	0,75	2,04	2,62	255
Pelagodinium_beii_MMETSP1338_43686	5,49	5,02	0,47	1,8	2,71	255
Pelagomonas_calceolata_35677	6,67	2,67	4	1,02	2,31	255
Pelagophyceae_sp._483370	7,38	6,67	0,71	0,78	1,84	255
Percolomonas_cosmopolitus_63605	7,88	4,35	3,53	0,35	1,77	255
Peridinium_aciculiferum_268820	6,04	5,33	0,71	1,33	2,63	255
Perkinsus_chesapeakei_321610	1,53	1,14	0,39	3,02	5,45	255
Perkinsus_marinus_1032511	1,22	1,14	0,08	3,14	5,64	255
Pessonnella_sp._MMETSP0420_498008	4,28	4,2	0,08	2,86	2,86	255
Phaeocystis_antarctica_33657	5,97	5,22	0,75	2,12	1,91	255
Phaeocystis_sp._33656	4,67	2,63	2,04	1,69	3,64	255
Phaeomonas_parva_MMETSP1163_124430	6,2	4,51	1,69	1,8	2	255
Picochlorum_sp._MMETSP1330_464278	6,54	5,76	0,78	1,49	1,97	255
Picocystis_salinarum_88271	8,74	6,78	1,96	0,27	0,99	255
Platyophrya_macrostroma_MMETSP0127_938130	5,02	3,84	1,18	2,27	2,71	255
Pleurochrysis_carterae_13221	5,45	5,1	0,35	1,88	2,67	255
Polarella_glacialis_89957	5,72	5,37	0,35	1,45	2,83	255
Polytomella_parva_MMETSP0052_51329	6,43	6,31	0,12	1,92	1,65	255
Porphyridium_aeruginum_MMETSP0313_2792	3,49	3,49	0	2,47	4,04	255
Prasinococcus_capsulatus_MMETSP0941_156131	7,77	7,14	0,63	1,22	1,01	255
Prasinoderma_singulare_MMETSP1315_676789	5,69	5,61	0,08	2,08	2,23	255
Proboscia_sp._216776	6,27	5,37	0,9	1,65	2,08	255
Prorocentrum_micans_MMETSP0251_2945	0,12	0,12	0	0,2	9,68	255
Prorocentrum_minimum_39449	5,92	4,43	1,49	1,41	2,67	255
Proteomonas_sulcata_MMETSP1049_77928	3,33	2,86	0,47	3,29	3,38	255
Protoceratium_reticulatum_MMETSP0228_133427	5,38	5,18	0,2	1,88	2,74	255
Protocruzia_adherens_MMETSP0216_223996	5,45	5,25	0,2	2,31	2,24	255
Prymnesium_parvum_97485	5,72	5,29	0,43	1,69	2,59	255

organism	completeness	single.copy	duplicate	fragmented	missing	total.genes
Prymnesium_polylepis_72548	6,04	5,61	0,43	1,92	2,04	255
Pseudokeronopsis_sp._MMETSP0211_311385	4,16	4,08	0,08	1,8	4,04	255
Pseudokeronopsis_sp._MMETSP1396_1634380	5,06	4,59	0,47	1,76	3,18	255
Pseudo-nitzschia_australis_44445	7,25	6,47	0,78	0,75	2	255
Pseudo-nitzschia_delicatissima_44447	4,86	4,82	0,04	2,43	2,71	255
Pseudo-nitzschia_fraudulenta_183588	4,86	4,35	0,51	1,92	3,22	255
Pseudo-nitzschia_heimii_MMETSP1423_476441	6,2	6,2	0	1,49	2,31	255
Pseudo-nitzschia_pungens_37318	6,74	5,41	1,33	1,18	2,08	255
Pseudopedinella_elastica_35684	5,72	5,41	0,31	2,31	1,97	255
Pteridomonas_danica_38822	3,73	3,65	0,08	1,33	4,94	255
Pterosperma_sp._MMETSP1438_1461541	4,35	4,35	0	2,67	2,98	255
Pycnococcus_provasolii_41880	5,41	4,86	0,55	1,73	2,86	255
Pycnococcus_sp._MMETSP1085_41879	7,84	7,8	0,04	1,02	1,14	255
Pyramimonas_parkeae_36894	7,88	7,49	0,39	1,02	1,1	255
Pyramimonas_sp._36882	7,53	3,53	4	1,45	1,02	255
Pyrodictum_bahamense_MMETSP0796_73915	6,4	6,16	0,24	1,02	2,58	255
Rhizochromulina_marina_MMETSP1173_1034831	6,47	6,27	0,2	1,37	2,16	255
Rhodella_violacea_38298	4,16	3,73	0,43	2,71	3,13	255
Rhodomonas_sp._302021	4,04	3,84	0,2	2,71	3,25	255
Rhodomonas_sp._37096	7,37	6,86	0,51	1,57	1,06	255
Rhodorus_marinus_101924	7,92	1,96	5,96	0,63	1,45	255
Rosalina_sp._MMETSP0190_37353	2,71	2,71	0	2,04	5,25	255
Sarcinocypris_sp._MMETSP1170_54406	1,77	1,69	0,08	1,18	7,05	255
Schizochytrium_aggregatum_4773	6,86	6,47	0,39	1,45	1,69	255
Schmidingerella_taraikaensis_626734	6,11	5,8	0,31	1,02	2,87	255
Scrippsiella_sp._54902	7,21	5,37	1,84	0,59	2,2	255
Scrippsiella_trochoidea_71861	5,84	4,9	0,94	1,29	2,87	255
Scyphosphaera_apsteinii_MMETSP1333_418940	5,92	5,92	0	1,65	2,43	255
Sexangularia_sp._MMETSP0437_1486928	4,51	4,51	0	2,16	3,33	255
Skeletonema_grethae_MMETSP0578_267565	3,65	3,53	0,12	2,04	4,31	255
Skeletonema_sp._2842	7,96	2,2	5,76	0,55	1,49	255
Stephanopyxis_turris_MMETSP0794_515487	4,82	3,8	1,02	2,39	2,79	255
Stichococcus_sp._MMETSP1473_1461547	6,83	6,59	0,24	1,49	1,68	255
Striatella_unipunctata_MMETSP0800_210618	3,1	2,98	0,12	2,94	3,96	255
Strombidinopsis_acuminata_MMETSP0126_141414	6,32	4,71	1,61	1,76	1,92	255
Strombidinopsis_sp._MMETSP0463_181622	6,55	5,96	0,59	1,06	2,39	255
Strombidium_inclinatum_MMETSP0208_197538	5,6	5,33	0,27	1,02	3,38	255
Strombidium_rassoulzadegani_MMETSP0449_1082188	5,68	5,41	0,27	0,78	3,54	255
Stygamoeba_regulata_MMETSP0447_1042828	8,51	5,53	2,98	0,86	0,63	255
Symbiodinium_kawagutii_104179	1,33	1,06	0,27	1,25	7,42	255
Symbiodinium_sp._230996	6,19	5,8	0,39	1,41	2,4	255
Symbiodinium_sp._2949	5,84	1,88	3,96	1,22	2,94	255
Synedropsis_cf._MMETSP1176_265551	4,71	4,67	0,04	2,12	3,17	255
Tetraselmis_chuii_MMETSP0491_63592	5,38	5,22	0,16	2,39	2,23	255
Tetraselmis_sp._MMETSP0419_582737	6,67	6,43	0,24	1,61	1,72	255
Tetraselmis_striata_3165	8,24	6,75	1,49	1,18	0,58	255
Thalassionema_frauenfeldii_MMETSP0786_186022	7,73	7,49	0,24	0,59	1,68	255
Thalassionema_nitzschoides_33649	7,81	7,65	0,16	0,63	1,56	255
Thalassiosira_aff._420261	6,94	5,96	0,98	1,18	1,88	255
Thalassiosira_gravida_420259	4,82	4,31	0,51	2,63	2,55	255
Thalassiosira_minuscula_385413	6,63	5,69	0,94	1,37	2	255
Thalassiosira_oceanica_159749	6,59	5,53	1,06	1,57	1,84	255
Thalassiosira_punctigera_MMETSP1067_265564	5,57	5,41	0,16	2,27	2,16	255
Thalassiosira_rotula_49265	6,35	5,88	0,47	1,57	2,08	255
Thalassiosira_sp._35127	7,33	3,53	3,8	1,02	1,65	255
Thalassiosira_weissflogii_67004	7,65	5,45	2,2	0,67	1,68	255
Thalassiothrix_antarctica_1049557	6,31	5,84	0,47	1,61	2,08	255
Thraustochytrium_sp._1112570	7,93	7,22	0,71	0,94	1,13	255
Tiarina_fusa_MMETSP0472_693140	6,59	4,55	2,04	1,88	1,53	255
Timspurckia_oligopyrenoides_MMETSP1172_708627	5,02	4,9	0,12	2,31	2,67	255
Togula_jolla_MMETSP0224_285029	5,56	5,29	0,27	1,53	2,91	255
Triceratium_dubium_MMETSP1175_1003142	2,43	2,31	0,12	2,31	5,26	255
Trichosphaerium_sp._MMETSP0405_498012	4,43	4,39	0,04	2,24	3,33	255
Tryblionella_compressa_913974	5,85	5,1	0,75	1,84	2,31	255

organism	completeness	single.copy	duplicate	fragmented	missing	total.genes
uncultured_ciliate_MMETSP1317_187299	1,29	1,25	0,04	1,69	7,02	255
Uronema_marinum_MMETSP0018_35107	2,86	2,82	0,04	1,96	5,18	255
Vannella_sp._95228	9,14	4,24	4,9	0,43	0,43	255
Vexillifera_sp._MMETSP0173_218659	5,8	5,76	0,04	2,39	1,81	255

Table S4: Data file 8 - Summary of BUSCO v4.0.5 completeness assessment based on 255 single-copy orthologous eukaryotic markers. The table reports the proportion of complete genes (completeness), complete single-copy genes (single.copy), duplicated genes (duplicate), fragmented genes (fragmented), and missing genes (missing), as well as the total number of markers evaluated (total.genes).

6. Transition: towards a large-scale complex algae analysis

Having established two complementary resources with controlled quality and broad taxonomic coverage: (i) the orthologous clans (orthoclans) built from curated reference proteomes (Chapter III) and (ii) the 172 consolidated algal MMETSP transcriptomes obtained after stringent contamination and redundancy control (Chapter IV), we next combine them into a single working dataset. In Chapter V, we will focus specifically on the tagged orthoclans (i.e., orthoclans identified as plastid-related based on targeting) and enrich them with MMETSP-derived sequences using the orthology enrichment tool that I contributed to develop with Pr. Denis Baurain: Forty-Two (Chapter II). This enrichment increases taxonomic sampling across algal lineages while preserving orthology, providing the foundation for the single-gene phylogenies and downstream transfer analyses presented in the remaining of the thesis. Thus, the main goal of Chapter V is to summarize inferred transfer events as organism-level donor α -diversity profiles, providing a quantitative metric to compare not only the number of contributing donor lineages, but also the relative distribution of their contributions across receiver lineages.

V. Chapter 5 - A Quantitative Phylogenomic Approach

Contributions

D. Baurain developed the code and performed the KEGG functional analyses (following eggNOG-mapper annotation). I designed the sets of events to be analyzed, reorganized pathways into KEGG-inspired categories, and derived all biological interpretations.

H. Brinkmann conducted an expert analysis of 242 single-gene trees to assess the performance of my transfer detection algorithms. He also assisted with species subsampling, based on my ribosomal phylogenomic tree. In addition, we had numerous insightful discussions related to my PhD topic.

All supplementary materials are available on Figshare : <https://figshare.com/projects/.../275595>

1. Introduction

In the General introduction, we depicted how complex algae form a network of photosynthetic organisms scattered across the eukaryotic tree. However, the exact mechanisms by which their plastids were acquired remain unclear. Contrary to expectations of secondary endosymbiotic models, a sizable part of plastid-targeted genes is not from the same origin as the plastid itself. For example, for CASH lineages (bearing complex red plastids), scenarios like the cryptic serial endosymbiosis or the Shopping Bag Endosymbiosis model provide a possible interpretation of gene mosaicism but focus on the endosymbiotic mechanism. Similarly, while the Rhodoplex hypothesis fits the single phylogenetic origin of those plastids, it does not explain gene mosaicism. Moreover, it implies multiple tertiary (or quaternary or even quinary) endosymbiotic events, whereas most complex algae do not bear nucleomorphs and only have three or four membranes surrounding their plastids.

To overcome the inconsistencies of those models, kleptoplasty was suggested as an additional mechanism for explaining plastid spread, not only in CASH lineages but also in euglenids. In line with the *shopping bag endosymbiosis model*, this hypothesis posits multiple transient interactions with prey of diverse origins but also proposes a rationale for the selective force driving the progressive accumulation of plastid-targeted genes during this pre-adaptation phase: to exploit and maintain functional kleptoplastids for increasingly longer periods of time, before ultimately reducing them into fully integrated plastids. In such a scenario, the phylogenetic diversity of nucleus-encoded plastid-targeted genes would be higher than predicted with endosymbiotic models, where genes mostly originate from a single source, the algal endosymbiont.

To test the idea, we assembled a high-quality dataset through careful contamination handling and controlled orthology definition yielding high-quality single-gene trees for which we designed and implemented an automated phylogenetic pipeline (PAGTI) for detecting transfers in a discovery-driven and taxonomy-aware fashion. By transforming individual transfer events to organism-centric α -diversities, this approach should allow us to quantify the relative contribution of endosymbiosis and kleptoplasty in the origin of complex algae, not only at a global scale but also for specific lineages of interest. Beyond α -diversities, individual taxonomic profiles of donor/receiver organisms belonging to the same donor/receiver lineages should help us to assess the congruence of the transfer signal, whereas multiple events occurring in the same orthoclan would point to convergent evolution, thus possibly indicating selection pressure for the corresponding function. To investigate these aspects, both event-level and orthoclan-level functional annotation will be considered in the framework of KEGG enzymes, modules and pathways.

1.2. Assembling a high-quality dataset

The primary objective of this PhD was to assemble a high-quality, taxonomically broad dataset suited for large-scale gene-transfer analyses in the context of plastid acquisition. Because the genes of interest (nucleus-encoded plastid-targeted sequences) are by nature transferable between lineages, the quality of the dataset is critical: residual contamination can produce false-positive transfer signals that are indistinguishable from genuine horizontal gene acquisitions. Dataset construction therefore proceeded in three controlled steps.

First, a curated set of 73 publicly available proteomes was compiled (Chapter III) and screened for contamination using Forty-Two's ribosomal marker pipeline (Chapter II), retaining only assemblies meeting strict purity and completeness criteria (Figure 23). These proteomes were used for orthology inference, producing a consistent set of orthogroups from which photosynthetic subtrees ("clans") were extracted. Because these clans result from pruning larger orthogroups, they are referred to as orthoclans throughout this chapter.

Second, orthoclans were filtered to retain only those containing at least one sequence tagged for plastid targeting, using four compartment/encoding tags: `cpcpt#` (plastid-encoded plastid-targeted), `nucpt#` (nucleus-encoded plastid-targeted), `nuppct#` (nucleus-encoded periplastid-targeted), and `nm#` (nucleomorph-encoded). This filtering yielded 1,788 plastid-related orthoclans forming the core dataset of this chapter. Orthoclans can yield several tags, as described and summarised in Figure 24, although most of them are unique to `nucpt#` (1,373).

Third, to improve taxonomic coverage beyond what curated proteomes alone could provide, these orthoclans were enriched with sequences from 172 consolidated MMETSP transcriptomes (Chapter IV, sampling described in Figure S8 of section 5.2) using Forty-Two's orthology-guided enrichment procedure (Chapter II), expanding the sampling across all major photosynthetic lineages (Figure 25). The following sections detail the criteria used to select orthoclans, classify transfer events and how we decided to analyse them, before presenting the results of the transfer inference and diversity analyses.

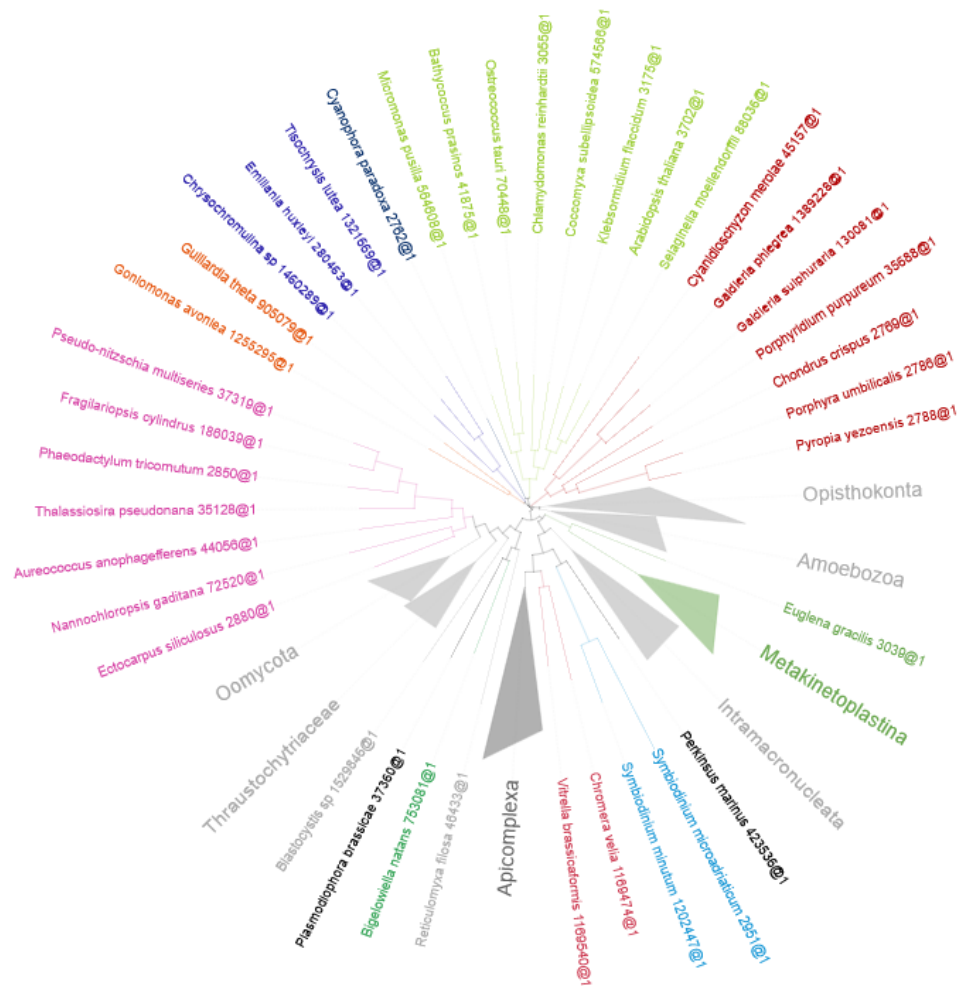


Figure 23: Phylogenetic tree (OrthoFinder species tree) illustrating the initial curated proteome sampling (73 species).

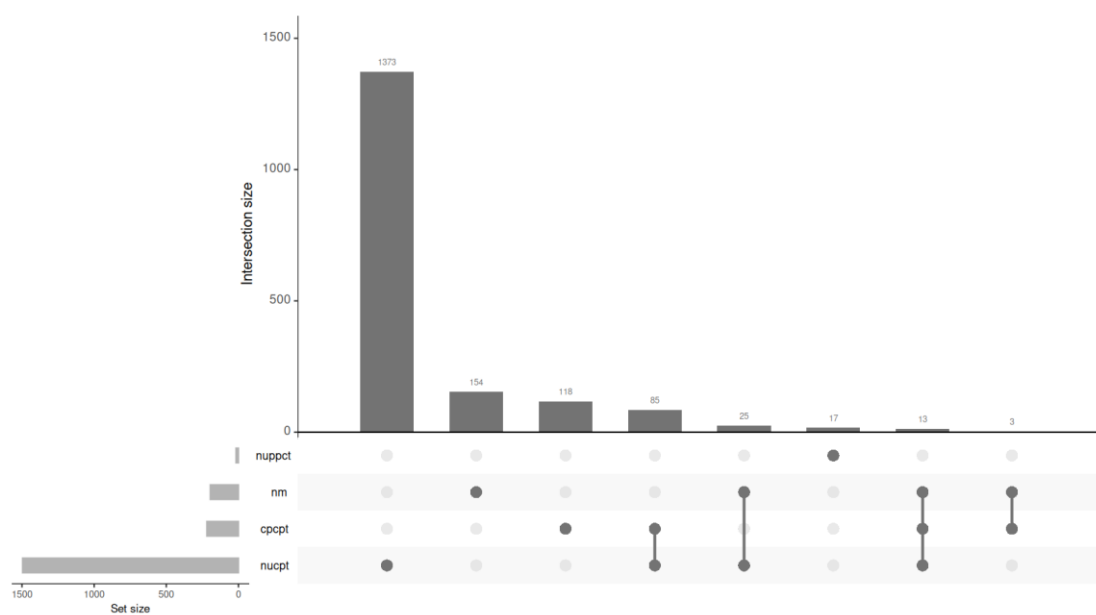


Figure 24: Summary overlaps among the four gene sets: nm# ($n = 195$), nucpt# ($n = 1,496$), nupct# ($n = 17$), and cpcpt# ($n = 219$), for a total of 1,788 unique orthoclans. Horizontal bars (left) indicate the size of each set. Vertical bars (top) indicate the size of each observed intersection, with the dot matrix below marking which sets contribute to each intersection.

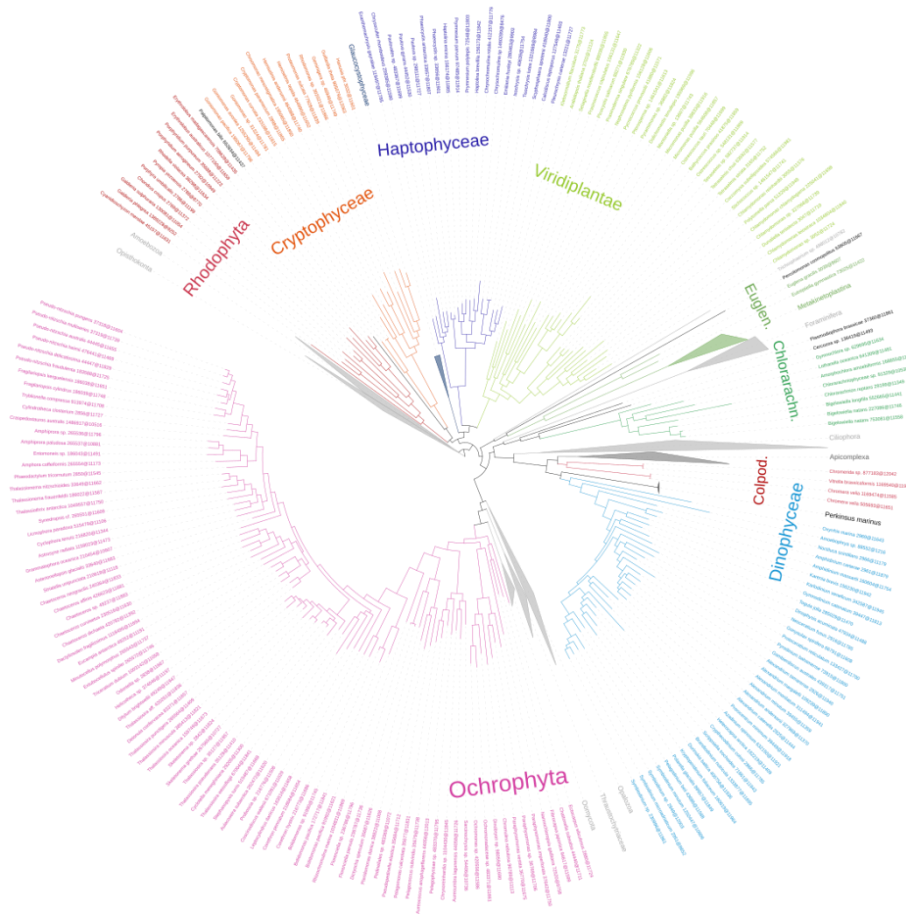


Figure 25: Phylogenetic species inferred from a supermatrix of 78 ribosomal gene markers with 293 species (73 curated proteomes enriched with MMETSP consolidated samples) and 12,224 positions.

1.2.1. Orthoclan selection

Transfer inference relied on orthoclans restricted to photosynthetic taxa and containing at least one tagged sequence. Orthoclans composed exclusively of cpcpt# sequences (i.e., plastid-encoded only) were not included in the main analysis, but those were retained and used as a control (i.e., expected signal for strictly plastid-encoded genes). All other orthoclans, including single-tag orthoclans (nucpt#, nuppct# or nm#) were retained. Mixed-tag orthoclans (Figure 24) were also retained because inference is performed at the event level and can, in principle, capture scenarios such as plastid-to-nucleus transfers (cpcpt# → nucpt#) or nested transfers (e.g., cpcpt# → nucpt# followed by nucpt# → nucpt#).

1.2.2. Event classification

Transfer events were classified according to the tag composition of donor and receiver sequences, using a rule-based procedure designed to distinguish plastid-encoded from nucleus-encoded plastid-associated signals. For the main analyses, nucpt#, nuppct#, and nm# categories were merged into a single operational class, hereafter referred to as “nucpt#”, whereas cpcpt# sequences were treated

separately as a plastid-encoded control set. The detailed classification procedure is described in the Methods section 5.4.

1.3. Gene transfer inference

1.3.1. Paraphyly-Aware Gene Transfer Inference (PAGTI)

Several computational approaches have been developed to infer horizontal or endosymbiotic gene transfer in photosynthetic eukaryotes. A first class of methods relies on sequence composition or similarity searches to identify genes with atypical evolutionary affinities (Doolittle et al., 2003; Zhaxybayeva and Doolittle, 2011). Among these, BLAST-based approaches flag candidate transfers when a gene shows stronger similarity to distant lineages than to close relatives of the receiver organism, as implemented for example in Alieness and HGTector (Rancurel et al., 2017; Zhu et al., 2014). These approaches are useful for screening large datasets, but they are sensitive to database composition and taxonomic sampling (Cornet and Baurain, 2022b; Cornet et al., 2024), and they do not always distinguish true transfers from missing data and differential gene loss. To address this, phylogeny-based approaches were developed to inspect the topology of single-gene trees. One early example is PhyloSort, which was designed to sort large collections of gene trees according to predefined monophyletic relationships and was applied to estimate endosymbiotic gene transfer in *Chlamydomonas* (Moustafa and Bhattacharya, 2008). This logic was later extended by PhySortR, which allows more flexible searches for exclusive or non-exclusive clades across large tree sets (Stephens et al., 2016). More recently, TreeSorter was used to classify plastid-targeted proteins in kareniacean dinoflagellates (Novák Vanclová et al., 2024). Other automated pipelines, such as AvP, similarly combine homolog retrieval, tree reconstruction and topological criteria to identify candidate HGTs at scale (Koutsovoulos et al., 2022). As with all HGT inference, however, these approaches remain interpretative: gene-tree incongruence can reflect transfer, but also incomplete sampling, hidden paralogy, gene-tree error or alternative evolutionary histories (Ravenhall et al., 2015).

In this thesis, we developed a pipeline called PAGTI (Paraphyly-Aware Gene Transfer Inference) that differs from these approaches in two main respects: it identifies transfer-compatible topologies based on paraphyletic donor assemblages rather than sister-group or best-hit criteria, and it systematically translates the inferred events into donor α -diversity profiles, enabling quantitative comparison across receiver lineages. The implementation of PAGTI is described in Methods (Section 5.3).

1.3.2. Illustrative topologies

For large-scale analysis and optimal reproducibility, gene transfer inference was automated through a Perl implementation. Figure 26 illustrates a typical example highlighting two key features of the algorithm: (i) the inference mechanism and (ii) its tolerance to “intruders”. The script first identifies a candidate receiver clan belonging to a major complex algal group (Table 1) (here Cryptophyceae). Although the subtree is formally non-monophyletic because of a single *Dinobryon* sp. sequence, the algorithm still treats Cryptophyceae as effectively monophyletic and proceeds to search the surrounding topology for a paraphyletic donor assemblage (here, Rhodophyta). In this configuration, *Dinobryon* sp. is considered an intruder and ignored for inference, because it is not an allowed donor lineage for Cryptophyceae and is represented by only one sequence. By contrast, if several Ochrophyta sequences were present in the same position, the topology would be interpreted as evidence for a nested transfer, as illustrated in Figure 27. Each inferred transfer event is then recorded at two complementary levels:

- lineage level: each donor/receiver event contributes one count to the corresponding donor/receiver lineage pair within a given tree (orthoclan).
- species level: the same event is decomposed by receiver species, so that each species in the receiver clan accumulates its own per-species event count.

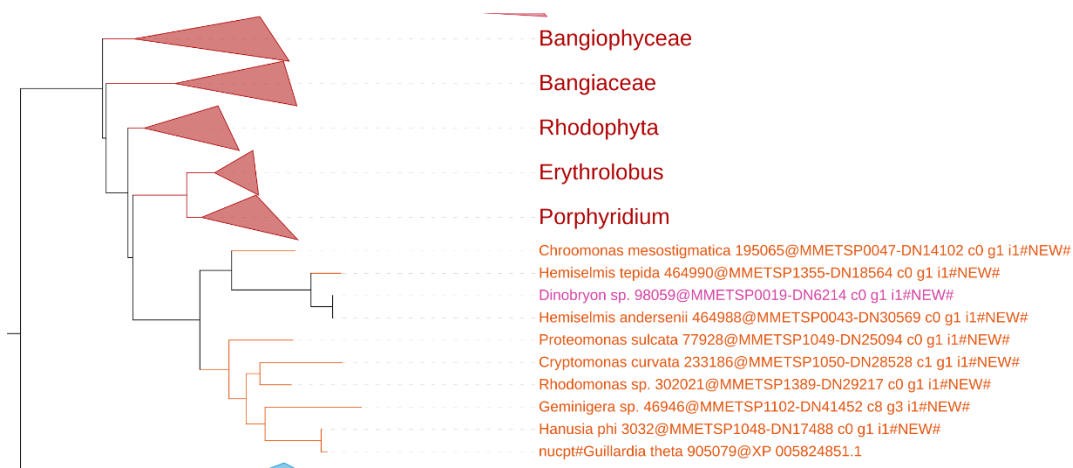


Figure 26: Tree of *PsbO* gene (OG0003862-0).

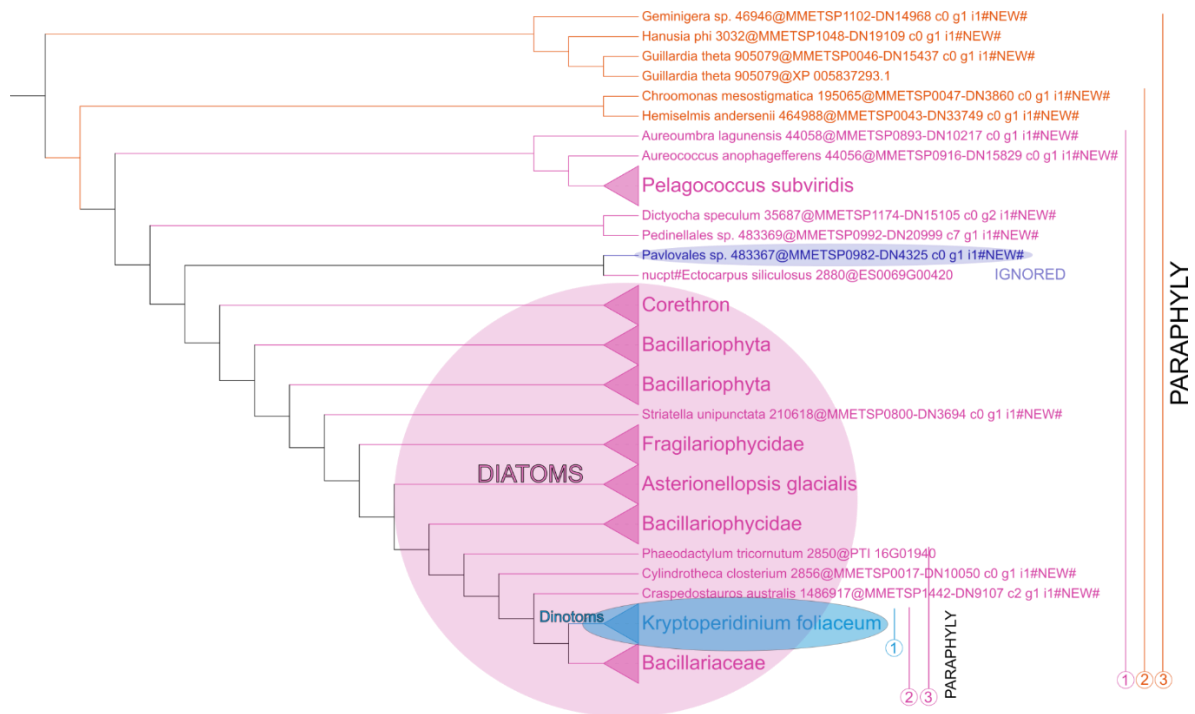


Figure 27: Nested gene transfer (OG0015365-0, DUF3598) illustrating the paraphyly detection algorithm. Two transfers are inferred.

1.3.3. Inferred transfer events

Parsing the 1,670 nucpt# gene trees yielded a total of 3,353 inferred transfer events across 1,126 orthoclans, whereas 544 orthoclans returned no event. As shown in Figure 28, the distribution of events per orthoclan is strongly right skewed: the majority of orthoclans contribute only one or two events, and the frequency drops rapidly beyond five events per orthoclan. A small number of orthoclans nonetheless accumulate up to 23 events.

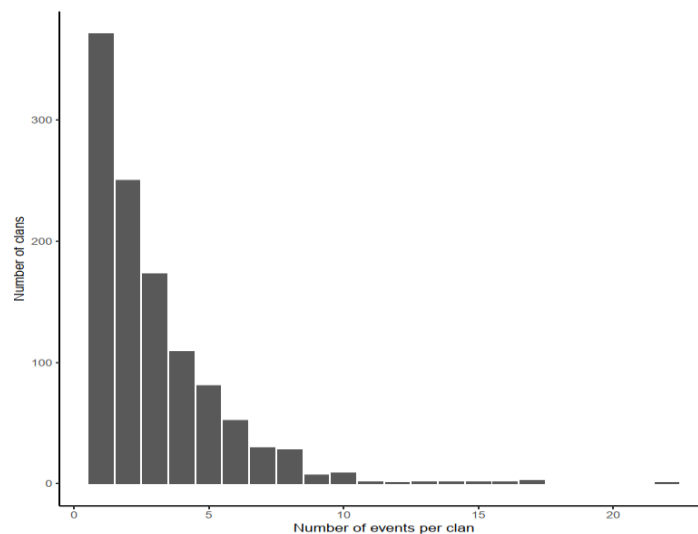


Figure 28: **Distribution of inferred transfer events per orthoclan.** Histogram showing the number of nucpt# orthoclans (y-axis; $n = 1,126$) as a function of the number of inferred transfer events they contain (x-axis; $n = 3,353$). Orthoclans with no inferred event ($n = 544$) are not shown.

1.3.4. Objective and strategies

These 3,353 transfer events form the empirical basis of the analyses presented in this chapter, which are organized around three complementary perspectives. First, the directionality and magnitude of gene fluxes between donor and receiver lineages are examined at a global scale (Section 3.1.1). Second, donor diversity is summarized at the species level using an entropy-based α -diversity metric, allowing us to test our kleptoplasty hypothesis by examining how *nucpt#* genes have been recruited across complex algal lineages (Section 3.1.2). Third, these global patterns are analysed through a series of integrative case studies, combining α -diversity, donor composition, and functional insights to reconstruct the evolutionary trajectory of specific lineages (Section 3.2).

To interpret these transfer events in functional terms, KEGG Orthology identifiers (KO-IDs, or K numbers) were used (Kanehisa and Goto, 2000). A KO-ID links a protein sequence to a defined molecular function. KEGG modules (M numbers) are compact functional units composed of one or several KO IDs, whereas KEGG pathway maps (map IDs) represent broader biological pathways within which these modules are embedded. In practice, this establishes the following hierarchy: sequence $\rightarrow$ KO ID $\rightarrow$ module $\rightarrow$ pathway map. KO-IDs were assigned to clan sequences using eggNOG-mapper (Cantalapiedra et al., 2021) and mapped to KEGG modules using *anvi'o*; (Eren et al., 2021) this allowed us to compare event sets in terms of KEGG pathways and KEGG-inspired categories (see Methods 5.7 for full details).

The KEGG functional library considered here comprised a total of 479 modules and we first quantified the extent to which this repertoire was addressable in our dataset. Coverage of the KEGG module repertoire was assessed at three increasingly restrictive levels: (i) across all 1,670 *nucpt#* orthoclans, (ii) restricted to orthoclans containing at least one inferred transfer event, and (iii) restricted to KO-IDs assigned directly to sequences involved in transfer events. When considering modules for which at least one KO-ID was recovered, the 1,670 *nucpt#* orthoclans covered 229 modules. Restricting the analysis to orthoclans with at least one inferred transfer event reduced this number to 206 modules, and focusing only on KO-IDs from sequences directly involved in transfer events further reduced it to 175 modules. In the reverse perspective, using the strict event-level strategy (iii), 572 events (from 256 orthoclans) could be assigned to 175 KEGG modules. Extending the annotation to the orthoclan-level (ii) increased coverage to 875 events (from 306 orthoclans), spanning 206 KEGG functional modules. We also quantified module recovery using a module completeness threshold of 0.5 in *anvi'o*. Under this criterion, the 1,670 *nucpt#* orthoclans covered 123 modules. This number decreased to 103 modules when only orthoclans with at least one inferred transfer event were considered, and to 81 modules when the analysis was restricted to KO-IDs from sequences directly involved in transfer events.

As shown in Figure 29, functional coverage decreases progressively across these levels, from 229 to 206 to 175 modules when considering any KO-ID match, and from 123 to 103 to 81 modules when further applying the module completeness threshold of 0.5. This nested structure defines the functional space within which downstream analyses are anchored. In this work, we settled on the most restrictive functional annotation strategy (i.e., event-level KO-IDs and module completeness threshold of 0.5). Consequently, only 35% of our KEGG ‘universe’ (81 modules out of 229) was addressable, and an even smaller fraction of our event sets (572 events out of 3,353, hence 17%) could actually contribute to functional insights. This might seem frustrating, but this is mostly due to the large number of unannotated areas in our trees. Besides, we had no choice to minimize over-interpretations.

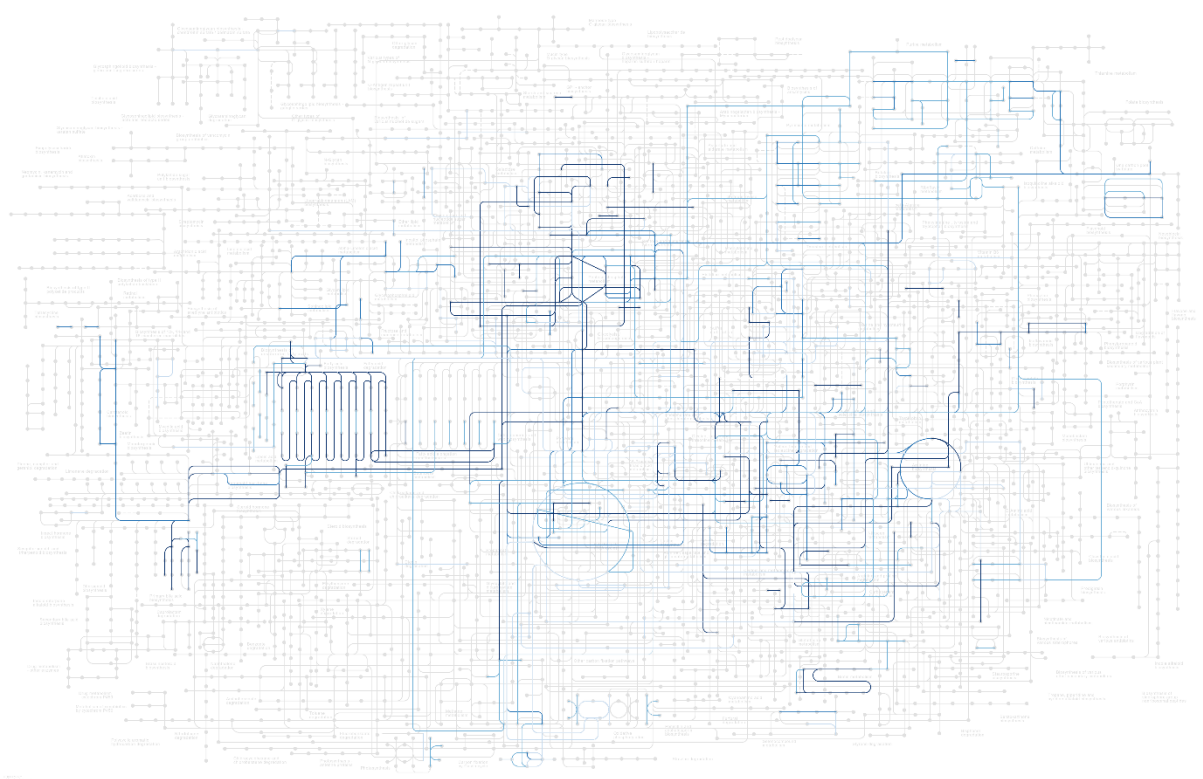


Figure 29: Coverage of the KEGG module repertoire by KO IDs recovered from nucpt# orthoclans and inferred transfer events. The grey background represents the full KEGG functional library (479 modules), whereas coloured elements indicate modules represented in our dataset. Light blue corresponds to modules recovered when considering KO IDs across all 1,670 nucpt# orthoclans, including orthoclans without inferred transfer events. Intermediate blue indicates modules recovered from orthoclans containing at least one inferred transfer event. Dark blue highlights the most specific subset, namely modules recovered from KO IDs assigned directly to sequences involved in inferred transfer events. Because these three categories are nested, increasing colour intensity reflects increasingly restrictive and event-specific levels of functional support.

2. Results

3.1. Global analysis

3.1.1. Large-scale gene fluxes

We first examined how these 3,353 events were distributed among major lineages by summarizing transfers between donor and receiver groups (Figure 30). At the broadest level, transfers are unevenly distributed among donor lineages, with a small number of groups accounting for the majority of inferred events (Figure 30B). In particular, ochrophytes dominate the donor pool, contributing 1,662 events (~50% of all transfers), followed by Viridiplantae (775 events) and Haptophyta (474 events), whereas all other donor lineages contribute comparatively fewer events. When considering the donor-receiver fluxes (Figure 30A), these transfers appear structured rather than randomly distributed, with particular donor lineages contributing preferentially to specific receiver groups; for example, Viridiplantae contribute mostly to complex green algae (Chlorarachniophyceae and euglenids) or rhodophytes, which contribute mostly to Cryptophyceae. Finally, the taxonomic diversity of donors also differs among lineages (Figure 30C): some donors contribute large numbers of events but affect a limited set of receiver groups, whereas others, despite contributing fewer events overall, are detected across a broader diversity of receiver lineages. This contrast is illustrated by Rhodophyta, which, despite contributing a relatively modest fraction of total transfers (231 events), are detected across eight receiver lineages, representing the broadest phylogenetic reach among donor groups.

These summaries provide a quantitative overview of how transfers are distributed among donor and receiver lineages and indicate that gene transfers among complex algal lineages do not follow a simple endosymbiotic-like scenario involving a single dominant donor repeatedly transferring genes to a receiver (Curtis et al., 2012). Instead, they reflect recurrent contributions from multiple lineages, resulting in a dense network of exchanges across the eukaryotic diversity (Figure 31). Within this network, some connections appear reciprocal, with transfers detected in both directions between particular groups (e.g., haptophytes and ochrophytes). Besides, some connections are consistent with known plastid relationships, whereas others connect groups with distinct plastid histories. In several cases, the magnitude of the inferred fluxes is substantial, particularly between ochrophytes and haptophytes (Dorrell et al., 2017), or between Viridiplantae and multiple receiver lineages. While part of this signal may reflect horizontal transfers occurring during different phases of the plastid acquisition process, including pre-adaptation, integration (or fixation), or lineage-specific remodelling (Hehenberger et al., 2019), the overall pattern shows that gene exchange among eukaryotic algal lineages has been more pervasive than predicted by simple endosymbiotic models. Together, these

observations suggest that the landscape of gene transfers among complex algae cannot be fully explained by plastid inheritance alone.

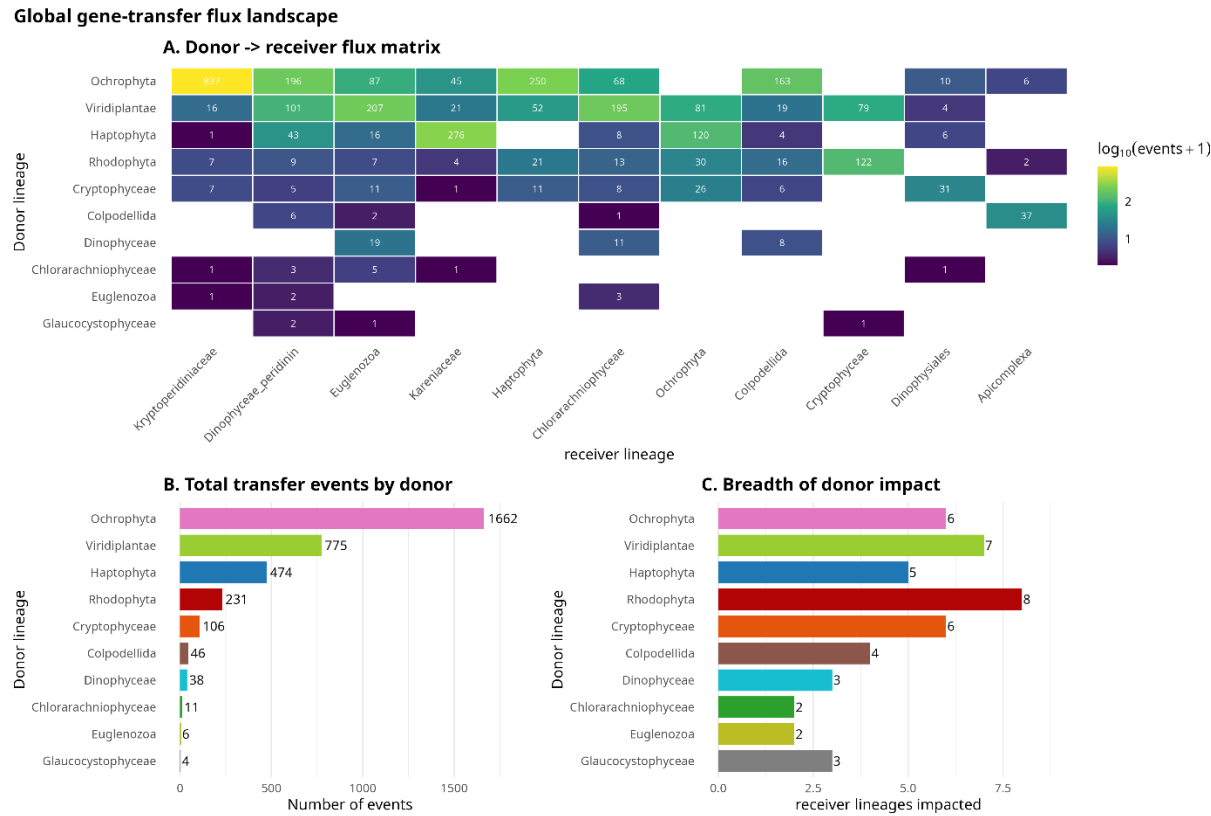


Figure 30: Global gene-transfer flux landscape across complex algal lineages. (A) Donor-receiver flux matrix showing the number of inferred transfer events between donor and receiver lineage pairs. Cell colour represents $\log_{10}(\text{event count} + 1)$, and numbers indicate raw counts. Dinoflagellates are subdivided into major plastid-related groups (peridinin-containing Dinophyceae, Karenia, Kryptoperidiniaceae, and Dinophysiales) to reflect their distinct plastid histories. (B) Total number of inferred transfer events contributed by each donor lineage. (C) Diversity of donor impact measured as the number of receiver lineages affected by each donor lineage (counted at the higher taxonomic level to avoid overcounting dinoflagellate subgroups).

In the following, rather than considering transfer events from single donors independently, we transformed all contributions within each individual receiver genome into a single measurement, in attempt to summarize the signal and identify (possibly lineage-specific or mechanism-specific) trends in our event dataset. Specifically, we quantified the diversity and balance of donor lineages at the receiver species level using an entropy-based α -diversity metric.

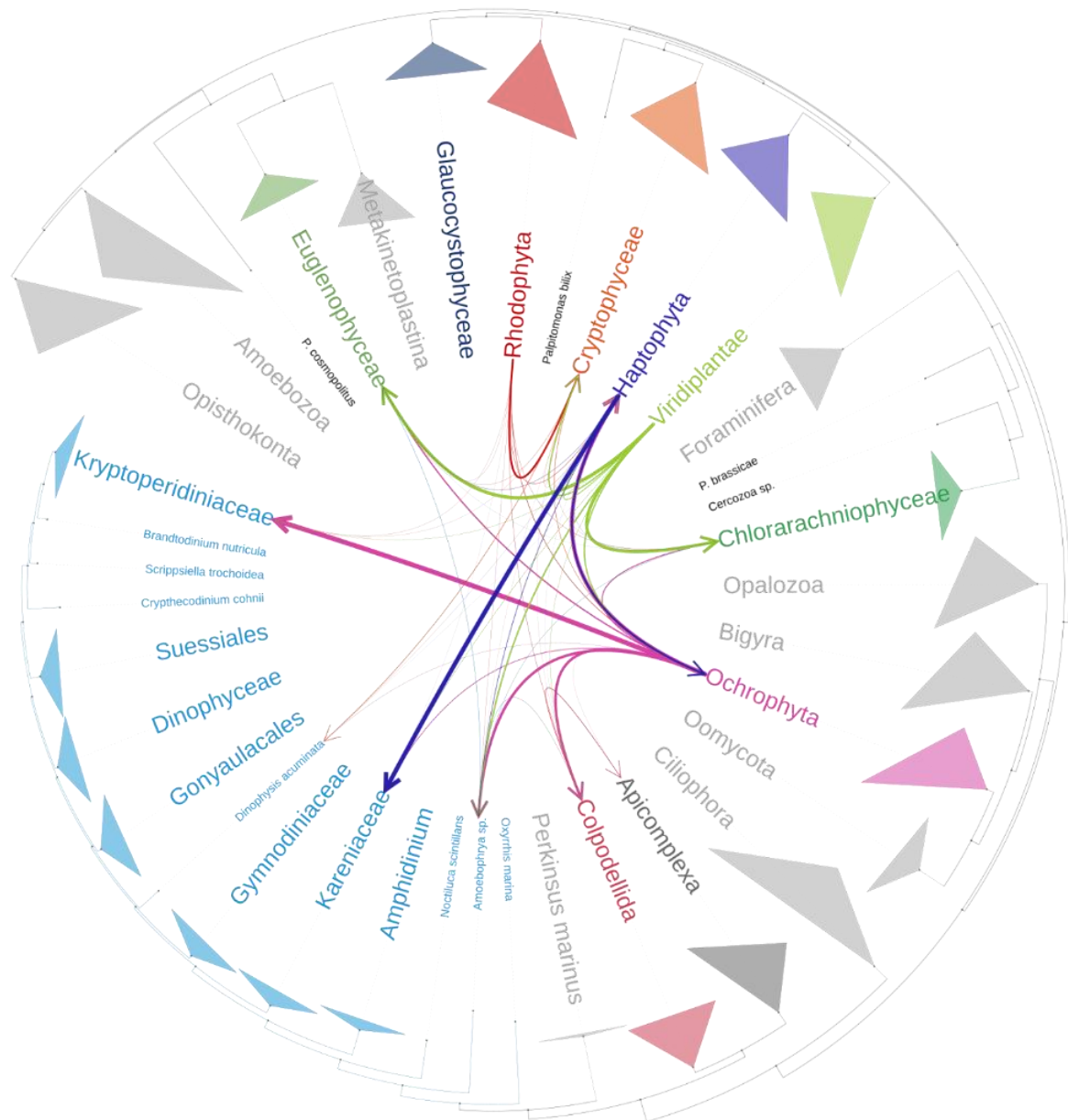


Figure 31: **Phylogenetic representation of inferred gene-transfer fluxes among eukaryotic algal lineages.** Circular phylogenetic tree inferred from a supermatrix of 78 ribosomal markers see Methods 5.9. Arrows represent inferred gene-transfer events between donor and receiver lineages, with arrow thickness proportional to the number of transfers detected between the corresponding lineage pairs. Dinoflagellates are subdivided into major plastid-related groups (peridinin-containing Dinophyceae, Kareniaceae, Kryptoperidiniaceae, and Dinophysiales) to reflect their distinct plastid histories.

3.1.2. Interpretation of the α -diversity metric

Using these 3,353 inferred events, we computed a Shannon-Wiener α -diversity (Shannon, 1948) metric for each species (receivers), which is used as a compositional summary of donors rather than as an estimate of ‘true’ ecological diversity. Because entropy-based diversity estimates are sensitive to low sampling, we applied a conservative minimum threshold of 30 inferred transfer events for an individual species to be taken into account (see Methods 5.5.4). This led to the exclusion of 13 species (Table 6), reducing the dataset from 164 to 151 species, excluding a total of 224 events (<7%), and primarily affecting ochrophytes, which are the most densely sampled lineage in the dataset ($n = 85$).

<i>receiver_label</i>	<i>receiver_org</i>	<i>org_transfers</i>	<i>alpha_div</i>
<i>Ochrophyta</i>	<i>Pteridomonas danica</i> _38822	8	1.106687
<i>Ochrophyta</i>	<i>Paraphysomonas vestita</i> _36770	8	1.045154
<i>Ochrophyta</i>	<i>Paraphysomonas</i> sp._36769	9	1.059079
<i>Ochrophyta</i>	<i>Paraphysomonas imperforata</i> _37642	10	1.238595
<i>Apicomplexa</i>	<i>Cryptosporidium parvum</i> _353152	11	0.415916
<i>Cryptophyceae</i>	<i>Goniomonas avonlea</i> _1255295	13	0.917802
<i>Ochrophyta</i>	<i>Sarcinochrysis</i> sp._54406	17	1.084742
<i>Ochrophyta</i>	<i>Ochromonadaceae</i> sp._483371	22	1.246164
<i>Ochrophyta</i>	<i>Chaetoceros curvisetus</i> _230516	23	1.030897
<i>Ochrophyta</i>	<i>Chromulina nebulosa</i> _96789	24	1.111545
<i>Ochrophyta</i>	<i>Fibrocapsa japonica</i> _94617	25	1.048957
<i>Ochrophyta</i>	<i>Ochromonas</i> sp._420556	26	1.313445
<i>Ochrophyta</i>	<i>Dinobryon</i> sp._98059	28	1.173162

Table 6: List of removed species by the minimum transfer threshold (<30) for computation of α -diversity.

3.1.2.1. Per species distribution

To assess whether species-level α -diversity values naturally cluster into discrete ranges, we analysed their distribution independently of their taxonomic affiliation (Figure 32). The overall distribution suggests the existence of a clear low-diversity cluster, while the transition between intermediate and high diversity is delimited by a smaller but nonetheless clear gap. Rather than setting arbitrary boundaries, we used statistical approaches to determine objective cutoffs. Among the different strategies tested, kernel density estimation (KDE) was selected because it provides a non-parametric and data-driven way to identify natural divisions within the distribution, based on the local minima of the density curve. This approach allowed us to separate lineages into low, intermediate, and high diversity groups according to the intrinsic shape of their α -diversity distribution.

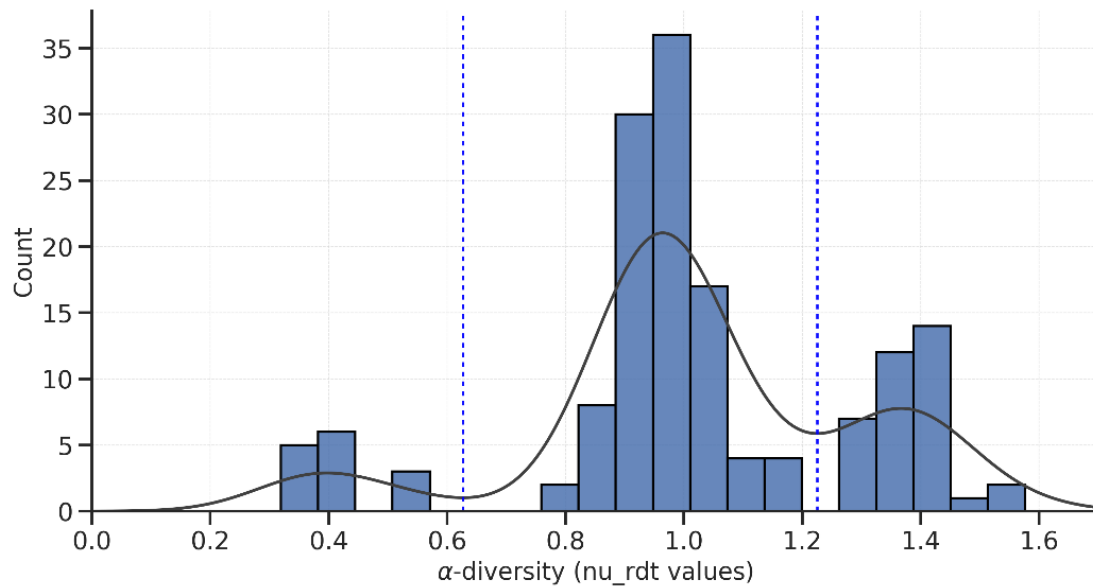


Figure 32: Histogram showing the distribution of species-level α -diversity values, pooling all lineages (anonymous species). The solid curve corresponds to the Gaussian kernel density estimate (KDE). Vertical dashed lines indicate KDE local minima, used as data-driven cutoffs delimiting α -diversity ranges along the distribution Low to intermediate cutoff: 0.67; intermediate to high cutoff: 1.21.

Because Shannon-Wiener α -diversity (H') blends donor richness and evenness, a given H' value can arise from different underlying compositions of donors. We therefore decomposed global H' patterns by jointly examining donor richness (S) and evenness (Pielou, 1966), to determine whether low, intermediate, and high diversity bins primarily reflect changes in the number of donors, shifts in dominance, or both (Figure 33). Globally, H' increases with donor richness, but most of the spread at a given richness is explained by evenness. Species with similar donor richness can display markedly different H' values depending on whether transfers are evenly distributed across donors (high J) or dominated by a single donor (low J). The isocurves highlight that “low-diversity” (low H') profiles can occur even at high richness when dominance is extreme, whereas the highest H' values require both high richness and high evenness. Whereas a high-dominance profile is intuitively pictured, it will be interesting to see how more even profiles looks like. To get a more intuitive view, we next rely on per-species treemaps that lay out, for each nuclear genome, the relative contribution of the different donor groups; for example, in *Guillardia theta* (Figure 34), a cryptophyte bearing a secondary red plastid with a nucleomorph.

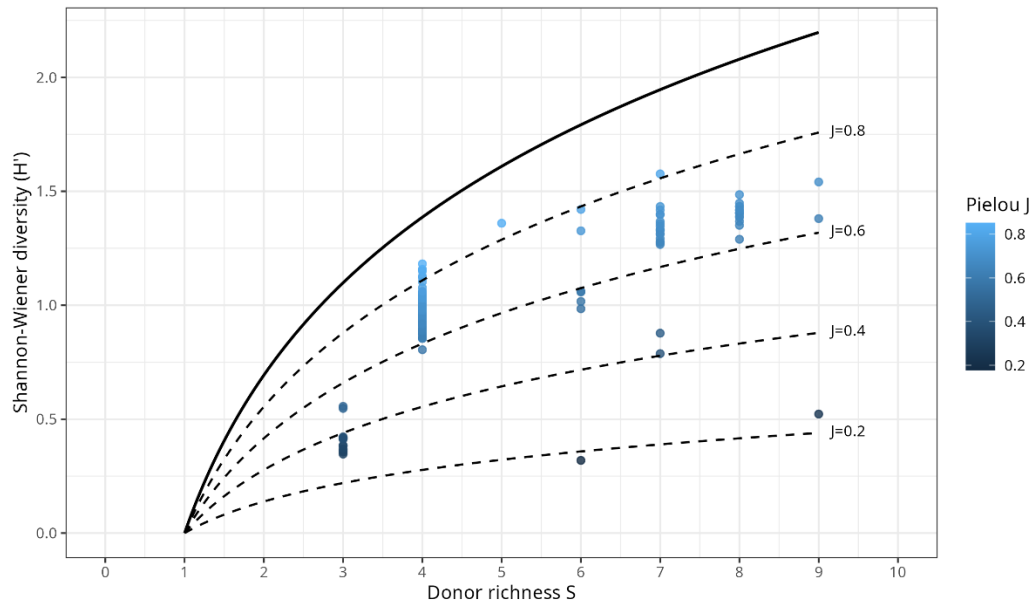


Figure 33: **Global relationship between donor richness and Shannon-Wiener α -diversity.** Shannon-Wiener α -diversity (H' ; *broad_div*) is plotted against donor richness (S) for species with ≥ 30 inferred events, with points coloured by Pielou evenness (J). The solid curve shows the theoretical maximum $H' = \ln(S)$ ($J=1$), and dashed curves indicate constant J isoclines ($H' = J \cdot \ln(S)$), highlighting that variation in H' reflects both donor richness and dominance/evenness.

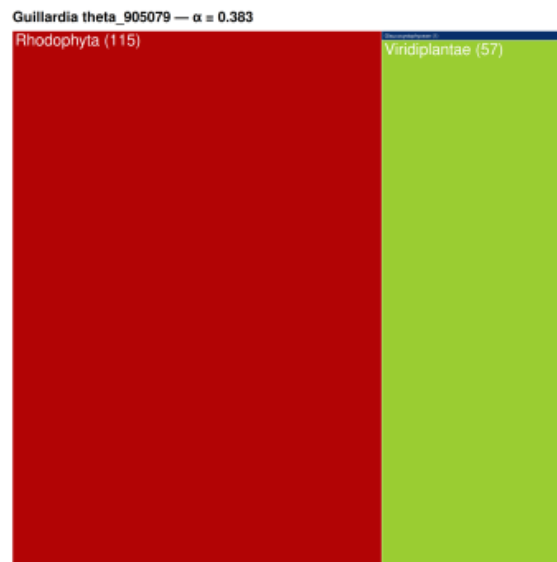


Figure 34: Inferred transfer events can be depicted as a per-species treemap (here *Guillardia theta*, Cryptophyceae).

To explore if there is one or more recurring patterns underlying low, intermediate and high diversity categories, we plotted all treemaps on the α -diversity axis; treemaps were binned according to the KDE cutoffs computed earlier (Figure 32). Figure 35 shows all the donor profiles identified; in most cases a single donor lineage dominates the profile, which is consistent with expectations if the majority of transfers originate from the lineage related to the plastid donor (see 3.1.2.3). Globally, this view confirms the earlier statement that increasing H' values display increasingly higher richness and evenness.

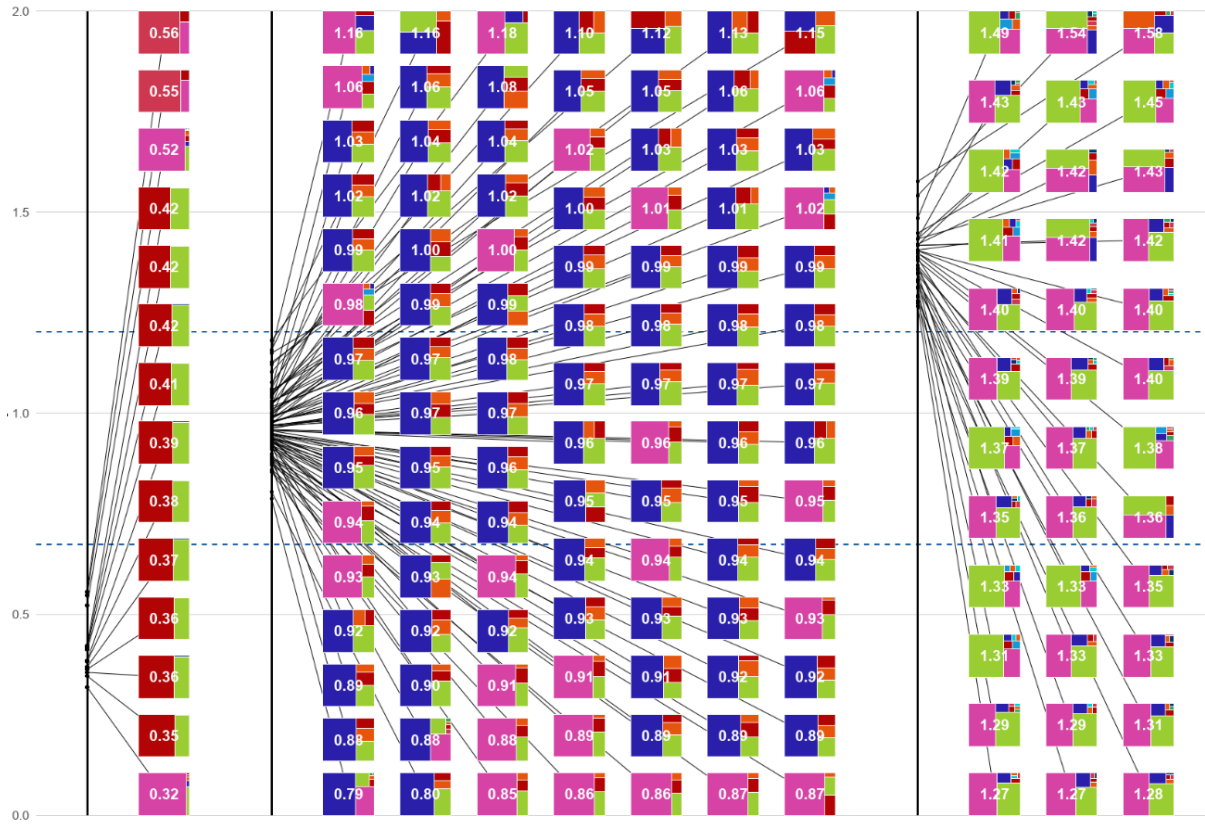


Figure 35: Treemap representations of α -diversity values for each species. The three panels show the three categories (low, intermediate, high) of the data binned according to the kernel density estimation (KDE).

While donor profiles do look similar within the same diversity range, some organisms contradict this observation. Indeed, donor profiles can be slightly different even for equivalent α -diversity values (Figure 36a); this is explained by the weight associated to each donor lineage (computed according to the lineage-specific features described in Methods section 5.5.5). Conversely, lineage-specific weights can lead to different α -diversity values even for similar profiles (Figure 36b).

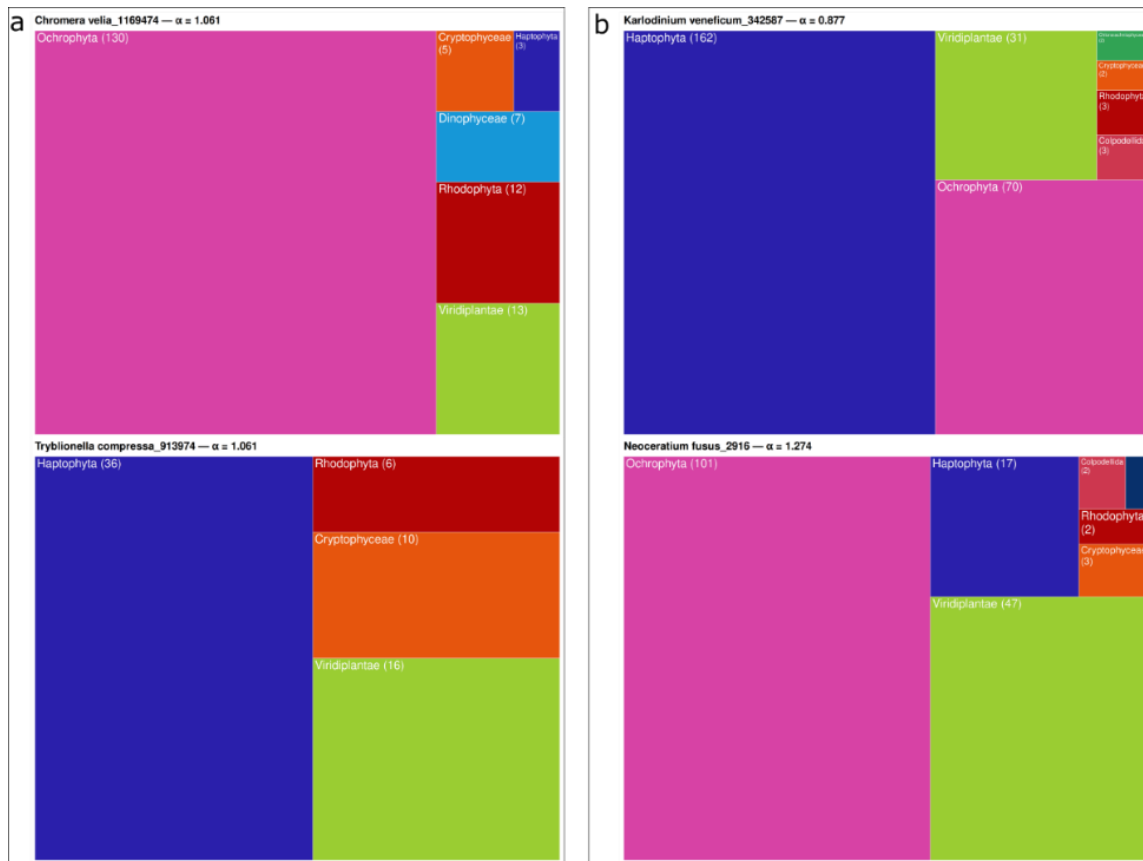


Figure 36: Donor profile VS α -diversity: (a) identical α -diversity values but different profiles and (b) similar profiles (i.e., recurring patterns) but different α -diversity values.

3.1.2.2. Per lineage distribution

Applying lineage membership on the overall species distribution (Figure 37) reveals that some species belonging to a given lineage tend to show highly similar α -diversity (e.g., Cryptophyceae) while some species suggest that subgroups might have followed a different evolutionary path, such as Dinophyceae, that occupy the three diversity bins (low, intermediate and high). Between those extreme cases, we see that others are also spread, but within the same diversity bin (intermediate), such as haptophytes and, more markedly, ochrophytes. That said, we cannot exclude that some species still suffer from completeness and/or contamination even though both issues have been carefully minimized in Chapters III and IV. Altogether, the observed distributions suggest different evolutionary scenarios across, and maybe within, the major complex algal groups.

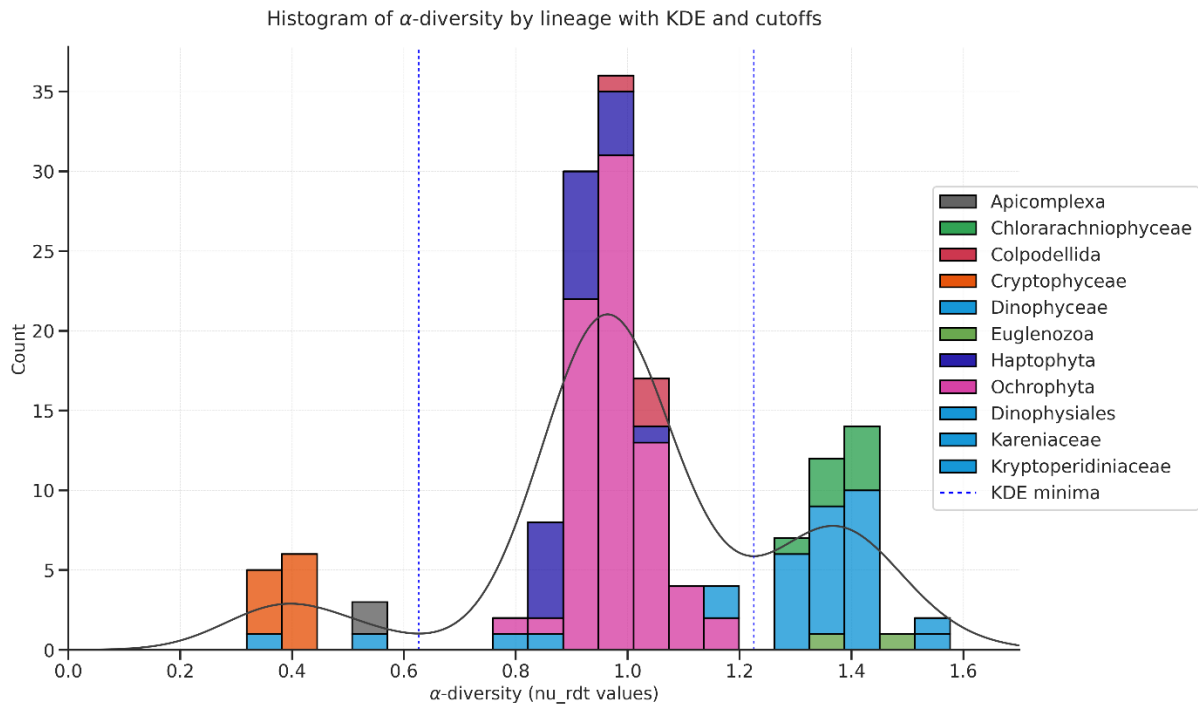


Figure 37: Distribution of individual species Shannon-Wiener α -diversity values according to the major complex algal group.

This becomes clearer when explicitly considering α -diversity distributions per lineage (Figure 38). Indeed, apart from Dinophyceae, most taxonomic groups occupy a narrow and almost exclusive range of Shannon-Wiener values (even within the same α -diversity category), suggesting that donor diversity primarily reflects shared ancestry and lineage-specific evolutionary history. The former cause is the target of this study and is the focal point of the heuristics implemented in PAGTI. In this figure, Dinophyceae refers specifically to peridinin dinoflagellates, which cluster almost exclusively in the high-diversity category. In contrast, dinoflagellate lineages with well-documented plastid replacement, in Kareniaceae (Nosenko et al., 2006), Kryptoperidiniaceae (Hehenberger et al., 2016a) or with known kleptoplastidic behaviour (Dinophysiales; Wisecaver and Hackett, 2010), were plotted separately, in order to study their expectedly distinct plastid histories. Unsurprisingly, they indeed occupy different α -diversity ranges (intermediate, low and high, respectively).

Applying the KDE-based classification at the lineage level reveals a structured distribution of α -diversity values across plastid-bearing groups. Cryptophyceae and Kryptoperidiniaceae fall almost exclusively within the low-diversity category, whereas Euglenophyceae, Dinophyceae, Dinophysiales, and Chlorarachniophyceae are concentrated in the high-diversity category. Between these two extremes, Haptophyta, Ochrophyta, Colpodellida, and Kareniaceae occupy the intermediate range. Their distributions are generally narrow and only moderately overlapping, indicating that α -diversity classes are not randomly distributed across lineages but instead tend to be lineage-specific. Overall,

these categories broadly correspond to documented differences in plastid history (Sibbald and Archibald, 2020), although some lineages depart from this general pattern (see Discussion, Section 1).

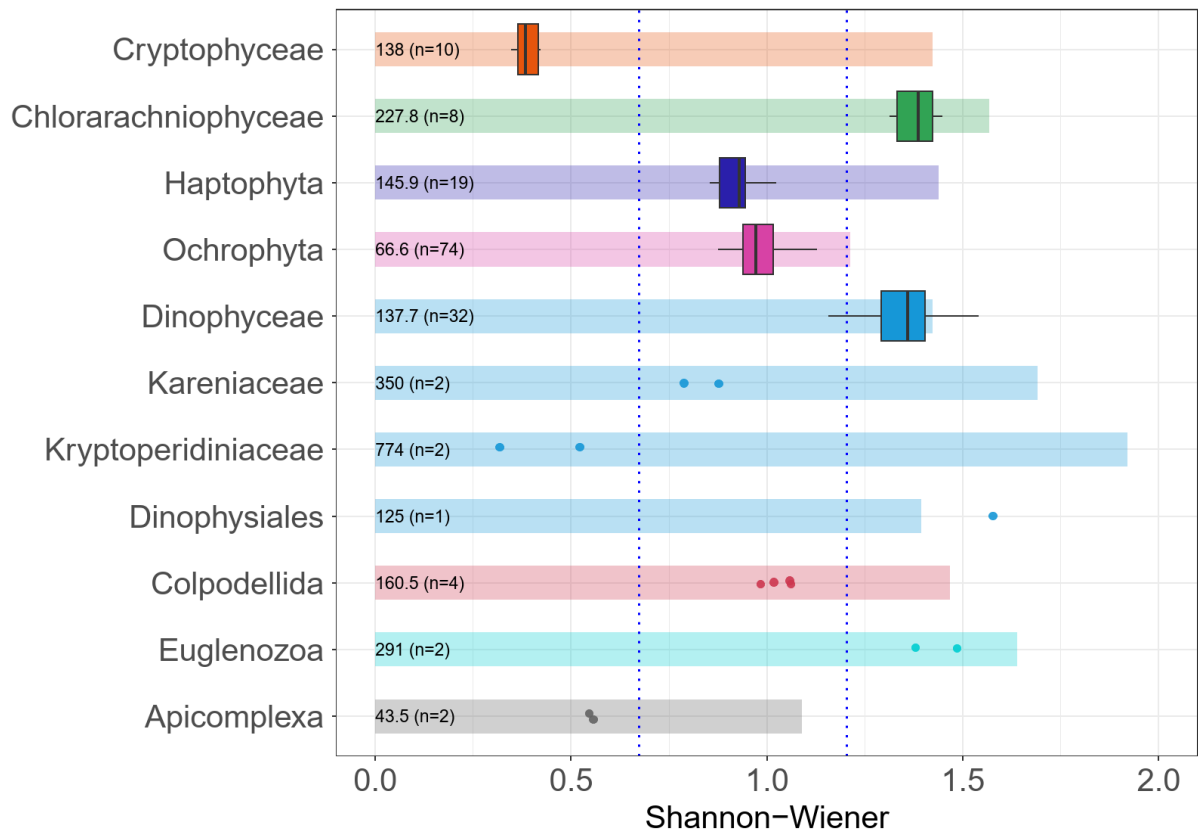


Figure 38: Species α -diversity distribution broken down by lineage. Blue dotted lines indicate KDE cutoffs. Pale bars in the background show mean transfer events per species, with n between parentheses indicating the number of species concerned by these events. Lineage populations ≤ 5 species were plotted as points whereas larger lineages are shown as boxplots.

3.1.2.3. Plastidial control: cpcpt# genes recover the expected donor signal

To verify that our donor inference procedure and the resulting α -diversity metric behave as expected on a single-origin dataset, we analysed plastid-encoded plastid-targeted genes (cpcpt#) as a positive control. Unlike nucleus-encoded plastid-targeted (nucpt#) genes, these genes are retained in the plastid genome and should therefore tightly reflect current plastid ancestry. Under this expectation, taxa bearing green plastids should be dominated by Viridiplantae, whereas taxa bearing red-derived complex plastids should be dominated by Rhodophyta. This prediction is consistent with the independent green plastid origins of euglenids and chlorarachniophytes (Jackson et al., 2018), and with the broadly accepted red-algal ancestry of plastids in cryptophytes, haptophytes and ochrophytes (Gould et al., 2015; Janouškovec et al., 2010; Ševčíková et al., 2015), even though the detailed history of plastid spread among most of these lineages remains debated (Ponce-Toledo et al., 2025; Strasser et al., 2021).

Because fewer transfer events were detected in this control set, we applied a lower minimum threshold of 20 inferred events per species (see Methods, section 5.5.4), therefore analysing 11 species (instead of 2 with the standard threshold of 30). In all retained species, the dominant inferred donor matched the expected plastid source. Rhodophyta was recovered as the major donor in all ochrophytes (*Thalassiosira pseudonana*, *Phaeodactylum tricornutum*, *Nannochloropsis gaditana* and *Ectocarpus siliculosus*), all haptophytes (*Tisochrysis lutea*, *Chrysochromulina* sp. and *Emiliana huxleyi*), and both cryptophytes (*Guillardia theta* and *Hemiselmis tepida*). Conversely, Viridiplantae dominated in the chlorarachniophyte *Bigelowiella natans* and in the euglenid *Euglena gracilis*. Across the full control set, the expected donor accounted for 253 of 295 inferred events (85.8%). By lineage, this represented 80.2% in ochrophytes, 80.3% in haptophytes, 95.0% in cryptophytes, 91.2% in *B. natans*, and 93.1% in *E. gracilis*.

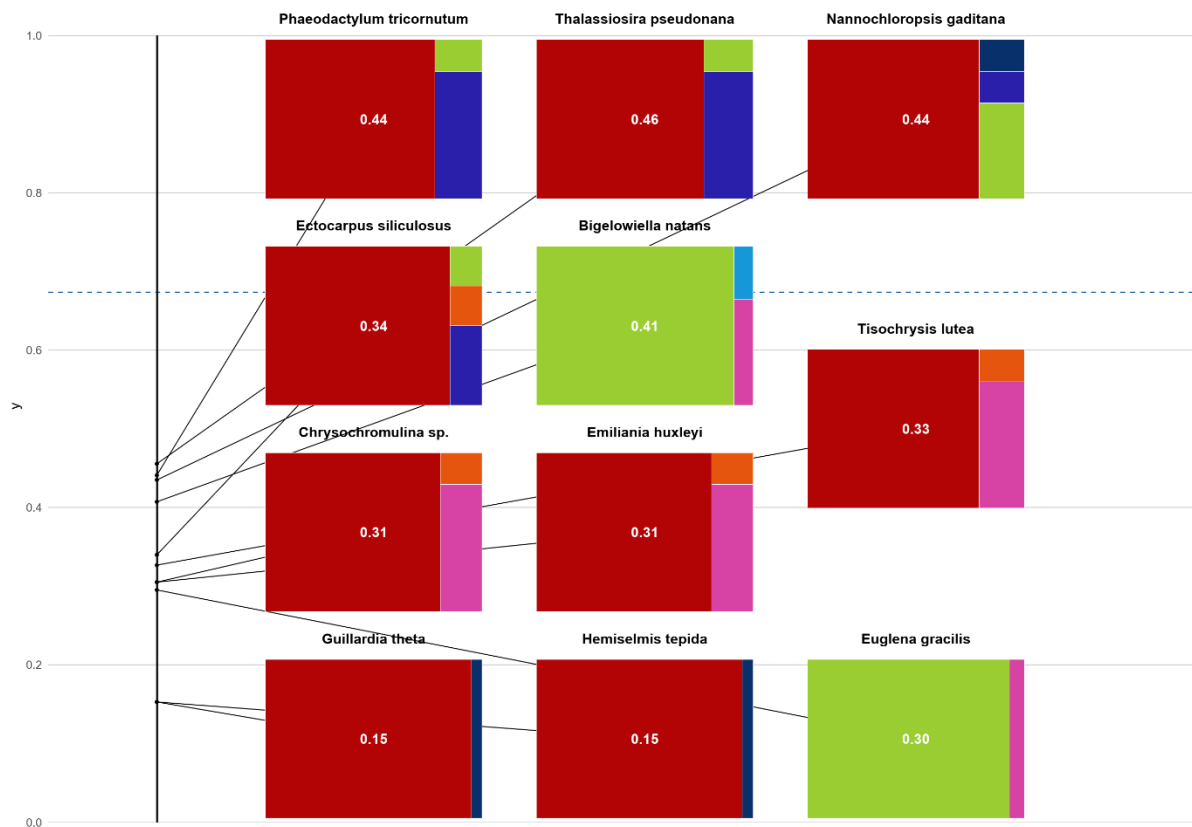


Figure 39: Treemap representations of cpcpt# gene α -diversity values for each species.

Overall, this control recovered the expected low-diversity signal centred on the current plastid source. This result supports the general validity of our whole inference framework and indicates that the broader and more mosaic donor profiles observed in nucpt# genes are unlikely to arise as a trivial artefact of the methods alone.

The global evolutionary interpretations of the events affecting those nucpt# genes will be covered in the biological discussion (Chapter 6, section 1), after further validation of our methods, in the methodological results (section 4.1.1) meant to assess the robustness of event inference to dataset variation and diverse parameterization of the different steps. Meanwhile, we will devote the next part to case studies about specific lineages for which we either hypothesized that kleptoplasty would have played a major role (i.e., euglenids and dinoflagellates) or observed intriguing patterns of reciprocal transfers (i.e., ochrophytes and haptophytes).

3.2. Case studies

3.2.1. Euglenids: a case study for a nucpt# core gene set associated with kleptoplast pre-adaptation?

3.2.1.1. Identification of a shared euglenid pre-adaptation nucpt# gene set

When this analysis was originally designed, genomic data from *Rapaza viridis* was not yet available. Recent studies have provided compelling evidence that the common ancestor of *Rapaza* and extant photosynthetic euglenids likely exhibited a kleptoplastic lifestyle, preceding the establishment of a permanent plastid (Karnkowska et al., 2023). These findings suggest that plastid acquisition in lineages such as *Euglena* and *Eutreptiella* may have involved a kleptoplastidic phase rather than direct endosymbiotic integration, in line with our working hypothesis. Building on these advances, the present analysis was extended to include the *Rapaza* genome (see Methods, section 5.8). Incorporating this new species allows nucpt# gene transfer events to be compared between *transient* kleptoplasts (*R. viridis*) and permanently integrated kleptoplasts (*Euglena* and *Eutreptiella*) in the same lineage and within a common phylogenomic comparative framework. This enables the identification of potential gene sets that are compatible with the different phases of plastid acquisition. The partitioning of nucpt# gene transfer events according to the presence across the three euglenid species considered in the receiver clan (*Rapaza*, *Euglena* and/or *Eutreptiella*) revealed three main event-affected gene categories: (i) species-specific events, (ii) events shared exclusively between *Euglena* and *Eutreptiella*, and (iii) events shared by *Rapaza* and at least one of other two species (Figure 40).

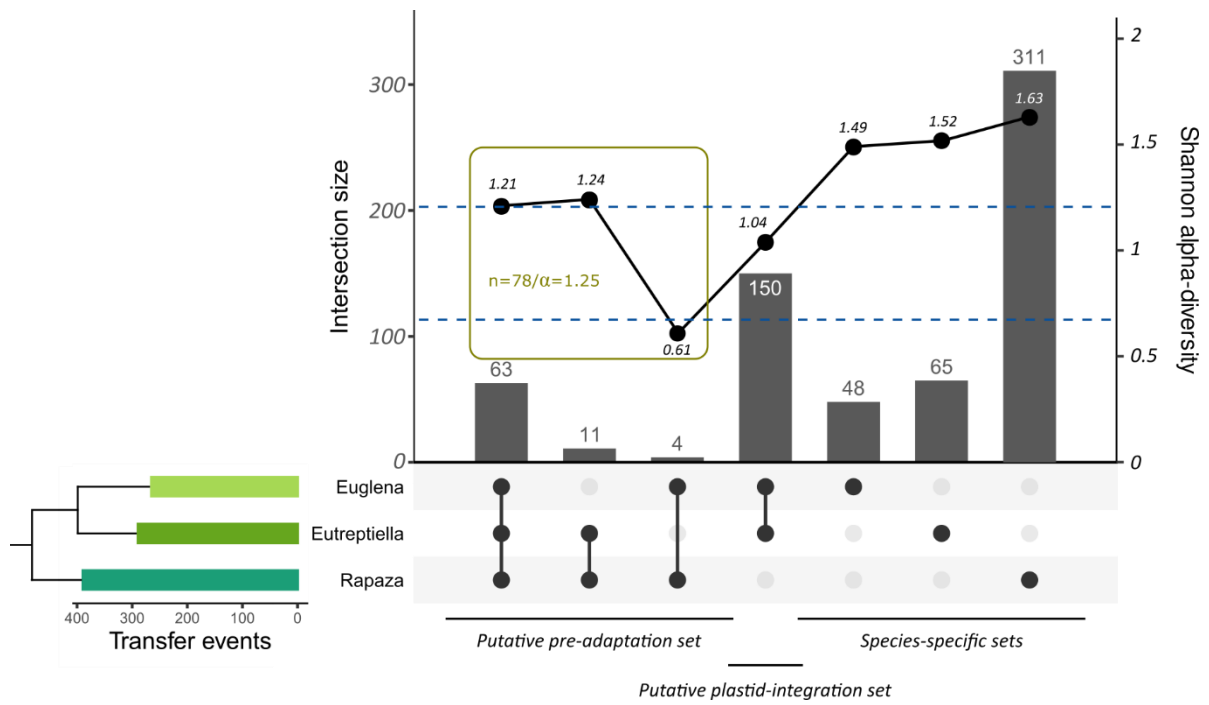


Figure 40: Distribution of plastid-targeted gene transfers across gene sets. The plot shows the number of inferred transfer events for each intersection between *Rapaza viridis*, *Euglena gracilis*, and *Eutreptiella* sp., defining the putative pre-adaptation, plastid-integration, and species-specific gene sets. Bar heights indicate the number of events per intersection, while the black line represents Shannon-Wiener α -diversity values computed for each set. KDE-derived thresholds (blue dashed lines) delimit low, intermediate, and high diversity regimes.

Orthoclans for which events were inferred and shared by *Rapaza* and *Euglena* and/or *Eutreptiella* (i.e., where the two or three species form a receiver clan) were defined as a putative pre-adaptation set ($n = 78$), as their presence across *Rapaza* and one of the other two species implies acquisition in stem branch prior to species divergence. While shared ancestry alone does not demonstrate a pre-adaptive role, this criterion is necessary for compatibility with such a scenario. Within this set, two sub-intersections were observed: 11 events shared between *Rapaza* and *Eutreptiella* but absent from *Euglena*, and 4 events shared between *Rapaza* and *Euglena* but absent from *Eutreptiella*. Given the phylogenetic position of *Rapaza*, these patterns are compatible with lineage-specific gene losses following divergence (or undetected events) rather than multiple independent acquisitions. Events shared exclusively between *Euglena* and *Eutreptiella* but absent from *Rapaza* were grouped into a putative plastid-integration set ($n = 150$), reflecting genetic features that potentially arose or were fixed after divergence from *Rapaza* and whose role was associated with plastid integration.

3.2.1.2. Looking for an α -diversity-driven signature

To evaluate whether elevated donor diversity preferentially originates from the putative pre-adaptation set, as predicted by our working hypothesis putting emphasis onto kleptoplasty, α -diversity was computed independently for each gene set (Figure 40). Under this expectation, the pre-adaptation set

should display higher donor diversity than both species-specific and plastid-integration sets. The putative pre-adaptation set indeed displayed a high α -diversity value ($\alpha = 1.25$); however, species-specific gene sets exhibited a significantly higher α -diversity (1.49, 1.52 and 1.63), while the plastid-integration set showed a markedly lower diversity ($\alpha = 1.04$). These results indicate that elevated donor diversity is not restricted to genes inferred to predate species divergence, but instead persists across different evolutionary phases, including early acquisition (pre-adaptation), subsequent plastid integration, and later lineage-specific remodelling. The lower α -diversity of the plastid-integration set ($\alpha = 1.04$) compared to the pre-adaptation set ($\alpha = 1.25$) reflects a combination of reduced donor richness and reduced evenness (Figure 41). Donor richness decreases slightly between sets (from 7 to 6), which lowers the theoretical maximum diversity ($\ln(S)$). In addition, Pielou's evenness (J) also declines (from 0.643 to 0.580), indicating a less even (i.e., more donor-dominated) profile in the plastid-integration set. Together, these results show that the drop in Shannon-Wiener α -diversity is not attributable to a single dominant donor alone, but instead to both a small contraction in the donor pool and a shift toward a less evenly distributed donor contribution. This reduction in evenness is consistent with the marked increase in the Ochrophyta contribution in the plastid-integration set (from 12.8 to 24%).

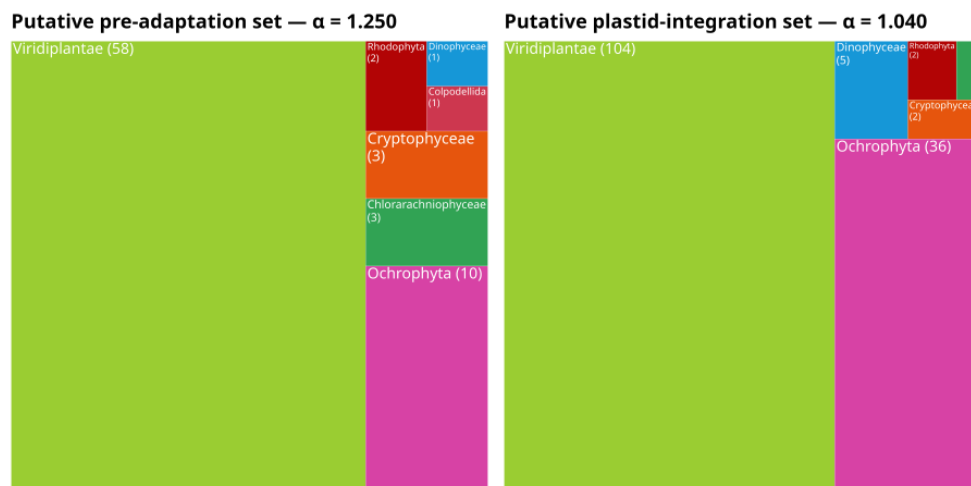


Figure 41: Comparison of donor contributions across the putative pre-adaptation set and the putative plastid-integration set.

3.2.1.3. Looking for a donor-driven differential signature

Before comparing the different contributors of each set, a quick sanity check is required to evaluate the contamination levels of the newly integrated species using our ribosomal marker-based contamination detection pipeline (see Section in Chapter II). Out of the 78 markers available, 68 were at least once assigned to Viridiplantae, including 62 specifically identified as *Tetraselmis*, thereby revealing a possible contamination. The significant and simultaneous detection of both universal and eukaryotic ribosomal sequences assigned to this lineage indicates that the signal does not result solely from retained

kleptoplasts of *Tetraselmis*, but rather reflects the presence of substantial nuclear prey-derived genetic material in the assembly, i.e., *Tetraselmis* (Yamaguchi et al., 2012; Karnkowska et al., 2023). Consequently, we will exercise caution when interpreting Viridiplantae-derived transfers.

Examining the species-specific transfers (Table 7) reveals a strong contribution from *Viridiplantae* to *Rapaza* (169). Among those transfers, 89 showed a topology where *Tetraselmis* sequences were the closest to *Rapaza viridis*; the remaining were affiliated to green algal groups outside Chlorodendrophyceae (*Tetraselmis*'s class) and deemed uncontaminated. To adopt a conservative strategy, we only discarded transfers affiliated to “Chlorodendrophyceae” (89), thus leaving 80 exploitable Viridiplantae-derived events and 142 events assigned to other algal groups in what we will call the *Rapaza*-clean set (Table 7). Unfortunately, potentially “true transfers” from *Tetraselmis* might have been discarded due to this choice.

<i>Donor_lineage</i>	<i>Rapaza</i>	<i>Euglena</i>	<i>Eutreptiella</i>
<i>Viridiplantae</i>	169 (80)	26	22
<i>Ochrophyta</i>	54	13	21
<i>Rhodophyta</i>	30	1	0
<i>Cryptophyceae</i>	23	1	6
<i>Dinophyceae</i>	18	1	12
<i>Haptophyta</i>	13	5	4
<i>Chlorarachniophyceae</i>	2	0	0
<i>Glaucocystophyceae</i>	1	0	0
<i>Colpodellida</i>	1	1	0
<i>Total</i>	311 (222)	48	65

Table 7: Species-specific donor counts. The numbers between parentheses indicate the donor count minus the suspected contaminated events.

Comparison of the species-specific sets shows that all euglenids receives substantial contributions from both Viridiplantae and ochrophytes. *Rapaza* also displays a markedly stronger contribution from all red-plastid lineages (Rhodophyta, Cryptophyceae, Dinophyceae and Haptophyta), dominated by rhodophytes (Table 7). This donor profile is exclusive to *Rapaza*, even compared to the pre-adaptation and plastid-integration sets, thus potentially reflecting *Rapaza viridis*' own plastid-integration path. Nonetheless, if our assumptions are correct; i.e., (i) that plastid pre-adaptation is a phase of opportunistic gene recruitment of many origins (from diverse prey), and (ii) that the plastid-integration phase would induce a narrowing of gene recruitment to the last prey (*Tetraselmis*), then the plastid-integration set should show a bias towards the now permanent plastid in *Eutreptiella* and *Euglena*, i.e., the plastid of *Pyramimonas* (prasinophytes) (Turmel et al., 2009; Jackson et al., 2018). Unfortunately, the detailed analysis of Viridiplantae-derived transfers revealed no significant bias towards *Pyramimonas* nor prasinophyte algae.

3.2.1.4. Looking for a function-driven differential signature

To further assess the evolutionary origin and timing of these acquisitions, we next examined orthology relationships among these events. Specifically, we contrasted (i) the putative pre-adaptation set, (ii) the putative plastid-integration set, and (iii) species-specific sets, as they reflect the relative order/timing of inferred acquisitions (early/shared vs more late/species-specific) (Figure 42).

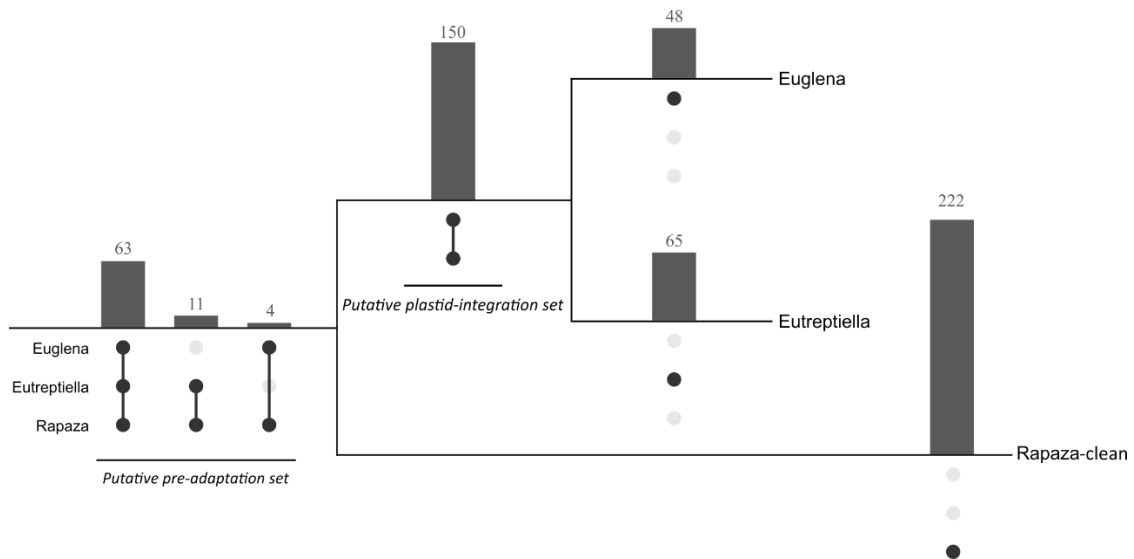


Figure 42: Identified gene sets on a schematic species tree.

At the orthoclan level, the three euglenid sets are largely disjoint. The pre-adaptation and plastid-integration sets share only 3 out of 218 unique orthoclans, the pre-adaptation and *Rapaza*-clean sets share 6 orthoclans out of 210, and the plastid-integration and *Rapaza*-clean sets share 14 orthoclans out of 269 (Figure 43). This indicates that each set is dominated by its own orthoclan repertoire, with only a very limited overlap between sets. Yet, orthologous gene disjunction alone is not sufficient to conclude to biological difference. Similar functions can be carried out by different, non-orthologous/paralogous genes, meaning that event sets may remain structurally distinct while still converging functionally. To address this possibility, we next compared the functional composition of the three sets.

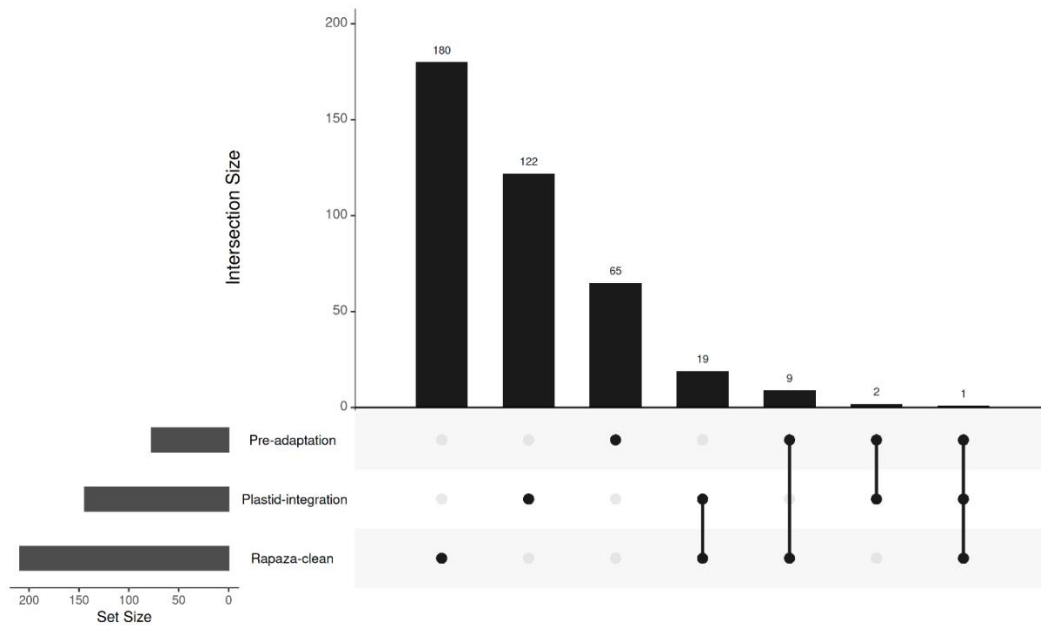


Figure 43: Upset representation of shared events by orthoclans. Shared orthoclans indicate events that occurred independently in the different gene transfer sets.

First, we compared the pre-adaptation set and the plastid-integration set (Figure 44). The differential pathway analysis shows a clear functional contrast between the two euglenid sets. Relative to the integration set, the pre-adaptation set is enriched in Energy metabolism (*Carbon fixation by Calvin cycle; Photosynthesis; Oxidative phosphorylation*), Carbohydrate metabolism (*Glycolysis / Gluconeogenesis; Citrate cycle (TCA cycle)*), Lipid metabolism (*Glycerolipid metabolism*), and Metabolism of terpenoids and polyketides (*Carotenoid biosynthesis*). By contrast, the integration set is enriched in Lipid metabolism (*Fatty acid biosynthesis; Fatty acid metabolism*), Glycan biosynthesis and metabolism (*N-Glycan biosynthesis; Various types of N-glycan biosynthesis*), Amino acid metabolism (*Arginine biosynthesis; 2-oxocarboxylic acid metabolism*), Nucleotide metabolism (*Nucleotide metabolism; Purine metabolism*), and Metabolism of cofactors and vitamins (*Porphyrin metabolism*). In summary, the pre-adaptation set is enriched in functions that sustain photosynthetic activity (light reactions, carbon fixation, photoprotection, and thylakoid membrane biogenesis), whereas the integration set is enriched in biosynthetic functions beyond photosynthesis (synthesis of fatty acids, amino acids, nucleotides, and cofactors).

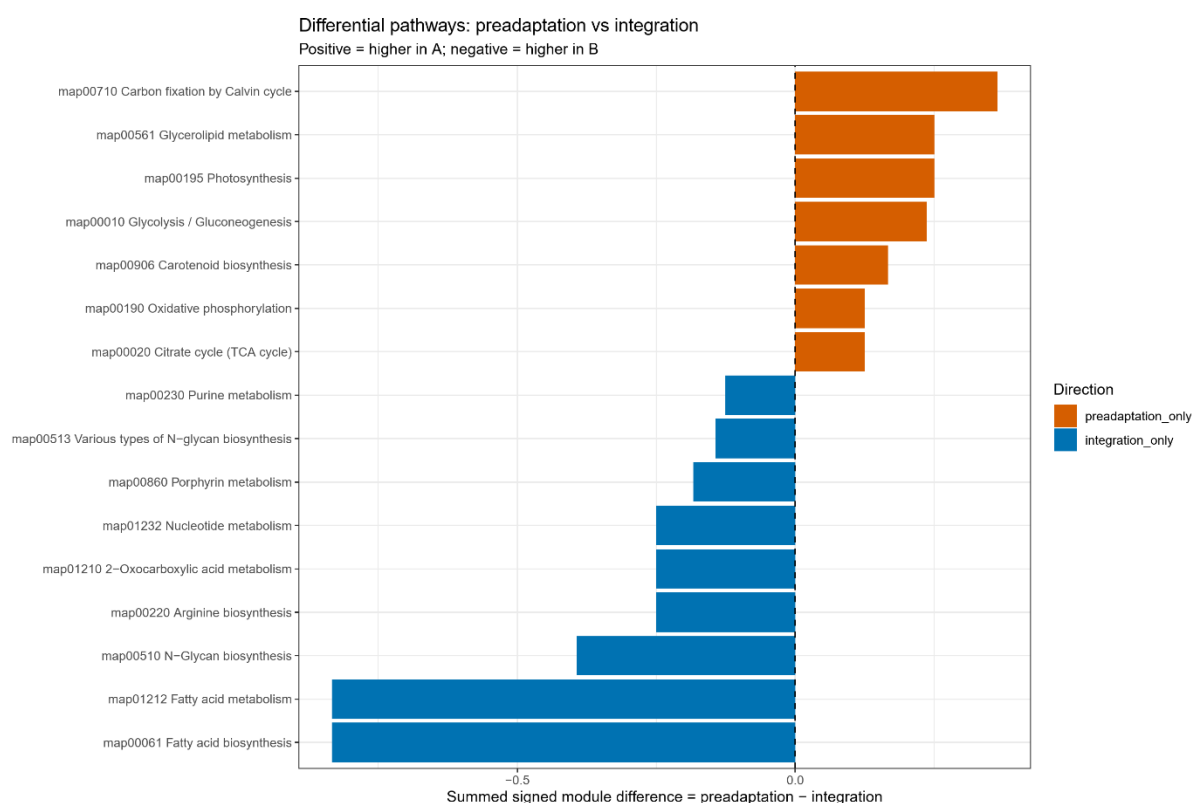


Figure 44: Differential pathways between the pre-adaptation (*preadaption_only*) and plastid-integration (*integration_only*) gene sets in euglenids.

Then, when comparing the *Rapaza*-clean set with the plastid-integration set (Figure 45), the functional partitioning is less clear. Relative to the integration set, the *Rapaza*-specific set is enriched in Metabolism of cofactors and vitamins (*Lipoic acid metabolism*; *One carbon pool by folate*), Carbohydrate metabolism (*Glycolysis / Gluconeogenesis*; *Citrate cycle (TCA cycle)*), Amino acid metabolism (*Valine, leucine and isoleucine biosynthesis*; *2-oxocarboxylic acid metabolism*; *Lysine biosynthesis*; *Glycine, serine and threonine metabolism*; *Cysteine and methionine metabolism*), and Energy metabolism (*Carbon fixation by Calvin cycle*). By contrast, the integration set is enriched in Metabolism of cofactors and vitamins (*Porphyrin metabolism*), Lipid metabolism (*Fatty acid biosynthesis*; *Fatty acid metabolism*), Nucleotide metabolism (*Nucleotide metabolism*; *Purine metabolism*), and Amino acid metabolism (*Arginine biosynthesis*). In summary, the *Rapaza*-specific set retains the photosynthesis-related signal observed in the pre-adaptation set (carbon fixation, glycolysis) but additionally includes amino acid and cofactor biosynthesis functions. The integration set, as in the previous comparison, remains enriched in fatty acid, nucleotide, and tetrapyrrole cofactor biosynthesis. The functional boundary between the two sets is less sharp than in the pre-adaptation versus integration comparison, consistent with an intermediate stage in which some biosynthetic functions have already been acquired but the full housekeeping repertoire has yet to be.

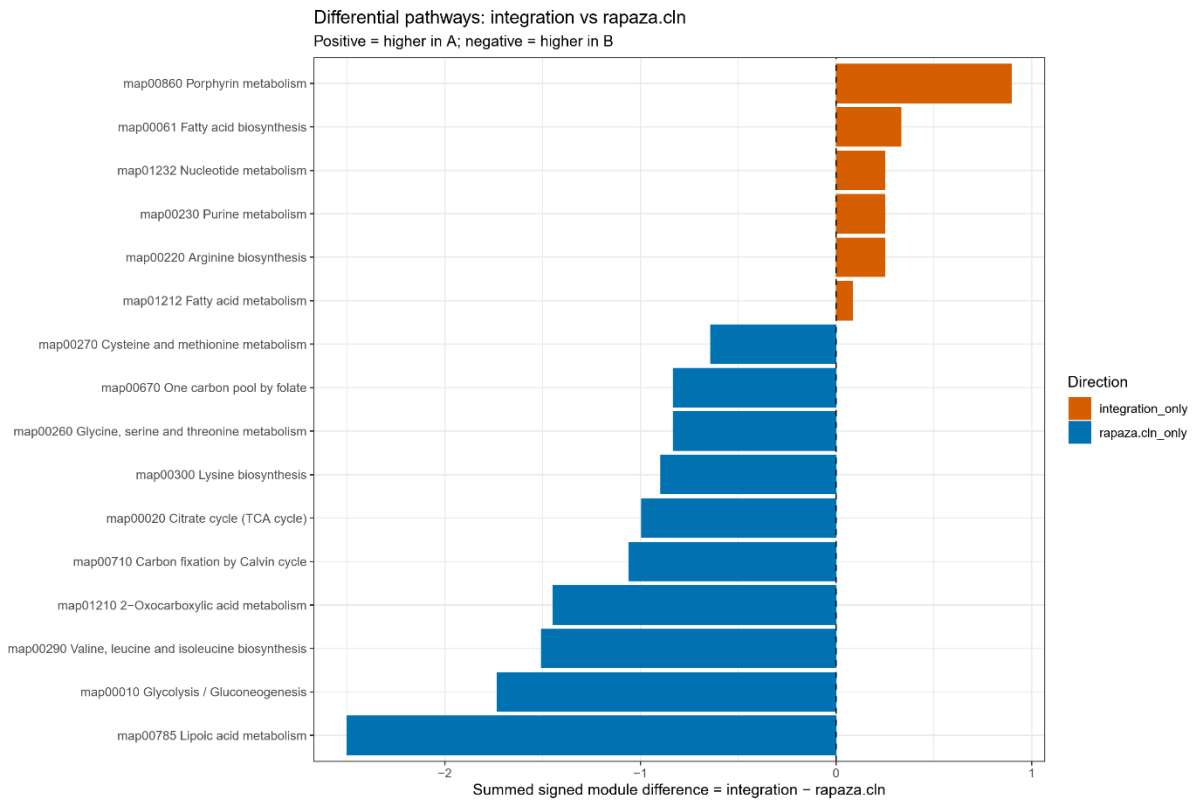


Figure 45: Differential pathways between the plastid-integration (integration_only) and Rapaza-clean (rapaza.cln_only) sets.

3.2.2. Peridinin dinoflagellates: a kleptoplastid-acquisition signature?

The three membrane-bound peridinin plastid is difficult to reconcile with the classical secondary endosymbiosis scenario predicting a four-membrane architecture. To account for this mechanistic difficulty, it was proposed that kleptoplastidic stages may have contributed to its origin (Bodyl, 2018; Moog and Maier, 2018). This possibility is particularly relevant in light of our hypothesis, in which elevated donor α -diversity is interpreted as a genetic signature compatible with kleptoplast-mediated plastid acquisition. Under this view, peridinin-containing dinoflagellates represent a natural candidate for comparison with euglenids. However, dinoflagellates are also a highly diversified group, with well-known recent plastid replacements and many species with demonstrated kleptoplastic activity (Waller and Kořený, 2017; Sibbald and Archibald, 2020). For that reason, we must consider four lineages when talking about Dinophyceae: peridinin dinoflagellates and three “(peridinin-)derived lineages”, i.e., Kryptoperidiniaceae (a.k.a. dinotoms), Kareniaceae (fucoxanthin dinoflagellates) and Dinophysiales (kleptoplastidic dinoflagellates). Distinguishing the derived lineages has also implications at the event-level: (i) peridinin-specific events are events where the receiver clan is composed of peridinin dinoflagellates (and eventually derived lineages that still have the ancestral peridinin gene), and (ii) lineage-specific events are events where only one of the three derived lineages has been identified as the exclusive receiver.

3.2.2.1. Identification of an ancestral core peridinin plastid event set

When trying to identify an ancestral core peridinin plastid set of genes, we were faced with issues in classifying 28 gene transfer events that were exclusively shared by non-peridinin lineages. Indeed, those transfer events (non-peridinin shared sets, Figure 46) had been lumped with those affecting peridinin dinoflagellates because PAGTI is configured to mechanically recognize non-peridinin lineages as “regular” (i.e., peridinin-containing) dinoflagellates when found together in a given receiver clan (LCA computation of the receiver clan; see above). These cases mainly include independent acquisitions that originate from the same donor lineage but different donor species (i.e., distinct events). Yet, sometimes, *Dinophysis* is involved in a nested gene transfer from Kareniaceae, as described in the work of Hongo et al., (2019). To resolve these issues, all candidate events were manually inspected at the phylogenetic tree level. This curation step led us to make those events as multiple and count for each receiver lineage independently, instead of single events affecting all lineages at once (i.e., shared events). This resulted in recording additional events in Dinophysiales (16 events), Kareniaceae (25), and Kryptoperidiniaceae (16). Moreover, among the 16 events reassigned to Dinophysiales, 6 events were interpreted as direct transfers from Kareniaceae (i.e., nested events).

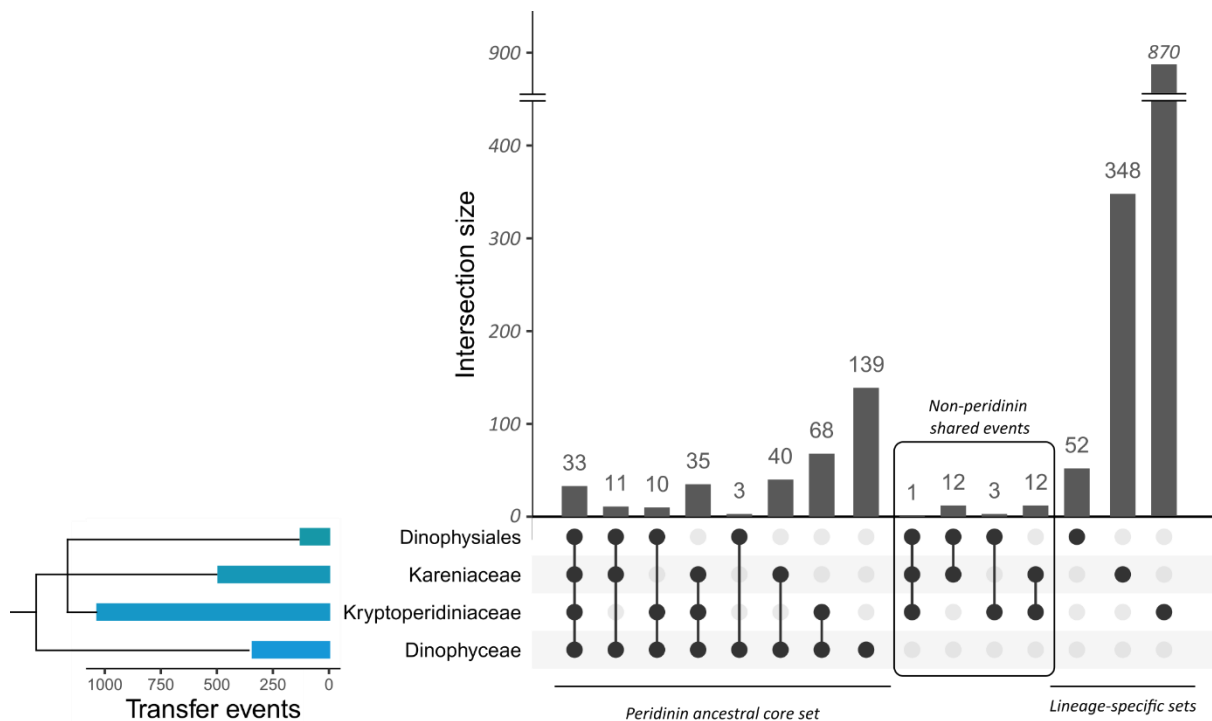


Figure 46: Upset representation of inferred transfer event counts across the compared sets. The horizontal bars on the left show the total number of events per set, while the vertical bars show the number of events recovered in each intersection.

In the main α -diversity figure (Figure 38), donor profiles for Kareniaceae, Kryptoperidiniaceae, and Dinophysiales were initially calculated from event sets that still included part of the ancestral peridinin-associated signal. After manual reclassification, α -diversity was recalculated using only the lineage-

specific signal of each derived dinoflagellate group (Table 8). This revision provides a biologically more meaningful estimate of donor diversity because it isolates the signal associated with the present plastid history of each lineage from the inherited peridinin background. In this corrected framework, Kareniaceae falls into the low α -diversity category (instead of intermediate initially), consistent with a much more polarized donor profile than initially suggested.

<i>Event set</i>	<i>Event count</i>	<i>Shannon-Wiener</i>
<i>Dinophysiales</i>	68	1.35
<i>Kareniaceae</i>	373	0.55
<i>Kryptoperidiniaceae</i>	886	0.29
<i>peridinin-ancestral-core</i>	339	1.50

Table 8: Updated α -diversity values after manual reclassification of events for derived dinoflagellate lineages. Values were recalculated using only lineage-specific donor signals, excluding the ancestral peridinin-associated background.

Aside from the reclassification, the ancestral peridinin core remains unchanged and comprises 339 transfer events. Of these, 139 are restricted to peridinin-bearing lineages, whereas 200 are still shared with lineages that later acquired a different plastid. This partition indicates that plastid replacement, whether permanent or transient (kleptoplasty), is associated with substantial remodelling of the ancestral peridinin-associated repertoire, but not with its complete disappearance. The persistence of this shared fraction is consistent with the idea that some ancestral peridinin-associated genes were retained across plastid replacement and, in some cases, may have been retargeted to the new plastid, as proposed for the plastid-targeted GAPDH isoform gapC1-pd in Kareniaceae (Nosenko et al., 2006; Patron et al., 2006).

Having clarified the event-level signal, we proceeded with the functional comparison of the identified gene transfer sets by examining their orthology relationships. This time, instead of using shared and specific events, we examine shared and specific orthoclans (for which a transfer was inferred), regardless of whether the event is shared or not (i.e., shared or independent origins) (Figure 47). The goal is to discover and characterize functional traits of each set.

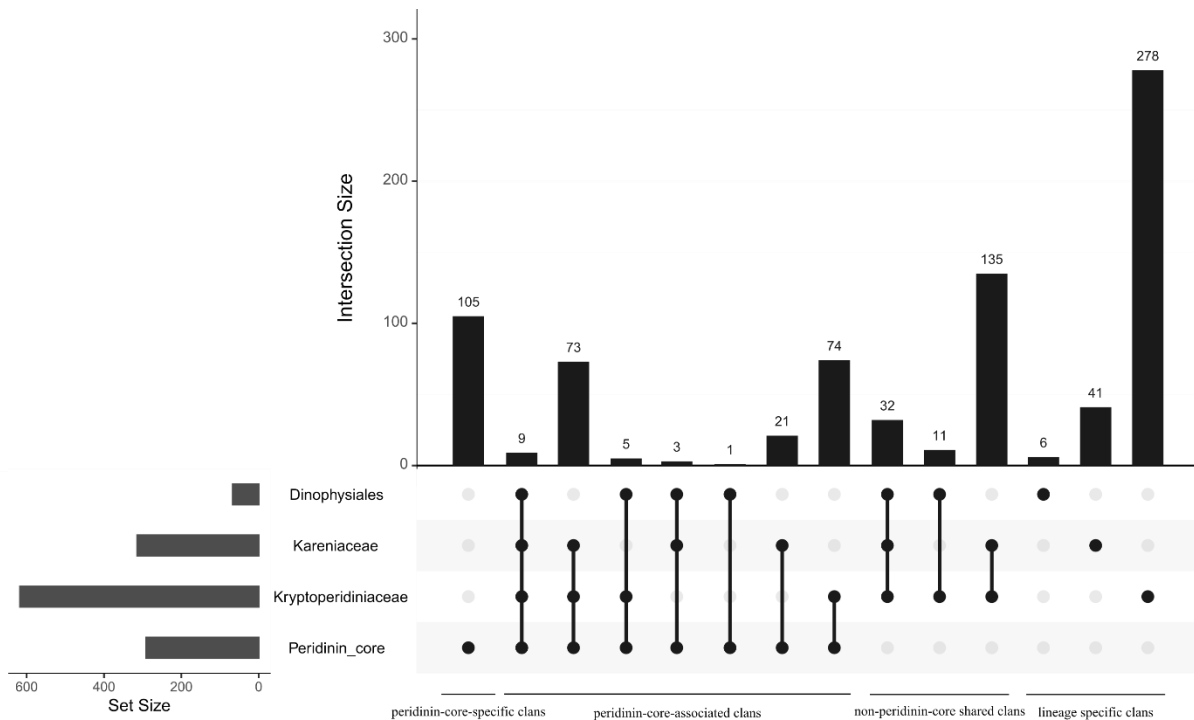


Figure 47: Upset representation describing the complex overlap of orthoclans among dinoflagellate lineages.

This new point of view suggests that dinoflagellate orthoclans can be partitioned into four categories. Peridinin-core-specific orthoclans (105) have transfer events that were detected only in peridinin dinoflagellates and are interpreted here as genes either lost or replaced by an homolog counterpart in the derived lineages. Peridinin-core-associated orthoclans (186) correspond to orthoclans where events are found both in peridinin dinoflagellates and in (some) derived lineages, consistent with gene retention or replacement within the same orthoclan. Non-peridinin-core shared orthoclans (178) correspond to orthoclans shared exclusively among non-peridinin lineages, consistent with shared remodelling through multiple independent acquisitions and/or nested recruitment. Finally, lineage-specific orthoclans correspond to orthoclans restricted to a single derived lineage, also consistent with remodelling, likely involving acquisitions and possibly compensatory replacement of unreplaced ancestral functions.

3.2.2.2. Functional characterization

The partitioning of orthoclans into two distinct peridinin-core sets allows us to address the question of propensity to gene replacement: there seems to be a core set deeply integrated in peridinin dinoflagellates (i.e., never replaced or lost in derived lineages) compared to the core that is replaced in other lineages; is there a functional explanation? To answer this question, we compared the KEGG pathway enrichment of the two sets (Figure 48).

The peridinin-core-specific set is enriched in Carbohydrate metabolism (*Starch and sucrose metabolism*; *Pentose phosphate pathway*; *Glycolysis / Gluconeogenesis*), Metabolism of cofactors and vitamins (*One carbon pool by folate*; *Riboflavin metabolism*), Metabolism of other amino acids (*Glutathione metabolism*), and several pathways assigned to Amino acid metabolism (*Arginine and proline metabolism*; *Glycine, serine and threonine metabolism*; *Histidine metabolism*), with as stronger bias towards central carbon metabolism(Figure 48).

By contrast, the peridinin-core-associated and replacement-compatible set is enriched in Metabolism of cofactors and vitamins (*Thiamine metabolism*; *Porphyrim metabolism*; *Biotin metabolism*; *Ubiquinone and other terpenoid-quinone biosynthesis*), Energy metabolism (*Sulfur metabolism*; *Sulfur cycle*), Metabolism of terpenoids and polyketides (*Terpenoid backbone biosynthesis*), and several pathways assigned to Amino acid metabolism (*Cysteine and methionine metabolism*; *Lysine biosynthesis*; *Phenylalanine, tyrosine and tryptophan biosynthesis*). with a stronger bias towards thiamine- and sulfur-related metabolism (Figure 48).

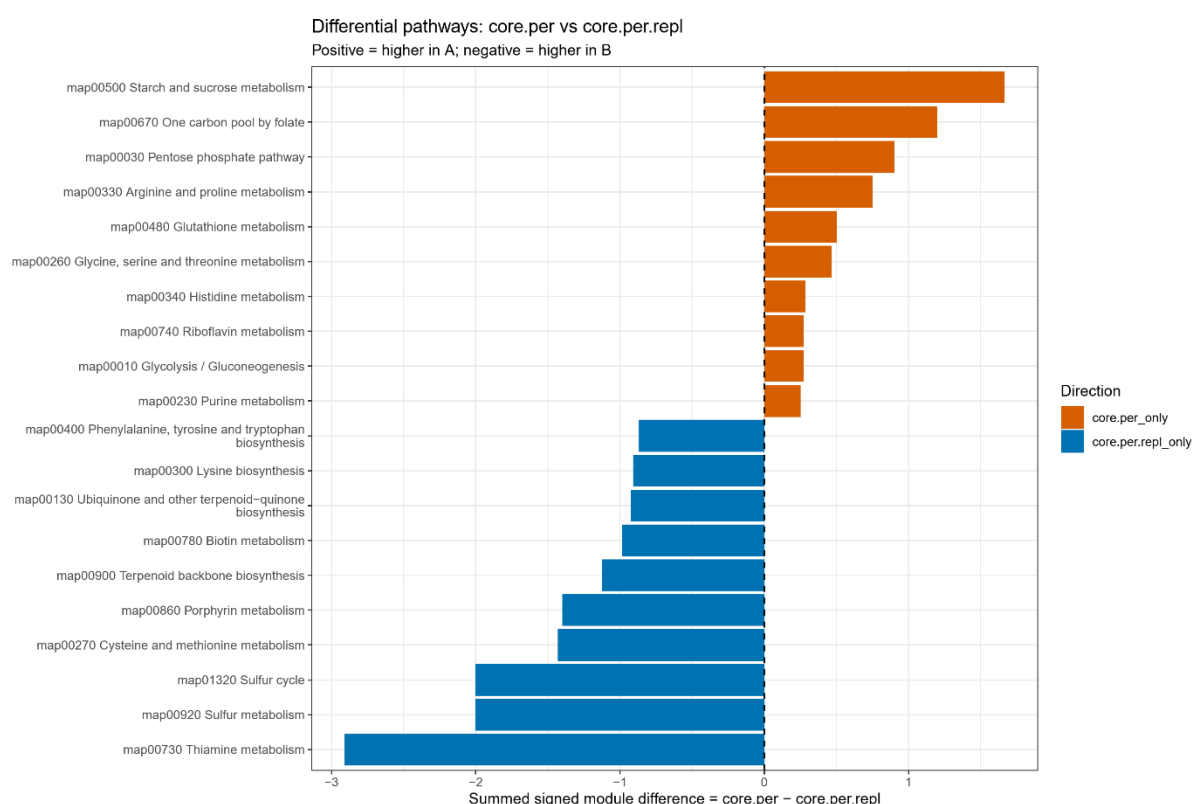


Figure 48: Differential pathways between peridinin-core-specific (core.per_only) vs peridinin-core-associated (core.per.repl_only).

This contrast may reflect two evolutionary fates of plastid-support genes. Peridinin-core-specific genes may represent functions tightly embedded in the ancestral peridinin plastid context, that may have been

lost, replaced, or rendered unnecessary during plastid replacement. In contrast, replacement-associated genes may represent a more portable layer of plastid-support functions that remained useful across plastid replacement events. Their stronger association with thiamine- and sulfur-related metabolism suggests that those functions may be among the ones that need to be retained or reconfigured to ensure compatibility with the newly acquired plastid.

3.2.2.3. Donor lineages

At the donor-lineage level, the ancestral peridinin core set of events is dominated by Ochrophyta, with sizeable contributions from Viridiplantae and, to a lesser extent, Haptophyta (Figure 49). Because Ochrophyta also emerge as a major donor lineage across our global event collection (i.e., beyond dinoflagellates), the composition and interpretation of the main signal (ochrophyte-derived) are examined in detail in section 3.2.4



Figure 49: Treemaps representing the union of donor contributions for peridinin dinoflagellates across all species.

3.2.3. Dinophysiales: non-permanent kleptoplasts

Having characterised the ancestral core set associated with the peridinin plastid, an integrated and permanent organelle, we now turn to the Dinophysiales, a dinoflagellate order that relies on non-permanent kleptoplasts. Most *Dinophysis* species harbour kleptoplasts of cryptophyte origin, acquired indirectly through predation on the ciliate *Mesodinium rubrum* (Wisecaver and Hackett, 2010). Unlike the peridinin plastid, these kleptoplasts must be continuously replenished from the environment. The Dinophysiales thus represent a contrasting case within the dinoflagellates, and comparing the gene-transfer signatures of these two groups may therefore help characterise signals associated with ongoing kleptoplastic interactions.

3.2.3.1. Enrichment with new *Dinophysis* species

At the time the dataset of this analysis was initially assembled, only a single *Dinophysis* species (*D. acuminata*) was represented. To better capture the diversity of kleptoplastic dinoflagellates within this genus, the present work was extended to include two additional species: *Dinophysis ovum* and *Dinophysis fortii*. For *D. ovum*, a transcriptome assembly was incorporated (Gaonkar and Campbell, 2023) whereas for *D. fortii* only a curated set of plastid-targeted nuclear genes (nucpt#) reported in the literature was available (58 proteins from Hongo et al., 2019).

Because of this limited sampling, *D. fortii* contributed very little information to the phylogenetic inference. Only four transfer events could be inferred in the Dinophysiales context, all of which were shared with *D. acuminata* and *D. ovum*. As a consequence, subsequent analyses were restricted to *D. acuminata* and *D. ovum*. The addition of *D. ovum* allowed the recovery of slightly more events (55, Table 9) compared to the initial analysis based solely on *D. acuminata* (52 events). Of these, 50 events were shared between the two species, while five remained unique to *D. acuminata*. The strong overlap between events inferred from independent *Dinophysis* species provides additional support for the reliability of the transfers identified in the original dataset.

Species	Events
<i>D. acuminata, fortii, ovum</i>	4
<i>D. acuminata, ovum</i>	46
<i>D. acuminata</i>	5
<i>D. fortii</i>	0
<i>D. ovum</i>	865
Total	920

Table 9: Shared events among *Dinophysis* species.

Unexpectedly, however, *D. ovum* alone yielded a very large number of inferred transfer events (865, Table 9), raising concerns about a potential contamination of its transcriptome assembly. That is why we performed an extensive contamination analysis in the following section using Forty-Two.

3.2.3.2. A perfect example of contaminated data

The contamination analysis revealed extensive contamination in the *D. ovum* assembly (Table 10), most prominently by *Paraphysomonas imperforata* (62/78 markers) and *Mesodinium rubrum* (22/78 markers). A large fraction of the markers assigned to Thalassiosirales (66/78) were further resolved as various *Skeletonema* species, including *Skeletonema menzelii*. Similarly, markers assigned to the genus *Geminigera* were largely attributable to *Geminigera cryophila*.

Contaminants	Ribosomal markers (78)
Thalassiosirales	66
<i>Paraphysomonas imperforata</i>	62
Geminigera	63
<i>Mesodinium rubrum</i>	22
Discosea	21

Table 10: Main contaminants of the *Dinophysis ovum* transcriptome assembly detected using Forty-Two and 78 eukaryotic ribosomal markers.

To further confirm the extent of contamination, we performed a comparative ribosomal-marker screening across all transcriptome assemblies generated in the same study (Gaonkar and Campbell, 2023), including *Dinophysis acuminata*, *Mesodinium rubrum* (two strains) and *Teleaulax amphioxeia* (Table 11). This broader comparison confirmed widespread contamination signatures, particularly involving *Paraphysomonas*, *Skeletonema*, *Geminigera*, *Discosea*, and *Vaucheriales*. It is important to note that *Teleaulax* sequences are absent from our ribosomal marker reference dataset. Consequently, reads originating from this cryptophyte lineage are likely assigned to the closest available taxon, *Geminigera*, explaining the strong *Geminigera* signal observed in our contamination screen.

Contaminants	<i>Dinophysis ovum</i>	<i>Dinophysis acuminata</i> DAVA01	<i>Mesodinium rubrum</i> DK2009	<i>Mesodinium rubrum</i> JAMR	<i>Teleaulax amphioxeia</i>
<i>Thalassiosirales</i>	66	15	0	0	0
<i>Paraphysomonas imperforata</i>	62	55	65	60	65
<i>Geminigera</i>	63	0	60	60	-
<i>Mesodinium rubrum</i>	22	0	-	-	0
<i>Discosea</i>	21	69	0	68	0
<i>Vaucheriales</i>	0	11	0	13	10

Table 11: Main contamination count for the five transcriptome assemblies from Gaonkar and Campbell, 2023. Values correspond to the number of ribosomal markers recovered out of 78.

Interestingly, the lineages detected as contaminants closely correspond to the major donor lineages identified with our PAGTI pipeline for *D. ovum*. Among the 865 inferred events, 618 were assigned to Ochrophyta, of which 506 were specifically associated with *Thalassiosirales* (including 426 cases closest to *Skeletonema*). Similarly, most Cryptophyceae-derived events (81/93) were associated with *Geminigera*, including 45 with *Teleaulax amphioxeia*, strongly suggesting that these signals originate from cross-contamination rather than genuine transfers. Consequently, the *D. ovum*-specific signal was excluded from downstream analyses, and the *Dinophysis* component of the dataset was restricted to the robust events recovered with *D. acuminata*.

<i>Lineage</i>	<i>Event count</i>	<i>Contaminant ribosomal sequences</i>
<i>Chlorarachniophyceae</i>	3	-
<i>Colpodellida</i>	4	-
<i>Cryptophyceae</i>	93	74
<i>Euglenozoa</i>	5	-
<i>Glaucocystophyceae</i>	1	3
<i>Haptophyta</i>	19	-
<i>Ochrophyta</i>	618	162
<i>Rhodophyta</i>	35	3
<i>Viridiplantae</i>	87	7

Table 12: High-level donor count for *Dinophysis ovum* and the corresponding contaminant ribosomal sequences found in the contamination detection analysis.

3.2.3.3. Orthoclan analysis

In this enriched version of the dataset, we were faced with the same issue with events exclusively shared by non-peridinin dinoflagellates. The same curation step as in the original dataset (identical reclassification procedure as in Section 3.2.2.1) was performed and led to 19 additional events (compared to 16 in the original dataset) involving Dinophysiales as receivers, amounting to a total of 74 gene transfer events in 66 unique orthoclans (7 orthoclans had 2 events, 1 orthoclan had 3 events). Among these, 18 orthoclans belong to the peridinin-core-associated set identified in Section 3.2.2.1.

The differential pathway analysis (Figure 50) compares the Dinophysiales-specific set with the peridinin-core-associated set. Relative to the peridinin-core-associated set, the Dinophysiales set is enriched in Energy metabolism (*Nitrogen metabolism*) and Glycan biosynthesis and metabolism (*Amino sugar and nucleotide sugar metabolism*; *Biosynthesis of nucleotide sugars*). By contrast, the peridinin-core-associated set is enriched in Metabolism of cofactors and vitamins (*Thiamine metabolism*; *Porphyrin metabolism*; *Biotin metabolism*), Energy metabolism (*Sulfur metabolism*; *Sulfur cycle*), Metabolism of terpenoids and polyketides (*Terpenoid backbone biosynthesis*), Lipid metabolism (*Fatty acid metabolism*), and Amino acid metabolism (*Cysteine and methionine metabolism*; *Lysine biosynthesis*; *2-Oxocarboxylic acid metabolism*).

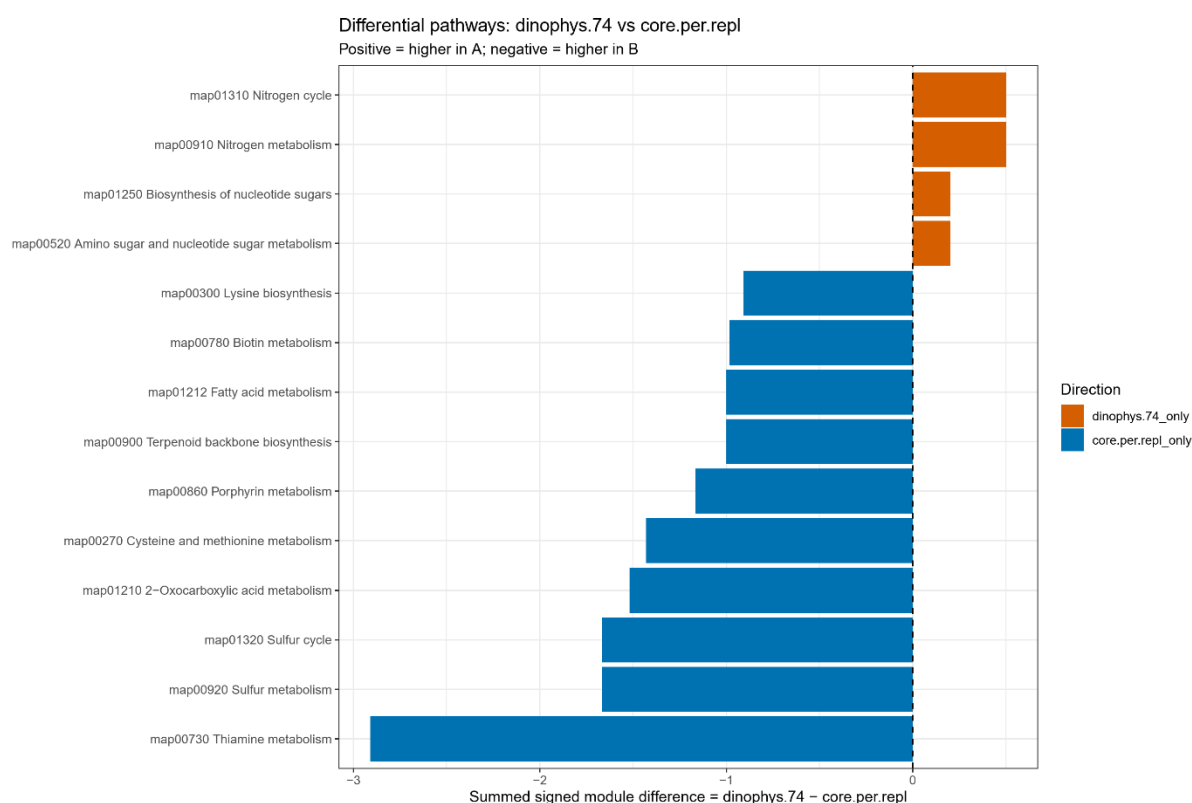


Figure 50: Differential pathways between the core-peridinin-associated (*core.per.repl_only*) and the *Dinophysiales*-specific (*dinophys.74_only*) sets.

This asymmetry is striking: the peridinin-core-associated set displays a broad and diversified functional profile, whereas the *Dinophysiales* signal is restricted to a narrow functional footprint. However, this functional signal should be interpreted with caution. The module-level analysis (anvi'o, pathwise strategy, threshold ≥ 0.5) recovered only four modules above the completeness threshold: assimilatory nitrate reduction (M00531), assimilatory sulfate reduction (M00176), and their corresponding signature modules (M00615, M00616). Moreover, all four modules are actually driven by a single event in a single orthoclan (OG0003300-2-1-65), encoding a ferredoxin-nitrite reductase (K00366) and/or a ferredoxin-dependent sulfite reductase (K00392). Both reactions are ferredoxin-dependent and are typically plastid-localised. Their recovery among *Dinophysiales*-associated transfers is functionally consistent with kleptoplast maintenance, but the reliance on a single orthoclan means that the differential pathway signal is largely anecdotal at this stage of this analysis.

3.2.4. Ochrophyta: a major donor lineage

As shown in the peridinin case study, ochrophytes already emerge as the dominant donor lineage within the ancestral peridinin core gene set, ahead of Viridiplantae and Haptophyta. However, this pattern is not restricted to peridinin dinoflagellates. At the scale of the full dataset, Ochrophyta constitute the major donor lineage overall with 1,662 out of 3,353 inferred transfer events (~50%, Table 13),

indicating that the peridinin signal is part of a broader and recurrent donor trend rather than an isolated case. However, this dominance is strongly influenced by the exceptionally large flux toward Kryptoperidiniaceae (837 events, ~25%), which harbour a diatom-derived plastid (Imanian et al., 2010). Even after excluding Kryptoperidiniaceae, Ochrophyta remains the main donor lineage with 825 events, followed by Viridiplantae with 775 events (~23%) and Haptophyta (~15%).

<i>Lineage</i>	<i>Event count as donors</i>
<i>Chlorarachniophyceae</i>	11
<i>Colpodellida</i>	46
<i>Cryptophyceae</i>	106
<i>Dinophyceae</i>	38
<i>Euglenozoa</i>	6
<i>Glaucocystophyceae</i>	4
<i>Haptophyta</i>	474
<i>Ochrophyta</i>	1662
<i>Rhodophyta</i>	231
<i>Viridiplantae</i>	775
<i>Total</i>	3353

Table 13: Number of inferred events per lineage (as donors); numbers are identical to Figure 30, panel B.

Interestingly, despite their large number of inferred transfers, ochrophytes affect a relatively limited number of receiver species (but a similar number of lineages), compared to other major donors such as Viridiplantae, Rhodophyta, Cryptophyceae and Haptophyta (Figure 51). This suggests that a fraction of the signal might be concentrated in a few sub-groups of specific receiver lineages. Reciprocally, the dominance of Ochrophyta as donors might not reflect a uniform contribution across receiver lineages.

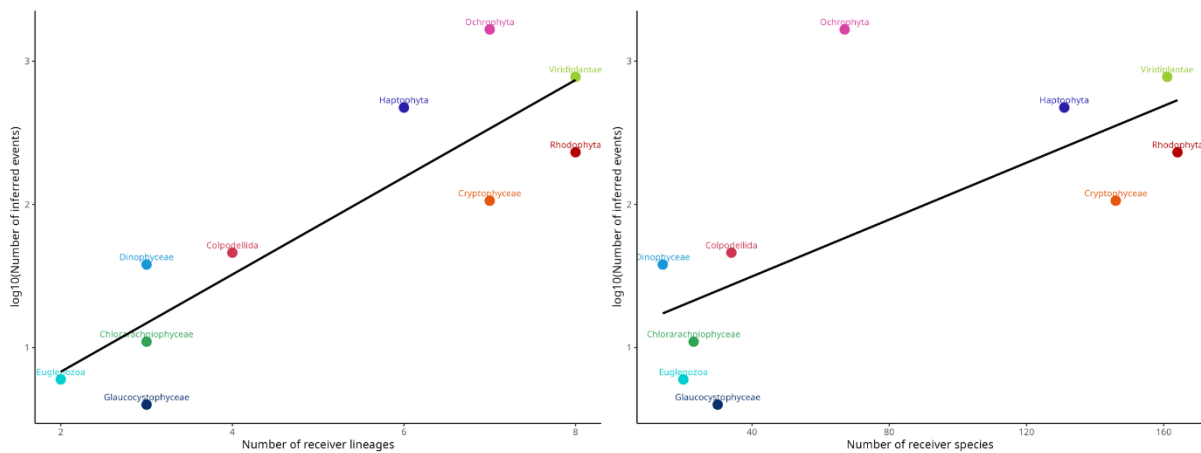


Figure 51: **Donor-centred view of inferred gene-transfer patterns.** Each point represents a donor lineage. The y-axis shows the log10-transformed number of inferred transfer events contributed by each donor lineage. This donor contribution is plotted against two measures of receiver diversity: **(a)** the number of receiver lineages reached and **(b)** the number of receiver species reached. The black line indicates the linear trend.

Importantly, donor assignments rely on a lowest common ancestor (LCA) approach applied to the set of donor sequences involved in each transfer event. As a result, an assignment to “Ochrophyta” does not necessarily imply that the donor originated from the ancestral ochrophyte lineage. Instead, such a broad label may reflect either a genuine signal spanning multiple ochrophyte subgroups or an insufficient phylogenetic resolution, leading to a partially unresolved taxonomic assignment.

In this regard, to investigate potential biases among ochrophyte donor subgroups, NCBI Taxonomy is not well suited. Although it represents the standard reference used in our laboratory for taxonomic annotation and LCA computation, it does not fully reflect the resolution proposed in recent phylogenomic studies. To obtain a finer view of ochrophyte donor diversity, we therefore adopted the recent taxonomic framework proposed by Di Franco et al., 2022 (Figure 52). Because *Pinguiphyceae* are absent from our dataset, we refer to Diatomista+P as Diatomista and Chrysista-P as Chrysista.

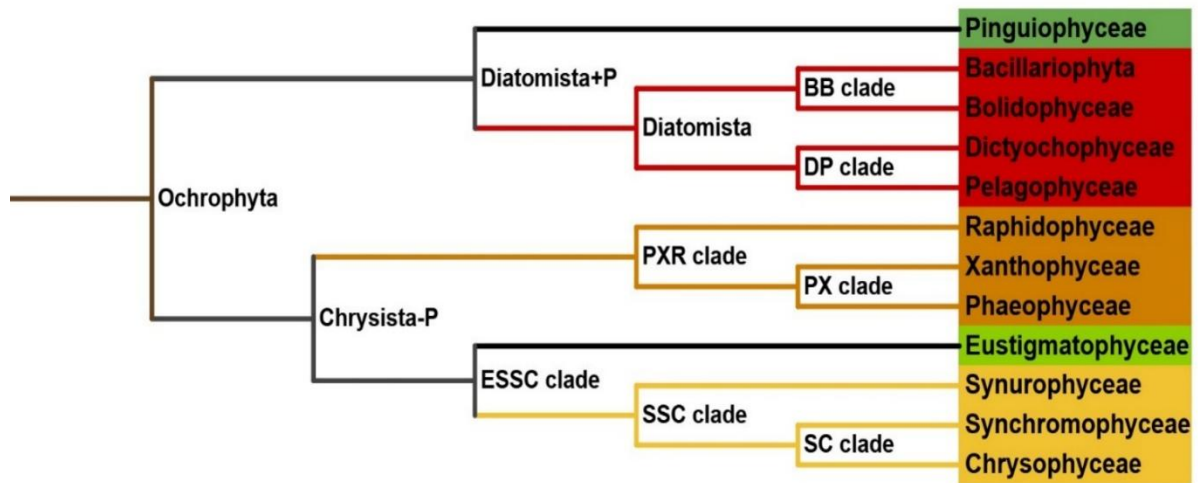


Figure 52: Cladogram representing the taxonomy of Ochrophyta, adapted from Di Franco et al., 2022.

3.2.4.1. Assessing contributions of ochrophyte subclades

Using Di Franco’s taxonomic framework to recompute the LCAs of donor sequences previously computed using NCBI Taxonomy, we notice that even at a broad level (classification into Chrysista and Diatomista), a substantial fraction of donors remains unresolved, meaning that the LCAs of donor sequences are distributed across multiple ochrophyte subclades, spanning both Diatomista and Chrysista. When considering only resolved assignments, most receiver lineages are dominated by

Diatomista donors. In contrast, Colpodellida show the opposite pattern, with Chrysista accounting for 77% of resolved assignments (Table 14).

receiver	total	Chrysista	Diatomista	Ochrophyta	resolved	resolved_ratio	Chrysista/Diatomista
Apicomplexa	6	1	0	5	1	0.17	1.00/0.00
Chlorarachniophyceae	68	12	22	34	34	0.5	0.35/ 0.65
Colpodellida	163	55	16	92	71	0.44	0.77 /0.23
Dinophyceae	196	32	94	70	126	0.64	0.25/ 0.75
Dinophysiales	10	0	8	2	8	0.8	0.00/ 1.00
Euglenozoa	87	9	43	35	52	0.6	0.17/ 0.83
Haptophyta	250	12	141	97	153	0.6	0.08/ 0.92
Kareniaaceae	45	5	31	9	36	0.8	0.14/ 0.86
Kryptoperidiniaceae	837	96	703	38	799	0.95	0.12/ 0.88

Table 14: **High-level composition of ochrophyte donor assignments across receiver lineages.** For each receiver lineage, the table reports the total number of inferred transfer events involving ochrophyte donors (total), partitioned into **Diatomista**, **Chrysista**, and unresolved **Ochrophyta** assignments. The column “resolved” indicates the number of events that could be assigned to either Diatomista or Chrysista, and “resolved_ratio” represents the proportion of resolved assignments relative to all ochrophyte donor events. The final column shows the relative contribution of Chrysista and Diatomista among resolved events (Chrysista/Diatomista). Most lineages are dominated by Diatomista donors, whereas **Colpodellida** show a strong relative enrichment in Chrysista-derived events.

To further examine ochrophyte donors, we analysed the composition of finely resolved assignments, i.e., solely events that could be attributed to specific subclades (Table 15). The apparent dominance of BB_clade donors in the reference pipeline is largely driven by Kryptoperidiniaceae and their diatom-derived plastids (Imanian et al., 2010; McEwan and Keeling, 2004). When this lineage is excluded, donor contributions become more balanced, with DP_clade donors dominating over BB_clade (120 VS 80) even though Bacillariophyta (BB_clade) is the most represented subclade among ochrophytes, with 57/85 species, excluding any potential sampling bias.

receiver	BB_clade	DP_clade	ESSC_clade	PXR_clade	resolved_fine
Apicomplexa	0	0	0	1	1
Chlorarachniophyceae	5	9	1	3	18
Colpodellida	4	9	9	7	29
Dinophyceae	18	29	21	4	72
Dinophysiales	7	1	0	0	8
Euglenozoa	19	11	0	1	31
Haptophyta	15	51	2	1	69
Kareniaaceae	12	12	4	0	28
Kryptoperidiniaceae	685	1	92	0	778
Total	765	123	129	17	1034
Total (-Kry)	80	121	37	17	256

Table 15: **Fine-scale distribution of resolved ochrophyte donor subclades across receiver lineages.** For each receiver lineage, the table reports the number of inferred transfer events attributed to four ochrophyte donor subclades: BB clade (primarily Bacillariophyta), DP clade (Dictyochophyceae + Pelagophyceae), ESSC clade (Eustigmatophyceae + Synchromophyceae + Synurophyceae + Chrysophyceae), and PXR clade. Only transfer events that could be confidently assigned to these specific subclades were considered (resolved_fine). Assignments at broader taxonomic levels (Diatomista,

Chrysis, or Ochrophyta) were treated as unresolved and excluded from the analysis. Totals are shown both including and excluding *Kryptoperidiniaceae*.

The observed related contributions (Figure 53, same data as in Table 15 but in percentage of contributions) confirm that the apparent dominance of ochrophyte donors does not correspond to a single donor lineage but rather reflects multiple independent signals originating from different subclades. In particular, *Kryptoperidiniaceae* display the strongest proportion of BB_clade donors, whereas Haptophyta show a stronger relative contribution from DP_clade donors (both subclades of Diatomista), consistent with previous studies (Dorrell et al., 2021, 2017). In contrast, Colpodellida display a mixed signal involving donors belonging to Chrysis (ESSC and PXR), consistent with the Chrysis-dominated profile observed at the broader taxonomic level.

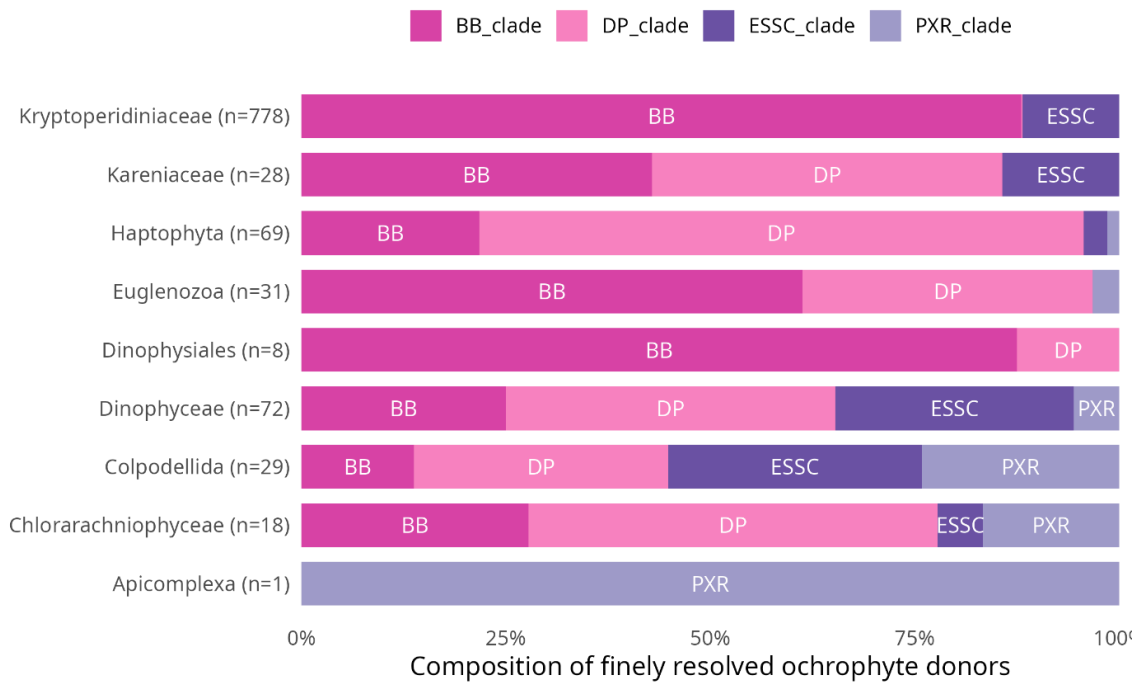


Figure 53: Fine-scale composition of resolved ochrophyte donor subclades across receiver lineages. The proportional contribution of four ochrophyte donor subclades (BB, DP, ESSC, and PXR) is shown for each receiver lineage. Only transfer events that could be assigned to these specific subclades were considered (resolved_fine); assignments at the Diatomista, Chrysis, or Ochrophyta levels were treated as unresolved and excluded.

In lineages with ochrophyte-derived plastids, the phylogenetic signal is quite clear. For example, in *Kryptoperidiniaceae*, the vast majority of events trace back to the BB clade (685/837), and more specifically to Bacillariophyta (652/685), as expected (Imanian et al., 2010). In contrast, Colpodellida show a distinct signal dominated by ESSC clade donors, with 20 sequences belonging to Eustigmatophyceae (*Nannochloropsis gaditana*), repeatedly recovered as the nearest donor taxon, which is consistent with previous findings proposing eustigmatophytes as the likely source of the colpodellid plastid, (Ševčíková et al., 2015; Füssy and Oborník, 2017), consistent with the fact that they share the absence of chlorophyll c (Oborník et al., 2012).

However, transfers attributed to ochrophyte donors are not restricted to lineages harbouring confirmed ochrophyte-derived plastids, i.e., Kryptoperidiniaceae and Colpodellida. Ochrophyte-derived contributions are also detected in lineages possessing complex red plastids of known non-ochrophyte origin like in Dinophysiales (cryptophyte origin, mainly from BB clade among 8 fine-resolved transfers) and Kareniaceae (haptophyte origin, mainly BB and DP clades among 28 fine-resolved transfers) but in a relatively moderate amount (10 and 45, respectively). Even complex green plastids lineages are concerned, with contributions amounting to 68 transfer events detected for chlorarachniophytes (from all subclades, among 18 fine-resolved transfers) and 87 for euglenids (mainly from BB and DP clades, among 31 fine-resolved transfers).

Yet, the strongest contributions from ochrophytes are in the unresolved red-plastid derived lineages, namely Haptophyta (250) and Dinophyceae (peridinin dinoflagellates, 196). On the one hand, Dinophyceae display the strongest absolute contribution from ESSC clade with 21 transfers (aside from Kryptoperidiniaceae), although it does not stand out relative to other contributing subclades, namely BB clade (18) and DP clade (29). On the other hand, haptophytes display the strongest relative contribution from DP clade donors (Dorrell et al., 2021, 2017) with 51/69 of fine-resolved events (69/250). However, close inspection of the nearest donor sequences does not suggest any association with a more specific member of the DP clade (33 events associated with Dictyochophyceae and 31 with Pelagophyceae).

3.2.4.2. Functional characterisation

Having shown that ochrophyte-derived transfers do not originate from a single donor source, we next ask whether these contributions are functionally similar or whether they cover a broad range of pathways across the major complex red-plastid receiver lineages (Kryptoperidiniaceae, Haptophyta, peridinin dinoflagellates, Kareniaceae and Colpodellida). The overall pathway composition of ochrophyte-derived transfers spans a wide functional range (Figure 54). Recurrent pathways are found in Nucleotide metabolism, Lipid metabolism, Carbohydrate metabolism, Amino acid metabolism, Metabolism of cofactors and vitamins, Metabolism of terpenoids and polyketides, and Energy metabolism, indicating that the ochrophyte signal is not restricted to a single functional class.

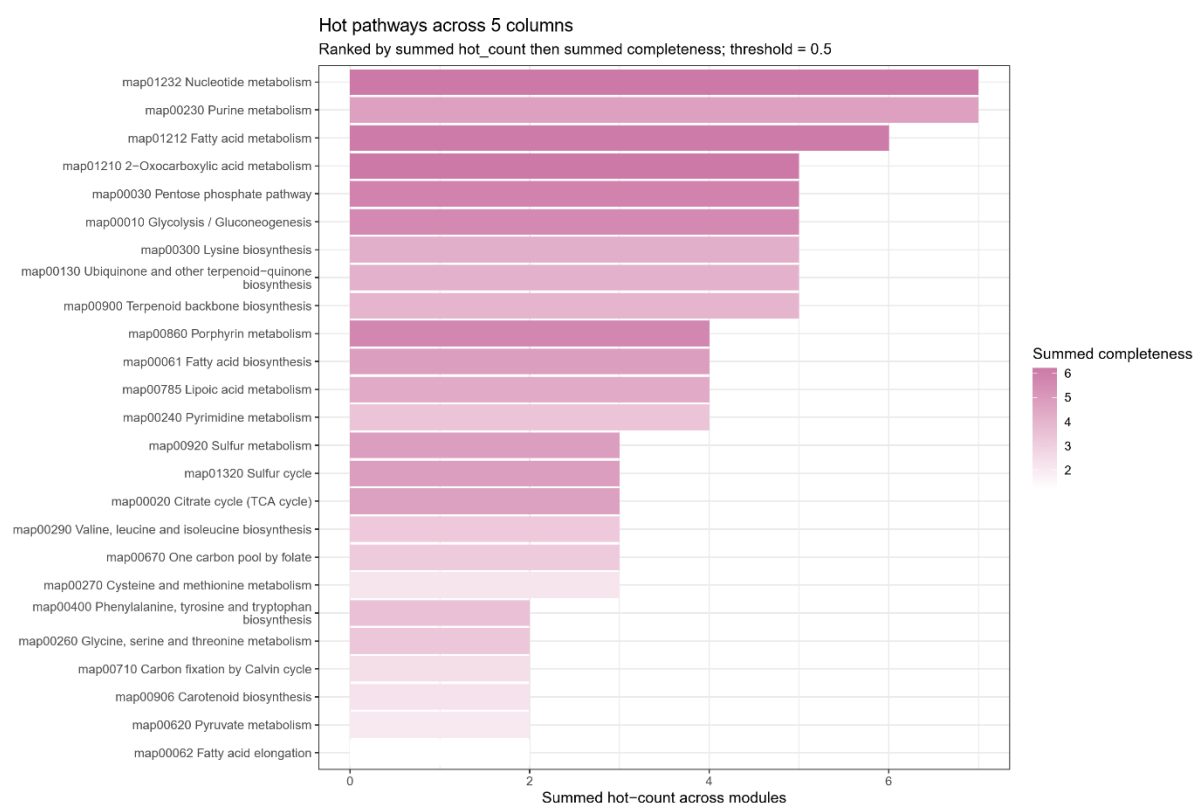


Figure 54: Highly shared pathways across the top 5 receiver complex red-plastid lineages.

However, the differential analysis shows that this global profile is not homogeneous across receivers (Figure 55). Compared with the other complex red-plastid lineages, Kryptoperidiniaceae are enriched in Lipid metabolism (*Fatty acid biosynthesis*; *Fatty acid metabolism*), Metabolism of cofactors and vitamins (*Lipoic acid metabolism*; *Porphyrin metabolism*; *Biotin metabolism*; *Pantothenate and CoA biosynthesis*), Amino acid metabolism (*Valine, leucine and isoleucine biosynthesis*; *Phenylalanine, tyrosine and tryptophan biosynthesis*; *Lysine biosynthesis*), and Carbohydrate metabolism (*Pyruvate metabolism*). By contrast, the non-Kryptoperidiniaceae receivers are enriched in Metabolism of cofactors and vitamins (*Thiamine metabolism*; *One carbon pool by folate*; *Ubiquinone and other terpenoid-quinone biosynthesis*), Glycan biosynthesis and metabolism (*Biosynthesis of nucleotide sugars*; *Amino sugar and nucleotide sugar metabolism*), Amino acid metabolism (*Arginine biosynthesis*; *Glycine, serine and threonine metabolism*), Nucleotide metabolism (*Purine metabolism*), Metabolism of other amino acids (*Glutathione metabolism*), and Carbohydrate metabolism (*Starch and sucrose metabolism*). Together, these results show that the functional profile of ochrophyte-derived transfers differs markedly between Kryptoperidiniaceae and the remaining receivers.

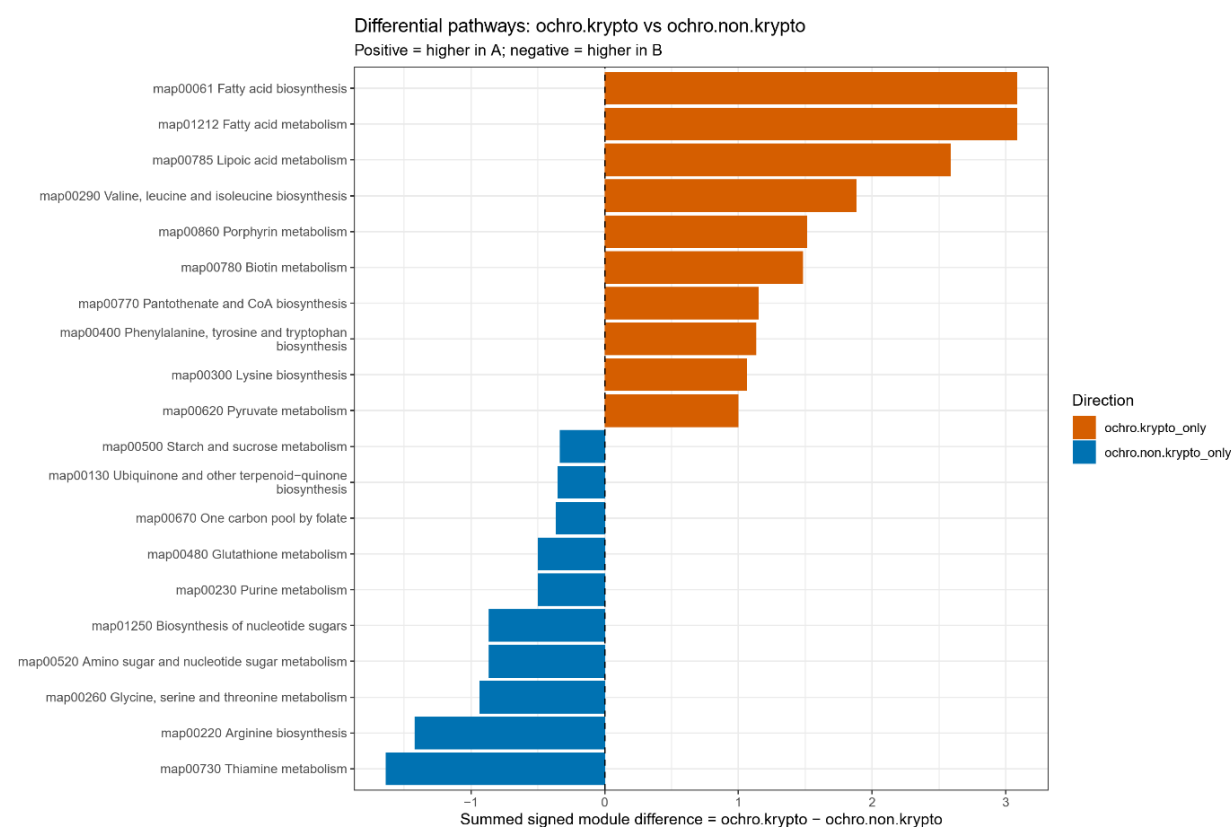


Figure 55: Differential pathways in *Kryptoperidiniaceae* (ochro.krypto_only) VS non-*Kryptoperidiniaceae* red-plastid receivers (ochro.non.krypto_only).

3.2.4.3. Ochrophyta and Haptophyta: hints to unresolved plastids origin?

The origin of the ochrophyte plastid remains unresolved, with phylogenomic studies alternatively suggesting a direct secondary endosymbiosis involving a red algal (Rhodophyte) endosymbiont (Delwiche, 1999), a relationship with plastids from other complex red lineages such as cryptophytes (Bodył et al., 2009a; Stiller et al., 2014b) or even haptophytes (Ponce-Toledo et al., 2025). The origin of the haptophyte plastid also remains unresolved, with phylogenomic studies alternatively suggesting a relationship with ochrophyte plastids (Stiller et al., 2014b), with cryptophyte plastids (Bodył et al., 2009a; Ponce-Toledo et al., 2025), or an independent acquisition from a red algal (rhodophyte) endosymbiont (Delwiche, 1999). The results presented here for haptophytes (as receivers) suggest a stronger relationship with ochrophytes, with 250/334 (~75%) transfers (Table 16).

	<i>Haptophyta</i>	<i>Pavlova</i>	<i>Prymnesiophyceae</i>	<i>Total</i>
<i>Cryptophyceae</i>	3	1	7	11
<i>Ochrophyta</i>	102	27	121	250
<i>Rhodophyta</i>	11	2	8	21
<i>Viridiplantae</i>	19	4	29	52

Table 16: Origin of inferred transfers for haptophytes as receivers. Breakdown by haptophyte sub-groups (*Pavlova* and *Prymnesiophyceae*) when resolution was available.

When breaking down haptophyte receivers into subclades, it seems that ochrophyte-derived transfers are more numerous into *Prymnesiophyceae* than in *Pavlova* (121 vs 27), but a substantial fraction of events (102) can only be assigned to Haptophyta without further resolution. This apparent bias towards *Prymnesiophyceae* is not specific to ochrophyte donors, as similar patterns are also observed for cryptophyte, rhodophyte, and Viridiplantae-derived events, suggesting that the bias may (at least partially) reflect the lower sampling of *Pavlova* in the dataset (5 vs 14 *Prymnesiophyceae*) which may cause some *Pavlova*-specific events to remain unresolved at the Haptophyta level (102 events). Nonetheless, taken at face value, these results are compatible with an ochrophyte contribution to the origin of the haptophyte plastid (Stiller et al., 2014b; Dorrell et al., 2017).

Interestingly, a reciprocal signal is observed: while haptophytes show a strong ochrophyte donor signal (250/334), ochrophytes themselves display a substantial contribution from haptophytes (120/257). One could argue in the favour of an inflated signal in haptophytes given the unbalanced sampling and especially the dominance of ochrophyte (85 species), but even with a more balanced sampling (36 species) the signal is still heavily dominant (229/351), revealing an apparently unbiased signal (Table 17, see section 4.1.1.2 for a full analysis using a balanced sampling). On the other hand, balancing the sampling shows a slightly different result overall for ochrophytes: while the contribution from haptophytes is still high (92), it is now more balanced and equivalent to other main donors (Rhodophyta: 80, Viridiplantae: 89). However, because a substantial fraction of transfers remains unresolved (~100 events in both directions), no clear subclade-to-subclade pattern could be identified, whether in the unbalanced or balanced sampling analyses.

Donor	Ochrophyta (85)	Ochrophyta (36)	Haptophyta (19)	Haptophyta (12)
<i>Cryptophyceae</i>	26	30	11	23
<i>Glaucocystophyceae</i>	0	1	0	2
<i>Haptophyta</i>	120	92	-	-
<i>Ochrophyta</i>	-	-	250	229
<i>Rhodophyta</i>	30	80	21	28
<i>Viridiplantae</i>	81	89	52	68

Table 17: Donors comparison for ochrophytes and haptophytes between the original sampling and a balanced sampling (bal_samp) version of the dataset.

Given the strong reciprocal signal between Ochrophyta and Haptophyta, we next ask whether transfers inferred in the two directions show the same functional profile, or whether each direction is associated with a distinct subset of pathways. The comparison (Figure 56) reveals that ochrophyte-derived transfers into Haptophyta are enriched in Carbohydrate metabolism (*Pentose phosphate pathway*; *Glycolysis / Gluconeogenesis*; *Citrate cycle (TCA cycle)*), Amino acid metabolism (*2-Oxocarboxylic acid metabolism*; *Lysine biosynthesis*; *Phenylalanine, tyrosine and tryptophan biosynthesis*), Glycan

biosynthesis and metabolism (*Biosynthesis of nucleotide sugars; Amino sugar and nucleotide sugar metabolism*), Metabolism of terpenoids and polyketides (*Terpenoid backbone biosynthesis*), and Metabolism of cofactors and vitamins (*Porphyrin metabolism*). By contrast, haptophyte-derived transfers into Ochrophyta are enriched in Metabolism of cofactors and vitamins (*Lipoic acid metabolism; Pantothenate and CoA biosynthesis; Biotin metabolism*), Amino acid metabolism (*Arginine and proline metabolism; Cysteine and methionine metabolism*), Carbohydrate metabolism (*Glyoxylate and dicarboxylate metabolism*), and Energy metabolism (*Methane metabolism*). In summary, ochrophyte-to-haptophyte transfers appear relatively more enriched in functions associated with central carbon metabolism and related biosynthetic pathways, whereas the reverse direction appears relatively more enriched in cofactor/vitamin and selected amino-acid pathways. This asymmetry suggests that reciprocal gene flow between Ochrophyta and Haptophyta affects partly distinct plastid-associated metabolic functions.

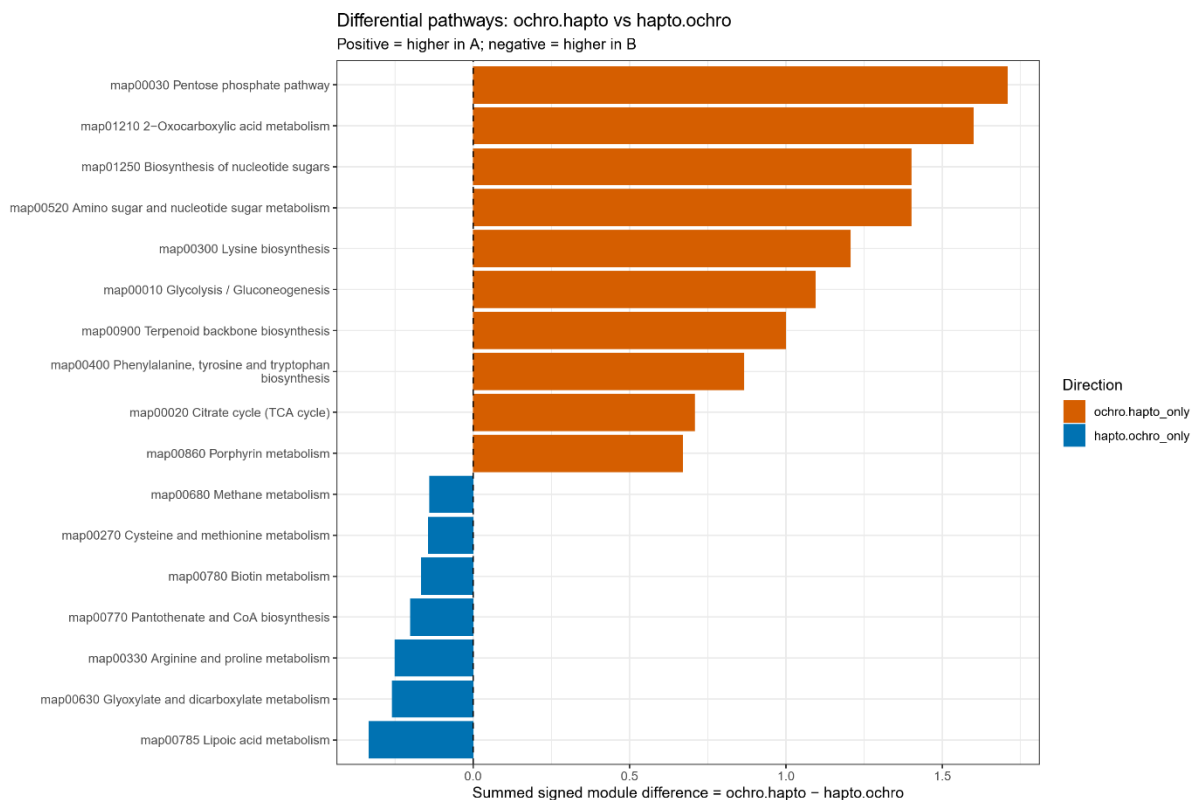


Figure 56: Differential pathways between ochrophyte-derived transfers into Haptophyta (ochro.hapto_only) and haptophyte-derived transfers into Ochrophyta (hapto.ochro_only).

4. Methodological results

4.1. How robust is the signal?

Aside from the choice of index for α -diversity measurement, which will be addressed in Section 4.1.2, there are many ways to perform the analysis due to the multiple parameters available at each layer of the PAGTI pipeline (Figure 57). To avoid confounding their effects, we adopted a single-parameter variation strategy: starting from the reference pipeline (nu_rdt), each parameter was changed individually while all others were held constant. The reference pipeline uses the following settings: (i) trees inferred under LG+C20+G with IQ-TREE; (ii) five age categories defined after published molecular-clock analyses (Jackson et al., 2018; Strassert et al., 2021); (iii) a maximum of two intruder taxa tolerated during event inference; (iv) no bootstrap support filter; and (v) the full taxonomic sampling (211 organisms, Figure 25).

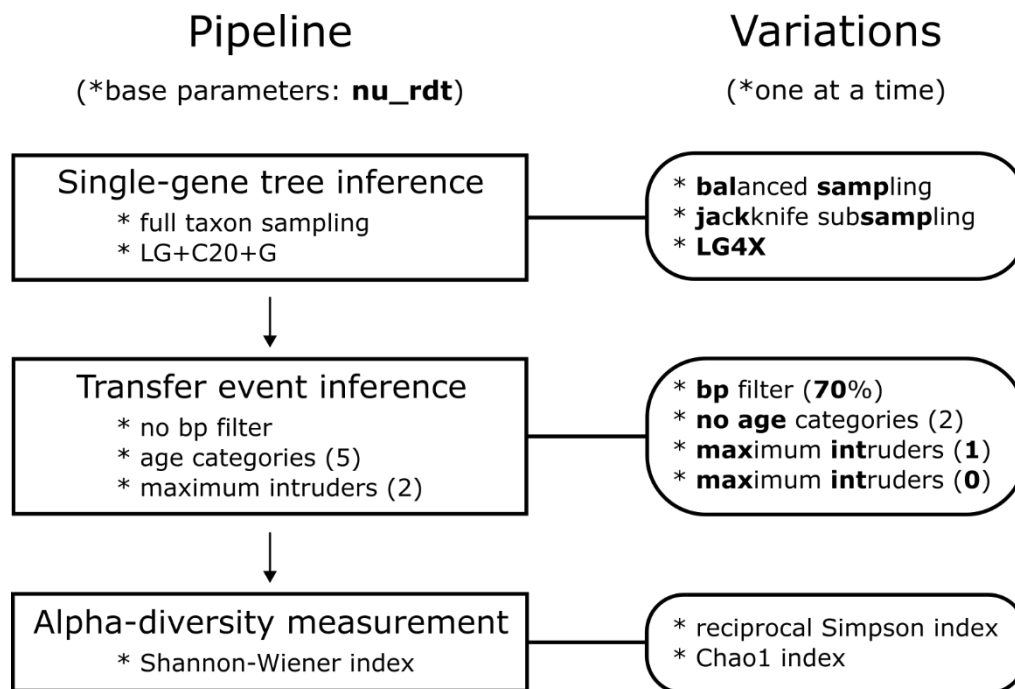


Figure 57: PAGTI pipeline and pipeline variations.

4.1.1. Pipeline robustness

Among the multiple possible variations, we focused on six that span the two main layers of the pipeline: single-gene tree inference and transfer event inference. Variations at the level of tree inference obviously required recomputing the single-gene trees, which yielded trees that potentially differed in topology from those of the reference pipeline. In contrast, variations at the level of event inference simply consisted in re-running PAGTI event detector under a different set of parameters. At the tree level, variations included a balanced taxonomic sampling (bal_samp) and a different model of sequence

evolution (LG4X). At the event level, we applied a filter only retaining events with a bootstrap support $\geq 70\%$ (bp70), investigated the effect of lowering the maximum number of intruders to one (maxint1) or zero (maxint0), and the consequences of reducing the number of age categories from five to two (noage). For each variation, we assessed the effect on α -diversity at three levels: (i) the magnitude and significance of the global shift in α -diversity values (scale), (ii) the preservation of the relative ordering of species and lineages (rank concordance), and (iii) the number of lineage-level class reassignments relative to the reference (class changes). A summary of these results is presented in Table 18, with species coverage and mean event counts in Table 19 and lineage-level event means in Table 20.

variation	Δmedian	ρ	class_changes
LG4X	+0.01 (ns)	0.95 (ns)	1
bp70	-0.04 (ns)	1.00 (ns)	0.5
maxint1	-0.03 (*)	0.97 (ns)	0
maxint0	-0.12 (**)	0.95 (ns)	0.5
noage	+0.13 (**)	0.93 (ns)	4
bal_samp	+0.07 (ns)	0.80 (ns)	1.5
jak_samp	+0.09 (ns)	0.95 (ns)	2.5

Table 18: Global robustness summary across pipeline variations. For each retained variation, the table reports the median shift in α -diversity values relative to the reference pipeline (Δ median; Wilcoxon signed-rank test), the Spearman rank-correlation coefficient with the reference ranking of lineage medians (ρ ; Fisher z-test of $H_0: \rho \geq 0.85$), the number of lineage α -diversity class changes (class_changes). The reference pipeline is nu_rdt. Significance codes: ns, not significant; *, $p < 0.05$; **, $p < 0.01$; ***, $p < 0.001$.

variation	n_species	coverage_vs_nu_rdt	mean_events_per_species	Δspecies	Δmean_events_per_species
nu_rdt	151	1.0	122.2	0	0.0
LG4X	156	1.0	106.9	+5	-15.3
bp70	139	0.9	81.7	-12	-40.5
maxint1	143	0.9	109.4	-8	-12.9
maxint0	81	0.5	109.4	-70	-12.8
noage	158	1.0	171.3	+7	+49.1
bal_samp	65	0.4	181.0	-86	+58.8

Table 19: Dataset size and event coverage across pipeline variations. For each retained variation, the table reports the number of species with enough inferred transfer events (≥ 30 transfers threshold) and thus retained for α -diversity analysis (n_species), the relative species coverage compared with the reference pipeline (coverage_vs_nu_rdt), the mean number of inferred events per species (mean_events_per_species), and the corresponding differences in species number (Δ species) and mean event count per species (Δ mean_events_per_species) relative to nu_rdt.

<i>Lineage</i>	<i>nu_rdt</i>	<i>LG4X</i>	<i>bp70</i>	<i>maxint1</i>	<i>maxint0</i>	<i>noage</i>	<i>bal_samp</i>
<i>Apicomplexa</i>	45	37	38	44	39	47	42
<i>Chlorarachniophyceae</i>	307	269	164	288	222	326	318
<i>Colpodellida</i>	216	179	128	191	125	257	85
<i>Cryptophyceae</i>	202	179	110	177	142	379	231
<i>Dinophyceae</i>	367	331	209	335	240	383	410
<i>Dinophysiales</i>	52	46	39	51	51	52	53
<i>Euglenozoa</i>	355	281	202	333	257	369	342
<i>Haptophyta</i>	334	296	210	291	179	437	350
<i>Kareniaceae</i>	348	307	275	348	323	348	322
<i>Kryptoperidiniaceae</i>	870	735	684	868	857	880	704
<i>Ochromytha</i>	257	235	153	219	137	416	292
<i>Total events</i>	3353	2895	2212	3145	2572	3894	3149

Table 20: Total number of inferred events per receiver lineage across pipeline variations. For each lineage, the table reports the total number of inferred events per lineage under the reference pipeline (*nu_rdt*) and under each variation.

4.1.1.1. Variations with limited effect on the signal

LG4X reconstruction model

Changing the model from LG+C20+G to LG4X produced no significant shift in α -diversity values (Δ median = +0.02; Table 18), preserved lineage rank concordance ($\rho = 0.84$), and resulted in only one lineage class change. This stability is reassuring, as LG4X and LG+C20+G represent two fundamentally different approaches to modelling among-site heterogeneity (the former using a free-rate mixture of four matrices, the latter combining a profile mixture with gamma-distributed rates). The near-identical α -diversity classifications obtained under both evolutionary models indicate that the transfer event signal is mostly driven by the underlying sequence data (including taxonomic sampling; see below) rather than by model-specific tree reconstruction. Species coverage remained virtually unchanged (156 vs. 151 species), and the mean number of events per species decreased only moderately (106.9 vs. 122.2; Table 19), indicating that LG4X trees yield slightly fewer inferred events overall but without altering the relative donor profile of each receiver lineages.

Bootstrap support filter (bp70)

Applying a bootstrap support filter of 70% (bp70) also preserved the classification, with one lineage class change and a moderate downward shift in α -diversity values (Δ median = -0.03; Table 18). Rank concordance was excellent ($\rho = 0.89$). As expected, this filter had a strong effect on the number of inferred events, reducing the mean event count per species by approximately 40 (from 122.2 to 81.7; Table 19) and excluding 12 species below the 30-events threshold. Nonetheless, the α -diversity patterns remained intact, indicating that the weakly supported events removed by this filter did not lead to systematic bias but instead were congruent with robust events and/or contributed noise. The lineage most affected in absolute terms was Kryptoperidiniaceae (from 870 to 684 total events; Table 20),

consistent with the large number of inferred events in this group, but even there, the α -diversity class was preserved.

Single-intruder tolerance (maxint1)

Reducing the maximum number of tolerated intruders from two to one (maxint1) was similarly conservative. No lineage class change was observed, and rank concordance was fully preserved ($\rho = 0.87$). The minimum transfer threshold (≥ 30) excluded slightly more species (143 species, instead of 151; coverage 0.9) and reduced the mean event count by roughly 13 events per species (Table 19). This confirms that the majority of the inferred transfer events in the reference pipeline do not depend on the tolerance of a second intruder taxon, and that most events are supported by clean paraphyletic topologies involving at most one misplaced sequence.

4.1.1.2. Variations with stronger effects

4.1.1.2.1. *Strict intruder exclusion (maxint0)*

Reducing the intruder tolerance to zero (maxint0), thereby requiring strictly monophyletic receiver clans with no misplaced sequences, had a dramatic effect on species retention: only 81 of the 151 species affected in the reference pipeline (54%) passed the 30-event threshold (Table 19). Two lineage class changes were observed, and the global α -diversity values were significantly shifted downward ($\Delta_{\text{median}} = -0.14$; Table 18). Crucially, however, rank concordance was only marginally affected ($\rho = 0.80$; Table 18), meaning that lineages largely retained the same relative ordering despite the reduced event counts.

The loss of species was not uniform across receiver lineages. The most severely affected group was Ochrophyta, for which the total event count dropped from 257 to 137 (Table 20), causing many ochrophyte species to fall below the 30-event threshold. This observation has a direct biological interpretation: ochrophyte transfer events are disproportionately embedded in complex topologies involving nested transfers or misplaced sequences, so that demanding strict receiver monophyly eliminates a large fraction of their signal. In contrast, lineages whose events are dominated by clean (and recent) paraphyletic donor-receiver relationships (e.g., Kryptoperidiniaceae: 857 total events under maxint0 vs. 870 under the reference) are far less affected. This result validates the biological relevance of the intruder tolerance parameter: it acts as a topological complexity mask, and its effect is predictably strongest in lineages known to be associated with complex or ambiguous donor configurations.

4.1.1.2.2. Removal of age-compatibility filtering (noage)

Removing age-compatibility rules, i.e., collapsing the five age categories into just two (Archaeplastida vs all complex algae), produced the strongest effect among all tested variations, with four lineage class changes and a highly significant upward shift in α -diversity values (Δ median = +0.38; Table 18, Figure 58). The mean number of events per species increased substantially (from 122.2 to 171.3; Table 19) because transfers that had previously been filtered out as chronologically impossible were now accepted. Despite this inflation, rank concordance was moderately affected ($\rho = 0.78$; Table 18), indicating that age filtering mainly rescales α -diversity rather than fundamentally reorganising lineage ordering. The effect was nonetheless uneven across lineages: older groups showed the largest increases, notably Cryptophyceae (median $\Delta \approx +0.72$), Ochrophyta (median $\Delta \approx +0.42$), and Haptophyta (median $\Delta \approx +0.32$), whereas younger lineages such as Euglenozoa (median $\Delta \approx +0.07$) and Dinophyceae (median $\Delta \approx +0.12$) were much less affected. This pattern is expected, because relaxing the filter disproportionately expands the donor repertoire of older receiver lineages by admitting donors from younger age classes.

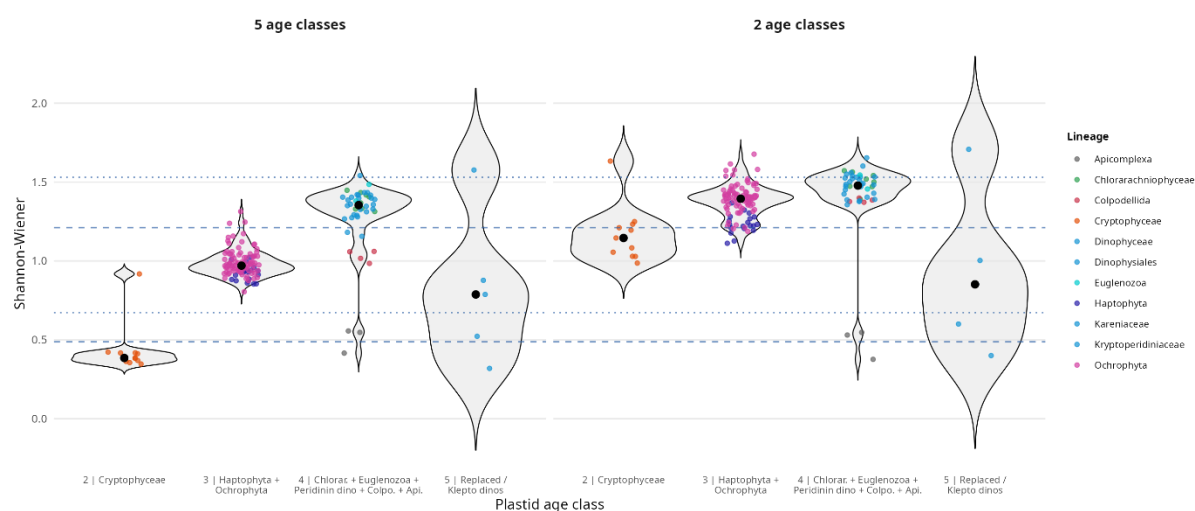


Figure 58: Effect of age-class scheme on species-level donor α -diversity across plastid-bearing lineages. Violin plots compare the distribution of Shannon-Wiener-based donor α -diversity values obtained under two transfer-compatibility schemes: a **five-age-class model** (left) and a simplified **two-age-class model** (right). Each point represents one species, coloured by receiver lineage, and black points indicate lineage medians. Blue dashed horizontal lines mark the KDE-derived Shannon-Wiener bin cutoffs defined in the main analysis.

The additional events recovered without age filtering likely reflect a mixture of genuine late transfers and artefacts caused by missing donor lineages. Late gene replacement after plastid establishment and receiver lineage diversification is biologically plausible, but the magnitude of the shift, especially in older lineages such as Cryptophyceae, suggests that it cannot fully explain the effect. A more parsimonious additional explanation is that, when the true donor is absent from the dataset, receiver sequences may cluster with the nearest sampled paraphyletic lineage, generating spurious donor assignments. Relaxing the age filter is expected to amplify this problem because some of these

misassigned events would otherwise be excluded as chronologically impossible. In that sense, the age-compatibility filter does more than enforce temporal plausibility: it also reduces sensitivity to missing-donor artefacts.

4.1.1.2.3. *Balanced taxonomic sampling (bal_samp)*

The balanced-sampling variation used a manually curated subset (with the help of Dr. Henner Brinkmann) of 65 species selected from a ribosomal phylogenomic tree to reduce the taxonomic imbalance present in the full sampling (Table 26; see Methods 5.10 in Chapter V). The pruning primarily targeted the most oversampled lineages: Ochrophyta was reduced from 85 to 21 species, Dinophyceae from 32 to 18, Viridiplantae from 29 to 15, and Haptophyta from 23 to 11; smaller lineages were left unchanged (Table 26). Unlike most other variations (except LG4X), which merely filter or reparametrized the reference analysis, *bal_samp* constitutes an independent analytical run on a different species set: the reduced taxonomic sampling was processed through the whole pipeline from tree inference to PAGTI transfer detection, producing gene trees that necessarily differ in topology from those of the reference pipeline. The total number of events per lineage remained broadly comparable to the reference for most lineages (Table 20), indicating that balanced sampling does not dramatically alter the volume of inferred transfers; the apparent increase in mean events per species (from 122.2 to 181.0; Table 19) is a mechanical effect of dividing a similar event pool by fewer species. Balanced sampling is the only variation that disrupted rank concordance ($\rho = 0.35$, $p < 0.001$; Table 18), likely because halving the species count not only can alter tree topologies but can also change the context in which transfers are inferred, many genuine “donor” species being absent from the reduced trees. Notably, while total event counts were stable, the donor profile shifted for the most heavily pruned lineage (Ochrophyta: Haptophyta contributions dropped from 120 to 92 while Rhodophyta rose from 30 to 80).

4.1.1.2.4. *Jackknife analysis of taxonomic sampling (jak_samp)*

To further assess the sensitivity of α -diversity estimates to taxonomic sampling of donor lineages, we performed a jackknife analysis in which all donor groups were subsampled to 11 species (groups with fewer than 11 species were left unchanged). Ten independent random subsamples were generated and, for each iteration, the PAGTI pipeline was re-run on the reduced species set. The results were compared to the reference pipeline by computing the shift in median Shannon-Wiener index per receiver lineage, and by assessing whether lineages crossed the KDE-derived class boundaries (Figure 59; Table 21).

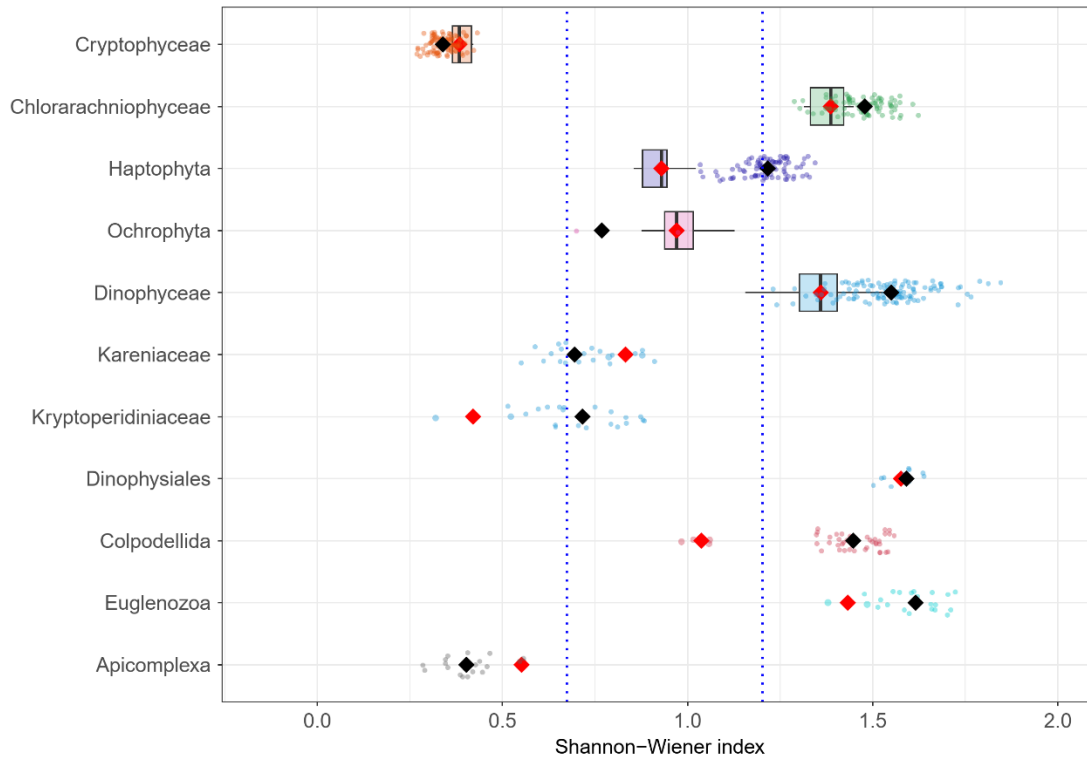


Figure 59: Species α -diversity distribution broken down by lineage across 10 jackknife subsamples. Blue dotted lines indicate KDE cutoffs. Lineage populations ≤ 5 species were plotted only as points, each point corresponding to one species on one subsample. Larger lineages are shown as boxplots, and each have 110 points (11 species in 10 subsamples). Jackknife medians are shown as black diamonds while red diamonds correspond to the original medians of the reference pipeline.

Lineage	Ref_med.	Ref_class	Jak_med.	Jak_class	% shift
Cryptophyceae	0.384	low	0.339	low	-11.6%
Chlorarachniophyceae	1.387	high	1.479	high	+6.6%
Haptophyta	0.929	mid	1.217	high	+30.9%
Ochrophyta	0.976	mid	0.768	mid	-21.3%
Dinophyceae	1.360	high	1.550	high	+14.0%
Kareniaceae	0.832	mid	0.695	mid	-16.5%
Kryptoperidiniaceae	0.421	low	0.716	mid	+70.2%
Dinophysiales	1.576	high	1.591	high	+0.9%
Colpodellida	1.037	mid	1.447	high	+39.5%
Euglenozoa	1.433	high	1.616	high	+12.8%
Apicomplexa	0.552	low	0.403	low	-27.0%

Table 21: Effect of jackknife subsampling on α -diversity estimates (Jak_med VS Ref_med) and classification (Jak_class VS Ref_class) for each receiver lineage. For each lineage, % shift is relative to the original median (Ref_med).

Eight of the eleven lineages retained their α -diversity class under subsampling. Cryptophyceae (-0.045) and Apicomplexa (-0.149) remained firmly in the low-diversity class. Chlorarachniophyceae (+0.092), Dinophyceae (+0.190), Dinophysiales (+0.015) and Euglenozoa (+0.183) all stayed in the high-diversity class. Ochrophyta (-0.208) and Kareniaceae (-0.137) remained intermediate, although Kareniaceae shifted close to the low/intermediate boundary (Table 21). These results confirm that the classification of the reference pipeline is robust for the majority of lineages (Figure 60). Three lineages,

however, shifted to a higher α -diversity class under jackknife subsampling. Kryptoperidiniaceae moved from low to intermediate (main median 0.421 $\rightarrow$ jackknife median 0.716, +70.2%), Colpodellida shifted from intermediate to high (1.037 $\rightarrow$ 1.447, +39.5%), and Haptophyta crossed from intermediate to high (0.929 $\rightarrow$ 1.217, +30.9%) (Table 21). In all three cases, reducing the donor pool increased the apparent diversity, consistent with a redistribution of relative donor proportions: when dominant donor groups are thinned (lower dominance), the relative weight of less common donors increases, raising the Shannon-Wiener index. However, Ochrophyta results should be interpreted with caution: only 3 of the 10 jackknife subsamples yielded species with enough transfer events to pass the detection threshold (lowered to 20), with at most 2 species per run, making the jackknife median unreliable for this lineage.

Notably, the inter-run dispersion remained low across all lineages (Figure 59), with no obvious cluster of points corresponding to a few specific runs (except maybe for Colpodellida), which indicate that the jackknife shifts are reliable (Table 21). In summary, the core classification holds for 8 of 11 lineages, while the upward shift observed in Kryptoperidiniaceae, Colpodellida and Haptophyta reflects a sensitivity to donor composition (Figure 60).

4.1.1.2.5. *Conclusions of the robustness analyses*

Overall, across the seven pipeline variations tested, no variation produced an α -diversity pattern fundamentally incompatible with the reference classification (Figure 60). The four variations with stronger effects (maxint0, noage, bal_samp, jak_samp) each altered the signal in a specific and interpretable way, but rank concordance was preserved except for bal_samp, where changes in event taxonomic contexts account for most of the disruption. The consistency of the signal across analytically diverse settings supports the conclusion that the observed donor α -diversity differences among receiver lineages reflect genuine biological heterogeneity rather than methodological artefacts.

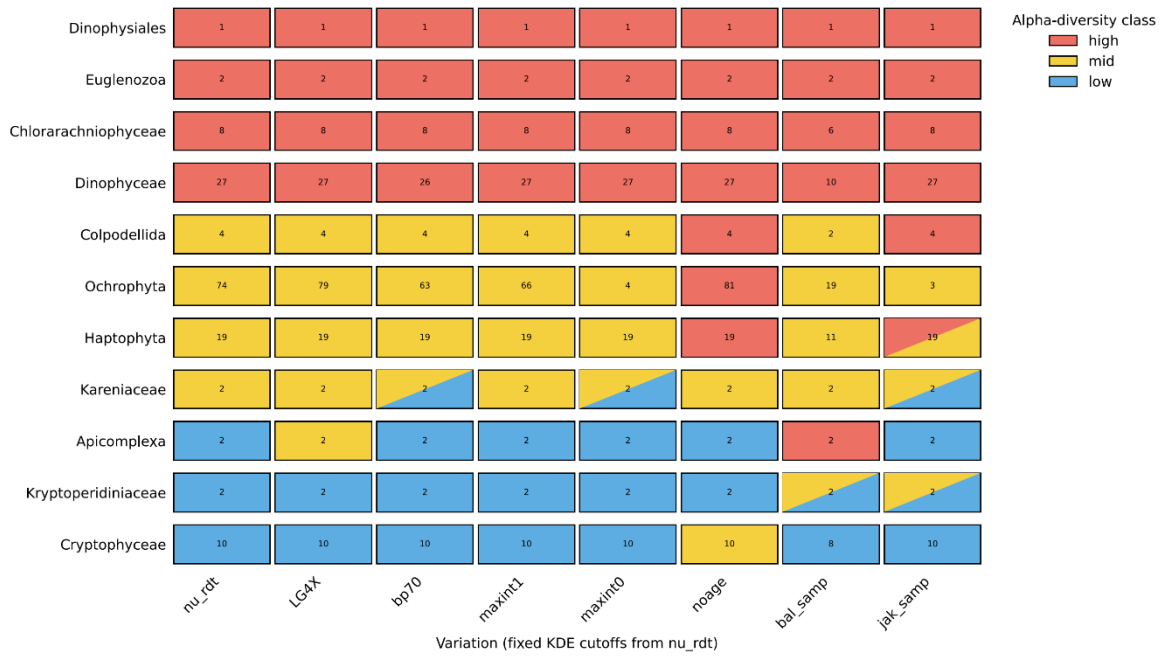


Figure 60: Lineage-level α -diversity classification across the seven pipeline variations (nu_rdt, LG4X, bp70, maxint1, maxint0, noage, bal_samp and jak_samp). The jackknife (jak_samp) corresponds to the 10 subsamplings summarised as one variation. Each cell summarizes the majority α -diversity class of species within a lineage using fixed KDE-derived cutoffs estimated once from the reference pipeline (nu_rdt) and then applied unchanged to all variations. Split cells indicate heterogeneous lineage composition, with the upper-left triangle marking the secondary class. Numbers within cells give the number of species contributing to each lineage $\times$ variation combination.

4.1.2. Does the signal depend on the choice of the α -diversity index?

The robustness analyses above all relied on Shannon-Wiener α -diversity. To assess whether the signal depends on the choice of index, we compared it with two alternatives that capture fundamentally different aspects of donor profiles (Figure 61, Figure 62, Figure 63): the reciprocal Simpson index ($1/D$), which reflects the effective number of dominant donors and down-weights rare contributors, and the Chao1 richness estimator, which extrapolates total donor richness from the frequency of singletons and doubletons (Jost, 2006). Shannon-Wiener integrates both richness and evenness, reciprocal Simpson is more strongly driven by dominance structure, whereas Chao1 ignores evenness and instead estimates the number of donor lineages that may remain undetected (Jost, 2006). The inclusion of Chao1 alongside reciprocal Simpson is deliberate: whereas reciprocal Simpson provides a genuine alternative to Shannon-Wiener (both are sensitive to dominance structure but weight it differently), Chao1 serves as a negative control because, in our view, it is not well suited for this type of data. In ecological surveys, Chao1 is well justified because singleton and doubleton counts arise from incompletely sampled but genuinely existing taxa (Chao et al., 2014). In the context of gene transfer inference, however, a singleton donor (a lineage contributing a single inferred event) may equally reflect a genuine rare interaction, a phylogenetic misattribution, or residual noise, making Chao1's extrapolations unreliable.

As a result, reciprocal Simpson and Shannon-Wiener yielded highly consistent patterns (Spearman $\rho = 0.973$; 94.9% exact bin matches; weighted $\kappa = 0.923$; Figure 61), with all discrepancies limited to adjacent-bin shifts (Figure 62). The only notable difference was an additional extreme high-diversity bin under reciprocal Simpson, exclusively occupied by *Dinophysis acuminata*. This near-perfect concordance between two indices that weight the frequency distribution differently implies that the α -diversity signal is primarily carried by the dominance structure of donor profiles, i.e., by the identity and relative contribution of the major donors, rather than by rare contributors.

As expected, Chao1 produced a markedly different result, the distribution is compressed with most taxa collapsing into low or intermediate bins (Spearman $\rho = 0.713$; 13.4% exact bin matches; weighted $\kappa = 0.258$; Figure 61, Figure 63). The instability this introduces is illustrated by the two Kryptoperidiniaceae species: *Kryptoperidinium foliaceum* (nine donor lineages, four singletons) was assigned to the highest Chao1 bin, whereas *Durinskia baltica* (six donor lineages, no singleton) clustered much lower, despite both species being classified similarly under Shannon-Wiener and reciprocal Simpson indices, where their dominance structures are comparable.

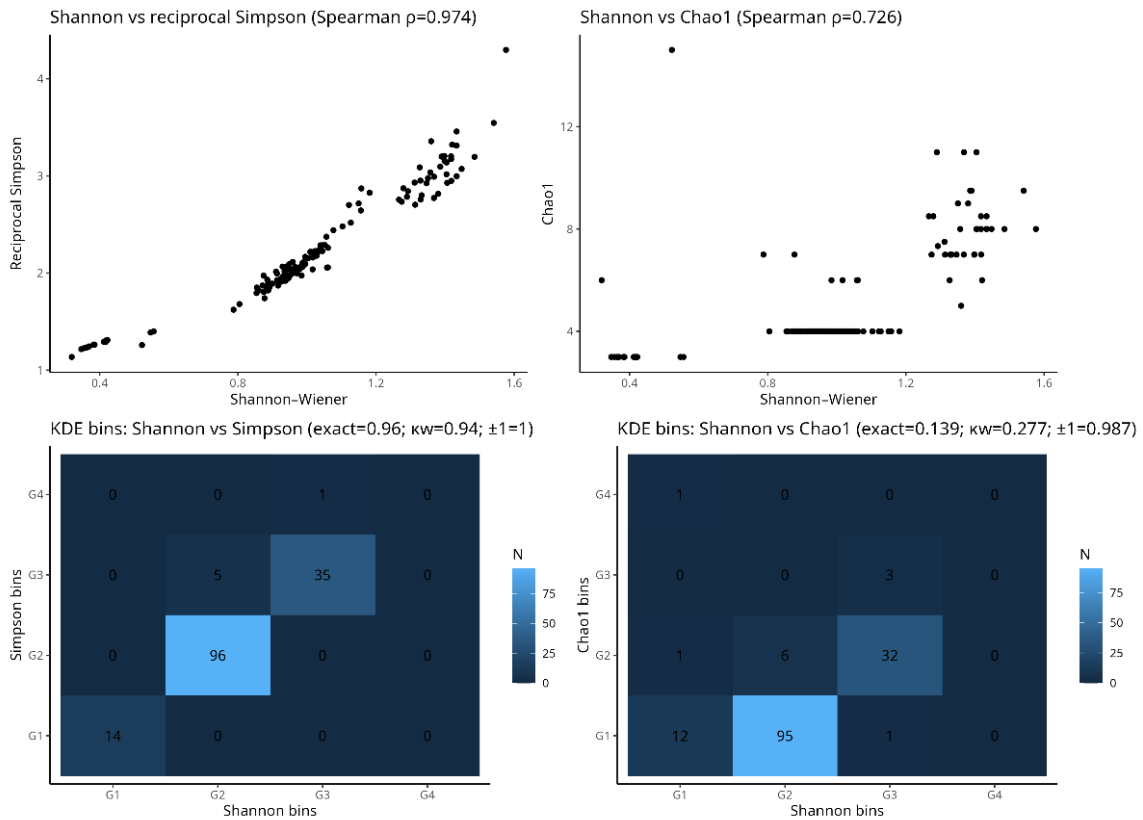


Figure 61: **Robustness of donor α -diversity patterns to index choice.** Comparison of species-level donor α -diversity values computed with Shannon-Wiener entropy against reciprocal Simpson (left) and Chao1 richness (right). Top panels show pairwise relationships across species. Bottom panels compare the corresponding KDE-based bin assignments (4 bins).

Together, these results show that the α -diversity signal is robust to index choice among indices sensitive to dominance structure. The donor diversity differences among lineages, as captured by Shannon-Wiener, therefore reflect differences in the number and relative importance of recurrent donor lineages rather than representing total donor richness over evolutionary time. Shannon-Wiener α -diversity was retained as the primary index because it captures precisely the component of donor diversity that the present pipeline can measure with confidence. Indeed, Shannon-Wiener entropy integrates the full donor distribution, capturing both richness and evenness, whereas the reciprocal Simpson index is dominated by the most frequent donors and relatively insensitive to rare contributors (Jost, 2006). Because rare donor lineages may carry biologically meaningful signal, reflecting occasional or transient interactions, Shannon-Wiener provides a better compromise for characterizing the overall structure of donor profiles.

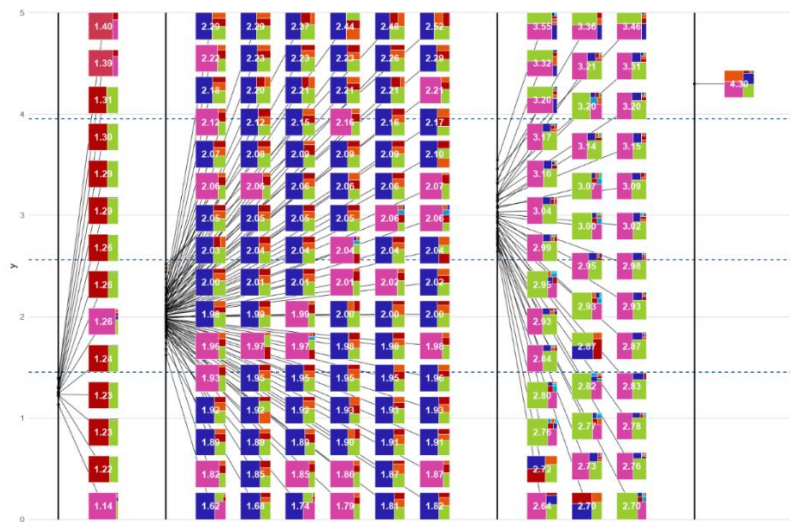
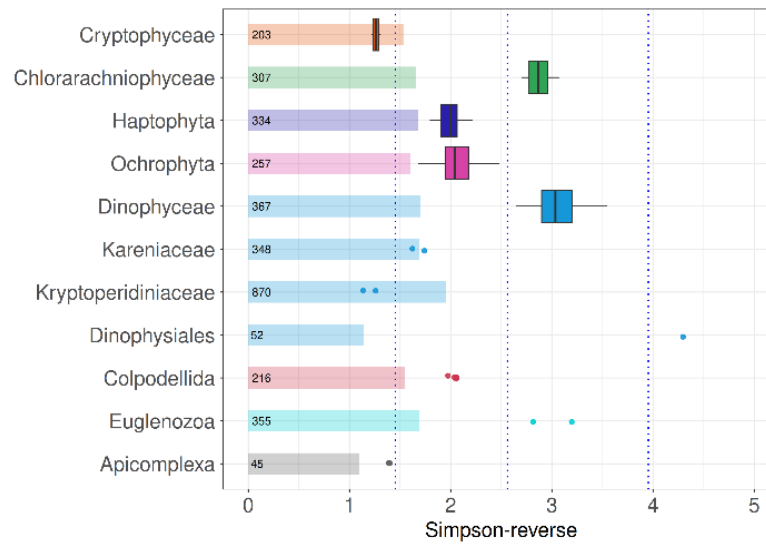
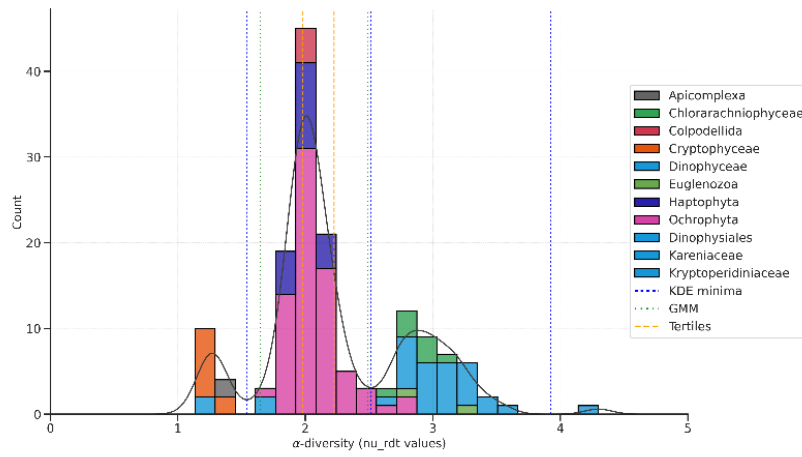


Figure 62: Distribution and binning of individual α -diversity values according to the reverse Simpson index. Compared to Shannon-Wiener distribution, a single species (*D. acuminata*) occupies an additional high-diversity bin.

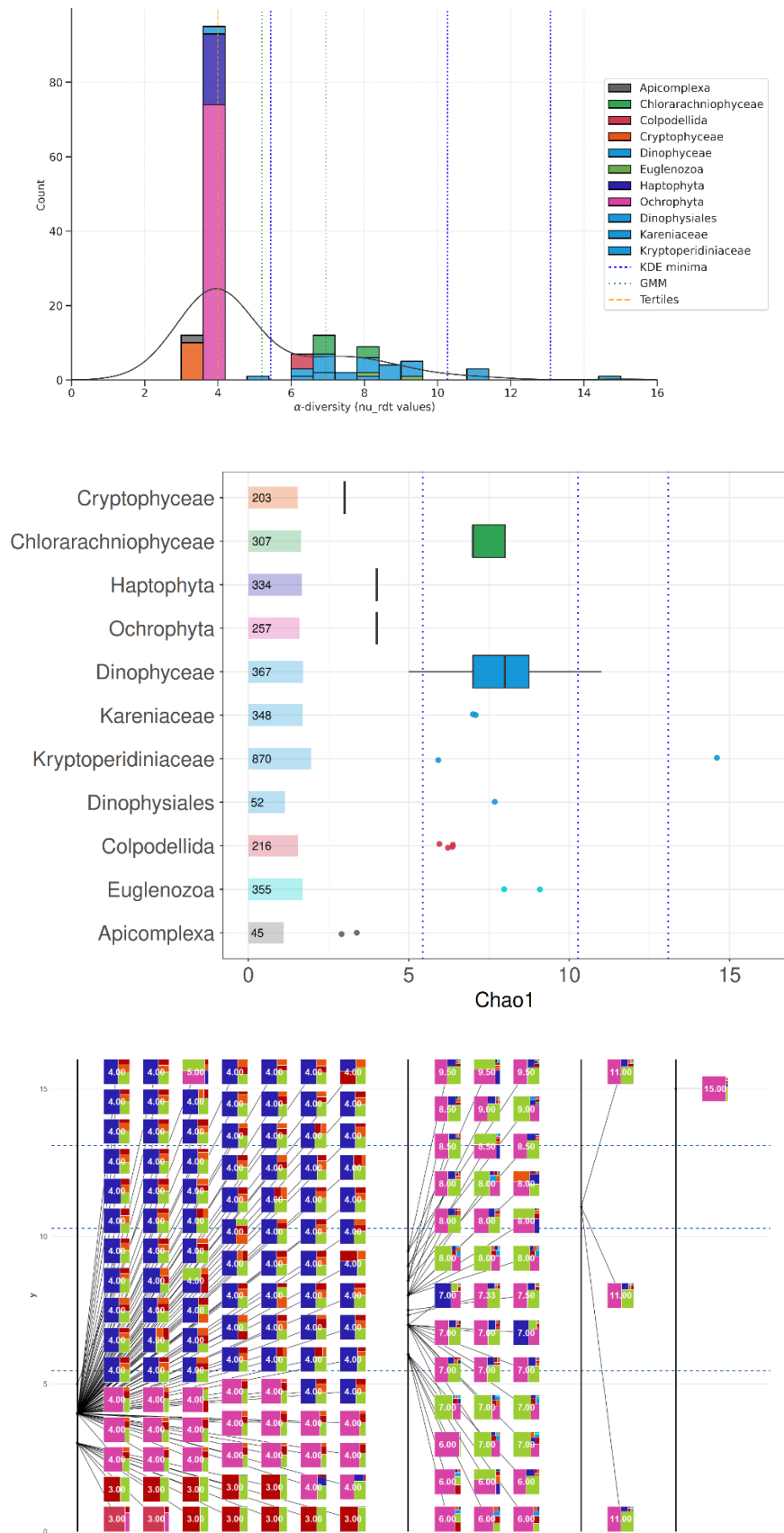


Figure 63: Distribution and binning of individual α -diversity values according to the Chao1 index. Compared to Shannon-Wiener and reciprocal Simpson indices, species are not well distributed among diversity bins; instead, they almost all fall in the low and intermediate bins.

4.1.3. Expert validation of transfer inferences

The pipeline variations above assess internal consistency. To provide an independent, external check, another professional phylogeneticist, expert in gene transfers in the context of complex algae (Dr. Henner Brinkmann), manually inspected all rhodophyte and Viridiplantae donor events inferred for Cryptophyceae, a lineage chosen because of its ancient plastid (age class 2) and well-characterised red algal endosymbiosis that make it both informative and challenging. For each event, the gene tree was examined and assessed on whether the topology supports the pipeline's donor attribution ("yes") or not ("no") across three levels of intruder tolerance (Table 22).

Origin	Yes(max2)	No(max2)	Yes(max1)	No(max1)	Yes(max0)	No(max0)
<i>Rhodophyta</i>	113	10	109	6	91	4
<i>Viridiplantae</i>	53	29	47	14	37	4

Table 22: Expert validation of donor inference in Cryptophyceae for rhodophyte- and Viridiplantae-derived events across three intruder-tolerance settings. Values indicate the number of events whose donor attribution was judged as supported (Yes) or unsupported (No) after manual inspection of the corresponding gene trees.

The results show a clear trend. For rhodophyte donor events, concordance was already high under the reference setting (113/123 endorsed, 91.9%) and rose to 109/115 (94.8%) under maxint1 and 91/95 (95.8%) under maxint0. The number of rejected events dropped from 10/123 (8.1%) to 4/95 (4.2%), at a cost of losing only 22 endorsed events (approximately 19%). The dominant red signal that defines the Cryptophyceae low-diversity profile is thus overwhelmingly supported by expert inspection at all stringency levels. The Viridiplantae donor events are more revealing. Under the settings of the reference pipeline, only 53/82 inferences were endorsed (64.6%). Under maxint1, concordance rose to 47/61 (77.0%), and under maxint0 to 37/41 (90.2%). In parallel, the rejection rate decreased from 29/82 (35.4%) to 4/41 (9.8%).

The expert annotations from Dr. Henner Brinkmann about each tree clarify why the two categories differ. Rejected rhodophyte events typically involve trees where cryptophyte sequences fall within a broad secondary-red cluster without forming a specific relationship with Rhodophyta sensu stricto, or cases where nucleomorph-encoded paralogues of *Guillardia theta* create misleading placements. Rejected Viridiplantae events show a distinct pattern: events were systematically rejected where cryptophytes were merely included in the entire green clade (presence of tolerated intruders); those where cryptophytes were nested within a specific green subclade were endorsed. The endorsed green events show placements sister to prasinophyte lineages (*Pyramimonas*, *Nephroselmis*, *Prasinoderma*, *Micromonas*); the rejected events reflect unresolved deep divergences, likely caused by missing bacterial donors (see Discussion; Limitations & perspectives; Section 5) whose sequences default to the nearest eukaryotic paraphyletic assemblage.

The transfer-inference algorithm itself is not at fault in these cases: it faithfully transcribes the tree topology; yet the topology is sometime artefactual. By reducing intruder tolerance, the algorithm masks out the topologically complex trees where these artefacts concentrate, retaining events supported by clean subclade placements. This confirms that the factors driving rejected inferences (ancient age, missing donors, and here nucleomorph paralogy) are the same factors identified by the robustness analysis (Section 4.1.1); the intruder tolerance parameter provides an effective quality filter by maximizing endorsed events and minimizing non-endorsed events. Nonetheless, this contrasts with the behaviour observed for Ochrophyta (Section 4.1.1), where stricter intruder tolerance has the opposite effect: it eliminates a large fraction of genuine events embedded in topologically complex trees, rather than preferentially removing artefactual ones. The fact that the same parameter setting improves concordance for Cryptophyceae while eroding signal for Ochrophyta illustrates the difficulty of tuning the pipeline in a way that performs optimally across all lineages in a large-scale analysis.

4.2. What does donor α -diversity actually measure?

4.2.1. Confounding factors

Two properties of receiver species could in principle inflate α -diversity estimates: genome size and taxonomic representation in the dataset.

Genome size

Species with larger gene repertoires present more sequences to the pipeline, which can mechanically inflate both event counts and donor diversity. At the species level, genome size does correlate with the number of inferred events (Spearman $\rho = 0.513$, $p = 2.14 \times 10^{-12}$; Figure 64), and the correlation persists at the lineage level ($\rho = 0.65$, $p = 0.032$; Figure 65).

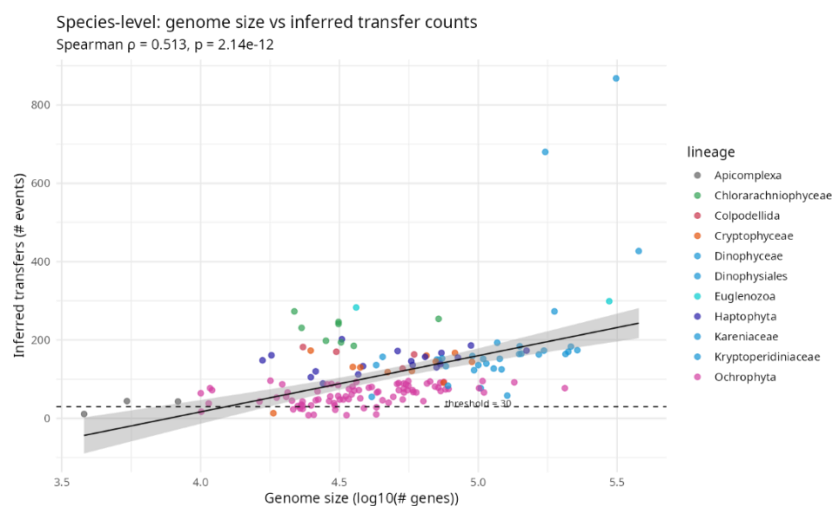


Figure 64: Species-level relationship between genome size and inferred transfer counts.

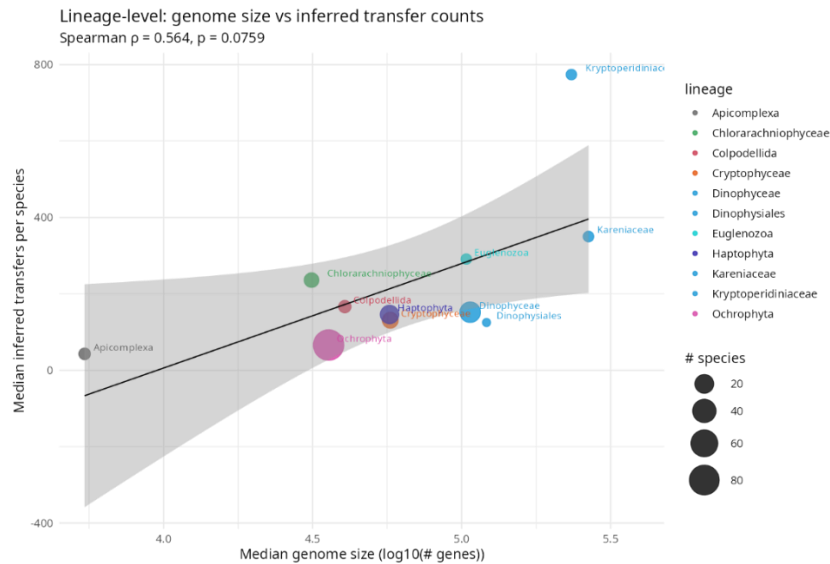


Figure 65: Lineage-level relationship between median genome size and median inferred transfer counts.

However, this correlation does not extend to α -diversity: at the species level, the relationship between genome size and Shannon-Wiener α -diversity is weak ($\rho = 0.197$, $p = 0.011$; Figure 66), and at the lineage level it vanishes entirely ($\rho = -0.009$, $p = 0.989$; Figure 67). The dissociation is both clear and reassuring: larger genomes yield more events, but those additional events are distributed among the same dominant donors rather than altering the donor profile. In other words, increasing genome size inflates the volume of the signal but not its composition.



Figure 66: Species-level relationship between genome size and inferred α -diversity.

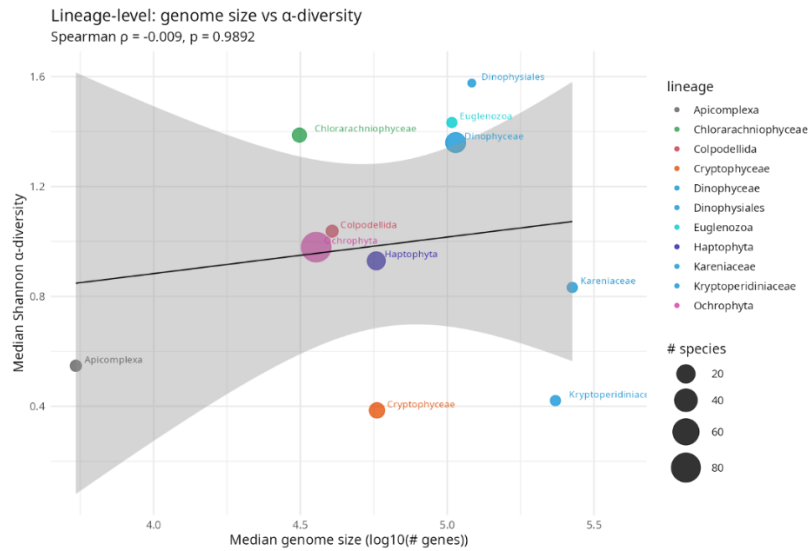


Figure 67: Lineage-level relationship between median genome size and median α -diversity.

Species representation

Lineages with more species in the dataset might appear to receive more diverse transfers simply because denser sampling would increase the odds of detection. Again, reassuringly, this is not the case: no significant correlation was found between the number of species per lineage and median α -diversity (Spearman $\rho = -0.500$, $p = 0.267$; Figure 68). Ochrophyta, the most densely sampled lineage (85 species), has among the lowest α -diversity values, while Euglenozoa (2 species) and Chlorarachniophyceae (8 species) rank among the highest in terms of α -diversity values.

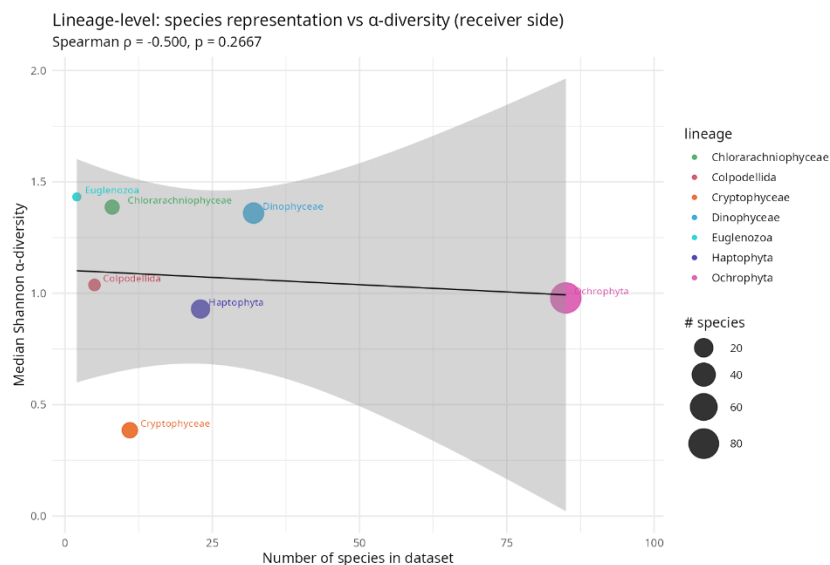


Figure 68: Relationship between lineage sampling density and median donor α -diversity across major complex algal lineages.

Donor-side bias

A complementary question is whether donor lineages that are more heavily represented in the dataset appear to donate more events simply because their sequences are more likely to form paraphyletic assemblages in gene trees, even if spurious. In this case, the correlation between the number of species per donor lineage and the number of donated events is strong (Spearman $\rho = 0.79$, $p = 0.007$; Figure 69), and it remains significant even after excluding Ochrophyta ($\rho = 0.71$, $p = 0.034$; Figure 70).

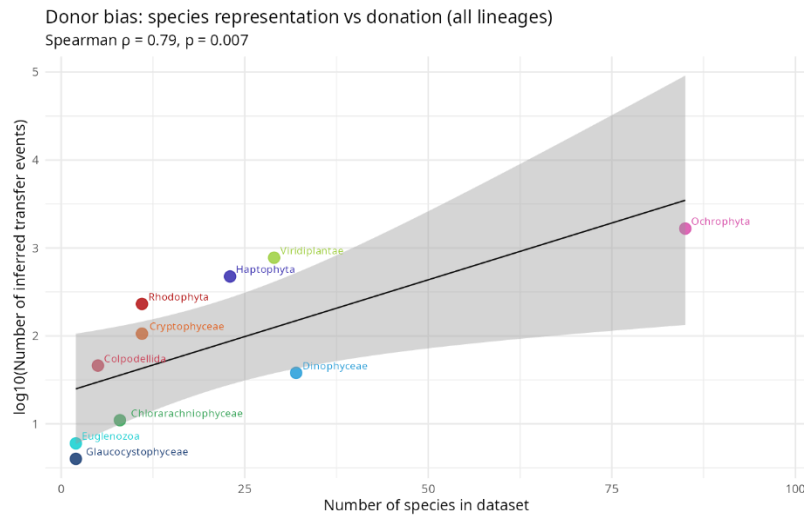


Figure 69: Relationship between lineage sampling density and donor richness across major plastid-bearing lineages.

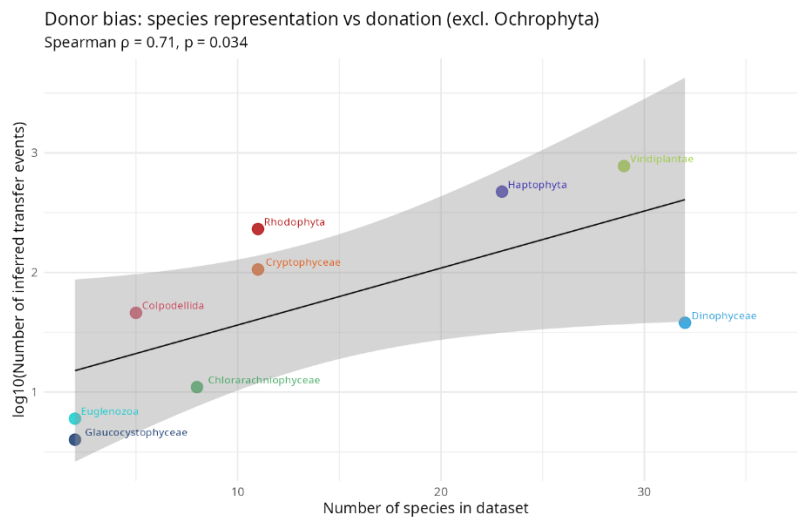


Figure 70: Relationship between lineage sampling density and donor richness across major plastid-bearing lineages (without Ochrophyta).

This correlation admits several non-exclusive interpretations. On the one hand, denser sampling of a donor lineage increases the probability that its species will appear in paraphyletic positions in gene trees, mechanically inflating the number of inferred donations, i.e., a sampling artefact. Moreover, the

most species-rich lineages in the dataset (Viridiplantae, Ochrophyta, Rhodophyta, Haptophyta) belong to the oldest age classes (classes 1 and 2), making them compatible as major donors. On the other hand, Dinophyceae are the second most species-rich group in the dataset (32 species) but from a higher age class (class 4). This mechanically limits the number of compatible receivers and may partly explain that it ranks comparatively low as a donor. Furthermore, plastid-targeted gene repertoire of Dinophyceae is already chimeric, both as documented by studies on plastid acquisition in dinoflagellates (Hongo et al., 2019; Novák Vanclová et al., 2024) and as reflected in our own data, where peridinin dinoflagellates display the highest donor α -diversity in the dataset ($\alpha = 1.50$, Table 8). This means that dinoflagellate sequences are less likely to form coherent donor clades in gene trees. Together, these observations suggest that the correlation between donor taxonomic sampling and event count reflects a combination of sampling density, age-class structure, and lineage-specific evolutionary history, none of which can be fully disentangled in the present dataset. To partially disentangle these factors, we used the balanced-sampling variation (bal_samp). By primarily pruning the most oversampled lineages (Ochrophyta from 85 to 21 species, Dinophyceae from 32 to 18, Viridiplantae from 29 to 15; Table 26), we reduced the contribution of sampling imbalance while altering tree topologies and event contexts. As discussed in the robustness section, ‘bal_samp’ is the only variation that disrupted rank concordance, and its main effect on Ochrophyta was to redistribute the donor profile (Haptophyta contributions decreased while Rhodophyta increased) rather than to change the overall volume of events (Section 5.2). This indicates that donor-side sampling influences “which” donors are resolved more than “how many” events are detected. However, most lineage classifications were preserved (Figure 60), suggesting that the biological component of the α -diversity signal is robust enough to survive subsampling.

5. Methods

5.1. Orthoclan enrichment

The first step of this section was to isolate plastid-targeted genes out of the 39,784 “photosynthetic” orthoclans obtained at the end of the previous section (Van Vlierberghe et al., 2021b) (Chapter II) using the tagged sequence identifiers relative to their encoding genome and cellular localization. As a result, a set of 1,788 plastid-related genes was defined and is described in Figure 24

Forty-Two was used in compress mode (see Chapter 2; section 2.3.4) and in PARA mode (see Chapter 2; section 2.3.3). For the PARA mode, Forty-Two requires files composed of representative sequences of unwanted out-paralogues. Such PARA files were generated when pruning photosynthetic orthoclans in Chapter III; for each bipartition (or split) generated when isolating a photosynthetic orthoclan from

a phylogenetic tree, two clans were generated: (i) one clan that was the actual photosynthetic orthoclan and (ii) the other clan, whose paralogous sequences populated the PARA file.

5.2. Phylogenetic inference

Enriched clans were aligned with MAFFT v7.453 (Katoh et al., 2002; Katoh and Standley, 2013) and used to compute single-gene trees under a LG+C20+G model with IQ-TREE v1.6.12 (Nguyen et al., 2015). The resulting unrooted phylogenetic trees were automatically rooted by applying a non-reversible model approach (NONREV) with their extracted (LG+C20+G) unrooted topology as input, using the capabilities of IQ-TREE v2.3.4 (Minh et al., 2020).

5.3. Gene transfer inference

5.3.1. Core principles

One of the main programmatic challenges of this project was to develop a script (PAGTI) able to perform gene transfer inference in a discovery-driven fashion and with a tolerance for complex tree topologies. Before proceeding with the description of PAGTI, here is bit of vocabulary used in the following parts of the manuscript. In a phylogenetic tree, a clade (monophyletic group) denotes an ancestor and all its descendants (Baum and Offner, 2008). Because clade terminology is formally rooted, Wilkinson et al. proposed the term clan as the unrooted analogue of a clade (Wilkinson et al., 2007), i.e., a set of taxa that can be recovered as one side of a split under an appropriate rooting. In an unrooted tree, each internal branch induces a bipartition (or split) of the taxa, obtained by removing that branch and separating the leaf set into two disjoint subsets, i.e., two clans (Wilkinson et al., 2007). Even though single-gene trees analysed in this study are rooted, the term clan has been preferred, both because they are the product of an initial pruning operation performed on the orthogroups obtained in Chapter III to isolate photosynthetic orthoclans and because the implemented algorithm will analyse one side of a bipartition at a time.

The parser has been written in Perl and relies mainly on the Bio::MUST::Core suite developed and maintained in our laboratory, which depends in part on the Bio::Phylo module (Vos et al., 2011). It takes as input the ‘.splits.nex’ file containing support values in percentage for all splits (bipartitions) and the ‘.treefile’ (the corresponding ML tree in NEWICK format) from IQ-TREE (Nguyen et al., 2015). The ML topology is used to filter out incompatible bipartitions, retaining only splits consistent with the ML tree for downstream analyses. All sequences in the tree are processed in turn, except those from Archaeplastida, for which we are not trying to infer the origins of plastid-targeted genes.

Starting from a leaf in the tree, the parser identifies all clans containing that sequence and sorts them by number of sequences in ascending order. Clans are processed in that order as long as the previous clan is included in the current one to let the subtree grow. This way, we can analyse the taxonomic neighbourhood of a given leaf step by step.

The first step is to identify the potential receiver group, i.e., the largest monophyletic clan containing the starting sequence and a minimum of two distinct organisms (including the starting sequence) to make inference more robust. We allowed some exceptions for specific taxonomic groups (Dinophysiales, Kareniaceae, Kryptoperidiniaceae and Euglenophyceae) given their lower sampling (1, 2, 2, 2 respectively). In these cases, only one organism is enough to infer a gene transfer. Then follows a complex analysis of the growing subtree containing the receiver clan where sequences are classified according to their taxonomic group. A given taxonomic group can be considered as a donor when receiver sequences form a monophyletic receiver clan nested within a paraphyletic assemblage of sequences assigned to the donor lineage, as shown in Figure 71a. This topological pattern is considered as a reliable signature of gene transfer, more stringent than a simple sister-group relationship.

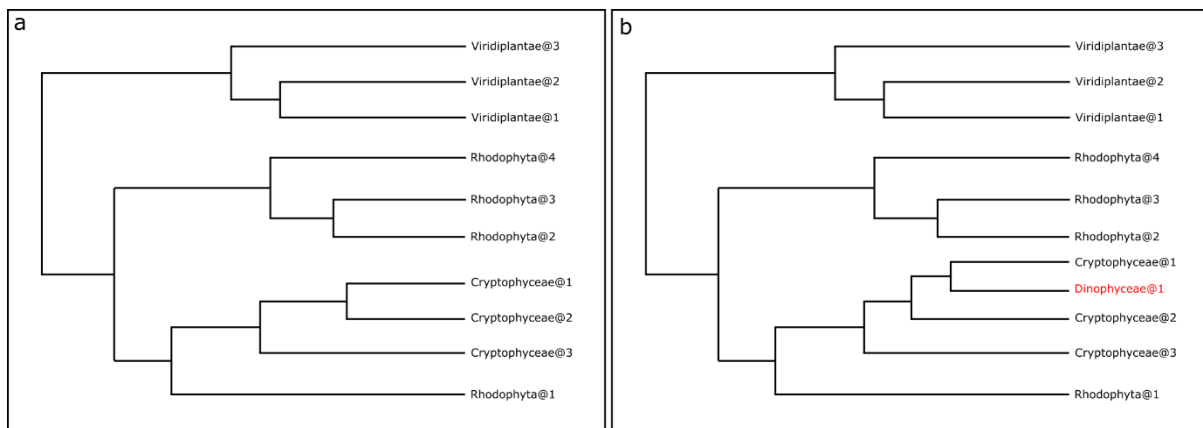


Figure 71: a) Typical case of paraphyletic relationship to infer a gene transfer; here *Cryptophyceae* form a monophyletic group nested inside a clan of *Rhodophyta*. b) Same topological arrangement as a) but with an additional sequence (*Dinophyceae*@1, in red) considered as an intruder and ignored for inference. Misplaced sequences regularly occur in the trees because of either low sequence quality, residual data contamination or phylogenetic artifacts such as LBA.

5.3.2. Complex topologies

Some topologies can break our algorithm flow and prevent the identification of the receiver clan. Hence, heuristics have been developed to address those complex cases where “intruder” organisms impair the simple transfer inference, as described in Figure 71a. We distinguish three cases: (i) single misplaced sequence (Figure 71b), (ii) dual-group gene transfer, i.e., a gene transfer from one lineage to two sister lineages (Figure 72a) and (iii) nested gene transfer, i.e., two successive gene transfers involving three taxonomic groups (Figure 72b). Although the latter resembles the “single intruder” topology, the

presence of several species of the same lineage cannot be dismissed on the account of artefacts or contaminations since the transfer is shared by several organisms of the same group.

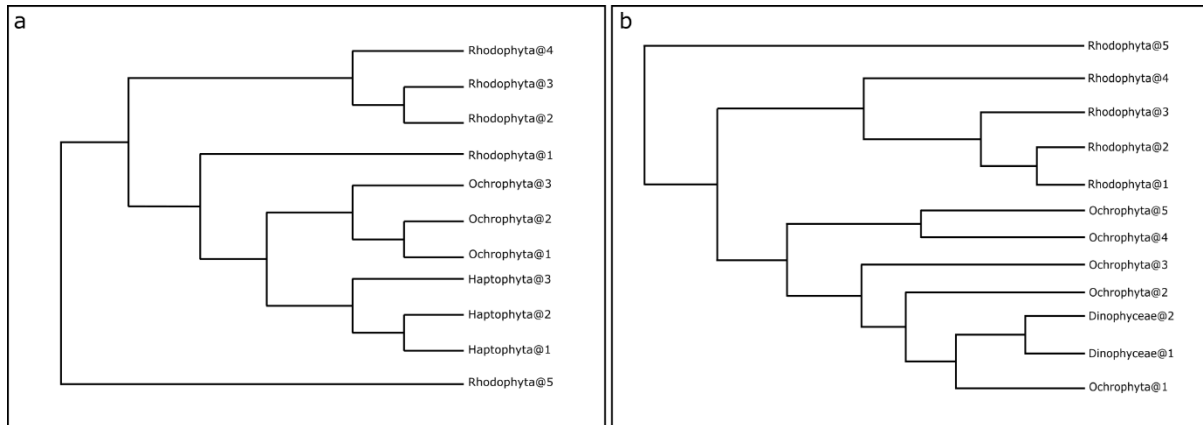


Figure 72: a) **Dual-group gene transfer:** Ochrophyta and Haptophyta display a sister lineage relationship while the whole group is inside Rhodophyta; this will be interpreted as two transfers: one from Rhodophyta to Ochrophyta and one from Rhodophyta to Haptophyta. b) **Nested gene transfer:** Using the base inference rules, the topology supports a transfer from **Ochrophyta to Dinophyceae**. Because Dinophyceae sequences are nested within the ochrophytes, they are not strictly monophyletic, which would normally prevent their assignment as receivers. Our algorithm, though, explicitly handles such nested configurations by treating Ochrophyta as the receiver clan despite the presence of “intruders”, thereby recovering a second, earlier transfer from Rhodophyta to Ochrophyta.

5.3.3. Early vs late transfers: transfer compatibility rules

The focus of this study is on “early” transfers, i.e., transfers plausibly associated with the establishment of the current plastid in a major photosynthetic lineage. Inevitably, “late” transfers, specific to smaller taxa, are also detected. Therefore, we implemented a safeguard informed by molecular-clock evidence. Strassert et al., 2021 (Strassert et al., 2021) provide a time-calibrated framework to evaluate whether serial endosymbiosis scenarios for complex red plastids are chronologically compatible (Figure 73), i.e., whether the relevant host stem lineages overlap during the intervals in which these plastids could have spread across distantly related eukaryotes. We used this framework to classify plastid acquisition events into five relative age classes with the aim to implement transfer-compatibility rules. Class 1 corresponds to Archaeplastida (primary plastids); Class 2 represents early red secondary plastid hosts (cryptophytes); Class 3 includes major red complex plastid lineages (notably ochrophytes and haptophytes); Class 4 groups other red-plastid derived lineages (peridinin dinoflagellates, colpodellids, and apicomplexans). Finally, Class 5 captures the most recent plastid acquisitions, including plastid replacements in Kareniaceae and Kryptoperidiniaceae, as well as transient plastids in kleptoplastidic dinoflagellates such as Dinophysiales. We also placed secondary green plastid lineages (euglenids and Chlorarachniophyceae) in Class 4, based on the dated eukaryotic framework of Strassert et al., 2021. Transfers are allowed from older to younger classes and within a class, providing a conservative safeguard against transfers that would be rather suggestive of later events, unrelated to plastid establishment.

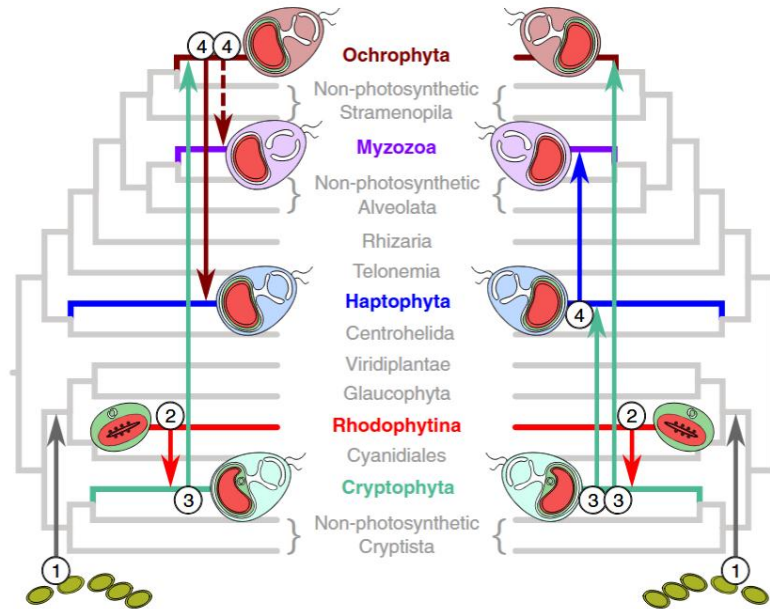


Figure 73: Topologies of compatible models for red-algal plastid origins as proposed by Stiller et al., 2014 (left) and Bodil, 2018 (right). Figure from Strassert et al., 2021.

Thus, the algorithm flow can be schematized as follows:

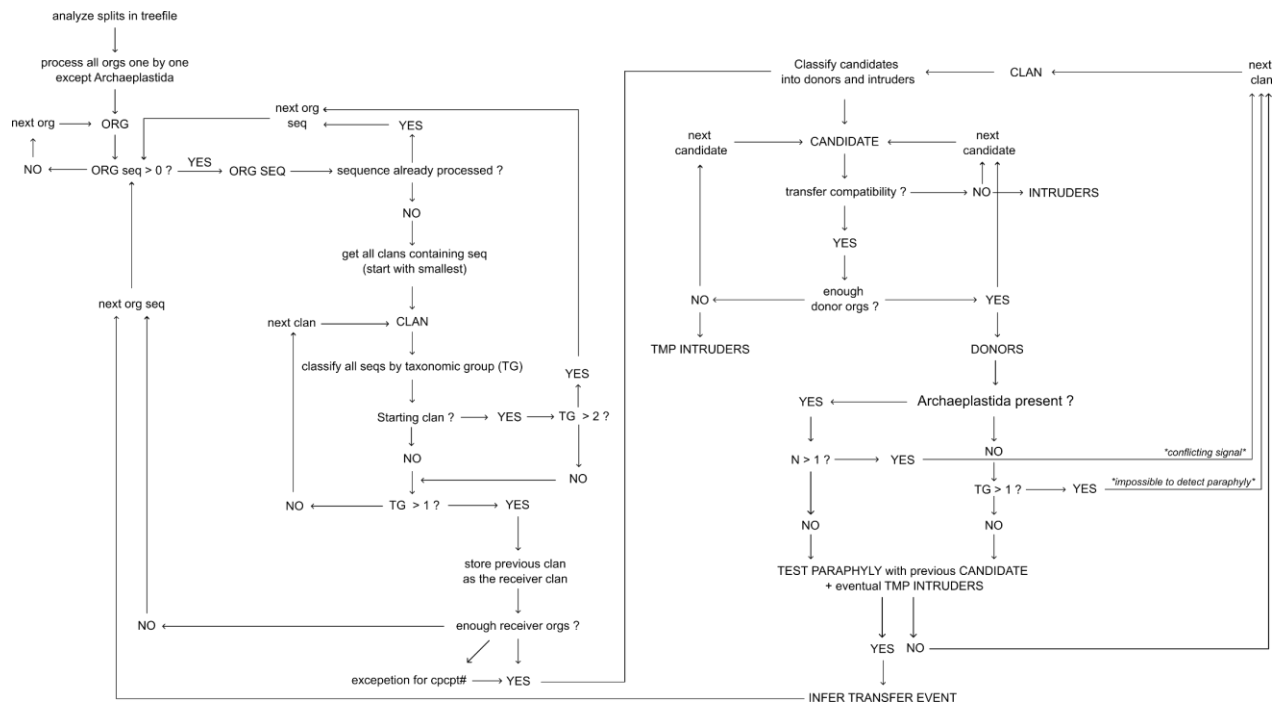


Figure 74: Flowchart of the PAGTI gene transfer inference algorithm.

At the end of the run, the parser generates a transfer report for each tree, precisely describing inferred transfer events. Each transfer event is labelled with a 'transfer id' because each event spans multiple lines, one for each of the receiver sequences; the structure is so for the script that will generate the per-species data.

5.4.Event classification

Transfer events are classified based on the tags carried by the donor and receiver sequences. Three levels of evidence were used in a hierarchical manner: (i) tags on receiver sequences, (ii) tags on donor sequences, and (iii) tags present elsewhere in the orthoclan, consulted only when both donor and receiver sequences are tagless.

5.4.1. Case 1: Tagged receiver (no ambiguity)

When at least one of the receiver sequences carries a tag (`nucpt#`, `nuppct#`, `nm#`, or `cpcpt#`), the event is considered unambiguous and classified directly according to the receiver tag.

5.4.2. Case 2: Tagless receiver, tagged donor

When receiver sequences are tagless but at least one of the donor sequences is tagged, we propagate the donor tag to the event (i.e., the receiver is treated as inheriting the donor status for classification) unless the donor is tagged `cpcpt#` (see “Case 4” below).

5.4.3. Case 3: Neither receiver nor donor is tagged

When neither receivers nor donors carry a tag, the decision relies on tags present elsewhere in the orthoclan:

- If the orthoclan contains only `cpcpt#` tags, then events from that orthoclan are excluded, because this configuration is dominated by plastid-encoded genes and tagless event classification is too uncertain.
- If the orthoclan contains a mixture of tags (e.g., `nucpt#` and `cpcpt#`) but the donor/receiver pair is tagless, we assign priority to non-`cpcpt#` tags to classify events. This is a conservative choice: if a plastid-encoded scenario truly occurred for the sequences involved, one would generally expect to observe at least a fraction of donors explicitly tagged as `cpcpt#`.

5.4.4. Case 4: Tagless receiver with a `cpcpt#` donor

The most problematic configuration corresponds to events where at least one of the donor sequences is `cpcpt#` and receiver sequences are tagless. In this context, it is not possible to robustly conclude whether the receiver is plastid-encoded (but undetected as such) or nucleus-encoded plastid-targeted. Therefore, these events are excluded from the main results.

5.5. α -diversity indices and weights

For each receiver species, we summarize donor-lineage diversity (i.e., how inferred transfer events are distributed across donor lineages) using standard α -diversity indices computed from donor relative contributions (p_i). Here, (p_i) is the relative contribution of donor lineage (i) (*weighted* counts, see below for weight computation), with ($\sum_i p_i = 1$). We used the Perl Bio::Community::Alpha module (Angly et al., 2014) to compute weighted α -diversity indices.

5.5.1. Shannon-Wiener (Shannon entropy)

$$H' = - \sum_{i=1}^S p_i \ln(p_i)$$

where (S) is the number of donor lineages observed. Shannon-Wiener captures both richness (how many donors) and evenness (how balanced their contributions are). It increases when (i) more donor lineages contribute and/or (ii) contributions become more evenly distributed.

5.5.2. Simpson (“reverse Simpson”)

$$D = \sum_{i=1}^S p_i^2$$

and the reverse form is:

$$D^{-1} = 1/D$$

The Simpson reverse index is strongly driven by dominance (large (p_i) values): it drops significantly when one (or a few) donors dominate, and it is less affected by very rare donors than Shannon-Wiener’s index.

5.5.3. Chao1 (richness estimator)

Chao1 estimates the total donor richness by correcting for incomplete sampling using rare classes:

$$\hat{S}_{(\text{Chao1})} = S_{(\text{obs})} + F_1^2/(2F_2)$$

where (S_{obs}) is the number of observed donors, F_1^2 the number of donors observed once (singletons), and (F_2) the number observed twice (doubletons).

In this study, “donor diversity” is the diversity of donor lineages contributing inferred gene transfers; “abundance” is the relative (weighted) contribution of each donor lineage. Shannon-Wiener provides a balanced summary (richness and evenness), reciprocal Simpson emphasizes dominance (few major

donors VS many), and Chao1 focuses on richness driven by rare donors. Because all these indices can be biased when the number of inferred events per species is low (especially richness), we applied a minimum-event threshold prior to computing donor α -diversity, as described in the next section.

5.5.4. Minimum species threshold

To ensure that Shannon-Wiener α -diversity estimates were computed from sufficiently sampled donor profiles, we assessed sampling adequacy using Good's sample coverage (Good, 1953) within a coverage-based rarefaction/extrapolation framework (Chao et al., 2014). For each species, donor lineages were treated as categories and transfer events as observations. We computed:

$$\hat{C} = 1 - \frac{f_1}{n}$$

where (n) is the total number of inferred transfer events for the species (sum of counts across donor lineages), and (f_1) is the number of singleton donor lineages (donors observed exactly once, i.e., donor count = 1). Intuitively, ($\hat{C}$) estimates the probability that an additional observation would belong to a donor lineage already observed, and thus it quantifies sampling completeness.

We then examined how often species reached high sampling completeness (coverage ≥ 0.9) across event-count strata (10-event bins). In our reference pipeline (nu_rdt), coverage was consistently ≥ 0.9 from 20 events onward (Table 23), indicating that Shannon-Wiener estimates are expected to be comparatively stable above this range. Nevertheless, to further limit residual small-sample effects and retain a conservative filter, we used a minimum threshold of 30 inferred transfer events per species for downstream α -diversity comparisons.

<i>Event count per species (n)</i>	<i>Proportion with coverage ≥ 0.9</i>
0-10	0.00
10-20	0.67
20-30	1.00
≥ 30	1.00

Table 23: Proportion of species reaching high sampling completeness (Good's coverage ≥ 0.9) across bins of inferred transfer event counts (n).

5.5.5. Lineage-specific weights

To compute weighted α -diversity for nucpt# genes across donor lineages, we combined three factors (Table 24): (i) the number of plastid-targeted genes (according to the literature), (ii) the average number of sequences per donor lineage per tree, and (iii) the average genome size (i.e., the number of protein-coding genes) per donor lineage. This approach balances contributions from each factor, reducing bias from extreme values. When combining these factors, we used average normalized weights to ensure

equal consideration of all factors, a common practice in diversity studies (Chao et al., 2010; Jost, 2007). This method provides a diversity metric preventing any single factor from dominating the calculation.

We normalized each parameter using min-max scaling, so all values range between 0 and 1:

$$\text{Norm}(x_i) = (x_i - \min(x)) / (\max(x) - \min(x))$$

Where:

- x_i = value for lineage i (e.g., number of plastid-targeted genes)
- $\min(x), \max(x)$ = minimum and maximum values across all lineages

The weight for each lineage is calculated as the average of the normalized values for the three chosen factors, where:

$$w_i = ((\text{Norm}(G_i) + \text{Norm}(S_i) + \text{Norm}(Z_i))) / 3$$

- G_i = number of plastid-targeted genes for lineage i
- S_i = average number of sequences per tree for lineage i
- Z_i = average genome size (number of genes) for lineage i

Lineage	G_i	S_i	Z_i
<i>Apicomplexa</i>	344	1.8	5841.7
<i>Chlorarachniophyceae</i>	395	3.1	21755.0
<i>Colpodellida</i>	263	2.5	27099.0
<i>Cryptophyceae</i>	628	3.6	21601.5
<i>Dinophyceae</i>	263	7.8	44066.0
<i>Euglenozoa</i>	390	1.4	17932.2
<i>Glaucocystophyceae</i>	274	1.3	24488.0
<i>Haptophyta</i>	167	5.6	22324.0
<i>Ochromphyta</i>	265	19.4	13762.9
<i>Rhodophyta</i>	307	4.5	8788.4
<i>Viridiplantae</i>	895	9.6	15588.6

Table 24: *Chlorarachniophyceae* and *Cryptophyceae* (Curtis et al., 2012) *Viridiplantae* (Terashima et al., 2010).

To avoid zero weights, a small constant ϵ (1% of the smallest non-zero normalized value) is added to each normalized value before averaging:

$$w_i = ((\text{Norm}(G_i) + \epsilon) + (\text{Norm}(S_i) + \epsilon) + (\text{Norm}(Z_i) + \epsilon)) / 3$$

5.6.KDE-based classification of α -diversity

Density estimates were computed using a Gaussian kernel with a bandwidth equivalent to R's default *bw.nrd0*. Class boundaries were defined by the local minima of the estimated density, thereby identifying natural separations in the distribution of α -diversity values. These cutoffs were derived from the global distribution of α -diversity across all species, ensuring that classes reflect intrinsic structure in the dataset rather than lineage-specific thresholds. Each species was then assigned to a category based on its α -diversity value.

5.7.KEGG module assignation

Functional annotations (KEGG Orthology identifiers or KO-IDs) were attached to tree leaves by sequence homology with *eggNOG-mapper* v2.1.12 (Cantalapiedra et al., 2021) (Chapter 4, Section 5.1.10). Those KO-IDs were then used to assign KEGG modules and KEGG pathways to specific sets of inferred transfer events, allowing us to compare these event sets functionally. KEGG module assignment was carried out using *anvi-estimate-metabolism* (Veseli et al., 2025) from the *anvi'o* v8 software package (Eren et al., 2021) and *KEGG_build_2023_09_22* (Veseli, 2023). The remaining of our annotation pipeline relied on custom code, including a large R script dedicated to the generation of figures highlighting differential or “hot” pathways between pairs (or sometimes more for hot pathways) of event sets. For interpretation, KEGG pathways were further grouped into higher-level (KEGG-inspired) categories more suited to our purposes (see Table 25).

KO-IDs were collected at two complementary levels: (i) the event level using annotations associated with donor and receiver sequences, and (ii) the orthoclan level by considering annotations aggregated across all members of the corresponding orthoclan. While the orthoclan level allows collecting KO-IDs even when event sequences are uncharacterized, in principle, it also bears the risk to pull down KO-IDs that annotate sequences that are only quite distantly related to the sequences involved in the event. However, this potential issue is likely to be mitigated by the fact that our orthoclans have first been split based on phylogenetic trees (see Methods in Chapter 3, section 5.1.9) and thus do not correspond to huge families with many paralogous sequences.

Event- or orthoclan-level collection of KO-IDs were then fed to *anvi-estimate-metabolism* (options `--enzymes-txt`, `--include-kos-not-in-kofam`, `--include-zeros`, and `--module-completion-threshold 0.5`) to compute KEGG module completeness. In *anvi'o*, module completeness can be estimated either from the proportion of required steps represented by the annotated enzymes (stepwise strategy, less granular) or by taking the most complete path across steps among all possible paths (pathwise strategy, more granular). A module is considered complete if its completeness score reaches or exceeds a user-defined

threshold; although the default threshold is 0.75, this parameter can be lowered, here to 0.5, to evaluate more partial functional reconstructions, since we only use KO-IDs from transfer events (or specific orthoclans) and not from the complete proteomes of the organisms under consideration. For the same reason, we opted for the pathwise strategy to compute module completeness.

Module completeness values were subsequently used as quantitative descriptors of functional potential in a custom R script. For pairwise comparisons between event sets, modules were classified according to the user-defined completeness threshold (here 0.5) into four categories: (i) present in both sets, (ii) present only in the first set, (iii) present only in the second set, or (iv) absent from both. Differential modules were identified based on the signed difference in completeness between the two sets, and the most contrasting modules were selected using a balanced approach that considers both directions of change. At the KEGG pathway level, module-level signals were aggregated by summing completeness differences or shared completeness values across all modules assigned to a given pathway, allowing the identification of pathways that are either differentially enriched or commonly represented between two event sets.

To identify “hot” pathways, module completeness values were combined across multiple (more than two) event sets. For each module, the number of sets in which completeness exceeded the threshold (“hot count”) and the summed completeness across sets were computed. These quantities were then aggregated at the pathway level to rank pathways according to the recurrence and intensity of their associated modules. This approach highlights KEGG pathways consistently supported across multiple event sets, independently of pairwise contrasts.

For visualization on the global KEGG map (Metabolic pathways: map01100), module assignments were projected back to KO-IDs by parsing module definitions. KO IDs were then assigned a quantitative score derived from module recurrence and mapped to discrete colour bins based on quantiles of this score. These bins were associated with a monotonic colour gradient, enabling direct rendering via the KEGG Mapper Color tool [<https://www.genome.jp/kegg/mapper/color.html>], where colour intensity reflects the recurrence of functional signals across sets.

<i>KO_map</i>	<i>Original_KEGG_category</i>	<i>Pathway</i>	<i>Reassigned_KEGG_category</i>
<i>map01210</i>	Global and overview maps	2-Oxocarboxylic acid metabolism	Amino acid metabolism
<i>map01212</i>	Global and overview maps	Fatty acid metabolism	Lipid metabolism
<i>map01232</i>	Global and overview maps	Nucleotide metabolism	Nucleotide metabolism
<i>map01250</i>	Global and overview maps	Biosynthesis of nucleotide sugars	Glycan biosynthesis and metabolism
<i>map01320</i>	Global and overview maps	Sulfur cycle	Energy metabolism
<i>map00010</i>	Carbohydrate metabolism	Glycolysis / Gluconeogenesis	Carbohydrate metabolism
<i>map00020</i>	Carbohydrate metabolism	Citrate cycle (TCA cycle)	Carbohydrate metabolism

<i>map00030</i>	Carbohydrate metabolism	Pentose phosphate pathway	Carbohydrate metabolism
<i>map00500</i>	Carbohydrate metabolism	Starch and sucrose metabolism	Carbohydrate metabolism
<i>map00620</i>	Carbohydrate metabolism	Pyruvate metabolism	Carbohydrate metabolism
<i>map00630</i>	Carbohydrate metabolism	Glyoxylate and dicarboxylate metabolism	Carbohydrate metabolism
<i>map00562</i>	Carbohydrate metabolism	Inositol phosphate metabolism	Carbohydrate metabolism
<i>map00190</i>	Energy metabolism	Oxidative phosphorylation	Energy metabolism
<i>map00195</i>	Energy metabolism	Photosynthesis	Energy metabolism
<i>map00710</i>	Energy metabolism	Carbon fixation by Calvin cycle	Energy metabolism
<i>map00720</i>	Energy metabolism	Other carbon fixation pathways	Energy metabolism
<i>map00680</i>	Energy metabolism	Methane metabolism	Energy metabolism
<i>map00910</i>	Energy metabolism	Nitrogen metabolism	Energy metabolism
<i>map00920</i>	Energy metabolism	Sulfur metabolism	Energy metabolism
<i>map00061</i>	Lipid metabolism	Fatty acid biosynthesis	Lipid metabolism
<i>map00062</i>	Lipid metabolism	Fatty acid elongation	Lipid metabolism
<i>map00561</i>	Lipid metabolism	Glycerolipid metabolism	Lipid metabolism
<i>map00564</i>	Lipid metabolism	Glycerophospholipid metabolism	Lipid metabolism
<i>map01040</i>	Lipid metabolism	Biosynthesis of unsaturated fatty acids	Lipid metabolism
<i>map00230</i>	Nucleotide metabolism	Purine metabolism	Nucleotide metabolism
<i>map00240</i>	Nucleotide metabolism	Pyrimidine metabolism	Nucleotide metabolism
<i>map00260</i>	Amino acid metabolism	Glycine, serine and threonine metabolism	Amino acid metabolism
<i>map00270</i>	Amino acid metabolism	Cysteine and methionine metabolism	Amino acid metabolism
<i>map00290</i>	Amino acid metabolism	Valine, leucine and isoleucine biosynthesis	Amino acid metabolism
<i>map00300</i>	Amino acid metabolism	Lysine biosynthesis	Amino acid metabolism
<i>map00220</i>	Amino acid metabolism	Arginine biosynthesis	Amino acid metabolism
<i>map00330</i>	Amino acid metabolism	Arginine and proline metabolism	Amino acid metabolism
<i>map00340</i>	Amino acid metabolism	Histidine metabolism	Amino acid metabolism
<i>map00400</i>	Amino acid metabolism	Phenylalanine, tyrosine and tryptophan biosynthesis	Amino acid metabolism
<i>map00480</i>	Metabolism of other amino acids	Glutathione metabolism	Metabolism of other amino acids
<i>map00520</i>	Glycan biosynthesis and metabolism	Amino sugar and nucleotide sugar metabolism	Glycan biosynthesis and metabolism
<i>map00510</i>	Glycan biosynthesis and metabolism	N-Glycan biosynthesis	Glycan biosynthesis and metabolism
<i>map00513</i>	Glycan biosynthesis and metabolism	Various types of N-glycan biosynthesis	Glycan biosynthesis and metabolism
<i>map00730</i>	Metabolism of cofactors and vitamins	Thiamine metabolism	Metabolism of cofactors and vitamins
<i>map00740</i>	Metabolism of cofactors and vitamins	Riboflavin metabolism	Metabolism of cofactors and vitamins
<i>map00770</i>	Metabolism of cofactors and vitamins	Pantothenate and CoA biosynthesis	Metabolism of cofactors and vitamins
<i>map00780</i>	Metabolism of cofactors and vitamins	Biotin metabolism	Metabolism of cofactors and vitamins
<i>map00785</i>	Metabolism of cofactors and vitamins	Lipoic acid metabolism	Metabolism of cofactors and vitamins
<i>map00670</i>	Metabolism of cofactors and vitamins	One carbon pool by folate	Metabolism of cofactors and vitamins
<i>map00860</i>	Metabolism of cofactors and vitamins	Porphyrin metabolism	Metabolism of cofactors and vitamins

<i>map00130</i>	Metabolism of cofactors and vitamins	Ubiquinone and other terpenoid-quinone biosynthesis	Metabolism of cofactors and vitamins
<i>map00900</i>	Metabolism of terpenoids and polyketides	Terpenoid backbone biosynthesis	Metabolism of terpenoids and polyketides
<i>map00906</i>	Metabolism of terpenoids and polyketides	Carotenoid biosynthesis	Metabolism of terpenoids and polyketides

Table 25: **Manual reassignment of selected KEGG map categories.** The table lists KEGG map IDs, pathway names, their original KEGG categories, and the reassigned categories used in this study when the default classification was considered too broad or functionally uninformative.

5.8. Updating the dataset: integrating new kleptoplastidic species

We expanded the dataset by adding the following kleptoplastidic species: *Rapaza viridis*, *Dinophysis ovum*, *Dinophysis fortii* using the Forty-Two enrichment protocol described in section “Orthoclan enrichment”. Furthermore, we added the cryptophyte algae *Teleaulax amphioxeia*, considered as the plastid donor for the *Dinophysis cf. acuminata* complex (Janson, 2004; Gaonkar and Campbell, 2023). The aforementioned species genomes were screened for contaminations with Forty-Two using 78 ribosomal protein markers and the protocol described in Van Vlierberghe et al., 2021a (Chapter 3). Given the cross-contamination suspicion on *D. ovum* and *T. amphioxeia*, the other assemblies (part of *Dinophysis*’ trophic chain) of the same study were screened for contamination as well, i.e., two *Mesodinium rubrum* strains (*JAMR* and *MBL-DK2009*) and *Dinophysis acuminata* strain DAVA01.

5.9. Ribosomal phylogenomic tree

78 ribosomal markers were enriched with our 73 proteomes sampling (Chapter II) and 224 transcriptomes sampling (Chapter III) with Forty-Two (Chapter I). The retrieved orthologous sequences were pruned from the ribosomal MSAs and concatenated using a pre-release version of scaphoid.pl from the Bio::MUST::Core module (`scaphoid.pl --level=strain --rank-by=dist-to-root --mode=discard`), resulting in a supermatrix of 293 species across 12,224 amino-acid positions.

5.10. Balanced sampling

We used the ribosomal phylogenomic tree as a guide to subsample our species, aiming at reducing redundancy of the larger groups without compromising diversity. We selected species based on their branch length (shortest branches were favoured) and expected position within their taxonomic groups as reported the literature (Table 26), resulting in a more balanced sampling.

<i>lineage</i>	<i>Full sampling</i>	<i>Balanced sampling</i>
<i>Ochrophyta</i>	85	21
<i>Dinophyceae</i>	32	18
<i>Viridiplantae</i>	29	15
<i>Haptophyta</i>	23	11
<i>Cryptophyceae</i>	11	9
<i>Rhodophyta</i>	11	11
<i>Chlorarachniophyceae</i>	8	6
<i>Colpodellida</i>	5	2
<i>Apicomplexa</i>	3	3
<i>Euglenophyceae</i>	2	2
<i>Glaucocystophyceae</i>	2	2

Table 26: Taxonomic sampling before and after balancing.

6. Transition: towards an α -diversity framework for plastid integration

The quantitative phylogenomic pipeline developed and applied in Chapter V produced a collection of inferred transfer events spanning all major complex algal lineages, from which organism-level donor α -diversity profiles were derived. These profiles revealed a structured gradient, ranging from single-donor dominance in lineages with well-characterised endosymbiosis to broad donor mixtures in lineages associated with kleptoplasty or complex acquisition histories. After assessing the robustness of the α -diversity signal to analytical choices, we discuss in Chapter VI what this gradient means biologically. We interpret the three observed α -diversity regimes in the light of known plastid histories, ask whether pre-adaptive gene sets carry a recognizable functional signature, and explore how trophic ecology and donor availability shape the diversity of gene recruitment. Finally, we assess whether these patterns can be formalized into a predictive framework for plastid integration.

VI. Chapter 6 - Discussion

1. The α -diversity framework

Donor α -diversity provides a single value capturing how broadly nucleus-encoded, plastid-targeted (nucpt#) genes have been recruited across algal sources. In this light, the α -diversity distributions and donor-pattern summaries (Figure 35, Figure 36, Figure 37) offer a first opportunity to test our working hypotheses. Schematically, lineages that have acquired their plastid through endosymbiosis are expected to show relatively limited donor profiles (low α -diversity), typically dominated by a single major donor lineage, whereas lineages shaped by diverse transient interactions, such as kleptoplasty, should display broader donor mixtures (high α -diversity). Intermediate values, by contrast, may arise under a range of evolutionary scenarios. We thus interpret α -diversity as genomic signature that can distinguish these two extreme modes of plastid acquisition.

1.1. Low α -diversity: single-donor dominance

In the low-diversity regime (left of the first KDE cutoff; $\alpha \approx 0.67$), Cryptophyceae consistently cluster among the lowest Shannon-Wiener values. This is expected for a textbook case of secondary endosymbiosis: cryptophytes retain a red algal nucleomorph, the remnant nucleus of the endosymbiont, providing direct evidence for a single-donor plastid origin (Archibald, 2007; Douglas et al., 2001). Their position therefore establishes a useful reference point for interpreting the other lineages.

Strikingly, both dinoflagellate lineages that underwent plastid replacement also map to the low-diversity range once their ancestral peridinin signal is accounted for. After reclassification (see Results, Table 8), Kryptoperidiniaceae ($\alpha = 0.29$) and Kareniaceae ($\alpha = 0.55$) both fall within the low-diversity regime, in marked contrast to the peridinin-containing core dinoflagellates ($\alpha = 1.50$), which retain a high-diversity signature. This pattern is consistent with a reset toward single-donor dominance following plastid replacement, through massive EGT (Timmis et al., 2004).

However, the case of Kryptoperidiniaceae calls for a more nuanced interpretation. Indeed, in dinotoms, the plastid retains the diatom-derived nucleus (Hehenberger et al., 2016b; Imanian and Keeling, 2007), which remains transcriptionally active with an expression profile heavily biased toward plastid-targeted functions and limited evidence for EGT to the dinoflagellate host nucleus (Yamada et al., 2025). The particularly low α -diversity observed in Kryptoperidiniaceae ($\alpha = 0.29$) may therefore not reflect massive EGT but rather the continued dominance of diatom-derived transcripts in currently available

transcriptomes, making it difficult to sort out the diatom-derived nucleus own gene expression from expression of genes resulting from genuine EGT events.

In Kareniaceae, the haptophyte-derived plastid replacement also converges on low diversity ($\alpha = 0.55$ after reclassification). In our dataset, the donor profile of individual Kareniaceae species (*Karenia brevis*, *Karlodinium veneficum*) is dominated by Haptophyta, consistent with the haptophyte origin of their current plastid. However, the integration process may not be uniform across the Kareniacean clade: Novák Vanclová et al. (Novák Vanclová et al., 2024) documented heterogeneous (different sources) plastid-associated protein repertoires among kareniacean species, consistent with a substantial peridinin-derived footprint (of multiple origins) re-targeted to the haptophyte plastid. In the present dataset, however, the Kareniaceae-specific signal after reclassification is strongly narrowed toward haptophyte donors, suggesting that, at the level of detectable transfer events, the replacement appears rather complete for the species analysed here (Figure 75).

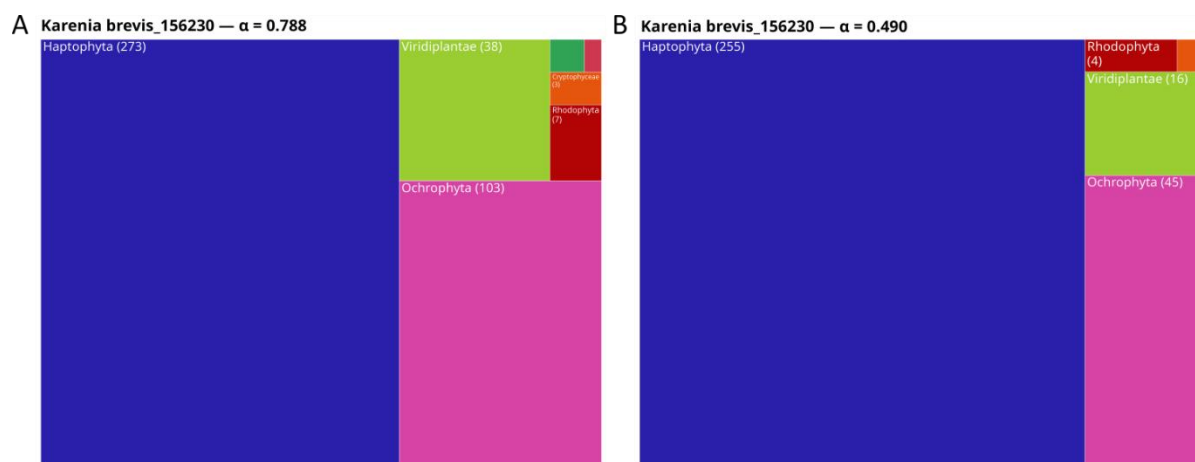


Figure 75: Donor profile of *Karenia brevis* with and without ancestral peridinin-associated transfer events. (A) Raw donor profile including all inferred transfer events. (B) Donor profile after removal of events shared with peridinin- dinoflagellates (106 events). Treemap areas are proportional to the number of inferred transfer events per donor lineage. The removal of the peridinin footprint reveals a shift from a moderately diversified donor profile ($\alpha = 0.788$) to a profile strongly dominated by Haptophyta ($\alpha = 0.55$), consistent with the haptophyte origin of the kareniacean replacement plastid.

Apicomplexa represent a particular case. Their placement in the low-diversity bin is consistent with the endosymbiotic origin of the apicoplast (Janouškovec et al., 2010; Woo et al., 2015). However, Apicomplexa and Chromerida are generally considered to share a common plastid ancestry (Fast et al., 2001; Janouškovec et al., 2010), yet they occupy different diversity regimes in our classification, since Chromerida display an intermediate α -diversity. This discrepancy likely reflects the extensive gene loss associated with obligate parasitism in Apicomplexa (Janouškovec et al., 2015b), which has reduced the number of plastid-related genes available for inference (but see Ralph et al., 2004), thereby compressing the apparent donor repertoire. Consistently, chromerids (*Chromera velia* and *Vitrella brassicaformis*)

retain a broader suite of plastid-targeted proteins than apicomplexans, supporting progressive gene loss along the parasitic branch (Woo et al., 2015; Füssy and Oborník, 2017). In addition, the accelerated sequence evolution typical of obligate parasites may further distort donor assignments: in our data, Colpodellida appear as the dominant donor in Apicomplexa (Figure 76), but this signal more likely reflects LBA (Long Branch Attraction) (Gouy et al., 2015) toward their closest photosynthetic relatives rather than genuine gene transfer.

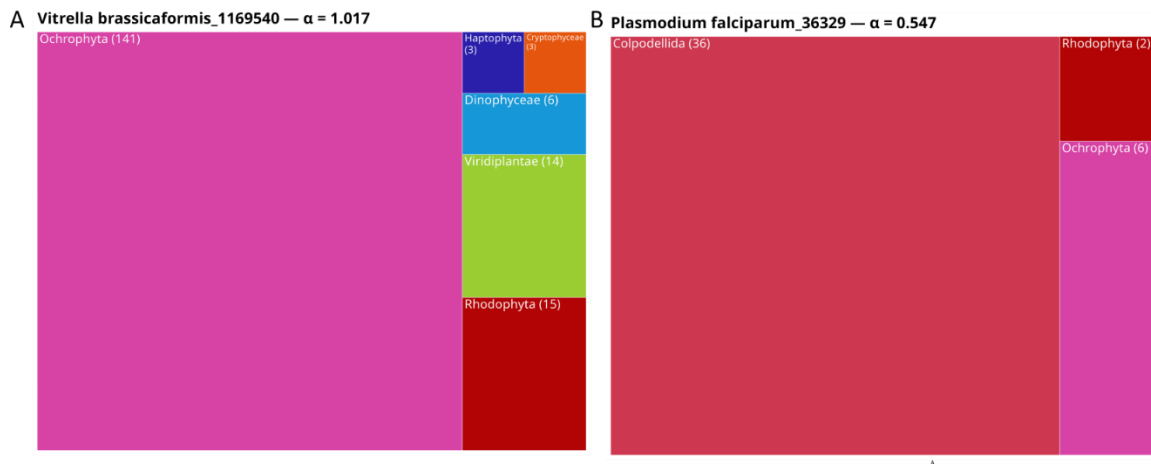


Figure 76: Comparison of donor profiles between a photosynthetic chromerid (*Vitrella brassicaformis*, $H' = 1.017$) and an apicomplexan parasite (*Plasmodium falciparum*, $\alpha = 0.547$). Treemap areas are proportional to the number of inferred transfer events per donor lineage. *Vitrella* displays a diversified donor profile dominated by Ochrophyta with contributions from Rhodophyta, Viridiplantae, Haptophyta, and Dinophyceae. In contrast, *Plasmodium* shows a strongly compressed profile dominated by Colpodellida, but most likely reflecting a vertical signal in this specific case (36/44 events).

1.2. Intermediate α -diversity: complex histories

Aside from Colpodellida, Haptophyta and Ochrophyta cluster in the intermediate diversity range. Their distributions are relatively narrow and overlap only moderately, indicating lineage-specific diversity levels rather than a single shared "intermediate" pattern.

For Haptophyta and Ochrophyta, an intermediate α -diversity regime is not consistent with a naive single-donor, single-event model but does not by itself discriminate among alternative scenarios. The chimeric gene repertoires documented in ochrophytes, with contributions from both red and green algal sources, are consistent with this interpretation (Dorrell et al., 2017). Whether this mosaicism reflects (cryptic) serial endosymbioses (Moustafa et al., 2009), integration with diverse kleptoplastic or endosymbiotic associations (Bodyl, 2018), or a combination of both, remains unresolved, and our analyses alone cannot discriminate among these scenarios. An alternative scenario would be that the observed mosaicism was inherited from a single endosymbiont that itself possessed a chimeric nucpt# repertoire. However, in our view, this scenario faces two difficulties. First, it requires massive endosymbiotic gene transfer from the endosymbiont nucleus to the host nucleus, a hypothesis that has been considerably mitigated in recent literature (Burki et al., 2014; Moreira and Deschamps, 2014;

Bernabeu and Gabaldón, 2026). Second, and more fundamentally, genes acquired from such a chimeric donor would most likely phylogenetically trace back to the immediate (chimeric) donor lineage, not to its own original sources. The α -diversity measured at the host level would therefore remain low, reflecting a single donor, unless the donor lineage is extinct or insufficiently sampled, in which case the signal might appear as a nested gene transfer configuration rather than as genuine donor diversity.

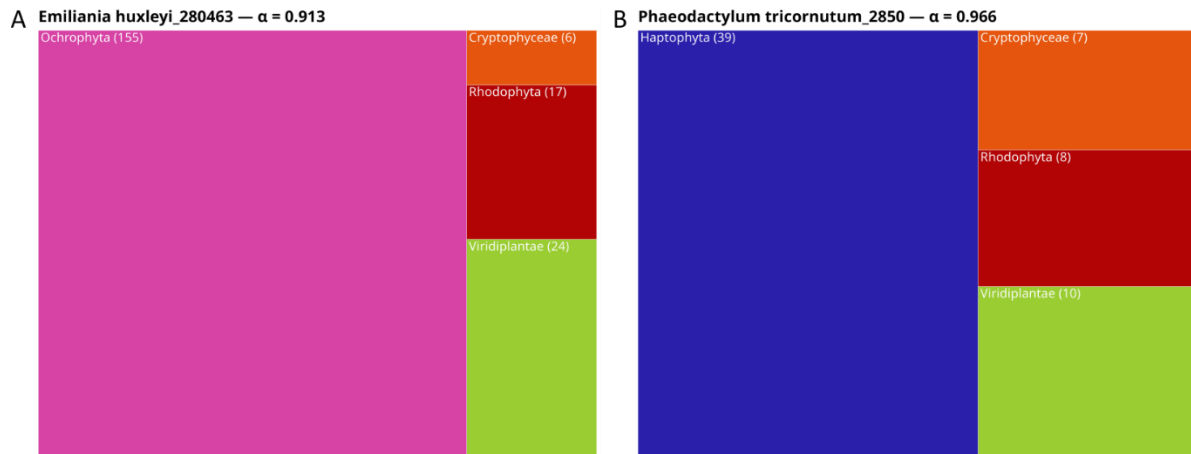


Figure 77: Donor profiles of two representative intermediate-diversity lineages: the haptophyte *Emiliana huxleyi* ($\alpha = 0.913$) and the ochrophyte *Phaeodactylum tricornutum* ($\alpha = 0.966$). Treemap areas are proportional to the number of inferred transfer events per donor lineage. Both species display multi-donor profiles with contributions from Ochrophyta, Haptophyta, Viridiplantae, Rhodophyta, and Cryptophyceae, consistent with chimeric gene repertoires rather than single-donor dominance. Notably, each species shows the other's lineage as its primary donor reflecting the strong reciprocal signal between these two lineages discussed in the ochrophyte case study (see Results, Section 3.2.4.3).

1.3. High α -diversity: wider prey range

Among red complex algae, high α -diversity characterizes Dinophysiales ($\alpha = 1.35$) and the peridinin-ancestral-core dinoflagellates ($\alpha = 1.50$) in our dataset (Figure 78). Dinophysiales are obligate kleptoplastidic organisms: they do not possess a permanent plastid but instead acquire and maintain transient kleptoplasts through a trophic chain involving the ciliate *Mesodinium rubrum* (Wisecaver and Hackett, 2010). Most species in the *D. acuminata* complex harbour kleptoplasts of cryptophyte origin, but the genus is not restricted to a single plastid source: *D. mitra* acquires haptophyte-derived kleptoplasts (Koike et al., 2005) and can retain a broad diversity of plastids, most of them haptophyte-related (Nishitani et al., 2012), demonstrating that Dinophysiales can exploit phylogenetically diverse photosynthetic prey. The exceptionally high α -diversity of peridinin-dinoflagellates combined with their three-membrane-bound plastid is consistent with a kleptoplastic acquisition, a hypothesis already proposed by Bodyl, (2018) and Moog and Maier, (2018). In both cases, the elevated donor diversity is consistent with the Shopping Bag models (Larkum et al., 2007; Howe et al., 2008b; Morozov and Galachyants, 2019), in which repeated transient associations with diverse algal prey are assumed to

provide opportunities for the host to accumulate plastid-targeted genes from multiple donors before, or without, stabilizing a permanent endosymbiosis.

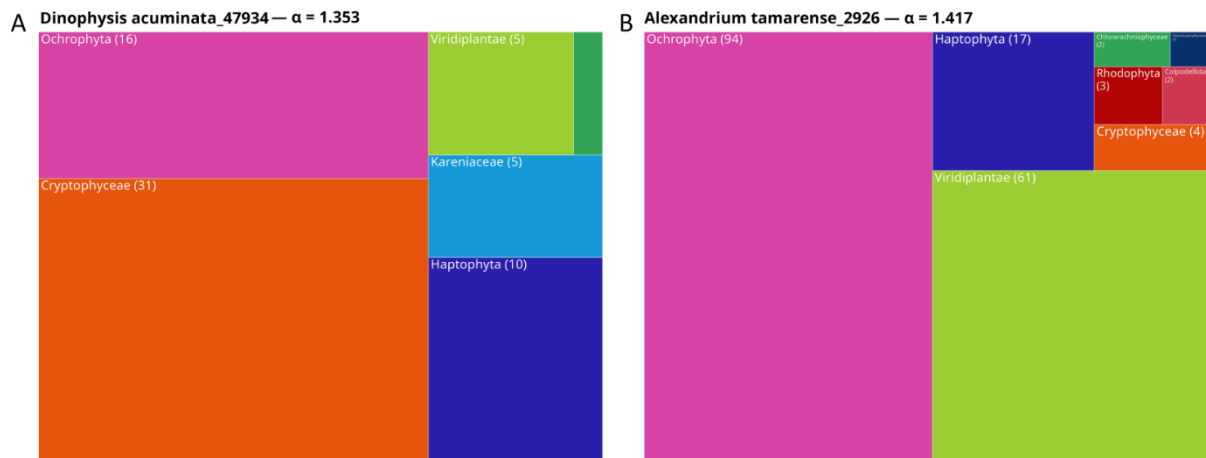


Figure 78: Donor profiles of two high-diversity dinoflagellate species: the kleptoplastidic *Dinophysis acuminata* ($\alpha = 1.353$) and the peridinin-plastid bearing *Alexandrium tamarense* ($\alpha = 1.417$). Treemap areas are proportional to the number of inferred transfer events per donor lineage. *Dinophysis acuminata* shows a profile dominated by Cryptophyceae (31 events), consistent with its known cryptophyte-derived kleptoplasts, alongside contributions from Ochrophyta, Haptophyta, Viridiplantae, and Karenia. *Alexandrium tamarense* displays a broadly diversified profile with Ochrophyta (94) and Viridiplantae (61) as dominant donors, followed by Haptophyta (17) ($\alpha = 1.50$).

Importantly, kleptoplasty-mediated acquisition implies a sequential relationship between gene accumulation and plastid stabilization: pre-adaptive gene recruitment occurs before the establishment of a permanent organelle, which is the central tenet of both the Shopping Bag models (Howe et al., 2008b; Larkum et al., 2007; Morozov and Galachyants, 2019) and the Red Carpet Hypothesis (Ponce-Toledo et al., 2019) proposed to justify mosaicism in an endosymbiotic context. Under this view, high α -diversity is not merely a consequence of a complex plastid history; it may instead represent the genomic footprint of the pre-adaptive phase itself. However, applying this logic to complex algae requires considering mechanisms that would be different from primary plastid origin. In primary endosymbiosis, the host interacted with a prokaryotic cyanobacterial partner. In secondary and higher-order contexts, by contrast, the host interacts with eukaryotic prey or endosymbionts that already contain their own plastids, additional membranes, and potentially other cellular compartments such as a nucleus or mitochondrion. This cellular complexity expands the range of possible mechanisms, including kleptoplasty, karyoklepty, myzocytosis (Bodyl, 2018) or modifications of phagocytic pathways allowing the retention of endosymbionts (Johnson, 2011).

Among green complex algae, the presence of Chlorarachniophyceae in the high-diversity regime is, at first glance, puzzling, because their nucleomorph provides unambiguous evidence for a bona fide secondary endosymbiosis, paralleling the situation in Cryptophyceae, which show low diversity. However, prior genomic studies have already revealed that a substantial fraction of chlorarachniophyte

plastid-targeted proteins do not trace back to the green algal endosymbiont but instead derive from diverse sources, including red algae and complex red algae (Archibald et al., 2003; Curtis et al., 2012; Yang et al., 2014) (Figure 79). This pattern is consistent with the Red Carpet Hypothesis, which posits that horizontally acquired red algal genes served as a molecular "welcome mat" facilitating the integration of the current green algal plastids in chlorarachniophytes and euglenids, without specifying an explicit mechanism (Ponce-Toledo et al., 2018). The contrast with Cryptophyceae may reflect differences in ecological and evolutionary context. One possible explanation is that chlorarachniophyte plastid acquisition occurred in a trophic landscape already enriched in complex red algal diversity, potentially broadening the pool of available gene donors relative to the older cryptophyte acquisition. Still, the selective pressure to do so remains questionable in our opinion, making chlorarachniophytes a really puzzling case without alternative hypothesis than endosymbiosis because of their nucleomorph. Euglenophyceae present a similar picture. The kleptoplastic capabilities recently documented in euglenozoan relatives (Karnkowska et al., 2023) support the hypothesis that the euglenid plastid was preceded by a kleptoplastic phase during which genes from multiple donors were recruited. Consistently, the gene partitioning analyses presented in our results (Chapter 5, section 3.2.1) show that euglenid *nucpt#* genes include contributions from diverse lineages beyond the green algal donor, in line with pre-adaptative models (Shopping Bag and Red Carpet models) (Figure 79).

Rather than reflecting a privileged proximity with the lineage that will ultimately provide the permanent plastid, this pattern suggests that recruitment from multiple sources operates throughout the acquisition process. This challenges the implicit expectation, rooted in primary endosymbiotic models where EGT from the organellar ancestor dominates the host gene repertoire (Reyes-Prieto et al., 2007), i.e., that the ultimate plastid donor should similarly prevail among *nucpt#* genes in complex algae. This extrapolation may not hold under kleptoplast-mediated acquisition, where the host is exposed to diverse prey over extended periods (Bodyl, 2018; Larkum et al., 2007). Under such a scenario, the plastid that is eventually retained may simply be the one for which the host has, by chance, accumulated the largest compatible gene set, becoming fortuitously pre-adapted to its integration, similar to what has been hypothesized in the Shopping Bag Endosymbiosis model for red complex algae (Morozov and Galachyants, 2019).

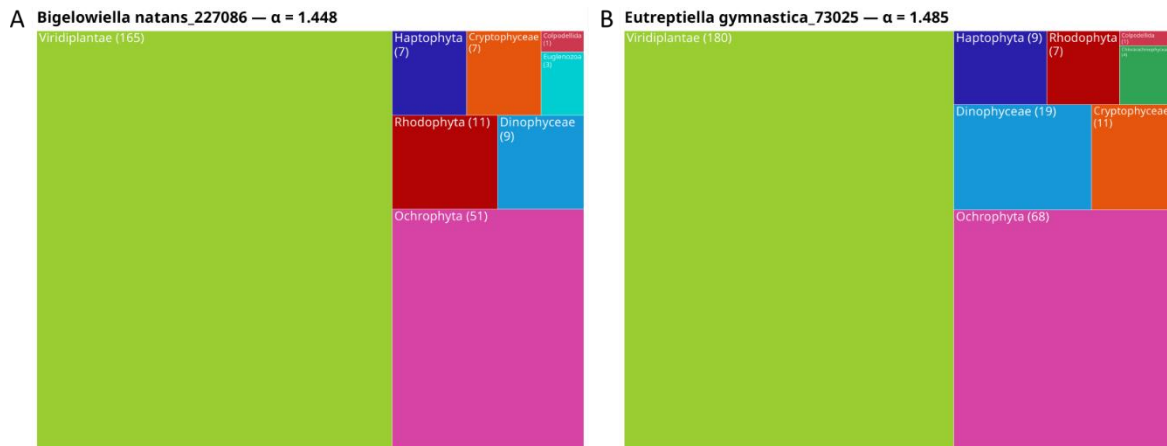


Figure 79: Donor profiles of two representative high-diversity green-plastid lineages: the chlorarachniophyte *Bigelowiella natans* ($\alpha = 1.448$) and the euglenid *Eutreptiella gymnastica* ($\alpha = 1.485$). Treemap areas are proportional to the number of inferred transfer events per donor lineage. Both species display remarkably similar profiles dominated by Viridiplantae, consistent with their current green algal plastid, but with substantial contributions from Ochrophyta, Rhodophyta, Haptophyta, Dinophyceae, and Cryptophyceae. This mixed red- and green-donor signature is consistent with earlier observations (Archibald et al., 2003) and the Red Carpet Hypothesis (Ponce-Toledo et al., 2018), which posits that red algal genes were acquired prior to the integration of the current green plastid.

1.4. Interpretative framework linking α -diversity to plastid integration stage

Before drawing broader conclusions, it is essential to be explicit about what α -diversity can and cannot establish on its own. A high α -diversity value is compatible with a kleptoplastic or pre-adaptive history, but does not prove one: serial endosymbioses, including cryptic ones (Moustafa et al., 2009) or phylogenetic artefacts could also contribute to multi-donor profiles but, more importantly, it could in theory reflect a endosymbiosis with multiple endosymbionts in the early phase of nucpt# acquisition by EGT. Conversely, low α -diversity does not prove a simple single-event endosymbiosis; it may also result from a reset toward single-donor dominance or from the erosion of an initially broader donor profile over evolutionary time. α -diversity is therefore best understood as a genomic signature consistent with specific evolutionary scenarios involving a diversity of prey with associated ERGT, and not a direct observation of the mechanism itself. For this reason, the biological interpretation of each lineage should rely on α -diversity patterns in combination with independent evidence, including nucleomorphs (Curtis et al., 2012), documented kleptoplasty, known plastid replacement events (Waller and Kořený, 2017), functional contrasts in gene content, and published comparative genomics (Curtis et al., 2012).

Within these interpretative limits, three broad regimes emerge from the data. Low α -diversity (Cryptophyceae, Kryptoperidiniaceae, Kareniaceae post-reclassification, Apicomplexa) reflects single-donor dominance, consistent either with the classical view of endosymbiotic integration or with a reset following plastid replacement. Intermediate α -diversity (Haptophyta, Ochrophyta, Colpodellida) reflects moderate donor mosaicism, compatible with more complex endosymbiotic histories involving diverse transient associations or serial events, although serial events tend to erase the previous signals as seen with plastid replacement in dinoflagellates (high to low diversity). High α -diversity

(Dinophysiales, peridinin dinoflagellates, Chlorarachniophyceae, Euglenophyceae) reflects broader donor mixtures, consistent with pre-adaptive gene accumulation through kleptoplastic (and karyokleptic) or other endosymbiotic interactions with broad prey diversity, as formalized in the Shopping Bag model (Howe et al., 2008b) and the Red Carpet Hypothesis (Ponce-Toledo et al., 2019), even if the selective pressure is difficult to conceive in the case of endosymbiosis (see Introduction; Section 5.2). The key difficulty is that α -diversity alone cannot always distinguish between kleptoplasty-compatible pre-adaptation and pre-adaptation that would have occurred during early endosymbiotic integration. Both scenarios may generate multi-donor nucpt# repertoires (Figure 16). However, because the cellular and selective contexts differ between transient kleptoplast retention and stable endosymbiotic integration, the functional composition of the genes retained under each scenario may also differ. Functional analyses therefore provide a complementary way to explore whether different contexts of pre-adaptation are associated with different functional signatures.

Yet, while a putative pre-adaptation gene set could be operationally identified in euglenids, defining an equivalent endosymbiotic pre-adaptation set is more difficult. Euglenids therefore provides a first test at the level of plastid-acquisition phases within a single lineage, by asking whether putative pre-adaptation and plastid-integration gene sets are associated with distinct functional repertoires.

2. How does pre-adaptation shape plastid integration?

2.1. The euglenid system: pre-adaptation identified under kleptoplastic selective pressure

Euglenids provide a natural experiment for dissecting the temporal sequence of gene accumulation during plastid acquisition (Larkum et al., 2007; Reyes-Prieto et al., 2007). *Rapaza viridis*, described as a mixotroph that feeds on green algal prey (*Tetraselmis*) and retains their plastids transiently (Yamaguchi et al., 2012), has not crossed the tipping point toward permanent plastid integration. Yet it already possesses a detectable set of nucpt# genes shared with *Euglena* and *Eutreptiella*. Because *Rapaza* is an active kleptoplast host, and because kleptoplasty generates direct selective pressure for maintaining stolen plastids (Bodół, 2018; Hehenberger et al., 2019), these shared genes represent the most concrete candidate evidence in our event collection for a pre-adaptive gene set assembled under kleptoplastic selective pressure. Notably, Karnkowska et al. 2023 showed that the plastid-targeting system of *Rapaza* is shared with other euglenids, indicating that the molecular infrastructure for kleptoplast exploitation predates the divergence of *Rapaza*. This is consistent with the view that nucleus-to-plastid protein trafficking is a central requirement that must be established early during plastid acquisition (Gould et al., 2015; Gross and Bhattacharya, 2011; Keeling, 2024).

The gene partitioning analysis described in the Results chapter (Section 3.2.1.1) identified two sets: a pre-adaptation set (nucpt# events shared between *Rapaza* and *Euglena/Eutreptiella*) and a plastid-integration set (events only shared by *Euglena* and *Eutreptiella*). This partitioning approximates the distinction between genes acquired before and after permanent plastid establishment. The differential pathway analysis (Figure 44) reveals a clear functional contrast between the two sets: the pre-adaptation set is thus enriched in functions directly related to the photosynthetic process and its (light reactions, carbon fixation, photoprotection, and thylakoid membrane biogenesis), whereas the integration set is enriched in biosynthetic housekeeping functions (synthesis of fatty acids, amino acids, nucleotides, and cofactors). The functional profile of the pre-adaptation set is consistent with the primary selective advantage of acquiring a plastid, namely access to photosynthates (Johnson, 2011; Stoecker et al., 2009) while the integration set contributes more to broader metabolic functions. The contrast between the two sets therefore provide a functional illustration of the gradual nature of plastid establishment, echoing the logic of the Shopping Bag model (Howe et al., 2008a; Larkum et al., 2007), but here at the functional level.

The recent demonstration by Maruyama et al. (Maruyama et al., 2023) that *Rapaza* acquired a functional nitrate reductase (*RvNaRL*) from a different source than other euglenids through horizontal gene transfer further illustrate the pre-adaptation mechanism operating in an extant kleptoplastic species: *RvNaRL* is essential for nitrate assimilation during kleptoplast-dependent photoautotrophy, and its loss abolishes growth on nitrate as a sole nitrogen source (Maruyama et al., 2023).

Notably, *RvNaRL* is a soluble enzyme rather than a component of a plastid-encoded multi-subunit complex. This observation connects to the slightly higher α -diversity of the pre-adaptation set (1.25 vs 1.04), which may reflect differences in the constraints governing the origin of acquired genes. Soluble enzymes may tolerate a broader range of donor origins than proteins that must physically assemble with plastid-encoded partners. By contrast, proteins incorporated into multi-subunit plastid complexes are expected to face stronger compatibility constraints, especially when they interact with components still encoded by the plastid genome. This idea is conceptually compatible with the COCOA hypothesis (Howe and Barbrook, 2024), which emphasizes the importance of assembly constraints in organellar gene retention, although COCOA does not directly predict donor origin. Under this interpretation, increasing plastid integration would progressively narrow the space of functionally acceptable donors, potentially contributing to the lower α -diversity observed in the integration set ($\alpha = 1.04$).

A recent study by Kashiyama et al. (2026) provided biochemical evidence that two nucleus-encoded proteins of *Rapaza viridis*, RvRbcS-like (RuBisCO small subunit-like) and RvRca-like (RuBisCO activase homologue), are translocated into kleptoplasts and are functionally required for photosynthesis. Their phylogenetic interpretation indicates that they were acquired during the ancestral euglenid plastid acquisition (the equivalent of our pre-adaptation set) and are now deployed in a different kleptoplast than their original source (Kashiyama et al., 2026). This constitutes direct evidence that the pre-adaptation gene set is functionally active in the kleptoplastic context.

Taken together, these results suggest that pre-adaptation shapes plastid integration by structuring it into at least two functional phases. Early acquisitions are biased toward plastid exploitation and maintenance (i.e., to increase kleptoplasts lifespan), whereas later acquisitions are biased toward the broader biosynthetic support (i.e., metabolic integration) required for long-term maintenance. In this view, pre-adaptation would not only be characterized by an unusually high donor diversity, but a functionally coherent stage in the transition from kleptoplast use to permanent plastid integration.

2.2. Parallel with peridinin dinoflagellates: a convergent functional partitioning?

Beyond euglenids, the peridinin dinoflagellate system provides a second, independent case in which two plastid-associated gene sets can be compared, here in the context of plastid replacement. As described in the Results chapter (Section 3.2.2.2; Figure 48), the peridinin-core orthoclangs were split into a peridinin-core-specific set (genes retained only in peridinin dinoflagellates and either lost or not detectably replaced in derived lineages) and a peridinin-core-associated set (genes found in both peridinin dinoflagellates and in some derived lineages, whether through continuous retention or independent re-recruitment within the same orthoclan). Importantly, the peridinin-core-associated set can be interpreted as "portable" genes, whose functions remained sufficiently necessary that, when a dinoflagellate replaced its peridinin plastid with a new one, related genes from the same orthoclan could be retained or recruited again. The peridinin-core-specific set, by contrast, comprises genes that were not replaced as identifiable members of the same orthoclan, either because derived lineages found alternative solutions or because their function became dispensable.

The differential pathway analysis (Figure 48) reveals no sharp functional contrast between these two sets, which appear more related to host integration: the peridinin-core-specific set is biased toward functions more closely associated with central carbon metabolism in the ancestral peridinin context, whereas the replacement-associated set is biased toward thiamine- and sulfur-related metabolism.

In euglenids, the contrast approximates a temporal distinction between a kleptoplasty-associated pre-adaptation set and a later plastid-integration set. In peridinin dinoflagellates, the contrast captures the different fates of ancestral peridinin-associated genes after plastid replacement: some are lost or replaced, whereas others are retained. The two systems are therefore not directly equivalent, but nevertheless point to a comparable functional problem: some plastid-related functions can remain useful across changes in plastid context, whereas others appear more tightly coupled to a specific plastid-host configuration.

What does this parallel would imply for the peridinin plastid? The euglenid pre-adaptation set was defined through its occurrence in a kleptoplastic organism, *Rapaza viridis*, thereby directly linking it to kleptoplast-associated selective pressure. By contrast, the peridinin-core-associated set was defined through its persistence across peridinin-derived lineages that underwent plastid replacement, thus capturing a form of functional portability. However, these two independently defined sets do not converge on a strict common functional logic, and we did not identify a clear enrichment in functions directly linked to plastid exploitation and maintenance that would support the hypothesis that the peridinin plastid was preceded by a kleptoplastic phase, as proposed by Bodyl (2018).

An alternative explanation is that some functions linked to metabolic integration are intrinsically more reusable across plastid transitions within red complex algae, independently of the historical route by which the plastid was acquired. Conversely, functions linked to plastid exploitation or maintenance may be more dependent on the specific plastid-host configuration and therefore more likely to be replaced by lineage-specific alternatives after plastid replacement. Nonetheless, recent evidence that plastid translocon components can be recycled during plastid replacement in dinoflagellates (Lewis et al., 2024), together with phylogenetic evidence that the SELMA translocation machinery was serially transmitted across CASH lineages (Ponce-Toledo et al., 2025), reinforces the idea that some plastid-associated systems are indeed unusually portable, but does not by itself resolve whether that portability was originally assembled under kleptoplastic selection.

2.3. Dinophysiales: the limits of large-scale functional inference

The Dinophysiales provide a third system in which to test the pre-adaptation framework, with the advantage that they are unambiguously kleptoplastic in the present day. However, the Dinophysiales case also illustrates the limits of pathway-level functional inference when the transfer event count is low. With only two informative species (*D. acuminata* and the decontaminated signal from *D. ovum*; see Section 4.1) and 72 associated orthoclans, the KEGG module-level analysis recovered only four modules above the completeness threshold. All four were populated by a single orthoclan (OG0003300-

2-1-65), encoding a ferredoxin-nitrite reductase (K00366) and/or a ferredoxin-dependent sulfite reductase (K00392). Importantly, this dual annotation may stem from ambiguity (due to structural similarity between these enzymes) rather than reflecting two distinctly transferred functions. In this case, the pathway level does not provide enough resolution for a comparative functional contrast analogous to those described for euglenids (Figure 44) or peridinin dinoflagellates (Figure 48).

At the level of individual KO assignments (Table 27), 26 annotations were recovered across 19 orthoclans. Setting aside three kinase/phosphatase annotations, likely reflecting mixed orthoclan composition (K06276, K08282, K15423, all on OG0000739-6-2-41), the remaining 23 assignments fall into five categories directly relevant to plastid function: central carbon metabolism and photosynthesis (*fructose-bisphosphate aldolase*, *glucose-6-phosphate isomerase*, *glucose-6-phosphate epimerase*, *acetyl-CoA synthetase*, *cytochrome b6-f iron-sulfur subunit*), pigment and terpenoid biosynthesis (*phytoene synthase*, *MEP pathway reductase*, *chlorophyll synthase*, *homogentisate phytyltransferase*), redox homeostasis and nutrient assimilation (*ferredoxin-dependent nitrite and sulfite reductases*, *ascorbate peroxidase*), plastid translation and protein homeostasis (*elongation factor G*, *peptide chain release factor 1*, *ribosomal protein L28*, *chaperone GrpE*, *phage shock protein A*), and porphyrin metabolism (*ferrochelatase*, *Mg-protoporphyrin O-methyltransferase*).

Notably, the categories that dominate the euglenid integration set (*fatty acid biosynthesis*, *amino acid metabolism*, *nucleotide metabolism*, *N-glycan biosynthesis*) are absent from this repertoire. Although this absence could simply reflect the small sample size, the contrast is qualitatively consistent with the pattern observed in the euglenid system: the Dinophysiales-associated transfers are biased toward plastid exploitation and maintenance functions rather than toward the broader biosynthetic support associated with long-term plastid metabolic integration.

Overall, while identified Dinophysiales events are too limited to support a formal functional comparison with the euglenid or peridinin systems, they are directionally consistent with the same functional logic: an early kleptoplast-associated regime dominated by exploitation rather than deeper metabolic integration (Wisecaver and Hackett, 2010). Dinophysiales, considering their scarce and contaminated data, should currently be treated with caution.

<i>Event id</i>	<i>KO id</i>	<i>Description</i>
OG0003300-2-1-65	K00366	ferredoxin-nitrite reductase
OG0003300-2-1-65	K00392	sulfite reductase (ferredoxin)
OG0000287-19-1-72	K00434	L-ascorbate peroxidase
OG0000739-6-1-40	K01623	fructose-bisphosphate aldolase, class I
OG0000739-6-2-41	K01623	fructose-bisphosphate aldolase, class I
OG0000739-6-3-63	K01623	fructose-bisphosphate aldolase, class I
OG0003884-1-1-69	K01623	fructose-bisphosphate aldolase, class I
OG0001803-3-1-8	K01772	protoporphyrin/coproporphyrin ferrochelatase
OG0000959-8-2-66	K01792	glucose-6-phosphate 1-epimerase
OG0000370-13-2-70	K01810	glucose-6-phosphate isomerase
OG0000395-9-1-59	K01895	acetyl-CoA synthetase
OG0003884-1-1-69	K02291	15-cis-phytoene synthase
OG0000592-8-2-57	K02355	elongation factor G
OG0002738-1-1-7	K02636	cytochrome b6-f complex iron-sulfur subunit
OG0000370-13-2-70	K02835	peptide chain release factor 1
OG0005369-0-1-71	K02902	large subunit ribosomal protein L28
OG0005253-0-1-61	K03428	magnesium-protoporphyrin O-methyltransferase
OG0004639-2-1-56	K03527	4-hydroxy-3-methylbut-2-en-1-yl diphosphate reductase
OG0001137-10-1-16	K03687	molecular chaperone GrpE
OG0004369-0-1-22	K03969	phage shock protein A
OG0005031-0-1-9	K04040	chlorophyll/bacteriochlorophyll a synthase
OG0000739-6-2-41	K06276	3-phosphoinositide-dependent protein kinase-1
OG0000739-6-2-41	K08282	non-specific serine/threonine protein kinase
OG0002929-1-1-3	K09833	homogentisate phytyltransferase / homogentisate geranylgeranyltransferase
OG0000739-6-2-41	K15423	serine/threonine-protein phosphatase 4 catalytic subunit
OG0003884-1-1-69	K17841	15-cis-phytoene synthase / lycopene beta-cyclase

Table 27:KEGG Orthology (KO) assignments recovered from the 74 Dinophysiales-associated transfer events. Each row corresponds to one KO annotation inferred for a specific transfer event in a specific orthoclan (event_id), KO accession and KO description.

3. What shapes high α -diversity in receivers?

The functional analyses in Section 2.1 established that pre-adaptation has a recognisable internal logic: early gene acquisitions serve plastid exploitation and maintenance, later ones serve deeper and potentially permanent metabolic integration. But the functional analysis does not address a more basic question: what determines which donors contribute genes in the first place? High α -diversity implies exposure to many donors, yet not all lineages, even those with kleptoplastic capabilities, are equally exposed.

3.1. Predators, intermediaries, and donor exposure

Dinophysiales illustrate how trophic intermediaries can amplify or constrain donor exposure. Most *Dinophysis* species do not feed directly on cryptophytes but instead prey on the kleptoplastidic ciliate *Mesodinium rubrum*, which itself acquires and maintains cryptophyte plastids and, in some cases, retains transcriptionally active cryptophyte nuclei (Johnson, 2011; Lasek-Nesselquist et al., 2015; Altenburger et al., 2020). This three-tiered chain (cryptophyte $\rightarrow$ *Mesodinium* $\rightarrow$ *Dinophysis*) means that the donor signal reaching *Dinophysis* is not simply that of the original prey, but a signal already filtered by the intermediary ciliate.

In this sense, *Mesodinium* does not act as a passive relay. Retained cryptophyte nuclei (kleptokaryon) can remain transcriptionally active, indicating that general plastid support may still be mediated by prey-derived nuclear machinery rather than by the ciliate host itself (Lasek-Nesselquist et al., 2015; Altenburger et al., 2020). The transfer opportunities available to *Dinophysis* are therefore shaped by what *Mesodinium* retains, maintains, and passes on. This may help explain why trophic intermediaries can expand donor exposure even when the final host himself specializes in a single prey.

The intermediary may also influence the diversity and quality of the donor pool. *Mesodinium* is known to exploit multiple cryptophyte species, and prey identity affects growth and photosynthetic performance in kleptoplastidic ciliates such as *Mesodinium chamaeleon* (Moeller et al., 2021). Trophic intermediaries therefore not only transmit plastids onward; they also filter which plastids remain functionally available, and for how long. In this sense, ciliates can act as trophic stepping stones, modulating both the accessibility and the quality of the plastid package entering downstream transfer networks.

This raises an intriguing question. If ciliates are among the most recurrent plastid thieves known, and in some cases even retain transcriptionally active prey nuclei, why do they still not appear to have crossed the threshold toward permanent plastid acquisition? One possible explanation is that ciliates may be evolutionarily specialized for recurrent predation and organelle turnover rather than for stable plastid retention (Johnson, 2011). In this view, systems such as *Mesodinium* represent advanced forms of dependence on stolen organelles, yet these still fall short of the host-encoded plastid support documented in more integrated kleptoplastidic systems such as the centrohelid *Meringosphaera* (Sørensen et al., 2023).

3.2. Ochrophytes as dominant donors in a reticulate transfer network

Across nearly all receiver lineages in the dataset, ochrophytes appear among the most frequent donors. This dominance has a straightforward ecological basis: diatoms are one of the most abundant and diverse eukaryotic phytoplankton in modern marine and freshwater environments, maximising co-occurrence with virtually all heterotrophic and mixotrophic eukaryotes that could serve as gene receivers (Dorrell et al., 2017; Jirsová and Wideman, 2024). Ecological availability directly translates into trophic exposure: organisms that feed on, are parasitised by, or simply coexist with ochrophytes have more opportunities to acquire ochrophyte-derived genes through HGT (Dorrell et al., 2021), whether facilitated by kleptoplasty or phagotrophy (Mansour and Anestis, 2021).

Yet ochrophytes are not merely donors; they are themselves receivers with a genuinely chimeric gene repertoire (Section 2.4) (Dorrell et al., 2021, 2017). The robustness analyses (Section 1.1) showed that the ochrophyte signal is the most sensitive to topological context: balanced sampling (bal_samp) redistributed the donor profile, Haptophyta contributions decreased while Rhodophyta increased, rather than changing the overall event volume. This confirms that donor-side sampling affects which donors are resolved more than how many events are inferred.

The parallel with Haptophyta is instructive; haptophytes occupy a similar intermediate-diversity regime, and similarly chimeric proteomes have been documented in both groups (Dorrell et al., 2017). Considering the phylogenomic evidence for independent plastid origins (Bodyl, 2018), their convergent α -diversity signatures may thus reflect a shared ecological context, i.e., both lineages recruited their plastid gene repertoires from the same diverse marine phytoplankton community, rather than a shared phylogenetic origin.

The duality of ochrophytes as both the most frequent donors and chimeric receivers illustrates a broader point: the gene transfer landscape is not a one-directional flow from donors to receivers but a reticulate network in which roles are not fixed (Dorrell et al., 2021, 2017). What Doolittle encapsulated as "you are what you eat" (Doolittle, 1998) applies at the level of the plastid proteome: the nucpt# gene repertoire of each lineage reflects the trophic and ecological interactions that have shaped its evolutionary history.

3.3. Availability matters: you eat what is there

The argument above has an important consequence: if trophic exposure structures donor diversity, then the same broad acquisition mechanism can produce different α -diversity outcomes depending on the ecological context. The contrast between Cryptophyceae and Chlorarachniophyceae illustrates this

point. Both lineages retain a nucleomorph, providing unambiguous evidence for secondary endosymbiosis (Irwin and Keeling, 2019; Sibbald and Archibald, 2020), yet they fall into opposite α -diversity regimes. Cryptophyceae ($\alpha = 0.39$, low diversity) acquired their red algal plastid early, at a time when the diversity of complex algal donors was presumably still limited (Strassert et al., 2021; Pietluch et al., 2022; Novak et al., 2024). By contrast, Chlorarachniophyceae ($\alpha = 1.39$, high diversity) acquired their green plastid later, in a trophic landscape already enriched in diverse complex algae (Pietluch et al., 2022; Sibbald and Archibald, 2020). Their high α -diversity is consistent with the Red Carpet Hypothesis, under which genes from earlier transient associations with red-lineage donors coexist with genes from the green endosymbiont (Ponce-Toledo et al., 2019).

More generally, this suggests that more recent plastid acquisitions may tend to show higher α -diversity because they may have arisen in a world containing a broader pool of potential algal donors (Pietluch et al., 2022; Ponce-Toledo et al., 2019; Sibbald and Archibald, 2020). This also clarifies an important limitation of the framework. High α -diversity should not be treated as an intrinsic property of kleptoplastic systems, because a kleptoplastidic lineage repeatedly exploiting a narrow prey spectrum could remain low-diversity. Conversely, low α -diversity does not exclude a transient pre-adaptive phase; it may simply reflect recruitment in a trophic landscape with limited donor availability or a plastid replacement (Ponce-Toledo et al., 2019; Sibbald and Archibald, 2020). In this sense, kleptoplasty provides the selective context for plastid-related gene recruitment, whereas ecological availability constrains how broad that recruited repertoire can become.

4. Toward a predictive framework for plastid integration

Sections 1 through 4 established that the α -diversity gradient is not artefactual (Methodological Results, Section 4), reflects genuine differences in plastid acquisition history (Discussion, Section 1), has a recognisable functional logic (Discussion, Section 2.1), and is shaped by trophic ecology (Discussion, Section 3). The question that remains is whether this pattern can be formalised into a predictive framework, one that not only describes the data but generates testable expectations about how α -diversity should behave under known biological conditions.

4.1. The gene reset as a natural perturbation

Plastid replacement in dinoflagellates provides an informative natural perturbation of the α -diversity signal. If α -diversity captures a biologically meaningful aspect of plastid acquisition, then replacing an established plastid should alter the donor profile in a directional way. This is what is observed: the peridinin-specific set of events lies in the high-diversity range, whereas both Kryptoperidiniaceae and

Karenia-specific sets fall in the low-diversity range after plastid replacement (Table 8). However, the stronger collapse in Kryptoperidiniaceae is plausibly linked to retention of a diatom-derived nucleus, which can continue to supply plastid-targeted functions directly (Hehenberger et al., 2014; Yamada et al., 2019, 2025). Nonetheless, in Karenia, this “reset” shows how α -diversity responds to plastid replacement.

4.2. Plastid integration as a genomic continuum

The Portable Plastid Hypothesis of Grzebyk et al., (2003) was formulated at a macroevolutionary scale, addressing why red plastids spread so successfully among eukaryotes (and not the green ones). The present results suggest a complementary genome-level perspective: donor α -diversity in nucpt# genes can be interpreted as a quantitative signature of plastid-associated genomic remodeling (Nosenko et al., 2006; Dorrell and Howe, 2015). Within this view, the three α -diversity regimes may represent different positions along a continuum rather than fixed categories. High-diversity regime is consistent with broad pre-adaptive recruitment under kleptoplastic transient trophic interactions; low-diversity with acquisition around a restricted donor pool or reset after plastid replacement; and mid-diversity with intermediate or mixed states. This interpretation remains tentative, but the gradient remains biologically informative and offers a coherent framework for comparing otherwise disparate plastid acquisition routes.

4.3. Testable predictions

This framework generates three immediate predictions. First, pre-adaptive gene sets should be functionally non-random and biased toward plastid exploitation and maintenance rather than deeper plastid metabolic integration (beyond photosynthesis). Second, the strength of the gene reset should scale with integration completeness, with more thoroughly integrated replacement plastids showing lower post-reset α -diversity. Third, newly characterized kleptoplastic lineages should tend to occupy a high-diversity regime or fall near it, unless ecological availability has strongly constrained the accessible donor pool. Taken together, these predictions do not establish a general law of plastid evolution, but they do show that the α -diversity gradient can be used as more than a descriptive pattern: it provides a comparative framework linking trophic exposure, functional pre-adaptation, and plastid integration.

5. Limitations & perspectives

Like any large-scale phylogenomic framework, the present analysis is constrained by both biological uncertainty and methodological simplification. These limitations do not invalidate the α -diversity signal recovered here, but they do define the scope within which it can be interpreted and point to the most important directions for future work.

First, the effort invested in curating, decontaminating, pooling and dereplicating MMETSP (Keeling et al., 2014) transcriptomes was not only useful for the analyses presented here, but also generated a reusable community resource. Its added value lies in reducing two major limitations of large protist transcriptome collections: cross-sample contamination and taxonomic redundancy. This relevance has already been recognized both through direct reuse in downstream studies, including plastid-focused analyses (Giustini et al., 2025); as well as applications outside a strictly photosynthetic context, such as surveys of eukaryotic transposons and salinity-associated proteins (Annenkova, 2024; Craig, 2022), and through methodological recommendations encouraging the use of reference transcriptome databases that have been screened for multi-species contamination and decontaminated (Cohen et al., 2022). In future iterations, this resource could be expanded and re-evaluated against newer reference initiatives, notably P10K database (Miao et al., 2020) (a data portal for the Protist 10,000 Genomes Project), which aims to provide genome resources across major protist clades, or more specific marine reference databases such as METdb (Niang et al., 2020), a marine microeukaryote transcriptome database built from MMETSP, Roscoff and Tara-derived resources, and eukaryotic protein resources such as EukProt (Richter et al., 2022), while remaining careful to apply comparable contamination screening and dereplication procedures before integration. This could greatly improve transfer inference by filling gaps in potentially missing donors and better define actual plastid-bearing lineage relationships, e.g., between cryptophytes and haptophytes. Such a candidate would definitely be leptophytes (Jamy et al., 2025), a deep-branching plastid lineage of nano-size algae known only from environmental metagenomes, whose plastid genome is recovered within a clade with cryptophytes and haptophytes, although their exact placement remains unresolved (Jamy et al., 2025). Additionally, early-branching non-photosynthetic relatives could be integrated, such as *Palpitomonas* (basal cryptophyte lineage; Yabuki et al., 2014), *Roombia* (katablepharids, sister group to cryptophytes; Okamoto et al., 2009), and *Telonemia* and *Centrohelea* (closely related to haptophytes; Burki et al., 2012). However, they would probably be most informative before transcriptome enrichment. They would provide deep homologous context around plastid-related orthoclans by improving the separation of ancient paralogues and refining orthoclan boundaries, such taxa could help discriminate between shared ancestral acquisitions and independent recruitment events that may currently appear collapsed within the same broad orthoclan.

Another limitation about donor representation is that cyanobacteria (and other bacteria) were excluded by design (not because they were not available) because an unbiased integration with respect to gene transfers would require us to represent all known prokaryote groups (Hambücker et al., 2026). We recognize that their absence (missing donors) can generate “phantom” transfers, in which the true donor is absent from the dataset and the receiver instead clusters with the nearest sampled paraphyletic lineage. This effect is likely to contribute to some of the additional events recovered under relaxed age filtering and especially on old lineages (like Cryptophyceae) that would be the most sensitive to this artifact. Relatedly, the age-category scheme used here should be viewed as a biologically motivated heuristic rather than a reconstruction of true plastid chronology, since it depends on uncertain molecular-clock estimates and debated scenarios of plastid acquisition. However, in a future validation step, a solution would be to classify phantom transfers with a simple but effective pipeline. Briefly, we could summarize an event to a HMMER profile (composed of the receiver clan sequences) to query a combined database of eukaryotes and prokaryotes, using the GTBD database (Parks et al., 2022). Events that firstly match prokaryotes would be candidates for phantom transfer. Such classification would also allow us to consider primary algae as receivers and eventually identify late transfers from complex algae.

Recent work on Kareniaceae further illustrates the limitations of donor inference. Although they are usually interpreted as a deeply integrated case of plastid replacement (Tengs et al., 2000), recent analyses of seven kareniacean species revealed highly chimeric plastid-targeted proteomes, a relatively limited shared haptophyte signal, and evidence that extant kareniacean plastids are not all derived from the same haptophyte lineage (Novák Vanclová et al., 2024). This group illustrates a key limitation of donor inference in ancient plastid acquisitions. In recently evolving lineages, distinct plastid donors may still be phylogenetically distinguishable (Novák Vanclová et al., 2024). Over longer evolutionary timescales, however, donor lineages may go extinct, remain unsampled, or be represented only by distant relatives. Genes originally acquired from different but related sources may therefore coalesce onto the closest sampled extant clade, producing an apparently simpler donor profile than the historical process that actually occurred. In this sense, Kareniaceae may provide a living analogue for processes that could have occurred during older plastid acquisitions but are now partly erased by extinction, sampling gaps and phylogenetic coalescence.

Another limitation concerns the algorithm implemented in PAGTI behind the event inference itself. Parameters such as intruder tolerance necessarily reflect a compromise rather than a universal optimum. Stricter settings improve concordance with expert opinion in some lineages, such as Cryptophyceae, but suppress a substantial part of the signal in others, such as Ochrophyta, where genuine events are more often embedded in complex trees. Likewise, plastid-targeting annotation remains imperfect in

many taxa, and KEGG-based functional reconstruction is limited by sparse KO-ID recovery. As a result, pathway-level interpretations are strongest in lineages with large and well-annotated event sets.

Dinophysiales illustrate this limitation particularly well. In the present dataset, the number of informative orthoclans and KO-ID assignments was too low to support a formal functional analysis comparable to those performed for euglenids or peridinin dinoflagellates. This results from both the scarcity of available species and the contamination-prone nature of *Dinophysis* transcriptomes, which complicates the distinction between genuine host-encoded plastid-support genes and prey- or intermediary-derived signal. In that sense, Dinophysiales currently mark one of the clearest limits of the functional framework developed here.

At the same time, Dinophysiales also emerge as a particularly promising direction for future work. Relative to the small number of nucpt# genes initially reported for *Dinophysis acuminata* and the broader but still functionally focused inventory later described for *D. fortii* and related taxa (Hongo et al., 2019; Wisecaver and Hackett, 2010), the present pipeline recovered a comparatively broad Dinophysiales-associated transfer signal. Importantly, this signal was dominated by cryptophyte-derived inferred events, in contrast to the stronger haptophyte/fucoxanthin emphasis highlighted by Hongo et al. (Hongo et al., 2019). Because these approaches quantify different objects, inferred transfer events here versus identified plastid-related gene products in the literature, this discrepancy cannot yet be interpreted directly. Nevertheless, it suggests that the plastid-associated nuclear repertoire of Dinophysiales may be broader, and perhaps more cryptophyte-influenced, than currently appreciated. Dinophysiales should therefore be considered a priority target for future analyses based on cleaner datasets, denser species sampling, and gene-by-gene phylogenetic validation.

As illustrated by Dinophysiales and *Rapaza*, contamination remains a persistent issue when studying gene transfer, and candidate transfers must be interpreted with caution. Unlike standard genomic analyses, it is not possible to decontaminate datasets *sensu stricto* when the objective is precisely to detect foreign sequences; one can only mitigate contamination by discarding the most affected sources, as performed in Chapter IV. This implies that every inferred transfer event must be confronted with the contamination profile of its source organism, as we did for freshly integrated transcriptomic data (*Rapaza viridis*, *Dinophysis ovum*, *Teleaulax amphioxeia*). This verification step could be automated within PAGTI: by reading the contamination reports previously generated by Forty-Two, PAGTI could, when inferring an event originating from a given source, flag it if that source is known to be contaminated by the donor lineage. Flagging is preferable to outright discarding, because contamination by a given lineage does not exclude that a genuine transfer also originates from that same lineage. Discriminating between the two would require additional evidence, such as branch length analysis: a

near-zero branch length indicates virtually identical proteins and thus points to genomic or transcriptomic contamination rather than genuine transfer.

Finally, another line of improvement concerns plastid-targeting annotations that were derived from published, peer-reviewed predictions rather than generated *de novo*. Specifically, we relied in part on the localisation calls produced by (Dorrell et al., 2017) using SignalP v3.0 (Dyrlov Bendtsen et al., 2004), ASAFind v2.0 (Gruber et al., 2015) and HECTAR (Gschloessl et al., 2008) to tag sequences in our reference proteomes (see Methods in Chapter III). This conservative approach was intended to limit the number of methodological variables in the early stages of pipeline development. Now that the phylogenomic pipeline has been refined and benchmarked across both datasets, incorporating an independent layer of *de novo* subcellular localisation prediction would be a natural next step. For instance, screening all plastid-associated orthoclans through a multi-label predictor such as DeepLoc 2.1 (Ødum et al., 2024), as suggested by R. Dorrell, which assigns localisation probabilities across ten compartments and additionally detects sorting signals, could provide an orthogonal line of evidence to corroborate or challenge the phylogenetically inferred transfers.

Finally, Viridiplantae were recovered as the second major donor lineage after Ochrophyta among red-complex algae. This raises the question of whether green-derived contributions are enriched in particular functional categories, or whether they instead represent a more diffuse background of transfers across plastid-associated pathways. Testing this would help determine whether the Viridiplantae are involved in a specific manner in red complex plastid lineages.

More generally, the framework developed in this work would benefit from improved annotation of plastid-targeted genes, denser and more balanced taxonomic sampling. In that sense, the present study should be regarded as a first quantitative approximation rather than a final account of plastid-associated gene transfer. Its main value lies in showing that donor α -diversity can be measured reproducibly, interpreted biologically, both in evolutionary and functional terms, and used to generate testable expectations across lineages with contrasting plastid histories.

6. Closing statement

Recent work seems to blur the boundary between kleptoplasty and canonical endosymbiosis. This is particularly intriguing in dinotoms, where closely related Kryptoperidiniaceae span a continuum of plastid integration states, from transient kleptoplastidy in *Durinskia capensis* to an early permanent state in *D. kwazulunatalensis* and more established permanent diatom-derived plastids in taxa such as *Durinskia baltica* and *Kryptoperidinium foliaceum* (Yamada et al., 2025). Comparative transcriptomic analyses show that, even in these permanent or near-permanent systems, the diatom nucleus can retain substantial transcriptional autonomy while displaying early signs of host influence, genomic reconfiguration and metabolic integration, supporting a mechanistic continuum from organelle theft to plastid permanence rather than a sharp boundary between kleptoplasty and endosymbiosis (Yamada et al., 2025).

These observations invite a reassessment of what constitutes a canonical endosymbiosis in dinoflagellates. The permanent diatom-derived plastid of *Kryptoperidinium foliaceum* (now named *Kryptoperidinium secundum*) and *Durinskia baltica* is widely treated as an established tertiary endosymbiont, yet several features complicate this interpretation. Closely related Kryptoperidiniaceae such as *D. capensis* are kleptoplastic, acquiring diatom organelles through cytoklepty (kleptoplasty and karyoklepty in the same host derived compartment) via myzocytosis (Yamada et al., 2019). Importantly, in all dinotoms, regardless of integration degree, the internalised diatom retains all of its organelles except the plasma membrane and cell wall (Yamada et al., 2025), a topology consistent with prey lysis during feeding rather than uptake of an intact symbiont. The retained diatom kleptokaryon is not a vestigial relic: it remains largely unreduced and transcriptionally autonomous (Yamada et al., 2025), and nucleus retention (karyoklepty) is also observed in the kleptoplastic *D. capensis*. A parallel argument applies to the peridinin plastid, which is bounded by three membranes, a count more consistent with kleptoplastic acquisition than with canonical endosymbiosis, a possibility explicitly raised by Bodyl (2018) and Moog and Maier (2018). Taken together, these cases suggest that some plastid acquisitions currently classified as endosymbiotic events may represent stabilised kleptoplastic associations, and that the boundary between organelle theft and endosymbiosis is less clear-cut than traditionally assumed.

VII. General conclusion

In this thesis, we developed a controlled phylogenomic framework to study plastid-related gene mosaicism in complex algae. Because this work ultimately aimed to study gene transfers, dataset construction was a central methodological issue. We sought to maximize taxonomic sampling so as to recover as many potential donor-receiver pairs as possible, while placing particular emphasis on data quality through explicit control of completeness and contamination. This concern was essential because, in transfer analyses, contaminant sequences do not merely add noise: they can be mistaken for genuine transfers and thus produce fundamentally misleading results, as illustrated by several striking cases in this manuscript.

A major strength of this work is that it makes the exploration of plastid-related gene histories scalable. Because many parameters can influence transfer inference, automation was not only a technical convenience but a methodological necessity: it allowed alternative analytical configurations to be tested while keeping the interpretation engine constant. In that sense, one of the main outcomes of this thesis is the methodology. Beyond the specific biological results presented here, it establishes a reusable system for querying plastid-related gene evolution under controlled and reproducible conditions. The rich orthoclan dataset and associated pipeline are not tied to a single hypothesis or a single taxonomic framework; they can easily be expanded with new species, as shown for *Rapaza viridis* and *Dinophysis* spp., reinterpreted under alternative taxonomic partitions, and queried from multiple complementary angles.

Applied to complex algae, this framework confirms that plastid-targeted gene repertoires often display substantial and structured donor mosaicism. This was the central biological question of the thesis: what can explain the widespread mosaic signal observed across complex algal lineages, including lineages that would initially seem more compatible with relatively simple endosymbiotic scenarios? The results indicate that no single, overly simple explanation is sufficient. Plastid-related gene evolution appears to reflect a plurality of processes, whose relative contributions likely vary among lineages. Donor diversity, donor affiliation, and functional patterns all suggest that plastid acquisition histories were often more composite than expected under a universal model.

Within this broader question, kleptoplasty served as the main interpretative entry point of this thesis. This choice was initially hypothesis-driven, but it proved heuristically productive. When this project began, kleptoplasty was already present in the literature as a serious conceptual possibility, notably in earlier work by A. Bodył, but it had not yet acquired the same degree of explicit formulation and empirical traction that it gained in later years. The core rationale remains that kleptoplasty provides a

biologically plausible context for strong selection on plastid-related gene acquisition and retention: a stolen plastid is transient and functionally constrained, so any host-encoded contribution that prolongs its maintenance may become selectively advantageous. This does not mean that kleptoplasty explains all cases, nor that endosymbiotic models should all be rejected. Rather, it suggests that kleptoplasty deserves to be considered a credible evolutionary mechanism contributing to plastid-related gene diversity, alongside more classical endosymbiotic processes for which the selective pressure remains difficult to conceive. In that sense, the initial hypothesis may have been too strong in some of its formulations: elevated donor diversity cannot be treated as a one-to-one signature of kleptoplasty. Yet it was directionally useful in revealing how widespread complex, potentially kleptoplasty-compatible signals may be.

At the same time, this work also makes clear that plastid evolution resists overly rigid interpretations. Yet this complexity does not weaken the framework developed here. On the contrary, it reinforces its value: the dataset persists, the tools remain usable, and the same analytical logic can be applied to future discoveries. Only a fraction of the accessible biological questions has been explored in this thesis. The case studies developed here, notably in euglenids, dinoflagellates, and ochrophyte-related signals, therefore represent not an endpoint, but a first set of demonstrations of what this analytical framework makes possible.

More generally, the main contribution of this thesis is not only to provide answers, but to provide a way to ask questions rigorously. It shows that plastid-related gene mosaicism can be studied quantitatively, phylogenetically, and comparatively at scale, and that this signal contains meaningful information about the evolutionary processes underlying plastid acquisition in complex algae.

VIII. Contributions

The vast majority of the code and analyses presented in this manuscript were developed by me. When this was not the case, they were produced in close collaboration with my PhD advisor, Prof. Denis Baurain. The experimental design (including control analyses), biological interpretations, and overall narrative were extensively discussed between us; however, I maintained a leading role and retained final responsibility for the manuscript. For transparency, I detail below the specific components for which I received substantial contributions.

Use of AI tools

During the preparation of this thesis, I used ChatGPT (OpenAI) as a support tool for grammatical correction, reformulation, logical structuring of arguments, and assistance with the use of R packages. I also used Consensus to identify relevant scientific references and extract specific information from the literature. The text/code/ images in this thesis are my own (unless otherwise specified). I have reviewed and edited the content as needed, and I take full responsibility for the content of the thesis.

Chapter 2 - Forty-Two

D. Baurain developed the core codebase, while I contributed substantially to the taxonomic methods implemented in the Bio::MUST::Core modules, as well as to unit testing. I also wrote scripts for run debriefing, preventing multiple integration of identical transcripts across MSAs, and generating YAML configuration files.

H. Philippe provided manually curated eukaryotic ribosomal MSAs. I performed taxonomic standardization (i.e., file and sequence renaming) to enable their use in our contamination detection framework.

Chapter 3 - Orthoclan assembly and characterization

This chapter was published with me as first author, reflecting my central role in its design and execution: **Van Vlierberghe, M.,** H. Philippe and D. Baurain (2021). *Broadly sampled orthologous groups of eukaryotic proteins for the phylogenetic study of plastid-bearing lineages*. BMC Research Notes 14(1): 143.

Chapter 4 - Decontamination and consolidation of MMETSP datasets

This chapter was also published with me as first author:

Van Vlierberghe, M., A. Di Franco, H. Philippe and D. Baurain (2021). *Decontamination, pooling and dereplication of the 678 samples of the Marine Microbial Eukaryote Transcriptome Sequencing Project*. BMC Research Notes 14(1): 306.

Chapter 5 - Main study

D. Baurain developed the code and performed the KEGG functional analyses (following eggNOG-mapper annotation). I designed the sets of events to be analyzed, reorganized pathways into KEGG-inspired categories, and derived all biological interpretations.

H. Brinkmann conducted an expert analysis of 242 single-gene trees to assess the performance of my transfer detection algorithms. He also assisted with species subsampling, based on my ribosomal phylogenomic tree. In addition, we had numerous insightful discussions related to my PhD topic.

Over the course of my PhD, I had the opportunity to collaborate with all members of the Unit of Eukaryotic Phylogenomics led by D. Baurain. I primarily contributed as a specialist in ortholog inference and enrichment, contamination detection, and Perl development. Below is a complete list of articles and preprints on which I am a co-author.

- Durieu, B., V. Savaglia, **M. Van Vlierberghe**, V. Lupo, D. Baurain, A. Wilmotte and L. Cornet (2025). *Out of Antarctica: new insights into Antarctic Subcluster 5.2 picocyanobacteria based on high-quality genomes*. bioRxiv: 2025.2012. 2017.694815.
- Giustini, C., D. Dal Bo, M. Storti, **M. Van Vlierberghe**, D. Baurain, P. Cardol, Y. Zhang, A. R. Fernie, D. Fitzpatrick, E. M. Aro, G. Allorete, P. Albanese, D. Tolleter, G. Curien and G. Finazzi (2025). *A Mitochondrially Related Plastidial Transporter Regulates Photosynthesis in the Diatom Phaeodactylum tricornutum*. Physiol Plant 177(6): e70640.
- Lupo, V., **M. Van Vlierberghe**, H. Vanderschuren, F. Kerff, D. Baurain and L. Cornet (2021). *Contamination in Reference Sequence Databases: Time for Divide-and-Rule Tactics*. Front Microbiol 12: 755101.
- Cordoba, J., E. Perez, **M. Van Vlierberghe**, A. R. Bertrand, V. Lupo, P. Cardol and D. Baurain (2021). *De Novo Transcriptome Meta-Assembly of the Mixotrophic Freshwater Microalga Euglena gracilis*. Genes (Basel) 12(6).

- Leonard, R. R., M. Leleu, **M. Van Vlierberghe**, L. Cornet, F. Kerff and D. Baurain (2021). *ToRQuEMaDA: tool for retrieving queried Eubacteria, metadata and dereplicating assemblies*. PeerJ 9: e11348.
- Cornet, L., L. Meunier, **M. Van Vlierberghe**, R. R. Leonard, B. Durieu, Y. Lara, A. Misztak, D. Sirjacobs, E. J. Javaux, H. Philippe, A. Wilmotte and D. Baurain (2018). *Consensus assessment of the contamination level of publicly available cyanobacterial genomes*. PLoS One 13(7): e0200323.

I also contributed to several additional projects, some of them extensively, although they have not yet reached the publication stage. The most notable is the work of M. Leleu on the MAT hypothesis for the primary endosymbiosis of plastids.

IX. REFERENCES

- Allen, J.F., 2017. The CoRR hypothesis for genes in organelles. *Journal of Theoretical Biology*, The origin of mitosing cells: 50th anniversary of a classic paper by Lynn Sagan (Margulis) 434, 50–57. <https://doi.org/10.1016/j.jtbi.2017.04.008>
- Allen, J.F., Raven, J.A., 1996. Free-radical-induced mutation vs redox regulation: Costs and benefits of genes in organelles. *J Mol Evol* 42, 482–492. <https://doi.org/10.1007/BF02352278>
- Altenburger, A., Blossom, H.E., Garcia-Cuetos, L., Jakobsen, H.H., Carstensen, J., Lundholm, N., Hansen, P.J., Moestrup, Ø., Haraguchi, L., 2020. Dimorphism in cryptophytes—The case of *Teleaulax amphioxeia*/*Plagioselmis prolunga* and its ecological implications. *Science Advances* 6, eabb1611. <https://doi.org/10.1126/sciadv.abb1611>
- Altenhoff, A.M., Dessimoz, C., 2009. Phylogenetic and Functional Assessment of Orthologs Inference Projects and Methods. *PLOS Computational Biology* 5, e1000262. <https://doi.org/10.1371/journal.pcbi.1000262>
- Altenhoff, A.M., Glover, N.M., Dessimoz, C., 2019. Inferring Orthology and Paralogy, in: Anisimova, M. (Ed.), *Evolutionary Genomics: Statistical and Computational Methods*. Springer, New York, NY, pp. 149–175. https://doi.org/10.1007/978-1-4939-9074-0_5
- Altenhoff, A.M., Studer, R.A., Robinson-Rechavi, M., Dessimoz, C., 2012. Resolving the Ortholog Conjecture: Orthologs Tend to Be Weakly, but Significantly, More Similar in Function than Paralogs. *PLOS Computational Biology* 8, e1002514. <https://doi.org/10.1371/journal.pcbi.1002514>
- Angly, F.E., Fields, C.J., Tyson, G.W., 2014. The Bio-Community Perl toolkit for microbial ecology. *Bioinformatics* 30, 1926–1927. <https://doi.org/10.1093/bioinformatics/btu130>
- Annenkova, N.V., 2024. Proteins Associated with Salinity Adaptation of the Dinoflagellates: Diversity and Potential Involvement in Species Evolution. *Diversity* 16, 739. <https://doi.org/10.3390/d16120739>
- Anschütz, A.-A., Flynn, K.J., Mitra, A., 2022. Acquired Phototrophy and Its Implications for Bloom Dynamics of the *Teleaulax*-*Mesodinium*-*Dinophysis*-Complex. *Front. Mar. Sci.* 8. <https://doi.org/10.3389/fmars.2021.799358>
- Archibald, J.M., 2015. Genomic perspectives on the birth and spread of plastids. *Proc. Natl. Acad. Sci. U.S.A.* 112, 10147–10153. <https://doi.org/10.1073/pnas.1421374112>
- Archibald, J.M., 2009. The Puzzle of Plastid Evolution. *Current Biology* 19, R81–R88. <https://doi.org/10.1016/j.cub.2008.11.067>
- Archibald, J.M., 2007. Nucleomorph genomes: structure, function, origin and evolution. *BioEssays* 29, 392–402. <https://doi.org/10.1002/bies.20551>
- Archibald, J.M., Keeling, P.J., 2002. Recycled plastids: a ‘green movement’ in eukaryotic evolution. *Trends in Genetics* 18, 577–584. [https://doi.org/10.1016/S0168-9525\(02\)02777-4](https://doi.org/10.1016/S0168-9525(02)02777-4)

Archibald, J.M., Rogers, M.B., Toop, M., Ishida, K., Keeling, P.J., 2003. Lateral gene transfer and the evolution of plastid-targeted proteins in the secondary plastid-containing alga *Bigeloviella natans*. *Proc. Natl. Acad. Sci. U.S.A.* 100, 7678–7683. <https://doi.org/10.1073/pnas.1230951100>

Azuma, T., Pánek, T., Tice, A.K., Kayama, M., Kobayashi, M., Miyashita, H., Suzuki, T., Yabuki, A., Brown, M.W., Kamikawa, R., 2022. An Enigmatic Stramenopile Sheds Light on Early Evolution in Ochrophyta Plastid Organogenesis. *Molecular Biology and Evolution* 39, msac065. <https://doi.org/10.1093/molbev/msac065>

Ball, S., Baurain, D., Leleu, M., Lafontaine, I., Cenci, U., Colleoni, C., Vallon, O., Greub, G., Weber, A., Bhattacharya, D., 2025a. None. *Comptes Rendus. Biologies* 348, 107–135. <https://doi.org/10.5802/crbio.176>

Ball, S., Baurain, D., Leleu, M., Lafontaine, I., Cenci, U., Colleoni, C., Vallon, O., Greub, G., Weber, A., Bhattacharya, D., 2025b. Obligate intracellular bacterial pathogens as major players in the metabolic integration of organelles. *Comptes Rendus. Biologies* 348, 107–135. <https://doi.org/10.5802/crbio.176>

Bapteste, E., Brinkmann, H., Lee, J.A., Moore, D.V., Sensen, C.W., Gordon, P., Duruflé, L., Gaasterland, T., Lopez, P., Müller, M., Philippe, H., 2002. The analysis of 100 genes supports the grouping of three highly divergent amoebae: *Dictyostelium*, *Entamoeba*, and *Mastigamoeba*. *Proceedings of the National Academy of Sciences* 99, 1414–1419. <https://doi.org/10.1073/pnas.032662799>

Barbrook, A.C., Howe, C.J., Nisbet, R.E.R., 2019. Breaking up is hard to do: the complexity of the dinoflagellate chloroplast genome. *Perspectives in Phycology* 31–37. <https://doi.org/10.1127/pip/2018/0084>

Barbrook, A.C., Howe, C.J., Purton, S., 2006. Why are plastid genomes retained in non-photosynthetic organisms? *Trends in Plant Science* 11, 101–108. <https://doi.org/10.1016/j.tplants.2005.12.004>

Baum, D.A., Offner, S., 2008. Phylogenics & Tree-Thinking. *ambt* 70, 222–229. [https://doi.org/10.1662/0002-7685\(2008\)70%5B222:PT%5D2.0.CO;2](https://doi.org/10.1662/0002-7685(2008)70%5B222:PT%5D2.0.CO;2)

Baurain, D., Brinkmann, H., Petersen, J., Rodriguez-Ezpeleta, N., Stechmann, A., Demoulin, V., Roger, A.J., Burger, G., Lang, B.F., Philippe, H., 2010. Phylogenomic Evidence for Separate Acquisition of Plastids in Cryptophytes, Haptophytes, and Stramenopiles. *Molecular Biology and Evolution* 27, 1698–1709. <https://doi.org/10.1093/molbev/msq059>

Baurain, D., Brinkmann, H., Philippe, H., 2007. Lack of Resolution in the Animal Phylogeny: Closely Spaced Cladogeneses or Undetected Systematic Errors? *Mol Biol Evol* 24, 6–9. <https://doi.org/10.1093/molbev/msl137>

Bernabeu, M., Gabaldón, T., 2026. Host retargeting predominated over gene transfer during early chromatophore integration in *Paulinella micropora*. <https://doi.org/10.64898/2026.04.14.718432>

- Bhattacharya, D., Archibald, J.M., Weber, A.P.M., Reyes-Prieto, A., 2007. How do endosymbionts become organelles? Understanding early events in plastid evolution. *BioEssays*. <https://doi.org/10.1002/bies.20671>
- Bodył, A., 2018. Did some red alga-derived plastids evolve via kleptoplastidy? A hypothesis. *Biological Reviews* 93, 201–222. <https://doi.org/10.1111/brv.12340>
- Bodył, A., Mackiewicz, P., Stiller, J.W., 2009a. Early steps in plastid evolution: current ideas and controversies. *BioEssays* 31, 1219–1232. <https://doi.org/10.1002/bies.200900073>
- Bodył, A., Moszczyński, K., 2006. Did the peridinin plastid evolve through tertiary endosymbiosis? A hypothesis. *European Journal of Phycology* 41, 435–448. <https://doi.org/10.1080/09670260600961080>
- Bodył, A., Stiller, J.W., Mackiewicz, P., 2009b. Chromalveolate plastids: direct descent or multiple endosymbioses? *Trends in Ecology & Evolution* 24, 119–121. <https://doi.org/10.1016/j.tree.2008.11.003>
- Bolte, K., Bullmann, L., Hempel, F., Bozarth, A., Zauner, S., Maier, U.-G., 2009. Protein Targeting into Secondary Plastids. *Journal of Eukaryotic Microbiology* 56, 9–15. <https://doi.org/10.1111/j.1550-7408.2008.00370.x>
- Boothby, T.C., Tenlen, J.R., Smith, F.W., Wang, J.R., Patanella, K.A., Osborne Nishimura, E., Tintori, S.C., Li, Q., Jones, C.D., Yandell, M., Messina, D.N., Glasscock, J., Goldstein, B., 2015. Evidence for extensive horizontal gene transfer from the draft genome of a tardigrade. *Proc. Natl. Acad. Sci. U.S.A.* 112, 15976–15981. <https://doi.org/10.1073/pnas.1510461112>
- Burki, F., Imanian, B., Hehenberger, E., Hirakawa, Y., Maruyama, S., Keeling, P.J., 2014. Endosymbiotic Gene Transfer in Tertiary Plastid-Containing Dinoflagellates. *Eukaryotic Cell* 13, 246–255. <https://doi.org/10.1128/ec.00299-13>
- Burki, F., Kaplan, M., Tikhonenkov, D.V., Zlatogursky, V., Minh, B.Q., Radaykina, L.V., Smirnov, A., Mylnikov, A.P., Keeling, P.J., 2016. Untangling the early diversification of eukaryotes: a phylogenomic study of the evolutionary origins of Centrohelida, Haptophyta and Cryptista. *Proc Biol Sci* 283, 20152802. <https://doi.org/10.1098/rspb.2015.2802>
- Burki, F., Okamoto, N., Pombert, J.-F., Keeling, P.J., 2012. The evolutionary history of haptophytes and cryptophytes: phylogenomic evidence for separate origins. *Proc. R. Soc. B.* 279, 2246–2254. <https://doi.org/10.1098/rspb.2011.2301>
- Burki, F., Roger, A.J., Brown, M.W., Simpson, A.G.B., 2020. The New Tree of Eukaryotes. *Trends in Ecology & Evolution* 35, 43–55. <https://doi.org/10.1016/j.tree.2019.08.008>
- Cantalapiedra, C.P., Hernández-Plaza, A., Letunic, I., Bork, P., Huerta-Cepas, J., 2021. eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. *Mol Biol Evol* 38, 5825–5829. <https://doi.org/10.1093/molbev/msab293>

Cavalier-Smith, T., 2003. Genomic reduction and evolution of novel genetic membranes and protein-targeting machinery in eukaryote-eukaryote chimaeras (meta-algae). *Phil. Trans. R. Soc. Lond. B* 358, 109–134. <https://doi.org/10.1098/rstb.2002.1194>

Cavalier-Smith, T., 1999. Principles of Protein and Lipid Targeting in Secondary Symbiogenesis: Euglenoid, Dinoflagellate, and Sporozoan Plastid Origins and the Eukaryote Family Tree ¹, ². *J Eukaryotic Microbiology* 46, 347–366. <https://doi.org/10.1111/j.1550-7408.1999.tb04614.x>

Cavalier-Smith, T., 1981. Eukaryote kingdoms: Seven or nine? *Biosystems* 14, 461–481. [https://doi.org/10.1016/0303-2647\(81\)90050-2](https://doi.org/10.1016/0303-2647(81)90050-2)

Chao, A., Chiu, C.-H., Jost, L., 2010. Phylogenetic diversity measures based on Hill numbers. *Philos Trans R Soc Lond B Biol Sci* 365, 3599–3609. <https://doi.org/10.1098/rstb.2010.0272>

Chao, A., Gotelli, N.J., Hsieh, T.C., Sander, E.L., Ma, K.H., Colwell, R.K., Ellison, A.M., 2014. Rarefaction and extrapolation with Hill numbers: a framework for sampling and estimation in species diversity studies. *Ecological Monographs* 84, 45–67. <https://doi.org/10.1890/13-0133.1>

Cho, A., Lax, G., Keeling, P.J., 2024. Phylogenomic analyses of ochrophytes (stramenopiles) with an emphasis on neglected lineages. *Molecular Phylogenetics and Evolution* 198, 108120. <https://doi.org/10.1016/j.ympev.2024.108120>

Cohen, N.R., Alexander, H., Krinos, A.I., Hu, S.K., Lampe, R.H., 2022. Marine Microeukaryote Metatranscriptomics: Sample Processing and Bioinformatic Workflow Recommendations for Ecological Applications. *Front. Mar. Sci.* 9, 867007. <https://doi.org/10.3389/fmars.2022.867007>

Cole, L.W., 2016. The Evolution of Per-cell Organelle Number. *Front. Cell Dev. Biol.* 4. <https://doi.org/10.3389/fcell.2016.00085>

Cordoba, J., Perez, E., Van Vlierberghe, M., Bertrand, A.R., Lupo, V., Cardol, P., Baurain, D., 2021. De Novo Transcriptome Meta-Assembly of the Mixotrophic Freshwater Microalga *Euglena gracilis*. *Genes* 12, 842. <https://doi.org/10.3390/genes12060842>

Cornet, L., Baurain, D., 2022a. Contamination detection in genomic data: more is not enough. *Genome Biol* 23, 60. <https://doi.org/10.1186/s13059-022-02619-9>

Cornet, L., Baurain, D., 2022b. Contamination detection in genomic data: more is not enough. *Genome Biol* 23, 60. <https://doi.org/10.1186/s13059-022-02619-9>

Cornet, L., Lupo, V., Declerck, S., Baurain, D., 2024. Evaluation of Genomic Contamination Detection Tools and Influence of Horizontal Gene Transfer on Their Efficiency through Contamination Simulations at Various Taxonomic Ranks. *Applied Microbiology* 4, 124–132. <https://doi.org/10.3390/applmicrobiol4010009>

Cornet, L., Magain, N., Baurain, D., Lutzoni, F., 2021. Exploring syntenic conservation across genomes for phylogenetic studies of organisms subjected to horizontal gene transfers: A case study with Cyanobacteria and cyanolichens. *Molecular Phylogenetics and Evolution* 162, 107100. <https://doi.org/10.1016/j.ympev.2021.107100>

Cornet, L., Meunier, L., Vlierberghe, M.V., Léonard, R.R., Durieu, B., Lara, Y., Misztak, A., Sirjacobs, D., Javaux, E.J., Philippe, H., Wilmotte, A., Baurain, D., 2018. Consensus assessment of the contamination level of publicly available cyanobacterial genomes. *PLOS ONE* 13, e0200323. <https://doi.org/10.1371/journal.pone.0200323>

Craig, R.J., 2022. Replitrans: a new group of eukaryotic transposons encoding HUH endonuclease. <https://doi.org/10.1101/2022.12.15.520654>

Cruz, S., Cartaxana, P., 2022. Kleptoplasty: Getting away with stolen chloroplasts. *PLOS Biology* 20, e3001857. <https://doi.org/10.1371/journal.pbio.3001857>

Curtis, B.A., Tanifuji, G., Burki, F., Gruber, A., Irimia, M., Maruyama, S., Arias, M.C., Ball, S.G., Gile, G.H., Hirakawa, Y., Hopkins, J.F., Kuo, A., Rensing, S.A., Schmutz, J., Symeonidi, A., Elias, M., Eveleigh, R.J.M., Herman, E.K., Klute, M.J., Nakayama, T., Oborník, M., Reyes-Prieto, A., Armbrust, E.V., Aves, S.J., Beiko, R.G., Coutinho, P., Dacks, J.B., Durnford, D.G., Fast, N.M., Green, B.R., Grisdale, C.J., Hempel, F., Henrissat, B., Höppner, M.P., Ishida, K.-I., Kim, E., Kořený, L., Kroth, P.G., Liu, Y., Malik, S.-B., Maier, U.G., McRose, D., Mock, T., Neilson, J.A.D., Onodera, N.T., Poole, A.M., Pritham, E.J., Richards, T.A., Rocap, G., Roy, S.W., Sarai, C., Schaack, S., Shirato, S., Slamovits, C.H., Spencer, D.F., Suzuki, S., Worden, A.Z., Zauner, S., Barry, K., Bell, C., Bharti, A.K., Crow, J.A., Grimwood, J., Kramer, R., Lindquist, E., Lucas, S., Salamov, A., McFadden, G.I., Lane, C.E., Keeling, P.J., Gray, M.W., Grigoriev, I.V., Archibald, J.M., 2012. Algal genomes reveal evolutionary mosaicism and the fate of nucleomorphs. *Nature* 492, 59–65. <https://doi.org/10.1038/nature11681>

Dalquen, D.A., Dessimoz, C., 2013. Bidirectional Best Hits Miss Many Orthologs in Duplication-Rich Clades such as Plants and Animals. *Genome Biol Evol* 5, 1800–1806. <https://doi.org/10.1093/gbe/evt132>

de Reviers, B., 2002. Biologie et phylogénie des algues.

Delsuc, F., Philippe, H., Tsagkogeorga, G., Simion, P., Tilak, M.-K., Turon, X., López-Legentil, S., Piette, J., Lemaire, P., Douzery, E.J.P., 2018. A phylogenomic framework and timescale for comparative studies of tunicates. *BMC Biol* 16, 39. <https://doi.org/10.1186/s12915-018-0499-2>

Delwiche, C.F., 1999. Tracing the Thread of Plastid Diversity through the Tapestry of Life. *The American Naturalist* 154, S164–S177. <https://doi.org/10.1086/303291>

Deschamps, P., Moreira, D., 2012. Reevaluating the Green Contribution to Diatom Genomes. *Genome Biology and Evolution* 4, 683–688. <https://doi.org/10.1093/gbe/evs053>

Di Franco, A., Baurain, D., Glöckner, G., Melkonian, M., Philippe, H., 2022. Lower Statistical Support with Larger Data Sets: Insights from the Ochrophyta Radiation. *Molecular Biology and Evolution* 39. <https://doi.org/10.1093/molbev/msab300>

Donoghue, P.C.J., Kay, C., Spang, A., Szöllösi, G., Nenarokova, A., Moody, E.R.R., Pisani, D., Williams, T.A., 2023. Defining eukaryotes to dissect eukaryogenesis. *Current Biology* 33, R919–R929. <https://doi.org/10.1016/j.cub.2023.07.048>

Doolittle, W.F., Boucher, Y., Nesbø, C.L., Douady, C.J., Andersson, J.O., Roger, A.J., 2003. How big is the iceberg of which organellar genes in nuclear genomes are but the tip? *Philos Trans R Soc Lond B Biol Sci* 358, 39–58. <https://doi.org/10.1098/rstb.2002.1185>

Dorrell, R.G., Gile, G., McCallum, G., Méheust, R., Baptiste, E.P., Klinger, C.M., Brillet-Guéguen, L., Freeman, K.D., Richter, D.J., Bowler, C., 2017. Chimeric origins of ochrophytes and haptophytes revealed through an ancient plastid proteome. *eLife* 6, e23717. <https://doi.org/10.7554/eLife.23717>

Dorrell, R.G., Howe, C.J., 2015. Integration of plastids with their hosts: Lessons learned from dinoflagellates. *Proc. Natl. Acad. Sci. U.S.A.* 112, 10247–10254. <https://doi.org/10.1073/pnas.1421380112>

Dorrell, R.G., Smith, A.G., 2011. Do Red and Green Make Brown?: Perspectives on Plastid Acquisitions within Chromalveolates. *Eukaryot Cell* 10, 856–868. <https://doi.org/10.1128/EC.00326-10>

Dorrell, R.G., Villain, A., Perez-Lamarque, B., Audren de Kerdrel, G., McCallum, G., Watson, A.K., Ait-Mohamed, O., Alberti, A., Corre, E., Frischkorn, K.R., Pierella Karlusich, J.J., Pelletier, E., Morlon, H., Bowler, C., Blanc, G., 2021. Phylogenomic fingerprinting of tempo and functions of horizontal gene transfer within ochrophytes. *Proceedings of the National Academy of Sciences* 118, e2009974118. <https://doi.org/10.1073/pnas.2009974118>

Douglas, S., Zauner, S., Fraunholz, M., Beaton, M., Penny, S., Deng, L.-T., Wu, X., Reith, M., Cavalier-Smith, T., Maier, U.-G., 2001. The highly reduced genome of an enslaved algal nucleus. *Nature* 410, 1091–1096. <https://doi.org/10.1038/35074092>

Dyall, S.D., Brown, M.T., Johnson, P.J., 2004. Ancient Invasions: From Endosymbionts to Organelles. *Science* 304, 253–257. <https://doi.org/10.1126/science.1094884>

Dyrlov Bendtsen, J., Nielsen, H., Von Heijne, G., Brunak, S., 2004. Improved Prediction of Signal Peptides: SignalP 3.0. *Journal of Molecular Biology* 340, 783–795. <https://doi.org/10.1016/j.jmb.2004.05.028>

Ebersberger, I., Strauss, S., Von Haeseler, A., 2009. HaMStR: Profile hidden markov model based search for orthologs in ESTs. *BMC Evol Biol* 9, 157. <https://doi.org/10.1186/1471-2148-9-157>

Eddy, S.R., 1998. Profile hidden Markov models. *Bioinformatics* 14, 755–763. <https://doi.org/10.1093/bioinformatics/14.9.755>

Elias, M., Archibald, J.M., 2009. Sizing up the genomic footprint of endosymbiosis. *BioEssays* 31, 1273–1279. <https://doi.org/10.1002/bies.200900117>

Emms, D.M., Kelly, S., 2015. OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biol* 16, 157. <https://doi.org/10.1186/s13059-015-0721-2>

Enright, A.J., 2002. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Research* 30, 1575–1584. <https://doi.org/10.1093/nar/30.7.1575>

Eren, A.M., Kiefl, E., Shaiber, A., Veseli, I., Miller, S.E., Schechter, M.S., Fink, I., Pan, J.N., Yousef, M., Fogarty, E.C., Trigodet, F., Watson, A.R., Esen, Ö.C., Moore, R.M., Clayssen, Q., Lee, M.D., Kivenson, V., Graham, E.D., Merrill, B.D., Karkman, A., Blankenberg, D., Eppley, J.M., Sjödin, A., Scott, J.J., Vázquez-Campos, X., McKay, L.J., McDaniel, E.A., Stevens, S.L.R., Anderson, R.E., Fuessel, J., Fernandez-Guerra, A., Maignien, L., Delmont, T.O., Willis, A.D., 2021. Community-led, integrated, reproducible multi-omics with anvi'o. *Nat Microbiol* 6, 3–6. <https://doi.org/10.1038/s41564-020-00834-3>

Fast, N.M., Kissinger, J.C., Roos, D.S., Keeling, P.J., 2001. Nuclear-Encoded, Plastid-Targeted Genes Suggest a Single Common Origin for Apicomplexan and Dinoflagellate Plastids. *Molecular Biology and Evolution* 18, 418–426. <https://doi.org/10.1093/oxfordjournals.molbev.a003818>

Felsner, G., Sommer, M.S., Gruenheit, N., Hempel, F., Moog, D., Zauner, S., Martin, W., Maier, U.G., 2011. ERAD Components in Organisms with Complex Red Plastids Suggest Recruitment of a Preexisting Protein Transport Pathway for the Periplastid Membrane. *Genome Biology and Evolution* 3, 140–150. <https://doi.org/10.1093/gbe/evq074>

Fitch, W.M., 1970. Distinguishing Homologous from Analogous Proteins. *Syst Biol* 19, 99–113. <https://doi.org/10.2307/2412448>

Ford Doolittle, W., 1998. You are what you eat: a gene transfer ratchet could account for bacterial genes in eukaryotic nuclear genomes. *Trends in Genetics* 14, 307–311. [https://doi.org/10.1016/S0168-9525\(98\)01494-2](https://doi.org/10.1016/S0168-9525(98)01494-2)

Füssy, Z., Oborník, M., 2017. Chromerids and Their Plastids, in: *Advances in Botanical Research*. Elsevier, pp. 187–218. <https://doi.org/10.1016/bs.abr.2017.07.001>

Gabaldón, T., Koonin, E.V., 2013. Functional and evolutionary implications of gene orthology. *Nat Rev Genet* 14, 360–366. <https://doi.org/10.1038/nrg3456>

Gabr, A., Grossman, A.R., Bhattacharya, D., 2020. *Paulinella*, a model for understanding plastid primary endosymbiosis. *Journal of Phycology* 56, 837–843. <https://doi.org/10.1111/jpy.13003>

Gaonkar, C.C., Campbell, L., 2023. De novo transcriptome assembly and gene annotation for the toxic dinoflagellate *Dinophysis*. *Sci Data* 10, 345. <https://doi.org/10.1038/s41597-023-02250-8>

García-Oliva, O., Wirtz, K., 2022. Size-dependent and -independent prey selection of dinoflagellates. *Mar Biol* 169, 122. <https://doi.org/10.1007/s00227-022-04102-2>

Garric, S., Ratin, M., Marie, D., Foulon, V., Probert, I., Rodriguez, F., Six, C., 2024. Impaired photoacclimation in a kleptoplastidic dinoflagellate reveals physiological limits of early stages of endosymbiosis. *Current Biology* 34, 3064–3076.e5. <https://doi.org/10.1016/j.cub.2024.05.066>

Gast, R.J., Moran, D.M., Beaudoin, D.J., Blythe, J.N., Dennett, M.R., Caron, D.A., 2006. ABUNDANCE OF A NOVEL DINOFLAGELLATE PHYLOTYPE IN THE ROSS SEA, ANTARCTICA. *Journal of Phycology* 42, 233–242. <https://doi.org/10.1111/j.1529-8817.2006.00183.x>

Gast, R.J., Moran, D.M., Dennett, M.R., Caron, D.A., 2007. Kleptoplasty in an Antarctic dinoflagellate: caught in evolutionary transition? *Environmental Microbiology* 9, 39–45. <https://doi.org/10.1111/j.1462-2920.2006.01109.x>

Gawryluk, R.M.R., Tikhonenkov, D.V., Hehenberger, E., Husnik, F., Mylnikov, A.P., Keeling, P.J., 2019. Non-photosynthetic predators are sister to red algae. *Nature* 572, 240–243. <https://doi.org/10.1038/s41586-019-1398-6>

Giustini, C., Dal Bo, D., Storti, M., Van Vlierberghe, M., Baurain, D., Cardol, P., Zhang, Y., Fernie, A.R., Fitzpatrick, D., Aro, E.-M., Alloreant, G., Albanese, P., Tolleter, D., Curien, G., Finazzi, G., 2025. A Mitochondrially Derived Plastidial Transporter Regulates Photosynthesis in the Diatom *Phaeodactylum tricornutum*. <https://doi.org/10.1101/2025.02.14.638225>

Glover, N., Dessimoz, C., Ebersberger, I., Forslund, S.K., Gabaldón, T., Huerta-Cepas, J., Martin, M.-J., Muffato, M., Patricio, M., Pereira, C., da Silva, A.S., Wang, Y., Sonnhammer, E., Thomas, P.D., Quest for Orthologs Consortium, 2019. Advances and Applications in the Quest for Orthologs. *Mol Biol Evol* 36, 2157–2164. <https://doi.org/10.1093/molbev/msz150>

Good, I.J., 1953. The population frequencies of species and the estimation of population parameter. *Biometrika* 40, 237–264. <https://doi.org/10.1093/biomet/40.3-4.237>

Gornik, S.G., Febrimarsa, Cassin, A.M., MacRae, J.I., Ramaprasad, A., Rchiad, Z., McConville, M.J., Bacic, A., McFadden, G.I., Pain, A., Waller, R.F., 2015. Endosymbiosis undone by stepwise elimination of the plastid in a parasitic dinoflagellate. *Proc. Natl. Acad. Sci. U.S.A.* 112, 5767–5772. <https://doi.org/10.1073/pnas.1423400112>

Gould, S.B., Maier, U.-G., Martin, W.F., 2015. Protein Import and the Origin of Red Complex Plastids. *Current Biology* 25, R515–R521. <https://doi.org/10.1016/j.cub.2015.04.033>

Gouy, R., Baurain, D., Philippe, H., 2015. Rooting the tree of life: the phylogenetic jury is still out. *Phil. Trans. R. Soc. B* 370, 20140329. <https://doi.org/10.1098/rstb.2014.0329>

Grosche, C., Hempel, F., Bolte, K., Abram, L., Maier, U.G., Zauner, S., 2014. Protein Import into Complex Plastids: Current Findings and Perspectives, in: Löffelhardt, W. (Ed.), *Endosymbiosis*. Springer, Vienna, pp. 215–231. https://doi.org/10.1007/978-3-7091-1303-5_12

Gross, J., Bhattacharya, D., 2011. Endosymbiont or host: who drove mitochondrial and plastid evolution? *Biol Direct* 6, 12. <https://doi.org/10.1186/1745-6150-6-12>

Gross, J., Bhattacharya, D., 2009. Mitochondrial and plastid evolution in eukaryotes: an outsiders' perspective. *Nat Rev Genet* 10, 495–505. <https://doi.org/10.1038/nrg2610>

Gruber, A., Roca, G., Kroth, P.G., Armbrust, E.V., Mock, T., 2015. Plastid proteome prediction for diatoms and other algae with secondary plastids of the red lineage. *Plant J* 81, 519–528. <https://doi.org/10.1111/tpj.12734>

Grzebyk, D., Schofield, O., Vetriani, C., Falkowski, P.G., 2003. THE MESOZOIC RADIATION OF EUKARYOTIC ALGAE: THE PORTABLE PLASTID HYPOTHESIS ¹. *Journal of Phycology* 39, 259–267. <https://doi.org/10.1046/j.1529-8817.2003.02082.x>

Gschloessl, B., Guermeur, Y., Cock, J.M., 2008. HECTAR: A method to predict subcellular targeting in heterokonts. *BMC Bioinformatics* 9, 393. <https://doi.org/10.1186/1471-2105-9-393>

Guiry, M.D., 2024. How many species of algae are there? A reprise. Four kingdoms, 14 phyla, 63 classes and still growing. *Journal of Phycology* 60, 214–228. <https://doi.org/10.1111/jpy.13431>

Hambücken, L., Sudianto, E., Verleyen, E., Saw, J.H., Baurain, D., Cornet, L., 2026. An early diverging SQR enzyme in Antarctic *Gloeobacterales* indicates sulfide tolerance in thylakoid-lacking cyanobacteria. *Microbiol Spectr* e00423-26. <https://doi.org/10.1128/spectrum.00423-26>

Hehenberger, E., Burki, F., Kolisko, M., Keeling, P.J., 2016a. Functional Relationship between a Dinoflagellate Host and Its Diatom Endosymbiont. *Mol Biol Evol* 33, 2376–2390. <https://doi.org/10.1093/molbev/msw109>

Hehenberger, E., Burki, F., Kolisko, M., Keeling, P.J., 2016b. Functional Relationship between a Dinoflagellate Host and Its Diatom Endosymbiont. *Mol Biol Evol* 33, 2376–2390. <https://doi.org/10.1093/molbev/msw109>

Hehenberger, E., Gast, R.J., Keeling, P.J., 2019. A kleptoplastidic dinoflagellate and the tipping point between transient and fully integrated plastid endosymbiosis. *Proc. Natl. Acad. Sci. U.S.A.* 116, 17934–17942. <https://doi.org/10.1073/pnas.1910121116>

Hehenberger, E., Imanian, B., Burki, F., Keeling, P.J., 2014. Evidence for the Retention of Two Evolutionary Distinct Plastids in Dinoflagellates with Diatom Endosymbionts. *Genome Biol Evol* 6, 2321–2334. <https://doi.org/10.1093/gbe/evu182>

Hempel, F., Bullmann, L., Lau, J., Zauner, S., Maier, U.G., 2009. ERAD-Derived Preprotein Transport across the Second Outermost Plastid Membrane of Diatoms. *Mol Biol Evol* 26, 1781–1790. <https://doi.org/10.1093/molbev/msp079>

Hof, R. van't, Kruijff, B. de, 1995. Characterization of the Import Process of a Transit Peptide into Chloroplasts (*). *Journal of Biological Chemistry* 270, 22368–22373. <https://doi.org/10.1074/jbc.270.38.22368>

Hongo, Y., Yabuki, A., Fujikura, K., Nagai, S., 2019. Genes functioned in kleptoplastids of Dinophysis are derived from haptophytes rather than from cryptophytes. *Sci Rep* 9, 9009. <https://doi.org/10.1038/s41598-019-45326-5>

Howe, C.J., Barbrook, A.C., 2024. Dinoflagellate chloroplasts as a model for extreme genome reduction and fragmentation in organelles – The COCOA principle for gene retention. *Protist* 175, 126048. <https://doi.org/10.1016/j.protis.2024.126048>

Howe, C.J., Barbrook, A.C., Nisbet, R.E.R., Lockhart, P.J., Larkum, A.W.D., 2008a. The origin of plastids. *Phil. Trans. R. Soc. B* 363, 2675–2685. <https://doi.org/10.1098/rstb.2008.0050>

Howe, C.J., Barbrook, A.C., Nisbet, R.E.R., Lockhart, P.J., Larkum, A.W.D., 2008b. The origin of plastids. *Phil. Trans. R. Soc. B* 363, 2675–2685. <https://doi.org/10.1098/rstb.2008.0050>

Huang, X., Madan, A., 1999. CAP3: A DNA Sequence Assembly Program. *Genome Res.* 9, 868–877. <https://doi.org/10.1101/gr.9.9.868>

Huson, D.H., Beier, S., Flade, I., Górska, A., El-Hadidi, M., Mitra, S., Ruscheweyh, H.-J., Tappu, R., 2016. MEGAN Community Edition - Interactive Exploration and Analysis of Large-Scale Microbiome Sequencing Data. *PLOS Computational Biology* 12, e1004957. <https://doi.org/10.1371/journal.pcbi.1004957>

Imanian, B., Keeling, P.J., 2007. The dinoflagellates *Durinskia baltica* and *Kryptoperidinium foliaceum* retain functionally overlapping mitochondria from two evolutionarily distinct lineages. *BMC Evol Biol* 7, 172. <https://doi.org/10.1186/1471-2148-7-172>

Imanian, B., Pombert, J.-F., Keeling, P.J., 2010. The Complete Plastid Genomes of the Two ‘Dinotoms’ *Durinskia baltica* and *Kryptoperidinium foliaceum*. *PLOS ONE* 5, e10711. <https://doi.org/10.1371/journal.pone.0010711>

Irisarri, I., Baurain, D., Brinkmann, H., Delsuc, F., Sire, J.-Y., Kupfer, A., Petersen, J., Jarek, M., Meyer, A., Vences, M., Philippe, H., 2017. Phylotranscriptomic consolidation of the jawed vertebrate timetree. *Nat Ecol Evol* 1, 1370–1378. <https://doi.org/10.1038/s41559-017-0240-5>

Irisarri, I., Strasser, J.F.H., Burki, F., 2021. Phylogenomic Insights into the Origin of Primary Plastids. *Systematic Biology* 71, 105–120. <https://doi.org/10.1093/sysbio/syab036>

Irwin, N.A.T., Keeling, P.J., 2019. Extensive Reduction of the Nuclear Pore Complex in Nucleomorphs. *Genome Biol Evol* 11, 678–687. <https://doi.org/10.1093/gbe/evz029>

Jackson, C., Knoll, A.H., Chan, C.X., Verbruggen, H., 2018. Plastid phylogenomics with broad taxon sampling further elucidates the distinct evolutionary origins and timing of secondary green plastids. *Sci Rep* 8, 1523. <https://doi.org/10.1038/s41598-017-18805-w>

Jacobson, D.M., Anderson, D.M., 1986. Thecate Heterophic Dinoflagellates: Feeding Behavior and Mechanisms. *Journal of Phycology* 22, 249–258. <https://doi.org/10.1111/j.1529-8817.1986.tb00021.x>

Jamy, M., Huber, T., Antoine, T., Ruscheweyh, H.-J., Paoli, L., Pelletier, E., O. Delmont, T., Burki, F., 2025. Identification of a deep-branching lineage of algae using environmental plastid genomes. *Nat Commun* 17, 662. <https://doi.org/10.1038/s41467-025-67401-4>

Janouškovec, J., Horák, A., Oborník, M., Lukeš, J., Keeling, P.J., 2010. A common red algal origin of the apicomplexan, dinoflagellate, and heterokont plastids. *Proc. Natl. Acad. Sci. U.S.A.* 107, 10949–10954. <https://doi.org/10.1073/pnas.1003335107>

Janouškovec, J., Tikhonenkov, D.V., Burki, F., Howe, A.T., Kolísko, M., Mylnikov, A.P., Keeling, P.J., 2015a. Factors mediating plastid dependency and the origins of parasitism in apicomplexans and their close relatives. *Proc. Natl. Acad. Sci. U.S.A.* 112, 10200–10207. <https://doi.org/10.1073/pnas.1423790112>

Janouškovec, J., Tikhonenkov, D.V., Burki, F., Howe, A.T., Kolísko, M., Mylnikov, A.P., Keeling, P.J., 2015b. Factors mediating plastid dependency and the origins of parasitism in apicomplexans and their close relatives. *Proc. Natl. Acad. Sci. U.S.A.* 112, 10200–10207. <https://doi.org/10.1073/pnas.1423790112>

Janson, S., 2004. Molecular evidence that plastids in the toxin-producing dinoflagellate genus *Dinophysis* originate from the free-living cryptophyte *Teleaulax amphioxiea*. *Environ Microbiol* 6, 1102–1106. <https://doi.org/10.1111/j.1462-2920.2004.00646.x>

Jauffrit, F., Penel, S., Delmotte, S., Rey, C., de Vienne, D.M., Gouy, M., Charrier, J.-P., Flandrois, J.-P., Brochier-Armanet, C., 2016. RiboDB Database: A Comprehensive Resource for Prokaryotic Systematics. *Mol Biol Evol* 33, 2170–2172. <https://doi.org/10.1093/molbev/msw088>

Jensen, L.J., Julien, P., Kuhn, M., von Mering, C., Muller, J., Doerks, T., Bork, P., 2008. eggNOG: automated construction and annotation of orthologous groups of genes. *Nucleic Acids Res* 36, D250–D254. <https://doi.org/10.1093/nar/gkm796>

Jesus, B., Jauffrais, T., Trampe, E.C.L., Goessling, J.W., Lekieffre, C., Meibom, A., Kühl, M., Geslin, E., 2022. Kleptoplast distribution, photosynthetic efficiency and sequestration mechanisms in intertidal benthic foraminifera. *ISME J* 16, 822–832. <https://doi.org/10.1038/s41396-021-01128-0>

Jirsová, D., Wideman, J.G., 2024. Integrated overview of stramenopile ecology, taxonomy, and heterotrophic origin. *ISME J* 18, wræ150. <https://doi.org/10.1093/ismejo/wrae150>

Johnson, M.D., 2011. The acquisition of phototrophy: adaptive strategies of hosting endosymbionts and organelles. *Photosynth Res* 107, 117–132. <https://doi.org/10.1007/s11120-010-9546-8>

Johnson, M.D., Oldach, D., Delwiche, C.F., Stoecker, D.K., 2007. Retention of transcriptionally active cryptophyte nuclei by the ciliate *Myrionecta rubra*. *Nature* 445, 426–428. <https://doi.org/10.1038/nature05496>

Jost, L., 2007. Partitioning Diversity into Independent Alpha and Beta Components. *Ecology* 88, 2427–2439. <https://doi.org/10.1890/06-1736.1>

Jost, L., 2006. Entropy and diversity. *Oikos* 113, 363–375. <https://doi.org/10.1111/j.2006.0030-1299.14714.x>

Kamikawa, R., Tanifuji, G., Kawachi, M., Miyashita, H., Hashimoto, T., Inagaki, Y., 2015. Plastid Genome-Based Phylogeny Pinpointed the Origin of the Green-Colored Plastid in the Dinoflagellate *Lepidodinium chlorophorum*. *Genome Biology and Evolution* 7, 1133–1140. <https://doi.org/10.1093/gbe/evv060>

Kanehisa, M., Goto, S., 2000. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* 28, 27–30. <https://doi.org/10.1093/nar/28.1.27>

Karnkowska, A., Yubuki, N., Maruyama, M., Yamaguchi, A., Kashiwayama, Y., Suzaki, T., Keeling, P.J., Hampl, V., Leander, B.S., 2023. Euglenozoan kleptoplasty illuminates the early evolution of

photoendosymbiosis. *Proc. Natl. Acad. Sci. U.S.A.* 120, e2220100120. <https://doi.org/10.1073/pnas.2220100120>

Kashiyama, Y., Maruyama, M., Nakazawa, M., Kagamoto, T., Imanishi, H., Yamamoto, S., Inoue, M., Onuma, R., Tanifuji, G., Ashida, H., Inada, N., Awai, K., Miyagishima, S., 2026. Transient molecular chimerism for exploiting xenogeneic organelles. *Nat Commun* 17, 2371. <https://doi.org/10.1038/s41467-026-70516-x>

Katoh, K., Misawa, K., Kuma, K., Miyata, T., 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research* 30, 3059–3066. <https://doi.org/10.1093/nar/gkf436>

Katoh, K., Standley, D.M., 2013. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Molecular Biology and Evolution* 30, 772–780. <https://doi.org/10.1093/molbev/mst010>

Katz, L.A., Leleu, M., Ani, G., Gawron, R., Cote-L’Heureux, A., 2025. Rethinking large-scale phylogenomics with EukPhylo v.1.0, a flexible toolkit to enable phylogeny-informed data curation and analyses of diverse eukaryotic lineages. *mBio* 16, e01770-25. <https://doi.org/10.1128/mbio.01770-25>

Keeling, P.J., 2024. Horizontal gene transfer in eukaryotes: aligning theory with data. *Nat Rev Genet* 25, 416–430. <https://doi.org/10.1038/s41576-023-00688-5>

Keeling, P.J., 2013. The Number, Speed, and Impact of Plastid Endosymbioses in Eukaryotic Evolution. *Annu. Rev. Plant Biol.* 64, 583–607. <https://doi.org/10.1146/annurev-arplant-050312-120144>

Keeling, P.J., 2010. The endosymbiotic origin, diversification and fate of plastids. *Phil. Trans. R. Soc. B* 365, 729–748. <https://doi.org/10.1098/rstb.2009.0103>

Keeling, P.J., 2009. Chromalveolates and the Evolution of Plastids by Secondary Endosymbiosis. *Journal of Eukaryotic Microbiology* 56, 1–8. <https://doi.org/10.1111/j.1550-7408.2008.00371.x>

Keeling, P.J., Burki, F., Wilcox, H.M., Allam, B., Allen, E.E., Amaral-Zettler, L.A., Armbrust, E.V., Archibald, J.M., Bharti, A.K., Bell, C.J., Beszteri, B., Bidle, K.D., Cameron, C.T., Campbell, L., Caron, D.A., Cattolico, R.A., Collier, J.L., Coyne, K., Davy, S.K., Deschamps, P., Dyhrman, S.T., Edvardsen, B., Gates, R.D., Gobler, C.J., Greenwood, S.J., Guida, S.M., Jacobi, J.L., Jakobsen, K.S., James, E.R., Jenkins, B., John, U., Johnson, M.D., Juhl, A.R., Kamp, A., Katz, L.A., Kiene, R., Kudryavtsev, A., Leander, B.S., Lin, S., Lovejoy, C., Lynn, D., Marchetti, A., McManus, G., Nedelcu, A.M., Menden-Deuer, S., Miceli, C., Mock, T., Montresor, M., Moran, M.A., Murray, S., Nadathur, G., Nagai, S., Ngam, P.B., Palenik, B., Pawlowski, J., Petroni, G., Piganeau, G., Posewitz, M.C., Rengefors, K., Romano, G., Rumpho, M.E., Ryneerson, T., Schilling, K.B., Schroeder, D.C., Simpson, A.G.B., Slamovits, C.H., Smith, D.R., Smith, G.J., Smith, S.R., Sosik, H.M., Stief, P., Theriot, E., Twary, S.N., Umale, P.E., Vaultot, D., Wawrik, B., Wheeler, G.L., Wilson, W.H., Xu, Y., Zingone, A., Worden, A.Z., 2014. The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP): Illuminating

the Functional Diversity of Eukaryotic Life in the Oceans through Transcriptome Sequencing. *PLoS Biol* 12, e1001889. <https://doi.org/10.1371/journal.pbio.1001889>

Kelly, S., 2021. The economics of organellar gene loss and endosymbiotic gene transfer. *Genome Biol* 22, 345. <https://doi.org/10.1186/s13059-021-02567-w>

Kim, M., Drumm, K., Daugbjerg, N., Hansen, P.J., 2017. Dynamics of Sequestered Cryptophyte Nuclei in *Mesodinium rubrum* during Starvation and Refeeding. *Front Microbiol* 8, 423. <https://doi.org/10.3389/fmicb.2017.00423>

Koike, Kazuhiko, Sekiguchi, H., Kobiyama, A., Takishita, K., Kawachi, M., Koike, Kanae, Ogata, T., 2005. A Novel Type of Kleptoplastidy in *Dinophysis* (Dinophyceae): Presence of Haptophyte-type Plastid in *Dinophysis mitra*. *Protist* 156, 225–237. <https://doi.org/10.1016/j.protis.2005.04.002>

Koonin, E.V., 2005. Orthologs, Paralogs, and Evolutionary Genomics1. *Annual Review of Genetics* 39, 309–338. <https://doi.org/10.1146/annurev.genet.39.073003.114725>

Koutsovoulos, G., Kumar, S., Laetsch, D.R., Stevens, L., Daub, J., Conlon, C., Maroon, H., Thomas, F., Aboobaker, A.A., Blaxter, M., 2016. No evidence for extensive horizontal gene transfer in the genome of the tardigrade *Hypsibius dujardini*. *Proc. Natl. Acad. Sci. U.S.A.* 113, 5053–5058. <https://doi.org/10.1073/pnas.1600338113>

Koutsovoulos, G.D., Noriot, S.G., Bailly-Bechet, M., Danchin, E.G.J., Rancurel, C., 2022. AvP: A software package for automatic phylogenetic detection of candidate horizontal gene transfers. *PLOS Computational Biology* 18, e1010686. <https://doi.org/10.1371/journal.pcbi.1010686>

Kristensen, D.M., Wolf, Y.I., Mushegian, A.R., Koonin, E.V., 2011. Computational methods for Gene Orthology inference. *Brief Bioinform* 12, 379–391. <https://doi.org/10.1093/bib/bbr030>

Kriventseva, E.V., Rahman, N., Espinosa, O., Zdobnov, E.M., 2008. OrthoDB: the hierarchical catalog of eukaryotic orthologs. *Nucleic Acids Res* 36, D271–D275. <https://doi.org/10.1093/nar/gkm845>

Kutschera, U., Niklas, K.J., 2008. Macroevolution via secondary endosymbiosis: a Neo-Goldschmidtian view of unicellular hopeful monsters and Darwin’s primordial intermediate form. *Theory Biosci.* 127, 277–289. <https://doi.org/10.1007/s12064-008-0046-8>

Lafond, M., Meghdari Miardan, M., Sankoff, D., 2018. Accurate prediction of orthologs in the presence of divergence after duplication. *Bioinformatics* 34, i366–i375. <https://doi.org/10.1093/bioinformatics/bty242>

Larkum, A.W.D., Lockhart, P.J., Howe, C.J., 2007. Shopping for plastids. *Trends in Plant Science* 12, 189–195. <https://doi.org/10.1016/j.tplants.2007.03.011>

Lasek-Nesselquist, E., Wisecaver, J.H., Hackett, J.D., Johnson, M.D., 2015. Insights into transcriptional changes that accompany organelle sequestration from the stolen nucleus of *Mesodinium rubrum*. *BMC Genomics* 16, 805. <https://doi.org/10.1186/s12864-015-2052-9>

Laurin-Lemay, S., Brinkmann, H., Philippe, H., 2012. Origin of land plants revisited in the light of sequence contamination and missing data. *Current Biology* 22, R593–R594. <https://doi.org/10.1016/j.cub.2012.06.013>

Lechner, M., Findeiß, S., Steiner, L., Marz, M., Stadler, P.F., Prohaska, S.J., 2011. Proteinortho: Detection of (Co-)orthologs in large-scale analysis. *BMC Bioinformatics* 12, 124. <https://doi.org/10.1186/1471-2105-12-124>

Lee, M.D., 2019. GToTree: a user-friendly workflow for phylogenomics. *Bioinformatics* 35, 4162–4164. <https://doi.org/10.1093/bioinformatics/btz188>

Léonard, R.R., Leleu, M., Vlierberghe, M.V., Cornet, L., Kerff, F., Baurain, D., 2021. ToRQuEMaDA: tool for retrieving queried Eubacteria, metadata and dereplicating assemblies. *PeerJ* 9, e11348. <https://doi.org/10.7717/peerj.11348>

Lewis, W.H., Paris, G., Beedessee, G., Koreny, L., Flores, V., Dendooven, T., Gallet, B., Yee, D., Lam, S., Decelle, J., Luisi, B.F., Waller, R.F., 2024. Plastid translocon recycling in dinoflagellates demonstrates the portability of complex plastids between hosts. <https://doi.org/10.1101/2024.07.06.602338>

Li, L., Stoeckert, C.J., Roos, D.S., 2003. OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes. *Genome Res.* 13, 2178–2189. <https://doi.org/10.1101/gr.1224503>

Linard, B., Ebersberger, I., McGlynn, S.E., Glover, N., Mochizuki, T., Patricio, M., Lecompte, O., Nevers, Y., Thomas, P.D., Gabaldón, T., Sonnhammer, E., Dessimoz, C., Uchiyama, I., QFO Consortium, 2021. Ten Years of Collaborative Progress in the Quest for Orthologs. *Mol Biol Evol* 38, 3033–3045. <https://doi.org/10.1093/molbev/msab098>

Lupo, V., Van Vlierberghe, M., Vanderschuren, H., Kerff, F., Baurain, D., Cornet, L., 2021. Contamination in Reference Sequence Databases: Time for Divide-and-Rule Tactics. *Front. Microbiol.* 12. <https://doi.org/10.3389/fmicb.2021.755101>

Mansour, J.S., Anestis, K., 2021. Eco-Evolutionary Perspectives on Mixoplankton. *Front. Mar. Sci.* 8. <https://doi.org/10.3389/fmars.2021.666160>

Martin, W., Herrmann, R.G., 1998. Gene Transfer from Organelles to the Nucleus: How Much, What Happens, and Why?1. *Plant Physiology* 118, 9–17. <https://doi.org/10.1104/pp.118.1.9>

Maruyama, M., Kagamoto, T., Matsumoto, Y., Onuma, R., Miyagishima, S., Tanifuji, G., Nakazawa, M., Kashiya, Y., 2023. Horizontally Acquired Nitrate Reductase Realized Kleptoplastic Photoautotrophy of *Rapaza viridis*. *Plant Cell Physiol* 64, 1082–1090. <https://doi.org/10.1093/pcp/pcad044>

Maselli, M., Anestis, K., Klemm, K., Hansen, P.J., John, U., 2021. Retention of Prey Genetic Material by the Kleptoplastidic Ciliate *Strombidium* cf. *basimorphum*. *Front. Microbiol.* 12. <https://doi.org/10.3389/fmicb.2021.694508>

Matsuo, E., Morita, K., Nakayama, T., Yazaki, E., Sarai, C., Takahashi, K., Iwataki, M., Inagaki, Y., 2022. Comparative Plastid Genomics of Green-Colored Dinoflagellates Unveils Parallel Genome Compaction and RNA Editing. *Front. Plant Sci.* 13. <https://doi.org/10.3389/fpls.2022.918543>

McEwan, M.L., Keeling, P.J., 2004. HSP90, Tubulin and Actin are Retained in the Tertiary Endosymbiont Genome of *Kryptoperidinium foliaceum*. *Journal of Eukaryotic Microbiology* 51, 651–659. <https://doi.org/10.1111/j.1550-7408.2004.tb00604.x>

McFadden, G.I., 2001. Primary and secondary endosymbiosis and the origin of plastids. *Journal of Phycology* 37, 951–959. <https://doi.org/10.1046/j.1529-8817.2001.01126.x>

McFadden, G.I., 1999. Endosymbiosis and evolution of the plant cell. *Current Opinion in Plant Biology* 2, 513–519. [https://doi.org/10.1016/S1369-5266\(99\)00025-4](https://doi.org/10.1016/S1369-5266(99)00025-4)

McManus, G.B., Liu, W., Cole, R.A., Biemesderfer, D., Mydosh, J.L., 2018. *Strombidium rassoulzadegani*: A Model Species for Chloroplast Retention in Oligotrich Ciliates. *Front. Mar. Sci.* 5. <https://doi.org/10.3389/fmars.2018.00205>

Miao, W., Song, L., Ba, S., Zhang, L., Guan, G., Zhang, Z., Ning, K., 2020. Protist 10,000 Genomes Project. *Innovation (Camb)* 1, 100058. <https://doi.org/10.1016/j.xinn.2020.100058>

Minh, B.Q., Schmidt, H.A., Chernomor, O., Schrempf, D., Woodhams, M.D., von Haeseler, A., Lanfear, R., 2020. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Molecular Biology and Evolution* 37, 1530–1534. <https://doi.org/10.1093/molbev/msaa015>

Minnhagen, S., Kim, M., Salomon, P.S., Yih, W., Granéli, E., Park, M.G., 2011. Active uptake of kleptoplastids by *Dinophysis caudata* from its ciliate prey *Myrionecta rubra*. *Aquatic Microbial Ecology* 62, 99–108. <https://doi.org/10.3354/ame01459>

Moeller, H.V., Hsu, V., Lepori-Bui, M., Mesrop, L.Y., Chinn, C., Johnson, M.D., 2021. Prey type constrains growth and photosynthetic capacity of the kleptoplastidic ciliate *Mesodinium chamaeleon* (Ciliophora). *Journal of Phycology* 57, 916–930. <https://doi.org/10.1111/jpy.13131>

Molina, J., Hazzouri, K.M., Nickrent, D., Geisler, M., Meyer, R.S., Pentony, M.M., Flowers, J.M., Pelser, P., Barcelona, J., Inovejas, S.A., Uy, I., Yuan, W., Wilkins, O., Michel, C.-I., LockLear, S., Concepcion, G.P., Purugganan, M.D., 2014. Possible Loss of the Chloroplast Genome in the Parasitic Flowering Plant *Rafflesia lagascae* (Rafflesiaceae). *Mol Biol Evol* 31, 793–803. <https://doi.org/10.1093/molbev/msu051>

Moog, D., Maier, U.G., 2018. Explaining the Origin of Three-Membrane-Bound Plastids in Dinoflagellates and Euglenophytes: Kleptoplastidy via Myzocytosis? *BioEssays* 40, 1700224. <https://doi.org/10.1002/bies.201700224>

Moore, R.B., Oborník, M., Janouškovec, J., Chrudimský, T., Vancová, M., Green, D.H., Wright, S.W., Davies, N.W., Bolch, C.J.S., Heimann, K., Šlapeta, J., Hoegh-Guldberg, O., Logsdon, J.M., Carter,

D.A., 2008. A photosynthetic alveolate closely related to apicomplexan parasites. *Nature* 451, 959–963. <https://doi.org/10.1038/nature06635>

Moreira, D., Deschamps, P., 2014. What Was the Real Contribution of Endosymbionts to the Eukaryotic Nucleus? Insights from Photosynthetic Eukaryotes. *Cold Spring Harb Perspect Biol* 6, a016014. <https://doi.org/10.1101/cshperspect.a016014>

Morozov, A.A., Galachyants, Y.P., 2019. Diatom genes originating from red and green algae: Implications for the secondary endosymbiosis models. *Marine Genomics* 45, 72–78. <https://doi.org/10.1016/j.margen.2019.02.003>

Moustafa, A., Beszteri, B., Maier, U.G., Bowler, C., Valentin, K., Bhattacharya, D., 2009. Genomic Footprints of a Cryptic Plastid Endosymbiosis in Diatoms. *Science* 324, 1724–1726. <https://doi.org/10.1126/science.1172983>

Moustafa, A., Bhattacharya, D., 2008. PhyloSort: a user-friendly phylogenetic sorting tool and its application to estimating the cyanobacterial contribution to the nuclear genome of *Chlamydomonas*. *BMC Evol Biol* 8, 6. <https://doi.org/10.1186/1471-2148-8-6>

Nassoury, N., Cappadocia, M., Morse, D., 2003. Plastid ultrastructure defines the protein import pathway in dinoflagellates. *J Cell Sci* 116, 2867–2874. <https://doi.org/10.1242/jcs.00517>

Natsidis, P., Kapli, P., Schiffer, P.H., Telford, M.J., 2021. Systematic errors in orthology inference and their effects on evolutionary analyses. *iScience* 24. <https://doi.org/10.1016/j.isci.2021.102110>

Nguyen, L.-T., Schmidt, H.A., von Haeseler, A., Minh, B.Q., 2015. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Molecular Biology and Evolution* 32, 268–274. <https://doi.org/10.1093/molbev/msu300>

Niang, G., Hoebeke, M., Meng, A., Liu, X., Scheremetjew, M., Finn, R., Pelletier, E., Corre, E., 2020. METdb: A GENOMIC REFERENCE DATABASE FOR MARINE SPECIES. <https://doi.org/10.7490/f1000research.1118000.1>

Nishitani, G., Nagai, S., Hayakawa, S., Kosaka, Y., Sakurada, K., Kamiyama, T., Gojobori, T., 2012. Multiple Plastids Collected by the Dinoflagellate *Dinophysis mitra* through Kleptoplastidy. *Appl Environ Microbiol* 78, 813–821. <https://doi.org/10.1128/AEM.06544-11>

Nosenko, T., Lidie, K.L., Van Dolah, F.M., Lindquist, E., Cheng, J.-F., US Department of Energy–Joint Genome Institute, Bhattacharya, D., 2006. Chimeric Plastid Proteome in the Florida “Red Tide” Dinoflagellate *Karenia brevis*. *Molecular Biology and Evolution* 23, 2026–2038. <https://doi.org/10.1093/molbev/msl074>

Novak, L.V.F., Muñoz-Gómez, S.A., Van Beveren, F., Ciobanu, M., Eme, L., López-García, P., Moreira, D., 2024. Nucleomorph phylogenomics suggests a deep and ancient origin of cryptophyte plastids within Rhodophyta. <https://doi.org/10.1101/2024.03.10.584144>

- Novák Vanclová, A.M., Nef, C., Füßy, Z., Vanclová, A., Liu, F., Bowler, C., Dorrell, R.G., 2024. New plastids, old proteins: repeated endosymbiotic acquisitions in kareniacean dinoflagellates. *EMBO Rep* 25, 1859–1885. <https://doi.org/10.1038/s44319-024-00103-y>
- Novák Vanclová, A.M.G., Dorrell, R.G., 2024. Complex Plastids Across the Eukaryotes: An Overview of Inherited and Convergently Evolved Characters, in: Schwartzbach, S.D., Kroth, P.G., Oborník, M. (Eds.), *Endosymbiotic Organelle Acquisition: Solutions to the Problem of Protein Localization and Membrane Passage*. Springer International Publishing, Cham, pp. 39–88. https://doi.org/10.1007/978-3-031-57446-7_2
- Nowack, E.C.M., Melkonian, M., 2010. Endosymbiotic associations within protists. *Phil. Trans. R. Soc. B* 365, 699–712. <https://doi.org/10.1098/rstb.2009.0188>
- Oborník, M., Lukeš, J., 2015. The Organellar Genomes of *Chromera* and *Vitrella*, the Phototrophic Relatives of Apicomplexan Parasites. *Annual Review of Microbiology* 69, 129–144. <https://doi.org/10.1146/annurev-micro-091014-104449>
- Oborník, M., Modrý, D., Lukeš, M., Černotíková-Stříbrná, E., Cihlář, J., Tesařová, M., Kotabová, E., Vanclová, M., Prášil, O., Lukeš, J., 2012. Morphology, Ultrastructure and Life Cycle of *Vitrella brassicaformis* n. sp., n. gen., a Novel Chromerid from the Great Barrier Reef. *Protist* 163, 306–323. <https://doi.org/10.1016/j.protis.2011.09.001>
- Ødum, M.T., Teufel, F., Thumhuri, V., Almagro Armenteros, J.J., Johansen, A.R., Winther, O., Nielsen, H., 2024. DeepLoc 2.1: multi-label membrane protein type prediction using protein language models. *Nucleic Acids Res* 52, W215–W220. <https://doi.org/10.1093/nar/gkae237>
- Ok, J.H., Jeong, H.J., Kang, H.C., Park, S.A., Eom, S.H., You, J.H., Lee, S.Y., 2021. Ecophysiology of the kleptoplastidic dinoflagellate *Shimiella gracilentia*: I. spatiotemporal distribution in Korean coastal waters and growth and ingestion rates. *Algae* 36, 263–283. <https://doi.org/10.4490/algae.2021.36.11.28>
- Okamoto, N., Chantangsi, C., Horák, A., Leander, B.S., Keeling, P.J., 2009. Molecular Phylogeny and Description of the Novel Katablepharid *Roombia truncata* gen. et sp. nov., and Establishment of the Hacrobia Taxon nov. *PLoS ONE* 4, e7080. <https://doi.org/10.1371/journal.pone.0007080>
- Orakov, A., Fullam, A., Coelho, L.P., Khedkar, S., Szklarczyk, D., Mende, D.R., Schmidt, T.S.B., Bork, P., 2021. GUNC: detection of chimerism and contamination in prokaryotic genomes. *Genome Biol* 22, 178. <https://doi.org/10.1186/s13059-021-02393-0>
- Overbeek, R., Fonstein, M., D'Souza, M., Pusch, G.D., Maltsev, N., 1999. The use of gene clusters to infer functional coupling. *Proceedings of the National Academy of Sciences* 96, 2896–2901. <https://doi.org/10.1073/pnas.96.6.2896>
- Paight, C., Johnson, M.D., Lasek-Nesselquist, E., Moeller, H.V., 2023. Cascading effects of prey identity on gene expression in a kleptoplastidic ciliate. *Journal of Eukaryotic Microbiology* 70, e12940. <https://doi.org/10.1111/jeu.12940>

Parks, D.H., Chuvochina, M., Rinke, C., Mussig, A.J., Chaumeil, P.-A., Hugenholtz, P., 2022. GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. *Nucleic Acids Research* 50, D785–D794. <https://doi.org/10.1093/nar/gkab776>

Parks, D.H., Imelfort, M., Skennerton, C.T., Hugenholtz, P., Tyson, G.W., 2015. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res* 25, 1043–1055. <https://doi.org/10.1101/gr.186072.114>

Patron, N.J., Waller, R.F., 2007. Transit peptide diversity and divergence: A global analysis of plastid targeting signals. *BioEssays* 29, 1048–1058. <https://doi.org/10.1002/bies.20638>

Patron, N.J., Waller, R.F., Keeling, P.J., 2006. A Tertiary Plastid Uses Genes from Two Endosymbionts. *Journal of Molecular Biology* 357, 1373–1382. <https://doi.org/10.1016/j.jmb.2006.01.084>

Petersen, J., Ludewig, A.-K., Michael, V., Bunk, B., Jarek, M., Baurain, D., Brinkmann, H., 2014. *Chromera velia*, Endosymbioses and the Rhodoplex Hypothesis—Plastid Evolution in Cryptophytes, Alveolates, Stramenopiles, and Haptophytes (CASH Lineages). *Genome Biology and Evolution* 6, 666–684. <https://doi.org/10.1093/gbe/evu043>

Petersen, M., Meusemann, K., Donath, A., Dowling, D., Liu, S., Peters, R.S., Podsiadlowski, L., Vasilikopoulos, A., Zhou, X., Misof, B., Niehuis, O., 2017. Orthograph: a versatile tool for mapping coding nucleotide sequences to clusters of orthologous genes. *BMC Bioinformatics* 18, 111. <https://doi.org/10.1186/s12859-017-1529-8>

Philippe, H., Brinkmann, H., Lavrov, D.V., Littlewood, D.T.J., Manuel, M., Wörheide, G., Baurain, D., 2011. Resolving Difficult Phylogenetic Questions: Why More Sequences Are Not Enough. *PLoS Biol* 9, e1000602. <https://doi.org/10.1371/journal.pbio.1000602>

Philippe, H., Poustka, A.J., Chiodin, M., Hoff, K.J., Dessimoz, C., Tomiczek, B., Schiffer, P.H., Müller, S., Domman, D., Horn, M., Kuhl, H., Timmermann, B., Satoh, N., Hikosaka-Katayama, T., Nakano, H., Rowe, M.L., Elphick, M.R., Thomas-Chollier, M., Hankeln, T., Mertes, F., Wallberg, A., Rast, J.P., Copley, R.R., Martinez, P., Telford, M.J., 2019. Mitigating Anticipated Effects of Systematic Errors Supports Sister-Group Relationship between Xenacoelomorpha and Ambulacraria. *Current Biology* 29, 1818–1826.e6. <https://doi.org/10.1016/j.cub.2019.04.009>

Philippe, H., Snell, E.A., Baptiste, E., Lopez, P., Holland, P.W.H., Casane, D., 2004. Phylogenomics of Eukaryotes: Impact of Missing Data on Large Alignments. *Mol Biol Evol* 21, 1740–1752. <https://doi.org/10.1093/molbev/msh182>

Philippe, H., Telford, M.J., 2006. Large-scale sequencing and the new animal phylogeny. *Trends in Ecology & Evolution* 21, 614–620. <https://doi.org/10.1016/j.tree.2006.08.004>

Pielou, E.C., 1966. The measurement of diversity in different types of biological collections. *Journal of Theoretical Biology* 13, 131–144. [https://doi.org/10.1016/0022-5193\(66\)90013-0](https://doi.org/10.1016/0022-5193(66)90013-0)

Pietluch, F., Mackiewicz, P., Ludwig, K., Gagat, P., 2024. A New Model and Dating for the Evolution of Complex Plastids of Red Alga Origin. *Genome Biol Evol* 16, evae192. <https://doi.org/10.1093/gbe/evae192>

Pietluch, F., Mackiewicz, P., Sidorczuk, K., Gagat, P., 2022. Dating the photosynthetic organelle evolution in Archaeplastida, *Paulinella* and secondary-plastid bearing lineages. <https://doi.org/10.1101/2022.07.01.497312>

Ponce-Toledo, R.I., David Moreira, Purificación López-García, Philippe Deschamps, 2025. Molecular Phylogeny of the SELMA Translocation Machinery Recounts the Evolution of Complex Photosynthetic Eukaryotes. <https://doi.org/https://doi.org/10.1093/molbev/msaf167>

Ponce-Toledo, R.I., Deschamps, P., López-García, P., Zivanovic, Y., Benzerara, K., Moreira, D., 2017. An early-branching freshwater cyanobacterium at the origin of plastids. *Curr Biol* 27, 386–391. <https://doi.org/10.1016/j.cub.2016.11.056>

Ponce-Toledo, R.I., López-García, P., Moreira, D., 2019. Horizontal and endosymbiotic gene transfer in early plastid evolution. *New Phytologist* 224, 618–624. <https://doi.org/10.1111/nph.15965>

Ponce-Toledo, R.I., Moreira, D., López-García, P., Deschamps, P., 2018. Secondary Plastids of Euglenids and Chlorarachniophytes Function with a Mix of Genes of Red and Green Algal Ancestry. *Molecular Biology and Evolution* 35, 2198–2204. <https://doi.org/10.1093/molbev/msy121>

Qiu, H., Lee, J.M., Yoon, H.S., Bhattacharya, D., 2017. Hypothesis: Gene-rich plastid genomes in red algae may be an outcome of nuclear genome reduction. *Journal of Phycology* 53, 715–719. <https://doi.org/10.1111/jpy.12514>

Quevarec, L., Bonnarde, R., Robaglia, C., Brasseur, G., 2026. Oxygen as a primary selective pressure for photosymbiosis evolution. *Current Biology* 36, 928-939.e2. <https://doi.org/10.1016/j.cub.2026.01.010>

Ralph, S.A., Van Dooren, G.G., Waller, R.F., Crawford, M.J., Fraunholz, M.J., Foth, B.J., Tonkin, C.J., Roos, D.S., McFadden, G.I., 2004. Metabolic maps and functions of the Plasmodium falciparum apicoplast. *Nat Rev Microbiol* 2, 203–216. <https://doi.org/10.1038/nrmicro843>

Rancurel, C., Legrand, L., Danchin, E.G.J., 2017. Alieness: Rapid Detection of Candidate Horizontal Gene Transfers across the Tree of Life. *Genes* 8, 248. <https://doi.org/10.3390/genes8100248>

Ravenhall, M., Škunca, N., Lassalle, F., Dessimoz, C., 2015. Inferring Horizontal Gene Transfer. *PLOS Computational Biology* 11, e1004095. <https://doi.org/10.1371/journal.pcbi.1004095>

Reyes-Prieto, A., Weber, A.P.M., Bhattacharya, D., 2007. The Origin and Establishment of the Plastid in Algae and Plants. *Annu. Rev. Genet.* 41, 147–168. <https://doi.org/10.1146/annurev.genet.41.110306.130134>

Rice, D.W., Palmer, J.D., 2006. An exceptional horizontal gene transfer in plastids: gene replacement by a distant bacterial paralog and evidence that haptophyte and cryptophyte plastids are sisters. *BMC Biol* 4, 31. <https://doi.org/10.1186/1741-7007-4-31>

Richter, D.J., Berney, C., Strasser, J.F.H., Poh, Y.-P., Herman, E.K., Muñoz-Gómez, S.A., Wideman, J.G., Burki, F., De Vargas, C., 2022. EukProt: A database of genome-scale predicted proteins across the diversity of eukaryotes. *Peer Community Journal* 2, e56. <https://doi.org/10.24072/pcjournal.173>

Rodríguez, A., Burgon, J.D., Lyra, M., Irisarri, I., Baurain, D., Blaustein, L., Göçmen, B., Künzel, S., Mable, B.K., Nolte, A.W., Veith, M., Steinfartz, S., Elmer, K.R., Philippe, H., Vences, M., 2017. Inferring the shallow phylogeny of true salamanders (*Salamandra*) by multiple phylogenomic approaches. *Molecular Phylogenetics and Evolution* 115, 16–26. <https://doi.org/10.1016/j.ympev.2017.07.009>

Rodriguez-Ezpeleta, N., Philippe, H., Brinkmann, H., Becker, B., Melkonian, M., 2006. Phylogenetic Analyses of Nuclear, Mitochondrial, and Plastid Multigene Data Sets Support the Placement of Mesostigma in the Streptophyta. *Molecular Biology and Evolution* 24, 723–731. <https://doi.org/10.1093/molbev/msl200>

Roger, A.J., Muñoz-Gómez, S.A., Kamikawa, R., 2017. The Origin and Diversification of Mitochondria. *Current Biology* 27, R1177–R1192. <https://doi.org/10.1016/j.cub.2017.09.015>

Rogers, M.B., Gilson, P.R., Su, V., McFadden, G.I., Keeling, P.J., 2007. The Complete Chloroplast Genome of the Chlorarachniophyte Bigelowiella natans: Evidence for Independent Origins of Chlorarachniophyte and Euglenid Secondary Endosymbionts. *Molecular Biology and Evolution* 24, 54–62. <https://doi.org/10.1093/molbev/msl129>

Roth, A.C., Gonnet, G.H., Dessimoz, C., 2008. Algorithm of OMA for large-scale orthology inference. *BMC Bioinformatics* 9, 518. <https://doi.org/10.1186/1471-2105-9-518>

Saary, P., Mitchell, A.L., Finn, R.D., 2020. Estimating the quality of eukaryotic genomes recovered from metagenomic analysis with EukCC. *Genome Biol* 21, 244. <https://doi.org/10.1186/s13059-020-02155-4>

Sanchez-Puerta, M.V., Bachvaroff, T.R., Delwiche, C.F., 2007. Sorting wheat from chaff in multi-gene analyses of chlorophyll c-containing plastids. *Molecular Phylogenetics and Evolution* 44, 885–897. <https://doi.org/10.1016/j.ympev.2007.03.003>

Sanchez-Puerta, M.V., Delwiche, C.F., 2008. A HYPOTHESIS FOR PLASTID EVOLUTION IN CHROMALVEOLATES. *Journal of Phycology* 44, 1097–1107. <https://doi.org/10.1111/j.1529-8817.2008.00559.x>

Sarai, C., Tanifuji, G., Nakayama, T., Kamikawa, R., Takahashi, K., Yazaki, E., Matsuo, E., Miyashita, H., Ishida, K., Iwataki, M., Inagaki, Y., 2020. Dinoflagellates with relic endosymbiont nuclei as models for elucidating organogenesis. *Proc. Natl. Acad. Sci. U.S.A.* 117, 5364–5375. <https://doi.org/10.1073/pnas.1911884117>

Schnepf, E., Deichgräber, G., 1984. “Myzocytosis”, a kind of endocytosis with implications to compartmentation in endosymbiosis. *Naturwissenschaften* 71, 218–219. <https://doi.org/10.1007/BF00490442>

- Schnepf, E., Elbrächter, M., 1992. Nutritional strategies in dinoflagellates. *European Journal of Protistology*. [https://doi.org/10.1016/S0932-4739\(11\)80315-9](https://doi.org/10.1016/S0932-4739(11)80315-9)
- Sellers, C.G., Gast, R.J., Sanders, R.W., 2014a. Selective feeding and foreign plastid retention in an Antarctic dinoflagellate. *Journal of Phycology* 50, 1081–1088. <https://doi.org/10.1111/jpy.12240>
- Sellers, C.G., Gast, R.J., Sanders, R.W., 2014b. Selective feeding and foreign plastid retention in an Antarctic dinoflagellate. *Journal of Phycology* 50, 1081–1088. <https://doi.org/10.1111/jpy.12240>
- Selosse, M.-A., Albert, B., Godelle, B., 2001. Reducing the genome size of organelles favours gene transfer to the nucleus. *Trends in Ecology & Evolution* 16, 135–141. [https://doi.org/10.1016/S0169-5347\(00\)02084-X](https://doi.org/10.1016/S0169-5347(00)02084-X)
- Ševčíková, T., Horák, A., Klimeš, V., Zbránková, V., Demir-Hilton, E., Sudek, S., Jenkins, J., Schmutz, J., Příbyl, P., Fousek, J., Vlček, Č., Lang, B.F., Oborník, M., Worden, A.Z., Eliáš, M., 2015. Updating algal evolutionary relationships through plastid genome sequencing: did alveolate plastids emerge through endosymbiosis of an ochrophyte? *Sci Rep* 5, 10134. <https://doi.org/10.1038/srep10134>
- Shannon, C.E., 1948. A Mathematical Theory of Communication. *Bell System Technical Journal* 27, 379–423. <https://doi.org/10.1002/j.1538-7305.1948.tb01338.x>
- Sibbald, S.J., Archibald, J.M., 2020. Genomic Insights into Plastid Evolution. *Genome Biology and Evolution* 12, 978–990. <https://doi.org/10.1093/gbe/evaa096>
- Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V., Zdobnov, E.M., 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 31, 3210–3212. <https://doi.org/10.1093/bioinformatics/btv351>
- Simion, P., Philippe, H., Baurain, D., Jager, M., Richter, D.J., Franco, A.D., Roure, B., Satoh, N., Quéinnec, É., Ereskovsky, A., Lapébie, P., Corre, E., Delsuc, F., King, N., Wörheide, G., Manuel, M., 2017. A Large and Consistent Phylogenomic Dataset Supports Sponges as the Sister Group to All Other Animals. *Current Biology* 27, 958–967. <https://doi.org/10.1016/j.cub.2017.02.031>
- Slater, G.S.C., Birney, E., 2005. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* 6, 31. <https://doi.org/10.1186/1471-2105-6-31>
- Smith, D.R., Crosby, K., Lee, R.W., 2011. Correlation between Nuclear Plastid DNA Abundance and Plastid Number Supports the Limited Transfer Window Hypothesis. *Genome Biology and Evolution* 3, 365–371. <https://doi.org/10.1093/gbe/evr001>
- Smith, D.R., Lee, R.W., 2014. A Plastid without a Genome: Evidence from the Nonphotosynthetic Green Algal Genus *Polytomella*. *Plant Physiol* 164, 1812–1819. <https://doi.org/10.1104/pp.113.233718>
- Soll, J., Schleiff, E., 2004. Protein import into chloroplasts. *Nat Rev Mol Cell Biol* 5, 198–208. <https://doi.org/10.1038/nrm1333>
- Sommer, M.S., Gould, S.B., Lehmann, P., Gruber, A., Przyborski, J.M., Maier, U.-G., 2007. Der1-mediated preprotein import into the periplastid compartment of chromalveolates? *Mol Biol Evol* 24, 918–928. <https://doi.org/10.1093/molbev/msm008>

- Sørensen, M.E.S., Zlatogursky, V.V., Onuț-Brännström, I., Walraven, A., Foster, R.A., Burki, F., 2023. A novel kleptoplastidic symbiosis revealed in the marine centrohelid *Meringosphaera* with evidence of genetic integration. *Current Biology* 33, 3571–3584.e6. <https://doi.org/10.1016/j.cub.2023.07.017>
- Sprecher, B.N., Johnson, M.D., 2025. Kleptoplasty: Letting the cart lead the horse. *Current Biology* 35, R414–R417. <https://doi.org/10.1016/j.cub.2025.04.068>
- Stamatakis, K., Vayenos, D., Kotakis, C., Gast, R.J., Papageorgiou, G.C., 2017. The extraordinary longevity of kleptoplasts derived from the Ross Sea haptophyte *Phaeocystis antarctica* within dinoflagellate host cells relates to the diminished role of the oxygen-evolving Photosystem II and to supplementary light harvesting by mycosporine-like amino acid/s. *Biochimica et Biophysica Acta (BBA) - Bioenergetics* 1858, 189–195. <https://doi.org/10.1016/j.bbabi.2016.12.002>
- Stephens, T.G., Bhattacharya, D., Ragan, M.A., Chan, C.X., 2016. PhySortR: a fast, flexible tool for sorting phylogenetic trees in R. *PeerJ* 4, e2038. <https://doi.org/10.7717/peerj.2038>
- Stiller, J.W., Schreiber, J., Yue, J., Guo, H., Ding, Q., Huang, J., 2014a. The evolution of photosynthesis in chromist algae through serial endosymbioses. *Nat Commun* 5, 5764. <https://doi.org/10.1038/ncomms6764>
- Stiller, J.W., Schreiber, J., Yue, J., Guo, H., Ding, Q., Huang, J., 2014b. The evolution of photosynthesis in chromist algae through serial endosymbioses. *Nat Commun* 5, 5764. <https://doi.org/10.1038/ncomms6764>
- Stoecker, D.K., Johnson, M.D., Vargas, C. de, Not, F., 2009. Acquired phototrophy in aquatic protists. *Aquatic Microbial Ecology* 57, 279–310. <https://doi.org/10.3354/ame01340>
- Stoecker, D.K., Silver, M.W., Michaels, A.E., Davis, L.H., 1988. Obligate mixotrophy in *Laboea strobila*, a ciliate which retains chloroplasts. *Mar. Biol.* 99, 415–423. <https://doi.org/10.1007/BF02112135>
- Strassert, J.F.H., Irisarri, I., Williams, T.A., Burki, F., 2021. A molecular timescale for eukaryote evolution with implications for the origin of red algal-derived plastids. *Nat Commun* 12, 1879. <https://doi.org/10.1038/s41467-021-22044-z>
- Sulli, C., Fang, Z., Muchhal, U., Schwartzbach, S.D., 1999. Topology of Euglena Chloroplast Protein Precursors within Endoplasmic Reticulum to Golgi to Chloroplast Transport Vesicles *. *Journal of Biological Chemistry* 274, 457–463. <https://doi.org/10.1074/jbc.274.1.457>
- Tatusov, R.L., Galperin, M.Y., Natale, D.A., Koonin, E.V., 2000. The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res* 28, 33–36. <https://doi.org/10.1093/nar/28.1.33>
- Tatusov, R.L., Koonin, E.V., Lipman, D.J., 1997. A Genomic Perspective on Protein Families. *Science* 278, 631–637. <https://doi.org/10.1126/science.278.5338.631>
- Teich, R., Zauner, S., Baurain, D., Brinkmann, H., Petersen, J., 2007. Origin and Distribution of Calvin Cycle Fructose and Sedoheptulose Bisphosphatases in Plantae and Complex Algae: A Single Secondary

Origin of Complex Red Plastids and Subsequent Propagation via Tertiary Endosymbioses. *Protist* 158, 263–276. <https://doi.org/10.1016/j.protis.2006.12.004>

Tekaia, F., 2016. Inferring Orthologs: Open Questions and Perspectives. *Genomics Insights* 9, GEI.S37925. <https://doi.org/10.4137/GEI.S37925>

Tengs, T., Dahlberg, O.J., Shalchian-Tabrizi, K., Klaveness, D., Rudi, K., Delwiche, C.F., Jakobsen, K.S., 2000. Phylogenetic Analyses Indicate that the 19'Hexanoyloxy-fucoxanthin-Containing Dinoflagellates Have Tertiary Plastids of Haptophyte Origin. *Molecular Biology and Evolution* 17, 718–729. <https://doi.org/10.1093/oxfordjournals.molbev.a026350>

Terpis, K.X., Salomaki, E.D., Barcytė, D., Pánek, T., Verbruggen, H., Kolisko, M., Bailey, J.C., Eliáš, M., Lane, C.E., 2024. Multiple plastid losses within photosynthetic stramenopiles revealed by comprehensive phylogenomics. <https://doi.org/10.1101/2024.02.03.578753>

Tice, A.K., Žihala, D., Pánek, T., Jones, R.E., Salomaki, E.D., Nenarokov, S., Burki, F., Eliáš, M., Eme, L., Roger, A.J., Rokas, A., Shen, X.-X., Strasser, J.F.H., Kolisko, M., Brown, M.W., 2021. PhyloFisher: A phylogenomic package for resolving eukaryotic relationships. *PLOS Biology* 19, e3001365. <https://doi.org/10.1371/journal.pbio.3001365>

Timmis, J.N., Ayliffe, M.A., Huang, C.Y., Martin, W., 2004. Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes. *Nat Rev Genet* 5, 123–135. <https://doi.org/10.1038/nrg1271>

Turmel, M., Gagnon, M.-C., O'Kelly, C.J., Otis, C., Lemieux, C., 2009. The Chloroplast Genomes of the Green Algae *Pyramimonas*, *Monomastix*, and *Pycnococcus* Shed New light on the Evolutionary History of Prasinophytes and the Origin of the Secondary Chloroplasts of Euglenids. *Molecular Biology and Evolution* 26, 631–648. <https://doi.org/10.1093/molbev/msn285>

Uzuka, A., Kobayashi, Y., Onuma, R., Hirooka, S., Kanesaki, Y., Yoshikawa, H., Fujiwara, T., Miyagishima, S., 2019. Responses of unicellular predators to cope with the phototoxicity of photosynthetic prey. *Nat Commun* 10, 5606. <https://doi.org/10.1038/s41467-019-13568-6>

Van Vlierberghe, M., Di Franco, A., Philippe, H., Baurain, D., 2021a. Decontamination, pooling and dereplication of the 678 samples of the Marine Microbial Eukaryote Transcriptome Sequencing Project. *BMC Res Notes* 14, 306. <https://doi.org/10.1186/s13104-021-05717-2>

Van Vlierberghe, M., Philippe, H., Baurain, D., 2021b. Broadly sampled orthologous groups of eukaryotic proteins for the phylogenetic study of plastid-bearing lineages. *BMC Res Notes* 14, 143. <https://doi.org/10.1186/s13104-021-05553-4>

Veseli, I., 2023. KEGG_build_2023-09-22. <https://doi.org/10.6084/M9.FIGSHARE.24183009>

Veseli, I., Chen, Y.T., Schechter, M.S., Vanni, C., Fogarty, E.C., Watson, A.R., Jabri, B., Blekhman, R., Willis, A.D., Yu, M.K., Fernández-Guerra, A., Füssel, J., Eren, A.M., 2025. Microbes with higher metabolic independence are enriched in human gut microbiomes under stress. *eLife* 12, RP89862. <https://doi.org/10.7554/eLife.89862>

Vilella, A.J., Severin, J., Ureta-Vidal, A., Heng, L., Durbin, R., Birney, E., 2009. EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res.* 19, 327–335. <https://doi.org/10.1101/gr.073585.107>

Vos, R.A., Caravas, J., Hartmann, K., Jensen, M.A., Miller, C., 2011. BIO::Phylo-phyloinformatic analysis using perl. *BMC Bioinformatics* 12, 63. <https://doi.org/10.1186/1471-2105-12-63>

Vosseberg, J., Van Hooff, J.J.E., Köstlbacher, S., Panagiotou, K., Tamarit, D., Ettema, T.J.G., 2024. The emerging view on the origin and early evolution of eukaryotic cells. *Nature* 633, 295–305. <https://doi.org/10.1038/s41586-024-07677-6>

Wall, D.P., Fraser, H.B., Hirsh, A.E., 2003. Detecting putative orthologs. *Bioinformatics* 19, 1710–1711. <https://doi.org/10.1093/bioinformatics/btg213>

Waller, R.F., Kořený, L., 2017. Plastid Complexity in Dinoflagellates: A Picture of Gains, Losses, Replacements and Revisions, in: *Advances in Botanical Research*. Elsevier, pp. 105–143. <https://doi.org/10.1016/bs.abr.2017.06.004>

Watanabe, M.M., Suda, S., Inouya, I., Sawaguchi, T., Chihara, M., 1990. LEPIDODINIUM VIRIDE GEN. ET SP. NOV. (GYMNODINIALES, DINOPHYTA), A GREEN DINOFLAGELLATE WITH A CHLOROPHYLL *A*- AND *B* -CONTAINING ENDOSYMBIONT^{1, 2}. *Journal of Phycology* 26, 741–751. <https://doi.org/10.1111/j.0022-3646.1990.00741.x>

Weatherby, K., Carter, D., 2013. *Chromera velia*, in: *Advances in Applied Microbiology*. Elsevier, pp. 119–144. <https://doi.org/10.1016/B978-0-12-407672-3.00004-6>

Wickett, N.J., Mirarab, S., Nguyen, N., Warnow, T., Carpenter, E., Matasci, N., Ayyampalayam, S., Barker, M.S., Burleigh, J.G., Gitzendanner, M.A., Ruhfel, B.R., Wafula, E., Der, J.P., Graham, S.W., Mathews, S., Melkonian, M., Soltis, D.E., Soltis, P.S., Miles, N.W., Rothfels, C.J., Pokorny, L., Shaw, A.J., DeGironimo, L., Stevenson, D.W., Surek, B., Villarreal, J.C., Roure, B., Philippe, H., dePamphilis, C.W., Chen, T., Deyholos, M.K., Baucom, R.S., Kutchan, T.M., Augustin, M.M., Wang, J., Zhang, Y., Tian, Z., Yan, Z., Wu, X., Sun, X., Wong, G.K.-S., Leebens-Mack, J., 2014. Phylotranscriptomic analysis of the origin and early diversification of land plants. *Proceedings of the National Academy of Sciences* 111, E4859–E4868. <https://doi.org/10.1073/pnas.1323926111>

Wietrzynski, W., Traverso, E., Wollman, F.-A., Wostrickoff, K., 2021. The state of oligomerization of Rubisco controls the rate of synthesis of the Rubisco large subunit in *Chlamydomonas reinhardtii*. *Plant Cell* 33, 1706–1727. <https://doi.org/10.1093/plcell/koab061>

Wilkinson, M., McInerney, J.O., Hirt, R.P., Foster, P.G., Embley, T.M., 2007. Of clades and clans: terms for phylogenetic relationships in unrooted trees. *Trends in Ecology & Evolution* 22, 114–115. <https://doi.org/10.1016/j.tree.2007.01.002>

Wisecaver, J.H., Hackett, J.D., 2010. Transcriptome analysis reveals nuclear-encoded proteins for the maintenance of temporary plastids in the dinoflagellate *Dinophysis acuminata*. *BMC Genomics* 11, 366. <https://doi.org/10.1186/1471-2164-11-366>

Wolf, Y.I., Koonin, E.V., 2012. A Tight Link between Orthologs and Bidirectional Best Hits in Bacterial and Archaeal Genomes. *Genome Biol Evol* 4, 1286–1294. <https://doi.org/10.1093/gbe/evs100>

Woo, Y.H., Ansari, H., Otto, T.D., Klinger, C.M., Kolisko, M., Michálek, J., Saxena, A., Shanmugam, D., Tayyrov, A., Veluchamy, A., Ali, S., Bernal, A., Del Campo, J., Cihlář, J., Flegontov, P., Gornik, S.G., Hajdušková, E., Horák, A., Janouškovec, J., Katris, N.J., Mast, F.D., Miranda-Saavedra, D., Mourier, T., Naeem, R., Nair, M., Panigrahi, A.K., Rawlings, N.D., Padron-Regalado, E., Ramaprasad, A., Samad, N., Tomčala, A., Wilkes, J., Neafsey, D.E., Doerig, C., Bowler, C., Keeling, P.J., Roos, D.S., Dacks, J.B., Templeton, T.J., Waller, R.F., Lukeš, J., Oborník, M., Pain, A., 2015. Chromerid genomes reveal the evolutionary path from photosynthetic algae to obligate intracellular parasites. *eLife* 4, e06974. <https://doi.org/10.7554/eLife.06974>

Wood, D.E., Lu, J., Langmead, B., 2019. Improved metagenomic analysis with Kraken 2. *Genome Biol* 20, 257. <https://doi.org/10.1186/s13059-019-1891-0>

Yabuki, A., Kamikawa, R., Ishikawa, S.A., Kolisko, M., Kim, E., Tanabe, A.S., Kume, K., Ishida, K., Inagaki, Y., 2014. *Palpitomonas bilix* represents a basal cryptist lineage: insight into the character evolution in Cryptista. *Sci Rep* 4, 4641. <https://doi.org/10.1038/srep04641>

Yamada, N., Bolton, J.J., Trobajo, R., Mann, D.G., Dąbek, P., Witkowski, A., Onuma, R., Horiguchi, T., Kroth, P.G., 2019. Discovery of a kleptoplastic ‘dinotom’ dinoflagellate and the unique nuclear dynamics of converting kleptoplastids to permanent plastids. *Sci Rep* 9, 10474. <https://doi.org/10.1038/s41598-019-46852-y>

Yamada, N., Dorrell, R.G., Cenci, U., Kroth, P.G., Lombard, V., Sprecher, B.N., 2025. How to upgrade stolen organelles into permanent plastids: A comparative transcriptomic perspective. *Proceedings of the National Academy of Sciences* 122, e2514821122. <https://doi.org/10.1073/pnas.2514821122>

Yamada, N., Lepetit, B., Mann, D.G., Sprecher, B.N., Buck, J.M., Bergmann, P., Kroth, P.G., Bolton, J.J., Dąbek, P., Witkowski, A., Kim, S.-Y., Trobajo, R., 2023. Prey preference in a kleptoplastic dinoflagellate is linked to photosynthetic performance. *ISME J* 17, 1578–1588. <https://doi.org/10.1038/s41396-023-01464-3>

Yamaguchi, A., Yubuki, N., Leander, B.S., 2012. Morphostasis in a novel eukaryote illuminates the evolutionary transition from phagotrophy to phototrophy: description of *Rapaza viridis* n. gen. et sp. (Euglenozoa, Euglenida). *BMC Evolutionary Biology* 12, 29. <https://doi.org/10.1186/1471-2148-12-29>

Yang, Y., Matsuzaki, M., Takahashi, F., Qu, L., Nozaki, H., 2014. Phylogenomic Analysis of “Red” Genes from Two Divergent Species of the “Green” Secondary Phototrophs, the Chlorarachniophytes, Suggests Multiple Horizontal Gene Transfers from the Red Lineage before the Divergence of Extant Chlorarachniophytes. *PLoS ONE* 9, e101158. <https://doi.org/10.1371/journal.pone.0101158>

Zhaxybayeva, O., Doolittle, W.F., 2011. Lateral gene transfer. *Current Biology* 21, R242–R246. <https://doi.org/10.1016/j.cub.2011.01.045>

- Zhu, G., Marchewka, M.J., Keithly, J.S., 2000. *Cryptosporidium parvum* appears to lack a plastid genome. *Microbiology* 146, 315–321. <https://doi.org/10.1099/00221287-146-2-315>
- Zhu, Q., Kosoy, M., Dittmar, K., 2014. HGTector: an automated method facilitating genome-wide discovery of putative horizontal gene transfers. *BMC Genomics* 15, 717. <https://doi.org/10.1186/1471-2164-15-717>