PROJECT TITLE

Genome dynamics following plastid endosymbiosis

AIMS AND BACKGROUND

Aims

This project aims to address significant gaps in our knowledge of the evolution of chloroplast endosymbiosis and the methods used to study them. It uses a combination of empirical and computational techniques on a carefully chosen case study. The specific goals are:

1. To develop a suitable case study by pinpointing the origin of chloroplasts of two algal groups (chlorarachniophytes and green dinoflagellates) with densely sampled genomic datasets.
2. To characterize in detail the genome dynamics following eukaryote-eukaryote endosymbiosis.
3. To determine the power and improve phyloinformatics approaches towards endosymbiosis research.

This project will deliver detailed, quantitative information about what happens to the pool of genomes in a cell following endosymbiosis. While novel and interesting in its own right, it also has broader methodological implications. The project builds on the information gained to assess the suitability of analysis techniques and develop improved methods. Finally, the project is set up to expedite the utilization of the acquired genomic data in collaborators' projects to broaden the research base and further augment its output.

Background

The evolutionary history of eukaryotic algae is driven by plastid endosymbiosis, a process in which a heterotrophic organism gains the capacity to perform photosynthesis by engulfing a photosynthetic organism that, over time, evolves to become an organelle. The acquisition of plastids through endosymbiosis is one of the most important events in the evolution of the eukaryotic cell. Besides supplying it with a continuous internal source of organic compounds, the compartmentalization of the cell and the cocktail of genes resulting from endosymbiosis offer an astounding capacity for adaptation and metabolic flexibility [1-3]. The ensuing increase in physiological and ecological potential has conveyed advantages that have made eukaryotic algae dominant life forms. They support the oceanic food chain, play crucial roles in various biogeochemical cycles, and are responsible for ca. 50% of primary production worldwide [4], which has also made them targets for biofuel production. Despite its global importance, many questions about endosymbiosis remain unanswered. This project addresses some of those questions using a combination of leading-edge empirical and computational methods.

The first eukaryotic alga originated from the merger of a heterotrophic eukaryote and a photosynthetic cyanobacterium that evolved to become a chloroplast stably integrated into the cell [5,6] (Fig. 1a). As descendants of the resulting eukaryotic alga diversified over time, the eukaryotic algae from various lineages became involved in secondary endosymbiosis events [5,6]. In secondary endosymbiosis, a heterotrophic eukaryote engulfs a photosynthetic eukaryote, resulting in a eukaryotic alga with a complex eukaryote-derived secondary plastid. At least seven eukaryote-eukaryote endosymbiosis events have occurred, involving heterotrophic hosts from many of the major eukaryotic groups and plastids derived from various other branches in the eukaryotic tree of life [6]. As a consequence, the capacity to perform photosynthesis has spread to many eukaryotic lineages, giving rise to the major algal groups (e.g. heterokonts, haptophytes, dinoflagellates, etc.) existing today.

Due to endosymbiosis, algal genomes contain a mixture of genes from various sources. After endosymbiosis, part of the plastid genome is lost, part is retained, and part is transferred to the host nucleus (endosymbiotic gene transfer; Fig. 1b). As a consequence, the plastid genome of eukaryotic algae resulting from the primary endosymbiosis is much reduced, while the nuclear genome has a large cyanobacterial component. Similarly, the genome of algal lineages resulting from secondary endosymbiosis is a mixture of genes from the host, the nucleus of the engulfed cell, and the plastid of the engulfed cell (i.e., 3 contributing genomes). The situation is more complex for tertiary endosymbioses (4 contributing genomes) or serial secondary endosymbioses, in which a secondary plastid is replaced by another plastid, leaving genomic remnants of the first endosymbiosis (3 genomes) complemented with genes from the new endosymbiosis (2 additional contributing genomes). The result is quite literally a hodgepodge of genes of various sources, which may explain the physiological flexibility and abundance of some algal lineages.

Since endosymbiosis is a process that occurred at some point in the past, it cannot be directly observed. Instead, one must infer potential endosymbiosis events from patterns observed in present-day genomes. More specifically, the plastid genome and the genes transferred to the nucleus contain information that can help identify the origin of the plastid. The inferences are generally based on phylogenetic analyses of genes and/or genomes. If a plastid gene is analysed along with the same gene from a range of free-living organisms, one can deduce that the organism that got engulfed during the endosymbiosis event was a relative of the free-living organism whose sequence is most closely related to that of the plastid. The same reasoning applies if the gene was transferred from the plastid to the host nucleus, with additional evidence potentially to be found in the plastid targeting signal of nuclear genes. For higher-order endosymbiosis events, genes from the nuclear genome of the engulfed plastid can also be transferred to the host nucleus, adding another layer of complexity in the analysis.

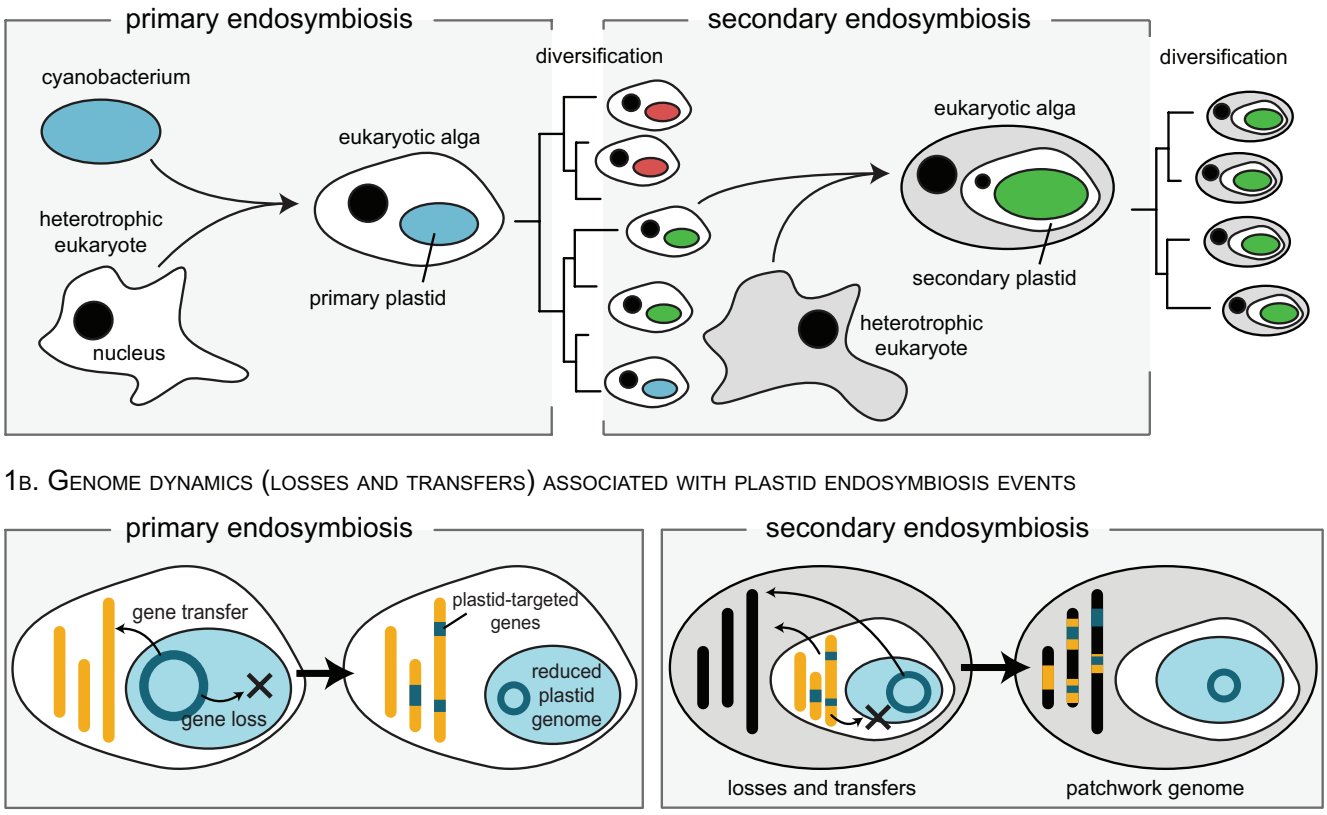

Fig 1. Graphical representation of the plastid endosymbiosis process and associated genome dynamics.

Problem statement

Even though the general concept of endosymbiosis is widely embraced by the scientific community, researchers disagree about how many plastid endosymbiosis events have taken place and which organisms were involved [6-8]. The reason for the controversies and uncertainties is that it is very difficult to make reliable inferences about events in the distant past from phylogenetic analyses of genes observed in present-day genomes. Phylogenetic analyses can yield biased results if taxa differ in their GC content and codon usage patterns, especially when inferences have to be made about ancient relationships based on sequences of just a handful of species [9-11]. Changes in GC and codon usage are likely to happen during endosymbiosis because of different usage patterns between plastid and host, and most plastid endosymbiosis events happened over 500 million or even a billion years ago, so there is certainly cause for concern [12-14]. In addition, there is disagreement about what exactly constitutes evidence for endosymbiosis in phylogenetic analyses, especially in situations where plastid loss or replacement has occurred [15]. Because of these problems, it remains difficult to present conclusive evidence and many open questions remain in endosymbiosis research. We will now dissect these problems in more detail and offer solutions that will form the aims of this project.

The first problem is our fragmentary knowledge of algal genomes. While this situation is improving thanks to several initiatives, algal genomes remain very sparsely sampled with respect to the immense phylogenetic diversity they have accumulated during 1.5 billion years of evolution. This fragmentary knowledge is an impediment to endosymbiosis research for three reasons. First, a good comparative dataset of gene sequences is

needed to pinpoint the origin of the plastids of the various eukaryotic algal groups. Only when close extant relatives of the organism that was engulfed are included, can reliable inferences about the exact origin of the plastid be made. Second, poor taxon sampling contributes greatly to the systematic biases that affect phylogenetic analyses [16,17]. Because the occurrence of past endosymbiosis events is inferred from them, obtaining unbiased phylogenetic trees is crucial. Third, the ability to study genome dynamics following secondary endosymbiosis (Fig. 2) relies on good comparative datasets of plastid and nuclear genes. Again, only when close extant relatives are present in these datasets, can reliable inferences be made about which endosymbiont genes were likely lost and which were transferred to the host nucleus, perhaps replacing some of the host's nuclear genes. In conclusion, the fragmentary knowledge of algal genomes currently hinders accurate inference of endosymbiosis events and the genome dynamics associated with them. Assembling a comparative genome dataset with a much denser and focused sampling of species will largely solve this problem.

The secondary endosymbiosis events involving green algae-derived plastids have received less attention than those involving red algae-derived plastids, probably because they have not given rise to highly dominant lineages. Yet they offer an outstanding model system to study the genome dynamics involved in secondary endosymbiosis in detail for three reasons. First, there are two independent endosymbiosis events leading to chlorarachniophytes and green dinoflagellates, offering natural replicates. Second, these events are not as old as the chromalveolate endosymbiosis, facilitating the recovery of genome dynamics signals. Third, they involve related plastid donors, making it easier and cheaper to assemble reference datasets. Phylogenetic studies have shown the host of the chlorarachniophytes to be a cercozoan amoeba [18] and that of the green dinoflagellates (*Lepidodinium*) a *Gymnodinium* dinoflagellate. While the host origin of *Lepidodinium* is fairly well established [19,20], that of the chlorarachniophytes is vague. In both cases, the plastids are of "core chlorophyte" origin, which is very vague because this lineage is rich in early-branching lineages [21,22]. While there is a good basis of genome information for chlorarachniophytes (*Bigelowiella*), dinoflagellates (*Symbiodinium*) and core chlorophytes (*Coccomyxa*, *Chlorella, Chlamydomonas*), much denser sampling will be needed to gain detailed insight in genome dynamics.

Endosymbiosis research also faces a more conceptual problem related to the interpretation of phylogenies. It is important to define more objective criteria to determine the origin of genes found in the nuclear genomes of eukaryotic algae [15]. When, in a phylogenetic analysis, a host nuclear gene sequence is not recovered among sequences of species related to the host but rather to sequences of an algal taxon, this is often taken as evidence of endosymbiotic gene transfer. This interpretation is problematic because such relationships can also arise by horizontal gene transfer in the absence of endosymbiosis and due to the fact that phylogenetic conflicts are not necessarily a consequence of gene transfers. This begs the question what finding genes of "algal origin" in a host genome based on gene trees means, and whether *a priori* criteria can be designed as to what constitutes evidence of endosymbiotic gene transfer. Considering these uncertainties in the interpretation of single-gene phylogenetic trees, an overarching question that should be asked is whether and to what extent the currently used phyloinformatics pipelines can infer endosymbiotic gene transfer.

Starting from these problems, we formulate the following working hypotheses for our project.

1. Host and plastid origins can be pinpointed with densely sampled genome data because both nuclear and plastid data will contribute to the inference.
2. Genome dynamics following endosymbiosis can be quantified from densely sampled genome data for events of intermediate age.
3. Existing single-gene phylogenomics pipelines to infer endosymbiosis will fail to yield accurate results for ancient endosymbiosis events as a consequence of deterioration of the signal in the data.
4. Modifications of the phyloinformatic pipelines informed by knowledge of genome dynamics will improve inferences.

Using these hypotheses as a starting point, our project will address important knowledge gaps in endosymbiosis research. The project goals defined at the start of the project proposal naturally follow from the problems outlined above and permit addressing these four working hypotheses.

RESEARCH PROJECT

Conceptual framework

The project takes advantage of 2nd and 3rd generation sequencing technologies to create densely sampled comparative genomic datasets that can be used to refine the origin of chlorarachniophyte and green

dinoflagellate plastids and track the genome dynamics associated with endosymbiosis. The project will also apply novel computational tools to investigate the suitability of classical methods used to infer endosymbiosis (hybrid simulation case-study approaches) and assess the merits and potential drawbacks of alternative methods (alignment-free phylogenomics). We will now address each of these aspects in more detail.

Study design for Goal 1 – Pinpoint endosymbiosis events with dense sampling

The first goal serves to develop a suitable case study. We have chosen to work with the endosymbiosis events leading up to the chlorarachniophytes and the green dinoflagellates. These form a suitable model system because there is a good base of background information (see above) and because they are of intermediate age, which facilitates estimation of genome dynamics parameters.

Eighteen algal and protist strains (see table below) will be grown in culture and sequenced using a combination of RNA-seq and WGS to obtain a draft genome sequence. The strains include green algae related to the plastid source, dinoflagellates (*Lepidodinium* and related *Gymnodinium*) and Cercozoa (chlorarachniophytes and related heterotrophs). The taxa have been selected to represent maximal phylogenetic diversity within the algal lineages in question, based on the literature and preliminary analyses by CI Verbruggen. In addition to these taxa, we will use taxa for which data is already available (several green algae, the dinoflagellate *Symbiodinium*, the chlorarachniophyte *Bigelowiella*; see above).

species	strain	group	purpose of strain
Pedinomonas minor	UTEX LB 1350	green alga	
Tetraselmis chuii	RCC128	green alga	
Acetabularia acetabulum	DI1	green alga	
Neocystis brevis	CAUP D 802	green alga	pinpoint plastid origins, establish
Leptosira terrestris	UTEX 333	green alga	genome content of plastid donor
Oocystis solitaria	SAG 83.80	green alga	prior to endosymbiosis, estimate
Koliella corcontica	SAG 24.84	green alga	background processes
Ostreobium quekettii	SBWA 26A	green alga	
Bryopsis hypnoides	West 4666	green alga	
Boodlea sp.10	FL 1101	green alga	
Lepidodinium viride	CTCC 17	dinoflagellate, green	establish genome dynamics
Lepidodinium chlorophorum	NIES 1868	dinoflagellate, green	**following** endosymbiosis event
Gymnodinium fuscum	CCMP 1677	dinoflagellate, peridinin	establish genome content of host
Gymnodinium catenatum	CCMP 1937	dinoflagellate, peridinin	**prior to** endosymbiosis event
Gymnochlora stellata	GUAM-1	Cercozoa, chlorarachniophyte	establish genome dynamics
Lotharella globosa	NEPCC 920	Cercozoa, chlorarachniophyte	**following** endosymbiosis event
Limnofila borokensis	ATCC50638	Cercozoa, heterotroph	establish genome content of host
Cercomonas vibrans	ATCC50530	Cercozoa, heterotroph	**prior to** endosymbiosis event

After strains have been grown to a reasonable density, they will be harvested and DNA and RNA extracted using protocols known to perform well for the group in question and sequenced with short-read (Illumina) and long-read (PacBio) technologies. We will multiplex 3 to 4 normalized RNA samples per Illumina lane. For *de novo* sequencing of plastid and nuclear genomes we will start with one lane of Illumina sequencing containing three libraries with different insert sizes. Based on the obtained coverage and assembly success, we may add additional Illumina lanes and/or PacBio SMRT cells. Our total sequencing budget is based on published estimates of genome sizes for green algae and dinoflagellates, a target average coverage of 100x (Illumina) and 5x (PacBio), and takes realistic levels of contamination of cultures with bacteria into account. Based on prior experience, we expect that 3-4 strains may not grow well enough in culture to obtain enough DNA. In such cases, we will perform whole genome amplification with the PicoPlex protocol prior to sequencing. Based on experience and published information, good assemblies of the coding regions (i.e., the data we need for further analyses) can be obtained even for genomes with large fractions of repetitive DNA and our preliminary data suggests that the minicircles that carry dinoflagellate chloroplast genes assemble well from Illumina data.

Raw sequencing data will be processed and filtered using CASAVA and further filtered for base-quality scores and/or normalised based on coverage biases to eliminate data redundancy and errors, and to limit sampling variation. We will make independent assemblies using the CLC Genomics Workbench, ABySS, Velvet and

MIRA, all of which we are routinely using. Tandem and interspersed repetitive elements and low-complexity regions will be identified using RepeatMasker. Gene prediction will be carried out with AUGUSTUS and GeneMark-E, using existing transcriptome data to guide our predictions. Predicted genes will be annotated by similarity search against sequence and protein-model (PRINTS, Pfam-A and -B, SCOP) databases, by domain prediction and Blast2GO, and manual curation. Genes and encoded proteins will be independently clustered into homologous sets using a Markov clustering algorithm and normalised all-versus-all BLAST scores.

The genomic data will be analysed with phylogenomic pipelines to pinpoint plastid origins. Two approaches will be used. First, the assembled plastid genomes (or minicircles) will be subjected to phylogenetic analysis to identify the closest relatives of the endosymbionts. These analyses will use the latest developments in model-based phylogenetic inference to avoid bias due to site- and tree-heterogeneity in evolutionary processes [23]. Second, the origin of the nuclear genes of the host will be determined. This will be achieved with model-based phylogenetic inference (ML with bootstrapping) and tree sorting pipelines [24]. Trees will be filtered based on reliable topological features to extract the origin of different genes and infer past endosymbiosis events [25]. In addition to this traditional method, we will use alignment-free phylogenomics [26] to capture signal otherwise lost to gappiness, recombination and shuffling. In addition, we will explore the use of spectral partitioning [27] to identify conflicting signals that could be due to endosymbiotic gene transfer and model-based techniques that specifically take horizontal gene transfer into account [28-31]. These techniques are still in development and some are designed with different goals in mind, but our datasets offer a perfect opportunity to investigate and advance their application in endosymbiosis research. They offer new ways to explore the conflicting signals that plague classical approaches and partial solutions to correctly infer the evolutionary history in the presence of such conflicts. The application of this array of analyses on our densely sampled datasets will help us pinpoint the origin of the plastids with the greatest degree of accuracy currently achievable.

Study design for Goal 2 – Quantify genome dynamics following eukaryote-eukaryote endosymbiosis

For this goal, the genomic datasets obtained under Goal 1 will be further analysed with the aim of characterizing and quantifying the changes that happen in the host and endosymbiont genomes during and following endosymbiosis. We will estimate the parameters of a model consisting of seven processes that affect the evolution of the pool of genomes in algal cells (Fig. 2).

1. Plastid engulfment. This adds two entire genomes to the pool of genomes of the host cell and occurs at a low rate.
2. Gene loss. An exponential decay model will be used to model gene loss following endosymbiosis while the background rate will be estimated from our big green algal dataset.
3. The mutation and fixation category pertains to the standard processes of molecular evolution (vertical inheritance).
4. Gene duplication. We will estimate the background rate of gene duplication using the big green algal dataset.
5. Recombination of domain-coding regions. Adopting our earlier approach in detecting within-gene recombination [32], we will quantify the correlation between recombined genic regions.
6. Import by horizontal gene transfer (HGT). By comparing our data with sequences from a wide sample of protists, we will attempt to quantify how commonly HGT occurs.
7. Transfer between organelles will also consist of exponential decline of gene transfer from plastid to nucleus following endosymbiosis plus a background rate.

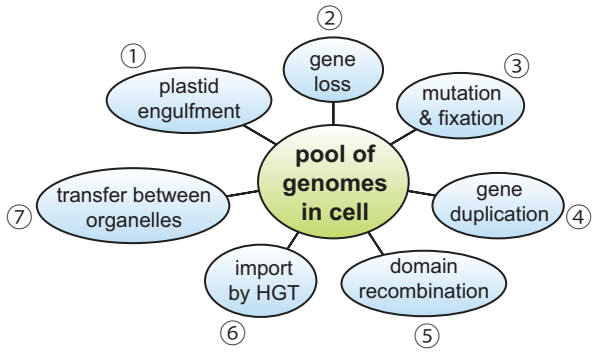

Fig 2. Conceptual model of the genome dynamics of algal cells. These seven processes affect the pool of genomes in the algal cell at each time step of the model. (HGT = horizontal gene transfer)

The procedure to estimate these rates of change consists of a few steps. Based on our densely sampled datasets, we will determine the presence-absence and/or frequency of genes in the sequenced organisms. We will calibrate the reference phylogenetic trees obtained for Goal 1 in geological time using existing temporal data [33-35]. Then we will optimize the rate parameters of models of evolutionary change in the presence-absence and/or frequency of genes to estimate the rate at which each of the processes in the model happen [36,37]. Mutation and fixation (process 3) is modelled with Markov models [38]. For gene loss and transfer following endosymbiosis (processes 2 & 7), exponential decay models have proven a good fit [39]. Most other processes (3, 4, 5, 6) are probably better modelled as constant-rate processes. Transfers from the plastid to the nucleus

will also be approached by analysing NUPTs, offering an alternative source of information. The characterisation of genome dynamics will be refined with details about the functional annotation of genes and whether the genes are transferred or lost. The dynamics of GC content, codon utilisation and intron size will be assessed with evolutionary models [40,41].

It is important to note that we chose not to develop and implement model-based procedures for joint estimation of all these parameters. Instead, we will infer parameters individually from the data. While we recognize the desirability of methods to jointly estimate all parameters, we do not consider this achievable in addition to our present goals. This project will however lay the conceptual foundation for a follow-up project of this nature involving phylogenetic methods developers.

Previous studies have rarely tried to quantify the rate of plastid genome reduction or other changes in the genome, and the exceptions have probably suffered from sparse taxon sampling. Our densely sampled dataset offers an exceptional opportunity to advance this field. Besides the interest of studying genome dynamics in and of itself, the quantitative information gathered in this experience will help parameterize the model used for the third goal.

Study design for Goal 3 – Improve and test inference methods in endosymbiosis research

This aspect of the project uses simulation experiments to examine under which conditions endosymbiosis events can be recovered from genomic data with phyloinformatics pipelines. Simulation is a powerful tool for testing the logical consistency of ideas, efficiency of methods, and reliability (bias and precision) of estimation methods. It will be used to achieve three interrelated goals. First, we will determine under which scenarios of genome dynamics an endosymbiosis event leaves a sufficiently large footprint to be detected and, equally importantly, under which conditions there is little hope of accurately inferring such events. Second, we will establish whether the currently used phyloinformatics approaches can reliably infer endosymbiosis events, in other words, assess these methods' efficiency. Third, we will develop, implement and test new pipelines that use alignment-free inference methods as well as model-based methods that should improve inferences.

As a first step, the comprehensive model of genome dynamics in Fig. 2 will be implemented. We will use simulation to generate genomic datasets of sets of species under various scenarios of plastid endosymbiosis (old and recent events, primary, secondary endosymbiosis, etc.). The studies of genome dynamics following endosymbiosis (Goal 2) will provide a good indication of realistic values of the model parameters and their variation between lineages. By using "hybrid simulation case studies", in which realistic parameter values inferred from case studies are varied within reasonable bounds, the simulations will apply to a broader range of theoretical endosymbiosis events. The resulting comparative genomic datasets will be analyzed using existing phyloinformatics pipelines and alternative methods (see under Goal 1 and below), permitting us to evaluate under which conditions the correct scenario can be detected from the footprints left in simulated genome data. Thanks to the hybrid approach, the simulations are run for a realistic range of parameter values and conclusions can be extrapolated to real-life data. We will follow up by taking inferences a step further and investigate whether it is feasible to reconstruct more complex scenarios of endosymbiosis. This will be useful to infer the confidence one may have in cryptic endosymbiosis events, which have been proposed [25,42] but not universally accepted [15]. It can also be used to evaluate more elaborate scenarios such as a series of nested endosymbioses [7] and potential mechanisms of endosymbiont establishment such as the targeting ratchet [6].

This dataset has great potential as a case study for a new approach that does not require prior multiple sequence alignment [26,43]. Alignment-free phylogenomic analyses based on *k*-mer spectra are highly scalable due to lower computational complexity and bypass the assumption of full-length contiguity of homologous sequences, which is unrealistic due to genome rearrangement, recombination and horizontal transfer. Among other things, they facilitate detecting recombination and horizontal transfer as unexpected *k*-mer spectra. Our preliminary analyses show that this approach is accurate, robust and scalable in application to synthetic and empirical data. This project offers an important testbed for further critical assessment and detailed testing of the behaviour of this approach in simulation studies, and will help determine their future role in phylogenomic pipelines.

REFERENCES – 1. Bowler C *et al.* (2010) *Annual Review of Marine Science* 2: 333-365. **2.** Allen AE *et al.* (2006) *Current Opinion in Plant Biology* 9: 264-273. **3.** Johnson MD (2011) *Photosynth Res* 107: 117-132. **4.** Falkowski PG *et al.* (2004) *Science* 305: 354-360. **5.** Delwiche CF (1999) *Am Nat* 154: S164-S177. **6.** Keeling PJ (2013) *Ann Rev Plant Biol* 64: 583-607. **7.** Baurain D *et al.* (2010) *Mol Biol Evol* 27: 1698-1709. **8.** Keeling PJ (2009) *J Eukar Microbiol* 56: 1-8. **9.** Bergsten J (2005) *Cladistics* 21: 163-193. **10.** Jermiin LS *et al.* (2004) *Syst Biol* 53: 638-643. **11.** Ho SYW, Jermiin LS (2004) *Syst Biol* 53: 623-637. **12.** Yoon HS *et al.* (2004) *Mol Biol Evol* 21: 809-818. **13.** Lockhart P *et al.* (2006) *Mol Biol Evol* 23: 40-45. **14.** Shalchian-Tabrizi K *et al.* (2006) *Mol Biol Evol* 23: 1504-1515. **15.** Stiller J (2011) *BMC Evol Biol* 11: 259. **16.** Hedtke SM *et al.* (2006) *Syst Biol* 55: 522-529. **17.** Zwickl DJ, Hillis DM (2002) *Syst Biol* 51: 588-598. **18.** Keeling PJ (2001) *Mol Biol Evol* 18: 1551-1557. **19.** Murray S *et al.* (2005) *Protist* 156: 269-286. **20.** Shalchian-Tabrizi K *et al.* (2006) *J Eukar Microbiol* 53: 217-224. **21.** Matsumoto T *et al.* (2010) *Protist* 162: 268-276. **22.** Rogers MB *et al.* (2007) *Mol Biol Evol* 24: 54-62. **23.** Verbruggen H, Theriot EC (2008) *Eur J Phycol* 43: 229–252. **24.** Moustafa A *et al.* In; 2010 16-18 Dec. 2010. pp. 103-107. **25.** Moustafa A *et al.* (2009) *Science* 324: 1724-1726. **26.** Chan CX, Ragan M (2013) *Biology Direct* 8: 3. **27.** Chen D *et al.* (2007) *Syst Biol* 56: 623 - 632. **28.** Akerborg O *et al.* (2009) *Proc Natl Acad Sci USA* 106: 5714-5719. **29.** Gorecki P *et al.* (2011) *BMC Bioinf* 12: S15. **30.** Bloomquist EW, Suchard MA (2010) *Syst Biol* 59: 27-41. **31.** Galtier N, Daubin V (2008) *Phil Trans Roy Soc B* 363: 4023-4029. **32.** Chan CX *et al.* (2009) *Genome Biol Evol* 1: 429-438. **33.** Hackett JD *et al.* (2007) In: Falkowski PG, Knoll AH, editors. Evolution of primary producers in the sea. pp. 109-132. **34.** Verbruggen H *et al.* (2009) *Mol Phylogenet Evol* 50: 642-653. **35.** Parfrey LW *et al.* (2011) *Proc Natl Acad Sci USA* 108: 13624-13629. **36.** Barker D *et al.* (2007) *Bioinformatics* 23: 14-20. **37.** Cocquyt E *et al.* (2009) *BMC Evol Biol* 9: 39. **38.** Yang Z (2006) Computational molecular evolution. 376 p. **39.** Khachane AN *et al.* (2007) *Mol Biol Evol* 24: 449-456. **40.** Cocquyt E (2009) Phylogeny and molecular evolution of green algae [PhD thesis]. 167 p. **41.** Butler MA, King AA (2004) *Am Nat* 164: 683-695. **42.** Chan CX *et al.* (2011) *PLoS One* 6: e29138. **43.** Domazet-Lošo M, Haubold B (2011) *Bioinformatics* 27: 1466-1472. **44.** Curtis BA *et al.* (2012) *Nature* 492: 59-65. **45.** Gilson PR *et al.* (2006) *Proc Natl Acad Sci USA* 103: 9566-9571.