PROJECT TITLE

Plastid endosymbiosis: a detailed study of genome dynamics

AIMS AND BACKGROUND

Eukaryotic algae have a most remarkable evolutionary history 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 in the ocean. They support the oceanic food chain, play crucial roles in various biogeochemical cycles, and are responsible for ca. 50% of all primary production worldwide [4], which has also made them targets for biofuel production. Despite the global importance of the process, many questions about endosymbiosis remain unanswered. This project addresses some of those questions using a combination of leading-edge empirical and computational methods.

Background

The first eukaryotic algae 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. 1). 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. Additional variants of this process occur. In tertiary endosymbiosis, a eukaryotic alga with a secondary plastid is engulfed by a heterotrophic eukaryote to become a tertiary plastid, as in some dinoflagellate algae. Plastids have also been drastically reduced and lost in some organisms, e.g. the parasites toxoplasmosis and malaria have a reduced plastid. In other cases, a plastid has been replaced entirely by a different plastid in a process called serial secondary endosymbiosis.

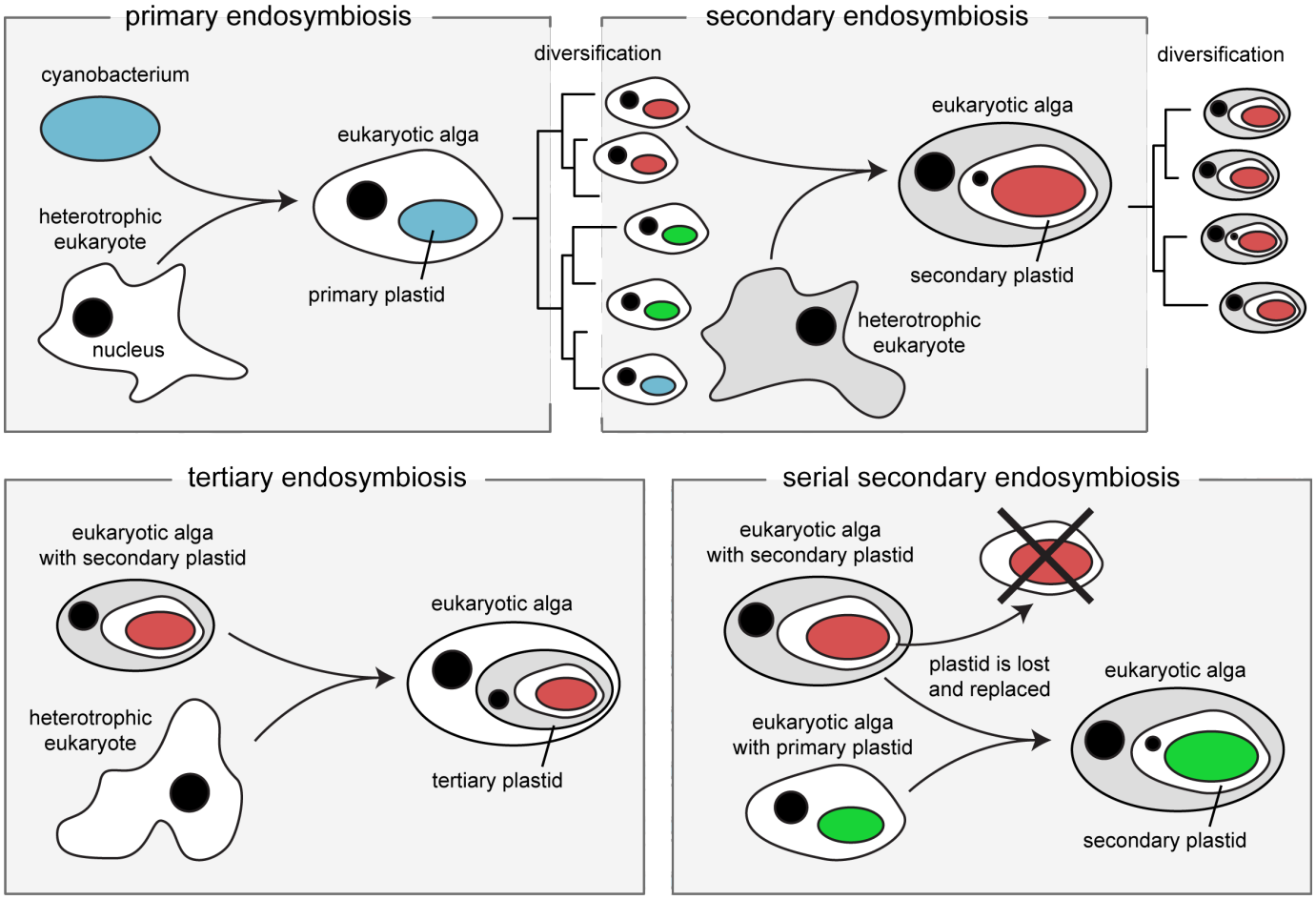

Fig. 1. Different types of plastid endosymbiosis.

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. 2). 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 analyzed 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.

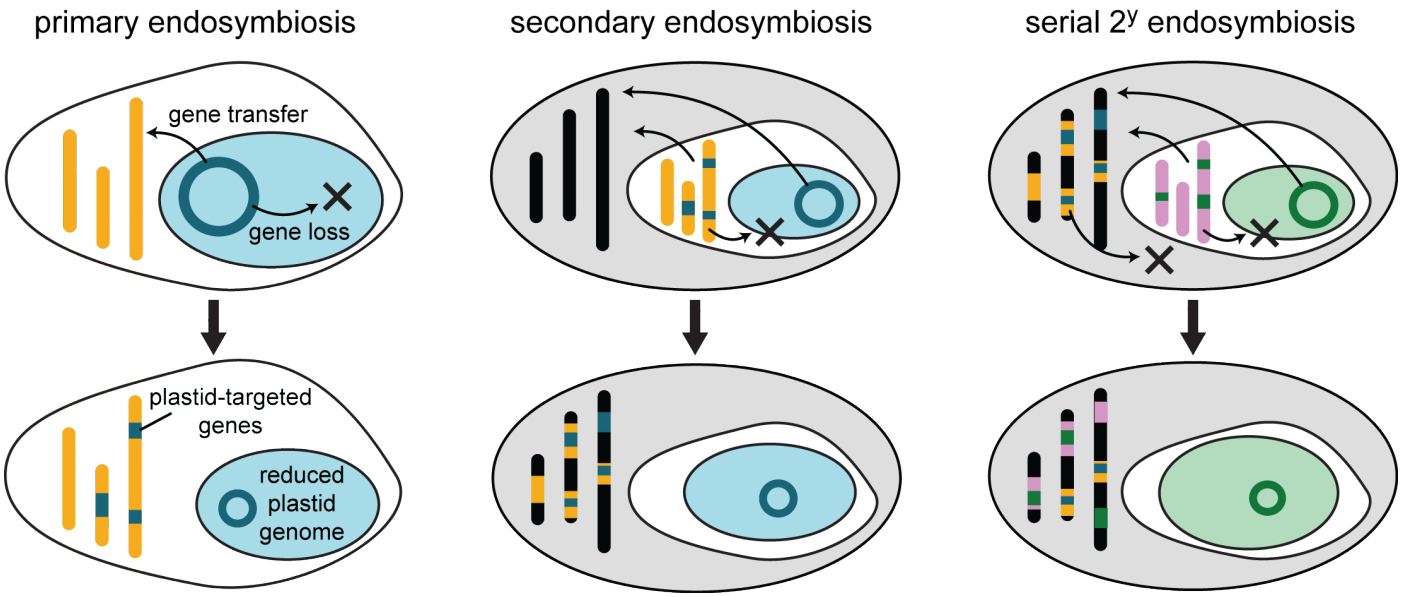

Fig. 2. Transfer and loss of genes during endosymbiosis lead to mosaic genomes.

Problem statement

Even though the general concept of endosymbiosis is widely embraced by the scientific community, researchers disagree about how many 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 very fragmentary knowledge of algal genomes. According to the Genomes Online Database (http://goo.gl/vRjri), 11 eukaryotic algal nuclear genomes have been completed and published. In addition to these nuclear genomes, a few dozen EST libraries and plastid genomes have been sequenced. This very 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 drawn. 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 both 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.

Endosymbiosis research also faces a second, more conceptual problem related to the interpretation of phylogenies. As explained in detail in a recent paper by John Stiller [15], more objective criteria are needed to determine the origin of genes found in the nuclear genomes of eukaryotic algae. 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. As Stiller points out, there are several problems with this interpretation. Besides bypassing the possibility that there is a true close relationship between the taxa that come out as being related in the analysis, one must take into consideration that such relationships can arise by horizontal gene transfer in the absence of endosymbiosis and 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.

Aims

This project aims to address some of these important shortfalls in algal evolutionary biology using a combination of empirical and computational techniques. The specific goals are:

1. Pinpoint the origin of plastids with densely sampled genomic datasets, focusing on the plastids of chlorarachniophytes and dinoflagellates.
2. Characterize the genome dynamics following eukaryote-eukaryote endosymbiosis in detail.
3. Determine the power of phyloinformatics approaches towards endosymbiosis research.

Besides focusing on these goals, we aim to expedite the utilization of the acquired genomic data in collaborators' projects to broaden the research base and augment the project output.

RESEARCH PROJECT

This project takes advantage of 2nd generation sequencing technologies to create densely sampled comparative genomic datasets that can be used to refine the origin of plastids and track the genome dynamics associated with endosymbiosis. The project will also use computational tools (hybrid simulation case-study approaches) to investigate the suitability of classical methods used to infer endosymbiosis. We will now address each of these aspects in more detail.

Goal 1: Pinpoint the origin of the plastid of chlorarachniophytes and dinoflagellate lineages.

The eukaryote-eukaryote endosymbiosis events selected for this goal include two tertiary endosymbioses, a secondary endosymbiosis and a presumed serial secondary endosymbiosis. Circa 22 algal strains will be grown in culture conditions (Table 1). Ingroup taxa have been selected to represent maximal phylogenetic diversity within the algal lineage in question. Outgroup taxa have been selected to represent diverse genera that are likely to be related to the endosymbiont based on preliminary analyses by CI Verbruggen. After they have grown to a reasonable density, cultures will receive treatments to maximize sequencing results. For plastid genome sequencing, strains will be put in low light for two days to increase the number of chloroplasts. For RNA-seq, subcultures will be subjected to a range of different conditions including alterations of the nutrient, light, temperature, salinity and day-night regimes. DNA and RNA will be extracted using standard protocols (modified CTAB protocol with chloroform-isoamyl alcohol extraction [18] for DNA and the Qiagen RNeasy plant kit for RNA).

Endosymbiosis event		ingroup taxa	outgroup taxa
1.	*Lepidodinium* dinoflagellates (serial secondary endosymbiosis) and chlorarachniophytes (secondary endosymbiosis)	*Lepidodinium* (N,P) *Bigelowiella* (–,–) *Chlorarachnion* (N,P)	*Pedinomonas* (N,P), *Tetraselmis* (N,P), *Acetabularia* (N,P), *Neocystis* (N,P), *Leptosira* (N,P), *Oocystis* (N,P), *Koliella* (N,P), *Chlorella* (–,–), *Ulva* (–,–), *Bryopsis* (–,–), *Boodlea* (–,P)
2.	peridinin-type dinoflagellates (putative tertiary endosymbiosis)	*Thoracosphaera* (N,P) *Amphidinium* (N,P) *Symbiodinium* (N,P)	*Nannochloropsis* (N,P), *Synura* (N,P), *Pinguicoccus* (N,P), *Ochromonas* (N,P)
3.	Kareniaceae dinoflagellates (tertiary endosymbiosis)	*Karlodinium* (N,P) *Karenia* (–,P)	*Pavlova* (N,–), *Prymnesium* (N,P), *Pavlomulina* (N,P)

Table 1. Tentative list of taxa that will be sequenced for this project. The letters N and/or P behind the name indicate what will be done for the taxon in question (N = nuclear gene sequencing, P = plastid sequencing). If either N or P is missing, it means that these data are already available for the taxon.

From the resulting RNA and DNA extracts, libraries will be prepared following standard Illumina protocols and sequenced on the Genome Analyzer IIx at Rutgers Univ. We will use the latest generation sequencing chemistry to obtain >9 Gbp of 150x150 bp reads per lane. DNA samples will be multiplexed (4 samples per lane), as previous experiments suggest that this still yields well over 100× coverage of the plastid genome. It is important to note that dinoflagellates do not have a genuine plastid genome but a series of plastid DNA minicircles [19,20]. Even though we anticipate that these will also be enriched in the dark treatment and will readily be recovered in multiplexed runs, this needs to be verified early on in the project. If they cannot readily be recovered from the sequencing results, dinoflagellate DNA runs will not be multiplexed to increase the coverage of plastid sequences. The cDNA synthesized from the RNA extracts will be normalized using the Evrogen Trimmer kit to maximize gene discovery. The resulting libraries will not be multiplexed to achieve sufficient coverage for *de novo* assembly of the nuclear transcriptome. We expect that at least 2-3 strains will not grow well enough in culture to obtain sufficiently large amounts of RNA. In such cases, total DNA will be extracted and amplified with the PicoPlex protocol. The resulting DNA will be sequenced on two GAIIx lanes and assembled into a draft genome (including plastid and nuclear genomes).

Reads will be assembled following the standard pipelines used at the Bhattacharya lab. These include Velvet/Oases, SOAPdenovo and Trans-ABySS as well as algorithms that are being developed in-house to tackle the coverage bias problems that result from multiple displacement amplification (MDA) in PicoPlex-amplified DNA. These algorithms use a Hidden Markov Model trained on MDA data from known genomes to distinguish between inherent MDA bias and genuine variation in DNA copy number. The assembled reads will then be integrated in the pico-PLAZA database. The computational pipelines will perform functional annotation using InterPro domains and Gene Ontology databases. Gene families are delimited and orthologs and paralogs are detected using protein clustering and phylogenetic approaches [21]. The fact that the data are integrated in a platform that also contains meticulously annotated genomes (e.g. *Arabidopsis, Chlamydomonas*, *Phaeodactylum*) will greatly improve the computational annotations for the new genomic data.

The comparative genomic data will be analyzed with automated phylogenomic pipelines. Two approaches will be used to refine our knowledge of endosymbiosis events. First, the assembled plastid genomes will be subjected to phylogenomic 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 [22]. Second, the origin of the nuclear genes of the host will be determined. This will be achieved with model-based phylogenetic inference and tree sorting pipelines [23]. Following automated alignment, the sets of orthologous proteins will be subjected to maximum likelihood phylogenetic inference with bootstrapping to determine statistical reliability. Subsequently, the obtained trees will be filtered based on reliable topological features to extract the origin of different genes and infer past endosymbiosis events [24]. In addition to this traditional (and criticized) method, we will explore the use of spectral partitioning [25] to identify conflicting signals that could be due to endosymbiotic gene transfer and model-based techniques that specifically take horizontal gene transfer into account [26-29]. These techniques are still in their infancy and 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.

Goal 2: Characterize the genome dynamics following endosymbiosis in detail.

For this goal, the comparative genomic dataset obtained under Goal 1 will be further analyzed using bioinformatics pipelines with the aim of characterizing and quantifying the changes that happen in the host and endosymbiont genomes during and following the endosymbiosis event. These changes include transfer of genes from the plastid to the host nucleus and loss of genes from the plastid and nuclear genomes, but also potential modifications of patterns of GC content and codon usage, among others. Previous work into plastid genome content has identified a trend of plastid genome reduction as genes are lost or transferred to the host nucleus. Whereas plastids resulting from recent endosymbiosis events retain large genes numbers [30-34], plastids from more ancient endosymbiosis events have lost more genes [35], and in some extreme cases nearly the entire plastid genome has been transferred and lost [36]. But previous studies have rarely tried to quantify the rate of plastid genome reduction or of any other genome dynamics and their conclusions have possibly suffered from sparse taxon sampling.

The goal here is to quantify the various changes in genome content that happen in association with endosymbiosis events. Based on our densely sampled datasets, ancestral genome composition will be inferred from datasets of the presence-absence of genes in extant organisms using model-based techniques [37]. Subsequently the rate of gene loss in acquired plastids will be inferred. We will fit exponential decay models which have proven a good fit for reductive genome evolution in mitochondria [38]. This basic characterization of genome reduction will be refined with details about the functional annotation of genes and whether the genes are transferred or lost. Using relaxed molecular clock techniques [39,40], we will also attempt to quantify the time lag between the first endosymbiotic gene transfers and the onset of plastid genome reduction, which may happen because protein import machinery has to be established before plastid genome reduction can proceed. The dynamics of GC content and codon utilization will be characterized with various evolutionary models [41,42].

Besides the interest of studying genome dynamics in and of itself, the quantitative information gathered in this experience will help parameterize the models used under the third goal.

Goal 3: Establish the power of phyloinformatics approaches towards 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, from the information gained, we will identify criteria that are useful indicators for endosymbiotic gene transfer.

As a first step, a comprehensive model of genome dynamics will be designed. Several recent papers lay the foundation for models incorporating genomic events such as gene duplication/loss and horizontal transmission in addition to vertical sequence evolution [26-28]. Elements of these models will be combined and refined to

allow for the full suite of evolutionary events needed to explain the genome dynamics associated with algal endosymbiosis.

Second, simulation will be used 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 coarse indication of what the realistic value of several parameters of this model could be and realistic amounts of between-lineage variation in evolutionary rates, GC content and codon usage will be incorporated. By using a "hybrid simulation case study" approach, 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 the current phyloinformatics pipelines [24,43] as well as some alternative methods [7,44], 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 to determine whether sensible inferences can be made from real-life data.

Third, we will take inferences a step further to investigate whether it may be feasible to reconstruct more complex scenarios of endosymbiosis. This will be especially useful to infer the confidence one may have in proposals of cryptic endosymbiosis events, which have been recently proposed [24,45] but not generally accepted [15]. It can also be used to evaluate more elaborate scenarios, such as a series of nested endosymbioses [7] to explain plastid origins in the chromalveolates, a hotly debated issue.

Finally, we will undertake a meticulous investigation of whether objective criteria can be defined to more reliably detect signatures of endosymbiotic gene transfer in genomes. While such criteria may appear evident for simple endosymbiosis events (primary, secondary), there is a need for better criteria for higher-level (tertiary) and complex (e.g. serial secondary, ancient cryptic) scenarios of endosymbiosis [15]. We will explore genome data simulated for such scenarios in an attempt to identify genomic imprints that are reliable predictors of their evolutionary history. Among these imprints may be differential phylogenetic signal in plastid genes, endosymbiotically transferred genes and host nuclear genes, as well as criteria based on divergence times of the various genomic compartments using relaxed molecular clock methods. If successful, this approach can contribute to elucidating various hotly debated issues in endosymbiosis research (e.g., chromalveolates, cryptic and higher-level endosymbioses). If unsuccessful, the method will add to the controversy in its conclusion that even under the relatively optimal case of simulated data, one may not be able to reliably infer the history of complex endosymbiosis scenarios.

Broadening the research base and augmenting project output through data integration

In addition to the goals outlined above, we aim to promote the use of the acquired genomic data by collaborators. Engaging world-class scientists in related fields to use our data will broaden the research base and add to the publication output of the project. We have come to collaboration agreements with Dr Olivier De Clerck (Ghent Univ.), who will use the data to study the genomic correlates of the development of multicellularity and siphonous cell types, with Dr Pavel Skaloud (Charles Univ. Prague) and Dr Frederik Leliaert (Ghent Univ.), who will use the green algal data to improve the resolution of the phylogeny of the core Chlorophyta, and with Dr Marek Elias (Univ. Ostrava), who will use the comparative genomic data to study the evolution of GTPases and endomembrane systems. CI Verbruggen will also incorporate the acquired data into the alignments that will be used to infer the phylogenetic trees that are needed to study the evolution of trace element utilization, an aspect of his Future Fellowship. Various other collaborators will be able to use the data to study the systematics and evolutionary biology of their taxa of interest.

In order to facilitate collaborator access to the generated information, all the obtained genomic data will be integrated in a central sequence repository (pico-PLAZA: http://goo.gl/1ELZy). Besides providing user-friendly access to data through a variety of downloads and querying options, pico-PLAZA offers an online workbench with an extensive toolbox for algal comparative genomics that greatly enhances the potential to perform evolutionary data mining and makes comparative genomics accessible even to people with limited knowledge of bioinformatics. Pico-PLAZA will combine the transcriptome data generated in this project with all other publicly available algal genomes for joint analysis.

This effort will strengthen international collaborations, increase the visibility of the project, and contribute to a positive and open image of Australian research in an international context. As such, the relatively minor

investment to facilitate user-friendly access to the data will yield many benefits and augment the outputs of the project.

Outcomes and timeline

Besides the direct outcomes detailed above, this project will form the basis for new developments in modeling genome dynamics, which can eventually lead to Bayesian approaches that simultaneously infer host and plastid trees by explicitly taking endosymbiosis and the ensuing genome dynamics into account. We will already develop much of the conceptual framework for such models in the context of our simulation study, and the empirical data will provide key information to further develop and parameterize the models.

To optimize the feasibility of the project and prevent delays, the actions for different endosymbiosis events will be interspersed (Table 2). With this design, data for the first event will be completed early in the 2nd year, allowing analyses for this event to start early on, while experimental work for the other events is being done. The computational work during the first year (in anticipation of the new data) will focus on conceptualizing the model of genome dynamics used for the simulations and preparing and testing the computational pipelines. Given that an RA1 will be hired to culture algae and perform molecular work and a PhD student for computational aspects, we consider this to be a feasible timeline.

	year 1				year 2				year 3	
hire RA1 and PhD student convene with collaborators										
culture algal strains	1	1	1,2	2	2,3	3	3			
harvest DNA/RNA		1	1	1,2	2	2	2,3	3	3	
sequence DNA/RNA			1	1	1	2	2	2,3	3	3
assemble, annotate, integrate data			1	1	1	2	2	2,3	3	3
prepare computational pipelines										
pinpoint plastid origins				1	1	1	2	2	3	
characterize genome dynamics					1	1	1	2	2	3
conceptualize models										
perform simulations, fine-tune models										
prepare publications										

Table 2. Anticipated timeline of completion of major tasks. Numbers indicate which endosymbiosis events will be worked on during the period in question (see Table 1 for explanation of numbers).

Significance and innovation

Besides the great importance of documenting eukaryotic algal genomes to better understand their bizarre evolutionary history, algal genomics can contribute to the knowledge base about aspects of algal biology that have significant societal relevance, such as harmful algal blooms, the role of algae in global biogeochemical cycles and the production of biofuel and other valuable compounds derived from algae.

Research on algal genomics and endosymbiosis tends to be high-impact research. Much of the recent work has been published in top journals like *Science*, *Nature*, *Current Biology* and *Molecular Biology and Evolution*. Novel techniques in evolutionary modeling of genomes are commonly published in highly ranked journals such as *Systematic Biology* and *PNAS*.

The current project advances the knowledge base by integrating the disciplines of phycology, genome sequencing and bioinformatics with molecular phylogenetics and computational simulation. State-of-the-art genome sequencing techniques are combined with bioinformatics to obtain annotated genome-scale datasets for > 20 algal species, some of which have societal relevance. The new knowledge can serve many purposes and has the potential to advance the biofuel industry as well as several fields of basic and applied research as described under "Broadening the research base and augmenting project output" above.

The new methods developed here will also advance knowledge. The simulation approach will permit us to explore the realm of the "knowable" in algal endosymbiosis research. Simulation is a very effective technique for learning about complex systems and evaluating analysis techniques. We intend to use it to investigate whether it is possible that endosymbiosis events may have left a sufficiently clear footprint in genomes to allow inferring the sequence of these ancient events from present-day data, i.e. whether the sequence of

endosymbiosis events is "knowable". This approach clearly has many applications in the field of evolutionary and comparative genomics.

The interdisciplinary approach used here is innovative in many ways. The development of novel models linking data from different genomes in a comprehensive evolutionary framework is an important innovation with applications in many other evolutionary questions.

In addition to developing these novel analytical advances, the project will use several state-of-the-art techniques in data acquisition and analysis: (1) High-throughput DNA sequencing: The genome sequencing will be entirely carried out using the latest generation of Illumina high-throughput sequencing. At present, this allows runs of paired-end 150 x 150 bp reads, yielding a total of 9 Gbp per lane or 72 Gbp per flow cell. (2) High-performance computing: Many analyses in the required bioinformatics, molecular phylogenetics and evolutionary modeling and simulation pipelines are computationally expensive. Fortunately, they are also highly parallelizable. The project will use recently established, modern super-computer facilities at all three host institutions to facilitate these computations.

RESEARCH ENVIRONMENT

The School of Botany at the University of Melbourne offers an outstanding environment to carry out this project. It has a record of excellence in plastid endosymbiosis research, including the research programs of Prof Geoff McFadden and Dr Ross Waller. Both have extensive expertise in the organisms studied for this project. McFadden was involved in the characterization of the secondary endosymbiotic event giving rise to the chlorarachniophytes, and in the sequencing of the plastid [46], the nucleomorph [47], and more recently the nuclear genome of the chlorarachniophyte alga *Bigelowiella natans*. Waller has characterized the use of genes from two endosymbionts in tertiary plastids and is a specialist of the peptides that are used to target proteins that are encoded in the nucleus to the plastid. Much of Waller's current work is on dinoflagellates, a focal taxon of this project. Both McFadden and Waller have also authored review papers about plastid evolution and associated molecular processes. Prof McFadden and Dr Waller have advised me about the elements of this project relevant to their expertise and will continue to do so for the duration of the project. The School of Botany has excellent algal culturing facilities, including a walk-in culture chamber in Dr Wetherbee's lab and QC2 certified labs for imported strains. Dr Verbruggen will start up his research unit in March 2012 and will have a fully operational lab to carry out the molecular work in advance of the start of this project in 2013. Should it be required, we also have access to a wide range of specialized molecular lab equipment through other labs in the department including the Plant Cell Biology Research Centre, as well as through the Bio21 institute. Dr Brendan Wintle and Dr Michael McCarthy are experienced mathematical modelers and will be consulted in relation to the design of the hybrid simulation case-study approach.

When the Bhattacharya lab moved to Rutgers University in 2009, it was able to establish its own genome facility with a second-generation Illumina GAIIx DNA sequencer that can generate >100 gigabases of data per week. The instrument supports de novo genome and transcriptome sequencing, analysis of small RNAs, methylomics, mRNAseq, and ChIpSeq all of which are being pursued by the lab and their collaborators using algae and other protist species of interest. This makes it the only algal evolution and genomics lab to have in-house high-throughput sequencing facilities, giving it an advantage over all other such labs, and making it an excellent partner for this project.

At Ghent University, bioinformatics is one of the five university-wide research priorities under the "Strategic Spearhead Research Policy". Dr Klaas Vandepoele was appointed as group leader in 2011 in the Flanders Institute of Biotechnology (Ghent University branch), which is renowned for its work on plant and protist genomics, among other things. The focus on bioinformatics at the university and the strong history of integration and distribution of genomic data in combination with the development of resources for comparative genomics within Vandepoele's group makes this a perfect match with the goals of this project.

The work described in this project also requires computational facilities to carry out bioinformatics and evolutionary analyses. All three institutions have excellent computing facilities. In Melbourne, we will have access to the VLSCI supercomputer facility (ca. 46 TFlops) and two dedicated servers in the Verbruggen lab. At Rutgers Univ., there are several servers dedicated to genome assembly and other bioinformatics applications. At Ghent Univ., great investments have recently been made in high-performance computing (Stevin cluster, 50 TFlops).

The University of Melbourne's research culture is highly supportive of our present and future collaborations with genome biologists and phycologists in Melbourne and beyond and our role in providing state of the art training in next-generation sequencing and bioinformatics to students.

The results of this project will be communicated to the scientific community and the wider public by various means. All collaborators have excellent track records of publishing in high-profile refereed journals, giving presentations at national and international meetings, and participating in workshops. The CI is a proponent of Open Science, and a dedicated page for the project will be set up on http://www.phycoweb.net to document the progress and research outputs and to publish press releases. I will also encourage the students working on the project to blog about their work. The generated sequence data will be available through pico-PLAZA and submitted to the classical sequence data repositories for maximal dissemination and openness.

ROLE OF PERSONNEL

Dr Heroen Verbruggen (CI) received a Future Fellowship from the ARC in 2011. His expertise includes molecular and evolutionary biology of algae, phylogenetic theory and bioinformatics. He is experienced in growing algal strains for physiological experimentation and sequencing projects, and he has coordinated and carried out bioinformatics for two second-generation sequencing projects. He will provide intellectual direction for the project, supervise the research assistant, graduate and undergraduate students, and conduct some computational analyses. He will contribute 0.25 FTE to this project. Verbruggen has an extensive network of contacts and collaborators that will facilitate any troubleshooting that may have to be done and that will use the generated data for their purposes.

Dr Debashish Bhattacharya (PI) has a long-standing interest in algal and protist evolution and genomics. He has worked extensively on the endosymbiotic origin of plastids. The Bhattacharya lab has a fully equipped genome facility with a second-generation sequencer and high-performance computing clusters. The lab has extensive experience in single cell genomics and genome assembly. Bhattacharya will facilitate the sequencing and assembly of reads for this project.

Dr Klaas Vandepoele (PI) has extensive experience with bioinformatics, especially in the fields of data integration, comparative genomics and transcriptional control. He has developed the PLAZA platform, which is an access point and workbench for plant comparative genomics, and forms the basis of pico-PLAZA. He is responsible for the integration of the generated data with other genomes and provide user-friendly solutions for accessing the data for downstream comparative analyses and other applications.

Research Assistant (grade 1) will primarily be responsible for the experimental work under the direction of the CI, and will also participate in student supervision. This project is ambitious, broad in scope, and utilizes cutting-edge techniques in DNA sequencing, assembly and analysis. It is therefore critical to source a research assistant who has prior experience with algal culturing and molecular biology techniques as well as basic skills in bioinformatics, and will be able to work semi-independently.

PhD and BSc (Hons)/MSc students: The CI will recruit a graduate student and two further students (BSc (Hons) or MSc) to the project for training in algal culturing, next-generation sequencing and bioinformatics. The PhD student to be hired on the project will have a background in bioinformatics and will be responsible for the computational simulation studies under the guidance of the CI.

External collaborators: Dr Pavel Skaloud (Charles Univ. Prague), Dr Marek Elias (Univ. Ostrava) and Dr Frederik Leliaert and Dr Olivier De Clerck (both Ghent Univ.) have ample experience researching and culturing the various algae used in this project. They will contribute their expertise and cultures of the taxa involved. They will have priority access to the genome data for their projects and will be involved in publications related to their area of expertise.

References

[1] Bowler C, et al. 2010. Oceanographic and biogeochemical insights from diatom genomes. *Annual Review of Marine Science* 2:333-365. **[2]** Allen AE, et al. 2006. An ecological and evolutionary context for integrated nitrogen metabolism and related signaling pathways in marine diatoms. *Current Opinion in Plant Biology* 9:264-273. **[3]** Johnson MD. 2011. The acquisition of phototrophy: adaptive strategies of hosting endosymbionts and organelles. *Photosynth Res* 107:117-132. **[4]** Falkowski PG, et al. 2004. The evolution of modern eukaryotic phytoplankton. *Science* 305:354-360. **[5]** Delwiche CF. 1999. Tracing the thread of plastid diversity through the tapestry of life. *Am Nat* 154:S164-S177. **[6]** Keeling PJ. 2010. The endosymbiotic origin, diversification and fate of plastids. *Phil Trans Roy Soc B - Biol Sci* 365:729-748. **[7]** Baurain D, et al. 2010. Phylogenomic evidence for separate

acquisition of plastids in cryptophytes, haptophytes, and stramenopiles. *Mol Biol Evol* 27:1698-1709. **[8]** Keeling PJ. 2009. Chromalveolates and the evolution of plastids by secondary endosymbiosis. *J Eukar Microbiol* 56:1-8. **[9]** Bergsten J. 2005. A review of long-branch attraction. *Cladistics* 21:163-193. **[10]** Jermiin LS, et al. 2004. The biasing effect of compositional heterogeneity on phylogenetic estimates may be underestimated. *Syst Biol* 53:638-643. **[11]** Ho SYW, Jermiin LS. 2004. Tracing the decay of the historical signal in biological sequence data. *Syst Biol* 53:623-637. **[12]** Yoon HS, et al. 2004. A molecular timeline for the origin of photosynthetic eukaryotes. *Mol Biol Evol* 21:809-818. **[13]** Lockhart P, et al. 2006. Heterotachy and tree building: A case study with plastids and eubacteria. *Mol Biol Evol* 23:40-45. **[14]** Shalchian-Tabrizi K, et al. 2006. Heterotachy processes in rhodophyte-derived secondhand plastid genes: Implications for addressing the origin and evolution of dinoflagellate plastids. *Mol Biol Evol* 23:1504-1515. **[15]** Stiller J. 2011. Experimental design and statistical rigor in phylogenomics of horizontal and endosymbiotic gene transfer. *BMC Evol Biol* 11:259. **[16]** Hedtke SM, et al. 2006. Resolution of phylogenetic conflict in large data sets by increased taxon sampling. *Syst Biol* 55:522-529. **[17]** Zwickl DJ, Hillis DM. 2002. Increased taxon sampling greatly reduces phylogenetic error. *Syst Biol* 51:588-598. **[18]** Doyle JJ, Doyle JL. 1987. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochem Bull* 19:11-15. **[19]** Laatsch T, et al. 2004. Plastid-derived single gene minicircles of the dinoflagellate *Ceratium horridum* are localized in the nucleus. *Mol Biol Evol* 21:1318-1322. **[20]** Zhang ZD, et al. 1999. Single gene circles in dinoflagellate chloroplast genomes. *Nature* 400:155-159. **[21]** Proost S, et al. 2009. PLAZA: a comparative genomics resource to study gene and genome evolution in plants. *Plant Cell* 21:3718-3731. **[22]** Verbruggen H, Theriot EC. 2008. Building trees of algae: some advances in phylogenetic and evolutionary analysis. *Eur J Phycol* 43:229–252. **[23]** Moustafa A, et al.2010. iTree: A high-throughput phylogenomic pipeline. In *Biomedical Engineering Conference (CIBEC), 2010 5th Cairo International*; *16-18 Dec. 2010*. 103-107. **[24]** Moustafa A, et al. 2009. Genomic footprints of a cryptic plastid endosymbiosis in diatoms. *Science* 324:1724-1726. **[25]** Chen D, et al. 2007. Spectral partitioning of phylogenetic data sets based on compatibility. *Syst Biol* 56:623 - 632. **[26]** Akerborg O, et al. 2009. Simultaneous Bayesian gene tree reconstruction and reconciliation analysis. *Proc Natl Acad Sci USA* 106:5714-5719. **[27]** Gorecki P, et al. 2011. Maximum likelihood models and algorithms for gene tree evolution with duplications and losses. *BMC Bioinf* 12:S15. **[28]** Bloomquist EW, Suchard MA. 2010. Unifying vertical and nonvertical evolution: A stochastic ARG-based framework. *Syst Biol* 59:27-41. **[29]** Galtier N, Daubin V. 2008. Dealing with incongruence in phylogenomic analyses. *Philosophical Transactions of the Royal Society B: Biological Sciences* 363:4023-4029. **[30]** Yoon HS, et al. 2006. Minimal plastid genome evolution in the *Paulinella* endosymbiont. *Curr Biol* 16:R670-R672. **[31]** Nowack ECM, et al. 2008. Chromatophore genome sequence of Paulinella sheds light on acquisition of photosynthesis by eukaryotes. *Curr Biol* 18:410-418. **[32]** Nowack ECM, et al. 2011. Endosymbiotic gene transfer and transcriptional regulation of transferred genes in *Paulinella chromatophora*. *Mol Biol Evol* 28:407-422. **[33]** Chan CX, et al. 2011. Plastid origin and evolution: new models provide insights into old problems. *Plant Phys* 155:1552-1560. **[34]** Imanian B, et al. 2010. The complete plastid genomes of the two "dinotoms" *Durinskia baltica* and *Kryptoperidinium foliaceum*. *PLoS One* 5:e10711. **[35]** Timmis JN, et al. 2004. Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes. *Nat Rev Genet* 5:123-135. **[36]** Hackett JD, et al. 2004. Migration of the plastid genome to the nucleus in a peridinin dinoflagellate. *Curr Biol* 14:213-218. **[37]** Barker D, et al. 2007. Constrained models of evolution lead to improved prediction of functional linkage from correlated gain and loss of genes. *Bioinformatics* 23:14-20. **[38]** Khachane AN, et al. 2007. Dynamics of reductive genome evolution in mitochondria and obligate intracellular microbes. *Mol Biol Evol* 24:449-456. **[39]** Drummond AJ, et al. 2006. Relaxed phylogenetics and dating with confidence. *PLoS Biol* 4:e88. **[40]** Huelsenbeck JP, et al. 2000. A compound Poisson process for relaxing the molecular clock. *Genetics* 154:1879-1892. **[41]** Cocquyt E. 2009. Phylogeny and molecular evolution of green algae. *PhD thesis.* Ghent University, Phycology Research Group. **[42]** Butler MA, King AA. 2004. Phylogenetic comparative analysis: a modeling approach for adaptive evolution. *Am Nat* 164:683-695. **[43]** Reyes-Prieto A, et al. 2010. Differential gene retention in plastids of common recent origin. *Mol Biol Evol* 27:1530-1537. **[44]** Stiller J, et al. 2009. Are algal genes in nonphotosynthetic protists evidence of historical plastid endosymbioses? *BMC Genomics* 10:484. **[45]** Chan CX, et al. 2011. Red and green algal origin of diatom membrane transporters: insights into environmental adaptation and cell evolution. *PLoS One* 6:e29138. **[46]** Rogers MB, et al. 2007. The complete chloroplast genome of the chlorarachniophyte *Bigelowiella natans*: Evidence for independent origins of chlorarachniophyte and euglenid secondary endosymbionts. *Mol Biol Evol* 24:54-62. **[47]** Gilson PR, et al. 2006. Complete nucleotide sequence of the chlorarachniophyte nucleomorph: Nature's smallest nucleus. *Proc Natl Acad Sci USA* 103:9566-9571.