THE PATTERN AND PROCESSES OF GENOME CHANGE IN ENDOSYMBIONTS OLD AND NEW

A thesis presented in partial fulfilment of the requirement for the degree of

Doctor of Philosophy

in Evolutionary Biology

Submitted by

Barbara Inge Karoline Schönfeld

2012

Institute of Molecular BioSciences

Massey University

New Zealand

Abstract

Bacterial endosymbionts are an important part of eukaryote evolution as they allow their hosts to exploit bacterial abilities. Plastids, the organelles that enable plants and eukaryotic algae to photosynthesise are ancient cyanobacterial endosymbionts. Since the initial symbiosis ~1.5 billion years ago the majority of their genes has been lost or transferred to their host's nucleus. This process has carried on independently in the different lineages following the diversification of the lineage.

I have compiled a comprehensive data set of fully sequenced plastid genomes to systematically study the frequency of gene transfers from the plastid to the nucleus across the different lineages. Following the reconstruction of the Plantae phylogenetic tree from plastid encoded proteins, gene loss events were reconstructed along its branches. My calculations show that gene losses have occurred at a relative high frequency and in a lineage specific way. This challenges the original idea that gene transfers from the organelle to the nucleus are rare and chance driven events.

Bacteria and eukaryotes continue to form endosymbioses and the study of these relationships produces valuable insights into the early stages of organelle evolution, bacterial metabolic pathways and metabolic regulation. They also allow us a glimpse into the ancient history of eukaryote evolution. For this reason, diatoms that have acquired cyanobacterial endosymbionts with the capability to fix molecular nitrogen were chosen to explore the potential and limitations of high-throughput sequencing technologies for investigating this type of relationship when DNA sequences are obtained from environmental samples and in the presence of bacterial contaminants. The results of this work confirmed the suitability of this relatively new technology to sequence mixed samples but also highlighted i) difficulties in sample preparation which can bias the composition of metagenomic samples obtained, and also ii) the varying suitability of different types of samples used in high-throughput sequencing.

In Gedenken an meinen Vater Günther Schönfeld

Acknowledgements

I would like to thank my supervisors Pete Lockhart and Lesley Collins for their support and guidance. They truly care and I will be forever grateful.

Many others have helped me along the way and should know that it is deeply appreciated:

Trish McLenachan and Phil Novis volunteered their time and expertise to help me in the lab and in the field, and have taught me much. Mike Steel and Tim White worked magic with numbers and scripts to make my results 14% more likely. Bill Martin provided inspiration, support, enthusiasm and helpful criticism when it was needed most. Uwe Maier's keen interest shaped the course of this project. David Penny and the assortment of brilliant people he has gathered around him over the years who have allowed me to share in the passion, the wonder, and the drinking of malt-based beverages, that is science.

The Institute of Molecular BioSciences and the Allan Wilson Centre for Molecular Ecology and Evolution gave the financial and practical support that made my work possible.

My warmest thanks have to go to all the friends and new family who have made my time in New Zealand the best of my life:

The Piripis and Ellicots who made me feel like family. The van Hoves and McComishs for being family. The flatmates, the crafters, the gamers, the foragers, the time travellers and the stargazers, whose friendship means the world to me.

Thank you, Bennet, for being there.

Finally and most importantly I would like to acknowledge my parents, my family and my good friends back home, on whose support I can always count and whose love is always with me.

Table of Contents

Abstract i

Acknowledgements iii

List of Figures ix

List of Tables xi

1 Introduction 1

1.1 Symbiosis 1

1.1.1 Endosymbiosis 2

1.1.2 Genomic changes in vertically transmitted endosymbionts 3

1.1.3 Endosymbionts and organelles 7

1.1.4 Cyanobacteria 14

1.1.5 Symbiotic nitrogen fixation 15

1.2 Diatoms 17

1.3 Sequencing with the Illumina Genome Analyzer 19

1.4 Data analysis 22

1.5 Aims and purposes 24

Bibliography 26

2 The frequency of gene loss events during plastid evolution 35

2.1 Abstract 35

2.2 Introduction 36

2.3 Methods 42

2.3.1 Gathering and processing of data 42

2.3.2 Data verification 43

2.3.3 Protein alignments 44

2.3.4 Phylogenetic Analyses 45

2.3.5 Mapping of gene loss events 47

2.4 Results 49

2.4.1 Quality of RefSeq data base entries 49

2.4.2 The gene presence/absence matrix 50

2.4.3 The plastid phylogeny 53

2.4.4 The prevalence of plastid gene transfers to the nucleus 55

2.4.5 Lineage specific gene loss and retention 59

2.5 Discussion 61
2.5.1 Quality of RefSeq data base entries 61
2.5.2 The presence/absence matrix 63
2.5.3 The plastid phylogeny 64
2.5.4 The prevalence of plastid gene transfers to the nucleus 72
2.5.5 Effects of uneven lineage sampling 78
2.6 Conclusion and outlook 80
Bibliography 82

***3* Sequencing the genome of the spheroid body of *Rhopalodia gibba* 93**
3.1 Abstract 93
3.2 Introduction 94
3.2.1 *Rhopalodia gibba* and its spheroid body 94
3.2.2 The cyanobacterium *Cyanothece sp*. ATCC 51142 95
3.2.3 The spheroid body of *Rhopalodia gibba* 96
3.3 Methods 98
3.3.1 Illumina sequencing 98
3.3.2 Assembly 98
3.3.3 Contig assembly and sequence comparison 99
3.3.4 Mapping and editing 100
3.3.5 Identification of bacterial contaminations 100
3.4 Results 101
3.4.1 Illumina sequencing 101
3.4.2 Assembly 101
3.4.3 Identification of misassemblies by mapping 103
3.4.4 Assembly quality 104
3.4.5 Identification of putative spheroid body sequences and contaminants 108
3.5 Discussion 111
3.5.1 Sequence assembly 111
3.5.2 Assembly quality and comparison to Sanger sequences 117
3.5.3 Contaminant sequences 119
3.6 Conclusion and outlook 121
Bibliography 123

4 High-throughput sequencing of an environmental sample of the diatom *Epithemia sorex* 127

4.1 Abstract 127
4.2 Introduction 128
4.3 Methods 131
4.3.1 Study species 131
4.3.2 Sample Selection 132
4.3.3 DNA extraction and Illumina sequencing 136
4.3.4 Mapping 137
4.3.5 Assemblies 138
4.3.6 BLAST analyses 139
4.3.7 Analyses of eukaryotic signature proteins (ESPs) 139
4.3.8 Sequencing of 16S and 18S rDNA genes 140
4.4 Results 141
4.4.1 DNA extraction and sequencing 141
4.4.2 Mapping 142
4.4.3 Assemblies 145
4.4.4 BLAST results 148
4.4.5 Eukaryotic signature proteins 150
4.4.6 16S and 18S rDNA sequencing 153
4.5 Discussion 155
4.5.1 Sample Quality 155
4.5.2 Sequence quality and assembly 157
4.5.3 Mapping 158
4.5.4 BLAST analyses with MEGAN 159
4.5.5 Analysis of Eukaryotic Signature Proteins 162
4.5.6 rDNA sequencing 164
4.6 Conclusions 168
Bibliography 169
5 Conclusion 173
Appendix

List of Figures

Figure 1-1 Diagram of the principles of Shotgun sequencing 19
Figure 1-2 Diagram of the Solexa (Illumina) sequencing process 21
Figure 1-3 Schematic of nodes and arcs of a de Brujin graph in Velvet 23
Figure 2-1 Schematic of the gene presence/absence matrix 50
Figure 2-2 Supertree of 124 Plantae species 55
Figure 2-3 Gene losses in crown groups 58
Figure 2-4 Summary Trees for the three possible root positions. 60
Figure 3-1 *Rhopalodia gibba* 94
Figure 3-2 The spheroid body of *Rhopalodia gibba* 96
Figure 3-3 Example of a long sequence assembly for contigs from different Velvet assemblies 102
Figure 3-4 Reads mapped to a misassembled repeat region, visualised in Tablet 103
Figure 3-5 Example of a miss-assembly 104
Figure 3-6 Example of a coverage pead of reads derived from a subsample of fosmid inserts 105
Figure 3-7 Mapping of the indexed subset of reads to Contig 2891 107
Figure 3-8 MEGAN visualisation of BLAST results for CG-rich contaminant seuquences 108
Figure 3-9 MEGAN visualisation of BLAST results for AT-rich contigs 109
Figure 3-10 Tablet screenshot of the read coverage of a 52 kb long contig that was Assembled from overlapping sequencing templates 114
Figure 3-11 Alignment of short read assemblies and Sanger sequences 116
Figure 4-1 Silicate shell of *Epithemia sorex* in valve view 131
Figure 4-2 *Epithemia sorex* sampling site 132
Figure 4-3 *Epithemia* cells in situ covering the stem of an aquatic plant 133
Figure 4-4 Environmental sample after enrichment for diatom cells 133
Figure 4-5 Overview of sampling sites on the south island of New Zealand. 134
Figure 4-6 Sampling sites south of Christchurch 135
Figure 4-7 Sampling sites around Arthur's Pass 135
Figure 4-8 Environmental sample whole DNA extraction (CTAB extraction protocol) 141
Figure 4-9 Mapping of paired end reads against *R. gibba* 16S rDNA 144

Figure 4-10 Mapping of paired end reads against *E. sorex* 18S rDNA 144
Figure 4-11 Mapping of paired end reads against the sequence of *PsbC* of *R. contorta* (chloroplast) 145
Figure 4-12 Contig size distribution of assembly results 147
Figure 4-13 Results of a BLASTn search against the nt database visualised in MEGAN 149
Figure 4-14 Results of a BLASTx search against all diatom sequences in GenBank visualised in MEGAN 150
Figure 4-15 Four examples of ESP gene trees 153
Figure 4-16 *Epithemia sorex* cells *in situ* and after incubation in RNAlater 156
Figure 4-17 Working principle of the Last Common Ancestor (LCA) algorithm used in MEGAN and the effects of poor database representation 161
Figure 4-17 The araphid, pennate diatom *Fragilaria ulna* 166

List of Tables

Table 2-1 Summary of results of Gene Loss reconstructions 56

Table 3-1 Parameters and statistics of Velvet assemblies 102

Table 3-2 Statistics of BWA mapping of indexed subset of reads to completely assembled fosmid inserts 106

Table 3-3 Average sequence identity between Sanger sequences and short read assemblies by alignment length. 107

Table 4-1 Primer sequences and expected product sizes 140

Table 4-2 Bowtie2 parameters and statistics for each mapping 143

Table 4-3 Summary of the assembly statistics 146

Table 4-4 Average coverage of contigs assembled from untrimmed reads using a k-mer of 25 147

Table 4-5 ESP containing contigs and their average coverage 151

Table 4-6 Summary of BLAST results for 16S and 18S rDNA sequences 154

1 Introduction

1.1 Symbiosis

Life as we know it evolved by adapting not only to the abiotic factors in the environment but also in response to the countless influences of other extant life forms as well. Organisms rarely live in isolation but in one way or another within communities and with other organisms. Thus, species inevitably co-evolve with their cohabitants. Interactions between species have been crucial in the emergence of major life forms and the generation of biological diversity, and these interactions need to be taken into account to understand evolutionary processes of adaptive radiation, lineage and ecological diversification (Darwin, 1859; Moran, 2006).

The term symbiosis was introduced by the German mycologist Heinrich Anton de Bary (1878) as "the living together of unlike organisms". Following this definition, symbiosis covers the whole range of relationships between species, including parasitism (where the parasite exploits its host), commensalism (where one species gains an advantage from another without harming it) and mutualism. Today the term is commonly used more narrowly and synonymous with mutualism, to describe those relationships in which both parties benefit. These benefits might aid in defence, produce a stable environment, provide access to nutrients or the utilisation of new metabolic pathways, to name just a few.

Symbiotic associations can be found between virtually all types of organisms and may take a myriad of forms. They range from temporary and optional associations to permanent obligate relationships, where at least one of the partners is not viable outside the symbiosis. Other distinguishing features are the mechanisms for achieving and maintaining the association, the types of benefits for the partners involved and the evolutionary history of the relationship. However, despite their differences, partners evolve to accommodate each other (Moran, 2006).

1.1.1 Endosymbiosis

An especially close symbiotic relationship is established when one partner (the endosymbiont) lives within the other (the host). Endosymbiotic relationships have had profound impacts on cellular evolution and diversity (Bodył et al., 2007), among others because the mitochondria and plastids of modern eukaryotes stem from endosymbionts, as first suggested by Mereschkowski (1905) in his endosymbiont theory (see Chapter 1.3). Endosymbiotic relationships are common across the tree of life (Bodył et al., 2007) but in most cases involve a prokaryotic symbiont and an eukaryotic host. The uni- or multicellular hosts provide stable and lush environments for the endosymbionts and in return benefit from metabolic abilities of the symbiont. Many metabolic functions like photosynthesis or nitrogen fixation are confined to prokaryotes and only accessible to eukaryotes via association with bacteria. Other metabolic pathways like the synthesis of some essential amino acids have been lost in some eukaryotes but can be provided by the symbionts (Kneip et al., 2007). As a consequence of this cooperation, the fitness of both host and symbiont are enhanced (Moran, 2006). Examples are gut bacteria that assist the digestion in herbivorous animals, the symbiotic bacteria that supply their aphid hosts with essential amino acids or nitrogen fixing endosymbionts in plants and algae.

In the case of unicellular hosts, the endosymbionts are localised either directly in the cytoplasm or in a vacuole surrounded by a host-derived membrane. Intracellular endosymbionts can either re-colonise the host in a process called endocytobiosis in or within every host generation or be vertically transmitted to the next host generation. Non-vertically transmitted symbionts may show morphological and physiological modifications to accommodate the host but they keep their autonomy and genomic changes do not occur. Vertically transmitted endosymbionts on the other hand have no more need to survive independently of the host and usually show reductive genome evolution and lose genes that are not essential for the symbiosis. Such a reduction of the endosymbiont genome results in an obligate and permanent relationship, because the symbiont becomes dependent on the metabolites of the host and can no longer survive free-living (Kneip et al., 2008).

1.1.2 Genomic changes in vertically transmitted endosymbionts

Genome shrinkage

Many examples of endosymbiotic relationships have been described so far and consequently a range of common characteristics have become apparent. One of the major consequences of an endosymbiotic lifestyle is progressive genome size reduction.

Bacterial genomes show a strong correlation between genome size and the number of genes present. A reduction in genome size therefore implies the loss of genes and of the associated gene functions (Moran and Mira, 2001). Usually the number of genes for basic cellular processes like protein biogenesis, DNA replication and repair, cell division, and primary metabolism, is genome size independent (Stover et al., 2000). It is the proportion of genes involved in transcription and its control, signal transduction, cell motility, secondary metabolism, and energy production and conversion that decreases with decreasing genome size. Larger genomes generally require a more complex regulation of gene expression and therewith an increased number of regulatory protein encoding genes (Bentley and Parkhill, 2004).

Genome size reduction is usually observed as a response to a simplified or more stable environment. This implies an ongoing selection pressure to reduce genome size. Free-living bacteria encounter uncertain and changing environments and need to be able to adapt their gene expression to changing conditions. This requires a tight and elaborate system of gene regulation. In a stable environment with easy access to nutrients, as an eukaryotic host can provide it, gene expression does not need to be tightly regulated or adaptable and many regulatory genes and alternative metabolic pathways become obsolete (Nilsson et al., 2005). Consequently those bacteria with the smallest genomes are intracellular pathogens and symbionts that maintain obligate associations with eukaryotic hosts (Ochman, 2005). Recently this was suggested by Koskiniemi et al. (2012) who demonstrated fitness gains after spontaneous genomic deletions in *Salmonella enterica*. Approximately 25% of the examined deletions caused an increase in fitness under one or several growth conditions. Deletions that substantially increased fitness were fixed in the population after 1000 generations.

The way in which reductive genome evolution progresses can be deduced from the observation of prokaryotes that are at different stages in the evolution of their endosymbiotic relationship. The possibilities for genetic exchange through conjugation or transformation are limited in vertically transmitted symbionts. They become isolated from the gene pool of the free-living population, and as a result their genomes are clonal or near-clonal (Moran, 2006). In the initial stage the genome size remains large as typified by *Mycobacterium tuberculosis*, which has only recently adapted its lifestyle, is no longer found outside of its host, and has little genetic variation but retains a larger genome (Lawrence et al., 2001). By losing their ability to recombine, endosymbionts become subject to "Muller's ratchet". This principle proposes that neutral as well as deleterious but non-lethal mutations accumulate more rapidly in asexually propagated genomes than in sexually propagated ones (Muller, 1964). The effects of genetic drift are also more intense in small populations and under relaxed functional constraints, both of which apply to endosymbiont evolution. Indeed, endosymbiont genes typically diverge rapidly and the genomes of obligate intracellular bacteria often show an accumulation of deleterious mutations when compared with free-living relatives (Kneip et al., 2008). When a mutation that incapacitates gene function is fixed in the population, the gene becomes a pseudogene. As a consequence pseudogenes accumulate in endosymbiotic genomes. This stage is typified by *Mycobacterium leprae*, which has exploited its intracellular lifestyle for a longer period of time and has accumulated >1100 pseudogenes .The acquisition of pseudogenes can be seen as the first stage in a larger-scale genome reduction because the mutational process in bacterial genomes is strongly biased toward deletions (Andersson and Andersson, 1999). This is thought to be the reason why bacterial genomes are compact and gene rich and do not accumulate large non-functional regions (Ochman, 2005). It is usually explained as a defence mechanism in bacterial genomes against invasions of self-replicating genetic elements. Ultimately, the bias favouring deletion over insertion would cull the numerous pseudogenes from the genome of the intracellular bacterium, resulting in genomes with moderate (as in the case of *Rickettsia*) and eventually very few (as in the cases of *Buchnera* and *Carsonella*) pseudogenes. In these organisms, the intracellular lifestyle was adopted more than 250 mya (million years ago) (Moran and Wernegreen, 2000) and although the process of gene loss is ongoing, most pseudogenes have already been eliminated (Lawrence et al., 2001).

The bias towards deletions can be even more pronounced in reduced genomes because often genes of the RecA DNA repair pathway that counteract deletions are themselves missing (Nilsson et al., 2005). The loss of the ability to recombine their DNA has thus been termed a "one way ticket to genome shrinkage" for obligate bacterial endosymbionts (Delmotte et al., 2006). As a side effect, this gene loss stabilises the symbiont's dependence on its host because the loss or inactivation of genes whose products are not required in the partnership might nonetheless be essential for an independent lifestyle. Consequently, intracellular obligate symbionts over time increasingly lose their autonomy (Kneip et al., 2008).

The early stages of symbioses are typically characterised by massive gene losses. Two mechanisms of gene loss are supported by several studies (Delmotte et al., 2006): i) deletions of large sets of contiguous genes including loci with unrelated functions and ii) single gene losses. Large-scale deletions are thought to be the result of chromosomal rearrangements due to recombination of repeat sequences. Repetitive regions are lost in the process and deleterious events become rarer over time (Moran and Mira, 2001) . Several other studies have found multiple events of single gene loss dispersed over the whole genome (Silva et al., 2001). Based on this observation and their own work on a pathogen, Dagan et al. (2006) proposed the "domino theory of gene death", a two step model for reductive genome evolution in bacteria. The process starts with a gradual corruption of genes over relatively long periods of time via Muller's ratchet, causing a gradual accumulation of pseudogenes. Eventually, a crucial gene within a complex pathway or network is rendered non-functional triggering a rapid "mass gene extinction" of co-dependent genes.

Both mechanisms imply that what genetic functions or pathways are lost in reductive evolution is to a large extent left to chance and that even apparently beneficial genes are lost this way. Many genes that have been found missing in reduced endosymbiotic genomes encode for information-processing functions, such as DNA repair, replication, transcription and translation – gene functions that are generally retained in bacterial genomes. Consistent with this, small genomes of eubacteria have been found to retain very different gene sets (Moran and Mira, 2001). Despite this, the level of conservation of a gene in a free-living bacteria still influences its propensity of being lost in an endosymbiotic lifestyle, as selection is not entirely withdrawn from endosymbiont evolution (Krylov et al., 2003; Delmotte et al., 2006). However, the importance of a gene

can change with the demise of functionally related genes. Successive losses of single genes are therefore not always independent events because the loss of a gene can influence the types of losses the organism can tolerate in the future (Delmotte et al., 2006). Compensatory mutations on the other hand can restore fitness and allow reiteration of the reductive process (Nilsson et al., 2005). Due to these factors - and of cause differential changes in the environment - genes identified as essential in one organism can become dispensable in another lineage over evolutionary time (Moran and Mira 2001).

But genomes not only shrink via the loss of genes. The chromosomal rearrangements and gene losses that accompany genome reduction can alter the boundaries of transcription units. Studies on the aphid endosymbionts of the genus *Buchnera* showed that promoter loss is frequent when a group of newly contiguous genes can be transcribed as a single transcriptional unit. The degeneration of Shine-Dalgarno sequences and loss of protein binding sites additionally contribute to the genome shrinkage and as a consequence only few regulatory proteins are retained (Shigenobu et al., 2000; Moran and Mira, 2001). This makes gene expression relatively independent of environmental conditions but in a stable host environment this may not be of much consequence.

A/T enrichment

The loss of genes is not the only effect of reductive evolution on the genomes of endosymbiotic bacteria. Accelerated genetic drift due to "Muller's ratchet" and the loss of DNA repair genes also lead to a marked A/T enrichment (Bentley and Parkhill, 2004). Every reduced genome appears to have lost some genes involved in DNA recombination and repair pathways, though the precise set discarded varies (Moran, 2002). Their absence allows for point mutations to accumulate. The most frequent random mutation occurring in cells is C to T (or G to A), due to the deamination of Cytosine to form Uracil, which is subsequently replicated as Thymidine (Glass et al., 2000). In bacteria the GC content can range from 72.1% (*Streptomyces coelicolor*) to 26.5% (*Wigglesworthia glossinidia*). The average for free-living bacteria has been estimated to be 49% compared with 38% in host obligates (Rocha, 2002). The mutational drift toward A/T is more pronounced in intergenic, non-coding sequences and in the third codon base position in coding sequences because changes at these positions are less likely to affect gene functions. However, the bias towards A/T

enrichment does eventually affect the amino acid composition of the proteins encoded in reduced genomes (Moran, 2003). A consequence of this may be another characteristic of small genome bacteria, the constitutive over-expression of heat shock proteins. Heat shock proteins can stabilise the conformation of proteins and in this way to a certain extent compensate for the accumulation of destabilising substitutions in the amino acid sequence (Moran, 2003).

1.1.3 Endosymbionts and organelles

Of the various organelles known in eukaryotic cells, the semi-autonomous mitochondria and plastids hold an exceptional position. They are the distinct locations of important metabolic functions, contain their own - though minuscule - genome and gene expression machinery, and are separated from the cytoplasm by two or more membranes with a sophisticated membrane transport machinery that mediates extensive metabolic interchange between the organelles and the cytoplasm. These striking characteristics are explained by the endosymbiont theory (Margulis, 1970; Gray et al., 1999). It states that these organelles originated as independent prokaryotic organisms, which were taken inside the cell as endosymbionts and subsequently merged with their host. The theory was first suggested by Konstantin Mereschkowsky in 1905 regarding chloroplasts. He based this idea on the observations that the division of chloroplasts in green plants closely resembles that of free-living cyanobacteria (Schimper, 1883) and that chloroplasts behave like independent, autonomous organisms that grow by division and adapt in number to the size of expanding leaves (Sachs, 1882)Ivan Wallin extended the idea of an endosymbiotic origin to mitochondria (Wallin, 1923) but the fact that they have the same staining properties as bacteria was already noted by Altmann (1890). Today the endosymbiont theory is widely accepted, due to supportive evidence from all areas of biology, including genomic data, protein structure and data on cell morphology and metabolic pathways (Gray and Doolitle, 1982).

The formation of new organelles or organisms by mergence of different individuals over the course of their co-evolution is known as symbiogenesis (Margulis and Fester, 1991). It is characterised by morphological and physiological modifications as well as genomic changes in both partners (Palenik, 2002). The endosymbiosis between the mitochondrial ancestor and its host may have been the trigger for the evolution of

eukaryotes. This event seems to be the origin of eukaryotes since all contemporary eukaryotes examined contain some genes from this symbiont (Embley et al., 2003). Thus, all eukaryotes once had mitochondria and where they are missing in extant taxa they have been lost.

The analysis of mitochondrial genomes indicates an α-proteobacterial ancestor, which merged with an (archea-like) host and whose genome has since been reduced due to gene loss and gene transfer to the host nucleus (Andersson et al., 1998; Martin and Müller, 1998). The origin of mitochondria was most likely a unique event that lies back more than 1.5 billion years (Martin and Russell, 2003). The main function of mitochondria is the ATP synthesis by oxidative phosphorylation, but they are also involved in many other processes in the eukaryotic cell, including calcium storage, regulation of the membrane potential and synthesis reactions (Pizzo and Pozzan, 2007; Rossier, 2006).

Plastids, the characteristic organelles of all photoautotrophic eukaryotes, likely originate from the merger of a cyanobacterial ancestor with a phagotrophic eukaryotic host cell (Bhattacharya and Medlin, 1995) and have since further spread by secondary and tertiary symbioses (Patron et al., 2006). The results of many studies seem to favour a monophyletic origin of plastids but the evidence is not unequivocal (Larkum et al., 2007). The plastid's main functions are the photosynthesis reactions and the subsequent synthesis of carbohydrates. Like in mitochondria the genomes of plastids are strongly reduced compared to cyanobacterial genomes and over time the majority of plastid genes have been transferred into the host nucleus (Delwiche and Palmer, 1996). Due to their bacterial origin, plastids and mitochondria have several common characteristics. These include a double membrane, their own translation machinery with bacterial 70S ribosomes and circular chromosomes. In both cases the symbiotic association was the only way for eukaryotes to acquire the purely bacterial metabolic pathways.

Another common feature of organelles is the strong genetic reduction due to gene loss, gene transfer to the host nucleus, and substitution with host genes. These processes resulted in a remarkable integration of organelles into the host organism. The majority of organelle proteins are now encoded in the host nucleus and translated in the host's

cytoplasm from where they are imported into their respective organelle to fulfil their function.

The endosymbiotic events leading to organelle formation took place so long ago, and the organelles are now so highly integrated, that the details of organelle evolution can hardly be unravelled. Moreover, due to the long time that passed, it cannot be assumed that the closest extent relatives of mitochondria or plastids are representative of the symbiotic ancestors involved. Organelle evolution involves not just massive gene loss, but a high degree of gene transfer and innovations to protein transport systems that are difficult to untangle without a clear idea of the ancestral conditions of the participant cells. It has for these reasons become common practise to look at younger symbioses with well known non-symbiotic relatives of the partners to infer the changes that go with symbiogenesis (Patron et al., 2006).

Based on the observations made on these younger symbioses the process is usually pictured similar to the following model: After establishing a stable endosymbiotic relationship, the endosymbiont presumably first lost genes that had become nonessential due to its intracellular life style. The bulk of the remaining essential genes were then transferred to the host's nuclear genome. Productive gene transfer requires the acquisition of a promotor. Following this, genes would exist in duplicate until the system evolved a targeting machinery to relocate the gene product into the proto-organelle. A nuclear copy of an organelle gene capable of producing a functional organelle protein would make the original gene redundant. It would be inactivated and eventually lost by mutation (Dyall et al., 2004; Schneider and Ebert, 2004). After the evolution of a protein import system genes are transferred into the nucleus by recombination into pre-existing promoter and/or transit peptide-coding regions. Organelle genes of equivalent function can replace pre-existing host genes. This process is known as endosymbiotic gene replacement and often results in hybrid pathways involving genes from both, host and endosymbiont (Timmis et al., 2004). Host genes can replace organelle genes in a similar way. In this case a host homologue of the organelle protein is imported and renders the organelle gene obsolete (Millen et al., 2001).

Other mechanisms for endosymbiotic gene transfer (EGT) into the nucleus have been suggested as well. The 'promiscuous hypothesis' of plastid evolution is based on the

observation that DNA can be transferred from plastid to nucleus surprisingly easily following plastid lysis (Sheppard et al., 2008). It is conceivable that organelle formation started with transient relationships similar to kleptoplasty, ending in lysis of the symbiont and incorporation of DNA into the host nucleus before a stable relationship was achieved (Larkum et al., 2007).

The recurring transfer of genes from the eventual organelles to the nuclear genome indicates a distinct selection pressure to that effect. Muller's ratchet has been suggested as one principle causing organelle to nucleus gene transfer as well as the massive gene loss in the endosymbiont genome. In organelles, as well as most maternally transmitted symbionts, the genome is clonal or near-clonal and is subject to different mutational processes (e.g. A/T bias) and a different population structure from the host's nuclear genome, makes the nucleus a favourable because more stable location for genes (Muller, 1964; Howe et al., 2000; Moran, 2006).

Experiments with selectable marker genes in mitochondria and chloroplasts have shown that transfers to the nucleus occur at comparatively high frequencies (Thorsness and Fox, 1990; Huang et al., 2003). It is an ongoing process although gene transfers to the nucleus seem to have reached a plateau in most eukaryotic groups. Thus at the same time a selection pressure to retain some genes in an organellar genome seems to exists.

Modern mitochondrial genomes range in size from 6 kbp in *Plasmodium falciparum* to 569 630 kbp in *Zea mays*. The median lies at ca. 30 kbp which equals about 1% of the estimated size of the ancestral bacterial genome. The number of genes is equally variable, though the represented gene functions are similar, with *Plasmodium* mitochondria encoding for only five genes, compared with several hundred protein coding genes in plant mitochondria (Gray et al., 2004). These differences are likely the result of differential migration of mitochondrial genes to the nucleus (Burger et al., 2000). Interestingly, mitochondrial genes show signs for selection for smaller genomes. The resident mitochondrial gene products are shorter than their homologs in α-proteobacteria. This correspondence between genome size and gene length has also often been observed in intracellular symbionts (Schneider and Ebert, 2004).

It has been estimated, that ca. 3000 genes were present in the cyanobacterium-like ancestor of plasmids (Prechtl and Maier, 2001). Typically only 100-200 genes are left in extant plastids (Martin et al., 2002). The similarity in gene content among contemporary plastid genomes is likely the result of immensely convergent evolution (Timmis et al., 2004). Some of the original genes have been lost altogether, but the majority have been relocated to the nucleus.

The loss of the bulk of their genomes necessitated the evolution of elaborate mechanisms for organelle biogenesis and metabolite exchange. In the process, symbionts acquired many host-derived properties and lost much of their eubacterial identity. The two critical events of organelle biogenesis - division and pre-protein translocation – are both driven by a combination of symbiont and host derived proteins (Dyall et al., 2004). Most transcription regulators of the endosymbiont were lost and replaced with host transcription factors. This change in gene regulation can be seen as the basis of nuclear control of plastid gene expression and must have been accompanied by various changes in regulatory elements and interactions. In plastids the replication system was replaced as well, most likely by duplication of the nuclear-coded mitochondrial replication system (Sato et al., 2003).

Even though organelles developed from endosymbionts and many recent microbial endosymbionts are developing in a very similar direction, a clear distinction between endosymbionts and organelles can be made following the definition by Cavalier-Smith and Lee (1985). The former is an organism, though not necessarily autonomous, that encodes all its proteins in its own genome, the latter is a cell compartment derived from an endosymbiotic bacterium with most of the required proteins encoded in the nucleus and then imported from the cytoplasm. Until recently only plastids and mitochondria (including organelles derived from mitochondria like hydrogenosomes e.g.) were known to pass the strict test of having well-developed protein import mechanisms.

The freshwater amoeba *Paulinella chromatophora* has long been known to carry two blue-green chromatophores within its theca (Lauterborn, 1895). It has since been established that these intracellular bodies are photosynthetic endosymbionts most closely related to *Synechococcus*-type cyanobacteria (Bodył et al., 2007). Moreover, indirect evidence has been found that endosymbiotic genes (*psaI* etc.) have not just

been transferred into the nucleus but that their gene products are re-imported into the endosymbiont. It therefore meets the definition of a true organelle. This makes the *Paulinella* cyanelle the only know photosynthetic organelle that is not derived from the ancient endosymbiosis that gave rise to the Arachaeplastidae but another endosymbiosis barely 60 mya (Nowack and Grossman, 2012).

Cavalier-Smith and Lee (1985), as many others, have reasoned that symbiogenesis must be a very rare event because it involves not just the loss of symbiont genes, but the transfer of symbiont genes into the nucleus of the host with subsequent relocation of the gene products into the organelles. This requires the evolution of a protein import system. Modern protein-targeting systems are very complex and dynamic because they not only trans-locate proteins with high accuracy but also serve to regulate protein abundance (Silva-Filho, 2003).

Organelle proteins are labelled with targeting sequences to indicate their appropriate location. These targeting sequences are often located at the N-terminus, loosely conserved, and composed of basic and hydrophobic amino acids. They are recognized by a complex protein import machinery. Proteins do not just need to be targeted to the organelle but to the correct compartment, four in mitochondria and six in chloroplasts (Lucattini et al., 2004). The presence of a second type of organelle in photosynthetic eukaryotes adds an additional level of complexity and required the coordinated evolution of both the mitochondrial and proto-plastid protein import machineries to avoid mis-targeting of potentially harmful proteins or to promote dual targeting of proteins shared by both organelles (Silva-Filho 2003). In secondary plastids, e.g., the apicoplast, the existence of two extra membranes necessitated the creation of a bipartite presequence, consisting of a signal peptide for entrance into the secretory pathway fused to a “traditional” plastid transit peptide for crossing the two inner plastid envelopes (Mcfadden, 1999). Once established, the targeting sequences could easily have spread into other gene sequences by exon shuffling or alternative splicing. This would also allow the introduction of non-organelle proteins into the evolving organelle (Lucattini et al., 2004).

But even though a very simple system would have stood at the beginning of protein import, it has long been assumed that the evolution of both, import system and targeting sequence attached to the right genes, would have been exceedingly unlikely

and therefore a unique event, as it presents much of a hen and egg problem (Cavalier-Smith and Lee, 1985; Theissen and Martin, 2006). This viewpoint has deeply affected the reconstruction of eukaryote phylogenies as it is widely assumed that it is much easier to lose an organelle than to gain one. By contrast, available empirical data demonstrate that endosymbionts are common and, once fully integrated into host-cell metabolism, are difficult or even impossible to lose (Waller and McFadden, 2005).

In recent years new evidence and ideas have been brought forward that indicate that the problem of developing a protein import system might have been overemphasized. Protein import evolved several times, in mitochondria, Plantae plastids, the Paulinella plastids, and presumably multiple times in secondary and tertiary plastids. This indicates that it is not as unlikely and rare an event as thought (Bodył and Moszczyński, 2006; Patron et al., 2006). With growing understanding of the import apparatuses it is becoming clear that their development are examples of evolutionary tinkering with pre-existing components in the host and the endosymbionts (Reumann et al., 2005; Cavalier-Smith, 2006).

Comparisons of the proteins of the organelle import machineries show that these are of mixed origin, reflecting the interoperation of symbiont and host on the way to a working metabolite and protein exchange across the membranes. Genes with regulatory functions on the other hand appear to be eukaryotic (Karlberg et al., 2000). This is consistent with the assumption that the assimilation of the bacterial genome provided the means for the eukaryote to control the metabolism of the endosymbiont (Lucattini et al., 2004).

The acquisition of the mitochondrion could have pre-adapted eukaryotic cells for additional endosymbiotic bacterial associations. The similarities between the protein import systems of both organelles and their interaction with the host organism might not just be due to similar starting conditions but the adoption of elements of the existing mitochondrial import system for the evolving plastid (Zhang and Glaser, 2002). Chloroplast transit peptides for example have some of the features of mitochondrial target peptides and many nuclear encoded proteins are imported into both organelles (Macasev et al., 2000). The presence of the plastid then again might ease the assimilation of other cyanobacterial endosymbionts, as many cyanobacteria genes are already present in the nucleus. This is likely to have played a role in the acquisition of

some tertiary plastids that have replaced their host's secondary plastids. In some dinoflagellates secondary and tertiary plastid seem to have existed next to each other for a period of time. The transit peptides of the tertiary plastid changed markedly, presumably as a reaction to the presence of another plastid. The tertiary plastid also adopted some secondary plastid proteins for use. It should be expected that it is easier to move a gene from the new organelle to the nucleus than to make a gene already present in the nucleus work in the heterologous environment of the new organelle, but the fact that these genes are already located and perhaps appropriately regulated in the host nucleus might have compensated for this (Patron et al., 2006).

The emergence of mitochondria and plastids were rare and significant events in the evolution of eukaryotes but secondary and tertiary plastids seem to have formed several times. We now know that the evolution of a plastid is being repeated in the Amoeba *Paulinella*. This makes it conceivable that more endosymbionts could make the transition to organelles, including those that fix Nitrogen in rhopalodian diatoms (Larkum et al., 2007). The metabolic capacity of the bacterial symbiont was the driving force for the evolution of the known organelles. The molecular fixation of nitrogen has the potential to do the same.

1.1.4 Cyanobacteria

Cyanobacteria, also known as blue-green algae, are a phylum of gram-negative bacteria that obtain their energy through photosynthesis. They are one of the oldest and genetically most diverse clades on earth with a fossil record dating back at least 2.8 billion years (Olson, 2006). The ability to perform oxygenic photosynthesis evolved in a common ancestor of the clade and made it the probably most successful lineage on earth. Cyanobacteria may also be the most influential lineage in the evolution of life on earth, not only because they developed oxygenic photosynthesis, which led to the oxidation of the earth's atmosphere and allowed for more efficient energy conversion via oxygenic respiration, but also because they made photosynthesis available to eukaryotes by entering into an endosymbiotic relationship with the common ancestor of the Plantae.

Today Cyanobacteria account for 20-30% of photosynthetic productivity on the planet and are one of the most important primary producer in marine ecosystems (Pisciotta et

al., 2010). Of all photosynthetic prokaryotes they are the only ones with thylakoids, highly differentiated internal membrane systems that facilitate photosynthesis (Liberton et al., 2009). Some species are also able to fix atmospheric nitrogen, the only oxygenic photosynthetic organisms capable of doing so (Haselkorn and Buikema, 1992). This is made possible by various evolved mechanisms, which protect the nitrogenase enzyme complex from oxygen. The ability to fix nitrogen in aerobic conditions and perform photosynthesis led to symbiotic relationships with a number of other groups of organisms such as fungi (lichens), diatoms, corals, pteridophytes (*Azolla*), angiosperms and others (Rai et al., 2000). The bacteria provide their hosts with the products of photosynthesis and/or nitrogen fixation, both processes that eukaryotes are incapable of. It was such a symbiotic relationship between a cyanobacterium and a eukaryote that gave rise to the photosynthetic organelles in plants and algae (Keeling, 2004).

1.1.5 Symbiotic nitrogen fixation

Nitrogen is an essential element of life's biochemistry. Among other things it is required to form amino and nucleic acids, photosynthetic pigments and many metabolic co-factors. It is also an important component of muramic acid in bacterial cell walls. Even though di-nitrogen (N_2) constitutes 78% of the earth's atmosphere, it is inaccessible for most organisms because of the stability of the triple bonded N_2 molecule. Non-biological processes fix nitrogen in biological available forms but at the same time the continual production of N_2 by the bacterial denitrification depletes ecosystems of this vital element. Thus although nitrogen as an element, is present in a nearly inexhaustible supply, its biologically available forms, either oxidized (NO-/NO_3) or reduced (NH-/organic N), are often the growth limiting factors in organic productivity (Karl et al., 2002). Under this selective pressure, some prokaryotes evolved the ability to fix molecular nitrogen, albeit in a very energy intensive reaction (Raymond et al., 2004).

All known nitrogen-fixing organisms (diazotrophs) are prokaryotes. The ability to fix nitrogen is paraphyletically distributed across the eubacterial and archaeal domains, indicating that the responsible enzyme system has been mobile through horizontal transmission. There are numerous instances of nitrogenase genes and operons being selectively lost, duplicated, horizontally transferred, and, in at least one significant case,

recruited into the photosynthetic apparatus pathway (Raymond et al., 2004). Eukaryotes on the other hand are not capable of fixing nitrogen but several symbiotic relationships between diazotrophic bacteria and eukaryotes make nitrogen fixation available to plants and algae (Prechtl et al., 2004).

The reduction of molecular nitrogen is catalysed by the same nitrogenase enzyme complex in all bacteria and the responsible *nif* genes are universal in nitrogen fixing organisms. *Nif* genes are generally only expressed under nitrogen fixing conditions and are found within one or several extensive, co-transcribed operons or regulons that not only encode the subunits of the functional nitrogenase protein but also code for an expansive suite of proteins involved with regulation, activation, metal transport, and cluster biosynthesis (Raymond et al., 2004).

The nitrogen fixation reaction is a strongly energy-dependent reduction of molecular nitrogen to two molecules of NH_3. The nitrogenase enzyme is extremely sensitive to oxygen (Karl et al., 2002). This means the enzyme needs to be protected from environmental oxygen as well as oxygen produced within the cell, for example by photosynthesis. Two main strategies have evolved to achieve this: spatial separation and temporal separation.

Spatial separation, for example in filamentous cyanobacteria, requires the specialisation of some cells (heterocysts) for nitrogen fixation (Wolk and Ernst, 1994). These cells show a reduced photosynthetic activity and form thicker cell walls that separate them from their neighbour cells (Murry and Wolk, 1989).

Temporal separation is used in unicellular diazotrophic cyanobacteria, performing the nitrogen fixing and photosynthetic reactions within the same cell. Oxygen producing reactions are confined to the light period, while nitrogen fixation and oxygen consuming reactions are performed during the dark period (Colón-López et al., 1997). The efficiency of this strategy is demonstrated by the fact that the photosynthetically active non-heterocyst cyanobacteria are estimated to be responsible for more than 50% of nitrogen fixation on Earth (Klipp, 2004). It also makes them interesting candidates for symbiotic relationships because every single cell is capable of nitrogen fixation.

1.2 Diatoms

Diatoms are the most species-rich group of unicellular algae, with more than 200 extant genera and - according to conservative estimates - tens of thousands of species (Droop et al., 1993). Diatoms are autotrophic and live in almost all kinds of aquatic and semi-aquatic environments that are exposed to light. Each cell possesses two or more yellow, olive or golden-brown photosynthetic chloroplasts, a central vacuole and a large central diploid nucleus. Diatoms lack flagella and pseudopodia but pennate diatoms can glide over substrates by producing a stream of mucus between the frustule and the surface. Planktonic centric diatoms are non-motile and rely on water movement to avoid sinking below the photic zone.

A distinctive characteristic of this group is the intricately patterned cell wall (frustule) composed of silica (hydrated silicon dioxide) and organic material. The frustule usually consists of two overlapping valves. In most species, when a diatom divides to produce two daughter cells, each cell keeps one of the two halves and grows a smaller half within it. This results in an unusual mode of cell-size reduction and restoration coupled to the sexual cycle. After each division cycle the average cell size in the population decreases. Once cells reach a minimum size, they reverse this decline by a cycle of sexual reproduction, resulting in a much larger cell called an auxospore, which then continues the size-diminishing divisions.

Frustules show a wide diversity in form and display species-specific nanostructures. Recent attention has focused on biosynthesis of these nanostructures as a paradigm for future silica nanotechnology (Parkinson and Gordon, 1999). Frustules are a convenient tool for classification because of the high diversity in shapes, though it isn't well understood how common homoplasies are in frustule structures. Traditionally, two types of diatoms are defined by shape, the radial symmetric (centric) and the bilateral symmetric (pennate) (Armstrong and Brasier, 2005). However, molecular evidence has determined that centric diatoms are paraphyletic and a more recent classification divides the diatoms into three classes: centric diatoms (Coscinodiscophyceae), pennate diatoms without a raphe (Fragilariophyceae), and pennate diatoms with a raphe (Bacillariophyceae) (Round et al., 1990; Williams and Kociolek, 2007).

Pennate diatoms dominate the freshwater, soil and epiphytic niches, while centric diatoms thrive as marine plankton. They constitute a major proportion of marine phytoplankton and are estimated to generate up to 40% of the primary production in the global ocean (Armbrust and Rynearson, 2006). Over all, diatoms are responsible for ~20% of global carbon fixation - thus playing an important role in the carbon, silica and nutrient budgets of the modern ocean.

Diatoms belong to the heterokontophytes. This group arose from a secondary endocytobiosis event when a red alga was engulfed by a phagotrophic eukaryote and subsequently reduced to a complex plastid surrounded by 4 outer membranes (Gibbs, 1979; Armbrust and Rynearson, 2006). As a result, nuclear-encoded diatom genes have several possible origins: nuclear or mitochondrial genomes of the eukaryotic host, and nuclear, plastid, or mitochondrial genomes of the red algal endosymbiont (Falkowski et al., 2004). To date only two diatom genomes have been fully sequenced, the marine centric *Thalassiosira pseudonana* and the pennate *Phaeodactylum tricornutum* (Armbrust and Rynearson, 2006). A comparison of their genomes finds that, despite diverging only 90 mya around 40% of their genes are not shared. More significantly, the analysis finds that the genomes of both species contain hundreds of genes from bacteria, many of which are shared between the species (Bowler et al., 2008). Some genera have symbiotic associations with extracellular or intracellular cyanobacteria, and it has been suggested that the hosts benefit from the nitrogen fixation capacity of their symbionts (Janson, 1999; Carpenter and Janson, 2000),

The involvement of red algae in the evolution of diatoms implys a more recent origin than for most other algae. It has been suggested that their origin may be related to the mass-extinction at the permian-triassic boundary (~250 mya), after which many marine niches were opened (Medlin et al., 1997). The radial centric forms likely arose sometime after this, with the first fossil record occurring at 180 mya (Rothpletz, 1896). The pennate forms are believed to have evolved from the multipolar centric diatoms and appear in the fossil record about 70 mya (Moshkovitz and Ehrlich, 1983). As a result of their silica content, diatoms have made a significant impression on the fossil record, with major deposits of fossil diatoms found from as far back as the early Cretaceous, and some rocks (e.g. diatomite, kieselguhr) being composed almost entirely of them.

1.3 Sequencing with the Illumina Genome Analyzer

Fast and accurate sequencing of large DNA molecules was made possible by the development of the chain-termination DNA sequencing reaction by Frederick Sanger in 1977. The introduction of fluorescent dye-terminators allowed for automated DNA sequence analysers, producing up to 0.5 megabases per day. This made the sequencing of entire genomes with Sanger sequencing feasible, though onerous and expensive. The method is ideally suited for targeted sequencing but the fact that it requires a sequence specific primer for the amplification reaction can also be considered as one of the main weaknesses of this method. For this reason large-scale sequencing projects like *de novo* sequencing of whole genomes usually require the creation of clone libraries. This approach is called shotgun sequencing (Anderson, 1981). Traditionally, the assembly of a genome without a reference is accomplished by aligning overlapping sequence between reads in an overlap-layout-consensus approach. Maximally overlapping segments are identified to build a consensus representation of the genome. The disadvantages of sequencing clone libraries include a high workload, and the possibility that some DNA sequences may be difficult to clone in some or all available bacterial strains, due to deleterious effects on the host bacterium.

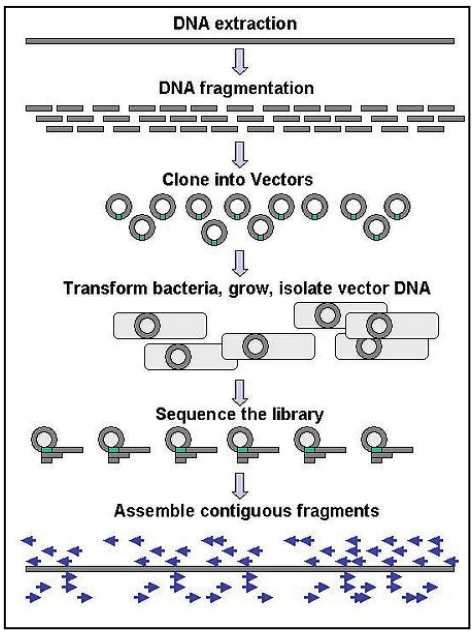


Figure 1-1 Diagram of the principles of Shotgun sequencing

Re-sequencing or targeted sequencing is used for determining variability in a DNA sequence. The resulting sequence is compared to a reference sequence to detect mutations or sequence rearrangements. The re-sequencing of whole genomes is as work and cost intensive as *de novo* sequencing and has therefore rarely been attempted using Sanger's method.

Sequencing a long stretch of DNA always involves taking numerous reads of DNA and subsequently assembling them into a contiguous sequence. The high-throughput sequencing technologies take this strategy a step further by massively parallelising the sequencing process, producing thousands or millions of short sequence reads at once. Despite the fact that the read-length is comparatively short, the overall throughput is enormous.

The Illumina Genome Analyzer and HiSEQ systems are the most widely adopted high-throughput sequencing platform to date. They use sequencing chemistry developed by Solexa (Bentley, 2006). The first Illumina Sequencers released to the market generated around 1.5 Giga base-pairs of sequence data per day in reads of up to 36 bps in length. The throughput and read length has since been steadily increased and to date has reached up to 60 Gbp per day and 150 bps read length. Further, the most recent bench top Illumina sequencers, although relatively low capacity (6.5 Gbp per day) can currently produce read lengths in excess of 250 bases. These reads can be used for a broad range of applications from *de novo* sequencing and re-sequencing to the generation of data for metagenomics and transcriptomics or the profiling of DNA methylation status.

The Illumina technology has an advantage over shotgun sequencing in that it does not require the preparation of a clone library, or knowledge of the target sequence. Instead, in genomic sequencing for example, a fragment library is created by shearing the sample DNA and ligating adapters to both ends of the fragments (see Figure 1-2). These fragments are subsequently bound to primers attached to a planar, optically transparent surface. The fragments are then amplified in a process known as bridge amplification. It produces millions of physically isolated clonal clusters, on an optical surface, facilitating signal detection.

The sequencing process uses DNA polymerase to incorporate reversibly fluorescent-labelled nucleotides that are complementary to the template sequence. After incorporation of each base a digital image is taken during laser excitation and the base

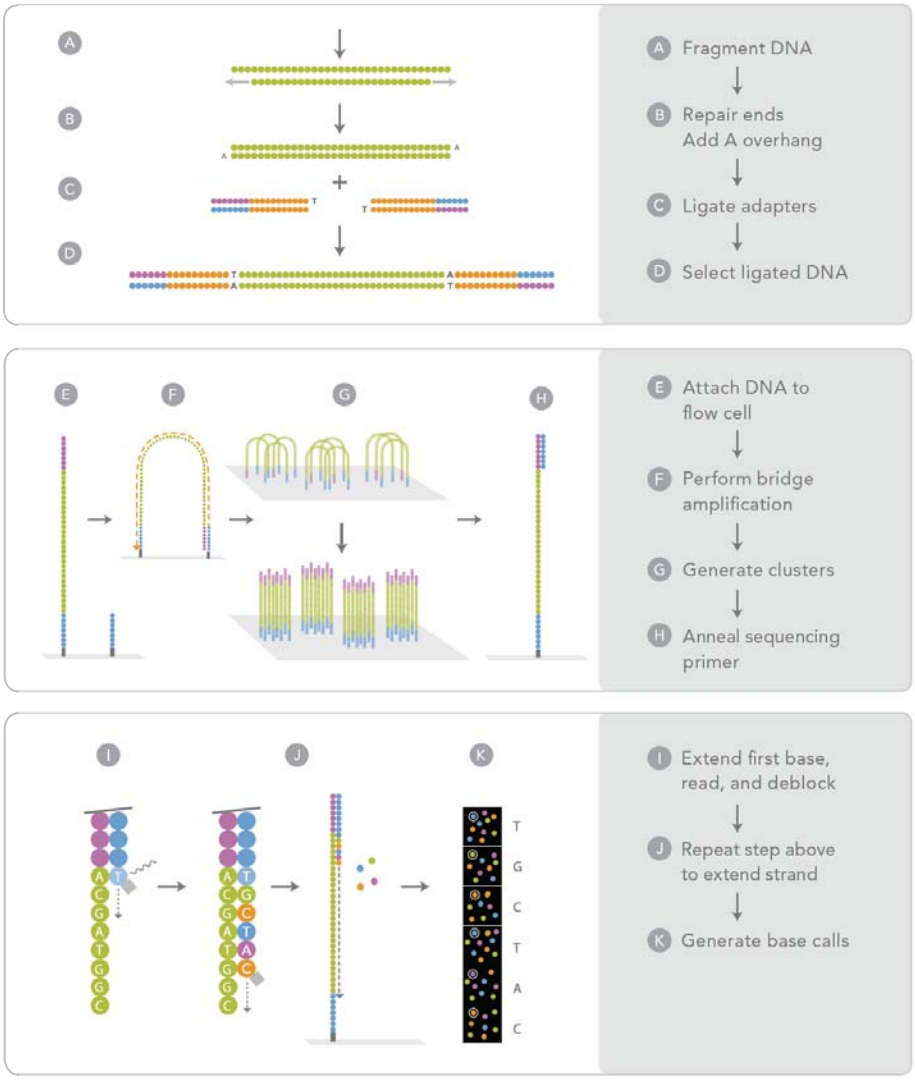


Figure 1-2 Diagram of the Solexa (Illumina) sequencing process

identified by the wavelength emitted by each cluster. This approach ensures high accuracy and avoidance of artifacts with homopolymeric repeats. The DNA sequences of all locations are determined in parallel from the digital images of the reaction surface. This requires extensive computational data analysis following the sequencing reaction itself.

Paired-end libraries can generate paired base reads separated by a known distance and known orientation. Paired reads generate greater genomic information than two single reads, because they add positional information. The fragment size can be chosen so that the two reads overlap to produce pseudo-long reads. This can help to overcome the analysis issues that short sequences create with sequence repeats and rearrangements.

This is of particular relevance for *de novo* assemblies of genomes and characterisations of copy number variations that are different from the reference sequence (Bentley, 2006).

1.4 Data analysis

The data produced by Illumina sequencers is markedly different than that generated by the Sanger sequencing technologies, and therefore requires a different type of data analysis. If a close reference sequence is available the reads can simply be aligned to it. If such a reference is not available, the reads need to be assembled. Very short reads are not well suited to the traditional overlap-layout-consensus approach for assembly. Because of their length, they must be produced in large quantities and at greater coverage depths. Whole genome sequencing projects using Sanger reads of about 750-800 bps in length require only seven to tenfold coverage. A much higher 50- to 100-fold coverage is required if shorter reads as produced by high-throughput sequencing are used (Salzberg et al., 2012). Even this much deeper coverage does not entirely compensate for the shorter read length and the resulting assemblies show poorer contiguity than long read assemblies as short reads make the assembly problem substantially harder (Gnerre et al., 2011; Nagarajan and Pop, 2009).

The sheer number of reads makes the traditional overlap graph, with one node per read, extremely large and impossible to compute. Short reads also only allow for short overlaps that lead to many more ambiguous connections in the assembly, even though increasing read length reduces this problem (Zerbino and Birney 2008). Instead, novel algorithmic approaches were developed to make optimal use of the Genome Analyzer data for *de novo* assembly. Most of them utilize graph theory, a well-explored area of applied mathematics and computer science. Graphs are mathematical structures used to model pair-wise relations between objects from a certain collection. The dutch mathematician Nicolaas de Bruijn used n-dimensional directed graphs to represent overlaps between sequences. The EULER assembler (Pevzner et al., 2001) was the first software to use de Bruijn graphs (see Figure 1-3).

The fundamental data structure in de Bruijn graph-based algorithms is based on k-mers (words of k nucleotides), not reads, thus high redundancy is handled by the graph

without affecting the number of nodes. In a first step the reads are hashed as k-mers. Increasing hash length decreases the possibility of two reads being incorrectly identified as overlapping. At the same time, increasing hash length also decreases total coverage, because the possibility of related overlapping reads left unidentified also increases. Conversely, a smaller hash length increases total coverage, but causes more reads to be incorrectly defined as overlapping. The ideal hash length value needs to be determined empirically for each dataset (Zerbino and Birney 2008).

K-mers that overlap over their whole length but with an offset of 1bp form nodes, each with a twin node made up of the reverse complement k-mers. Nodes are connected by arcs if the k-mers at the end of both nodes overlap. Nodes and connecting arcs are then mapped as paths, traversing the graph. Whenever a node has only one outgoing arc that points to another node that has only one ingoing arc, the two nodes are merged to a contig (continuous sequence), thus simplifying the graph (Zerbino and Birney 2008).

Additional heuristics are employed for error removal, usually based on coverage information and topological features of the graph to identify sequencing errors. Errors at the end of reads create short tips that can easily be removed. Internal read errors produce bulges and cloning errors erroneous connections. These three features are consecutively identified and removed. The short read assembly is limited by the intrinsic repeat structure of the genome being sequenced.

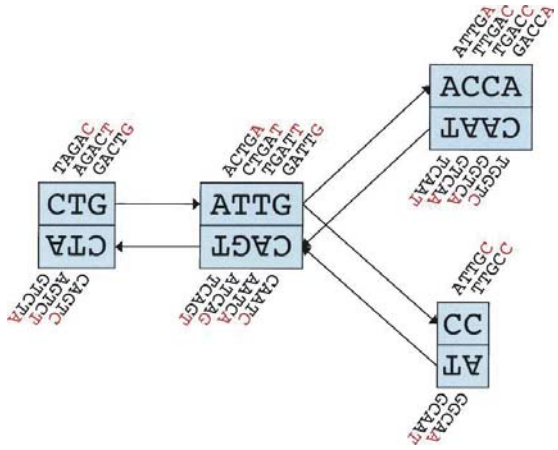


Figure 1-3 Schematic of nodes and arcs of a de Brujin graph in Velvet

Each node, represented by a single rectangle, represents a series of overlapping *k*-mers (in this case, *k* = 5). Each node is paired with its reverse complement twin-node. Arcs are represented as arrows between nodes. The last *k*-mer of an arc's origin overlaps with the first of its destination.

Figure modified from (Zerbino and Birney 2008).

The introduction of high-throughput sequencing technologies thus created novel assembly problems. While the theoretical approach of de Bruijn graph assemblers is

very appropriate for the problem, the actual properties of the data produced by the new technologies were initially unknown and had to be established empirically. The same applies to the effects of contaminations and template mixtures on the assembly process. The various assemblers that were developed for short read data have for this reason been steadily improved since their introduction. There still is much room for improvement as most assemblies still prove too hard to be solved with short reads alone. To date hybrid approaches combining the high sequence output of the short read sequencers with the accuracy, specificity and contiguity of long reads produce the most economical and reliable assemblies (Loman et al., 2012).

1.5 Aims and purposes

Endosymbioses have played an important part in the evolution of life on earth and continue to do so. They allow prokaryotes and eukaryotes to combine their attributes and facilitate adaption to new ecologic niches. An understanding of symbiotic relationships, the role they play, the forces that act upon them and the processes their evolution follows, is indispensable for our understanding of life.

Even though endosymbioses are the focus of an active field of research, many fundamental questions raised by the endosymbiont theory as well as observations on more recent endosymbiotic relationships remain unanswered. Questions concerning the evolutionary pressures that govern the retention and loss of genes in the endosymbiont genome are some of the most salient examples. While it is well established that a strong selection pressure exists to transfer genes to the host's nucleus, it is not at all clear what the nature of this pressure is or what has prevented it from making modern organelle genomes obsolete.

Recent research efforts have seen massive increase in available sequence data for model as well as non-model organisms. This trend has been expedited by the introduction of high-throughput sequencing technologies. The new technologies make an unprecedented wealth of sequence data available to us and open up new possibilities for the study of endosymbiotic relationships. My aim in this study was to utilize the wealth of plastid genome data produced by the scientific community to systematically investigate the processes of gene loss from the genome of an ancient

endosymbiont, and to evaluate the capabilities and of the new Illumina sequencing technology for the investigation of young endosymbioses.

In Chapter 2 I detail my analyses of the gene content of 186 photosynthetic eukaryotes for which fully sequenced plastid genomes are available. My aim was to provide estimates of the number of gene transfers from plastid genome to the nuclear genome that have occurred during the evolutionary history of these taxa.

In Chapter 3 I describe my contributions to the efforts to sequence the nitrogen-fixing endosymbiont of the diatom *Rhopalodia gibba*. It was the aim of this project to establish the suitability of high-throughput sequencing data for the sequencing of a fosmid library and to develop computational methods for the identification and removal of bacterial contaminations from the sequence assemblies.

In Chapter 4 I report on my attempt to sequence the symbiont-host system of another rhopalodian diatom, *Epithemia sorex*, from an environmental sample to enable comparative analyses with the sequences obtained for the *R. gibba* endosymbiont in Chapter 3.

Bibliography

Altmann, R. (1890). Elementarorganismen und ihre Beziehungen zu den Zellen., Leipzig: Veit & Co.

Anderson, S. (1981). Shotgun DNA sequencing using cloned DNase I-generated fragments. Nucleic acids research. **9**:3015–3027.

Andersson, J.O. and Andersson, S.G. (1999). Insights into the evolutionary process of genome degradation. Current opinion in genetics & development. **9**:664–71.

Andersson, S.G., Zomorodipour, A., Andersson, J.O., Sicheritz-Pontén, T., Alsmark, U.C., Podowski, R.M., Näslund, A.K., Eriksson, A.S., Winkler, H.H. and Kurland, C.G. (1998). The genome sequence of Rickettsia prowazekii and the origin of mitochondria. Nature. **396**:133–40.

Armbrust, E.T. and Rynearson, T.A. (2006). Chapter 14 - Genomic insights into diatom evolution and metabolism. In L. A. Katz, ed. *Genomics and Evolution of Microbial Eukaryotes*. Oxford: Oxford University Press.

Armstrong, H. and Brasier, M.D. (2005). Microfossils, Oxford: Blackwell.

Bary, H. de (1878). Vortrag: Über Symbiose. In *51. Versammlung Deutscher Naturforscher und Ärzte in Cassel.* Kassel: Baier & Lewalter.

Bentley, D.R. (2006). Whole-genome re-sequencing. Current opinion in genetics & development. **16**:545–52

Bentley, S.D. and Parkhill, J. (2004). Comparative genomic structure of prokaryotes. Annual review of genetics. **38**:771–92

Bhattacharya, D. and Medlin, L. (1995). the Phylogeny of Plastids: a Review Based on Comparisons of Small-Subunit Ribosomal Rna Coding Regions. Journal of Phycology. **31**:489–498.

Bodył, A., Mackiewicz, P. and Stiller, J.W. (2007). The intracellular cyanobacteria of Paulinella chromatophora: endosymbionts or organelles? Trends in microbiology. **15**:295–6.

Bodył, A. and Moszczyński, K. (2006). Did the peridinin plastid evolve through tertiary endosymbiosis? A hypothesis. European Journal of Phycology. **41**:435–448.

Bowler, C. et al. (2008). The Phaeodactylum genome reveals the evolutionary history of diatom genomes. Nature. **456**:239–244.

Burger, G, Zhu, Y., Littlejohn, T.G., Greenwood, S.J., Schnare, M.N., Lang, B F and Gray, M W (2000). Complete sequence of the mitochondrial genome of Tetrahymena

pyriformis and comparison with Paramecium aurelia mitochondrial DNA. Journal of molecular biology. **297**:365–80.

Carpenter, E.J. and Janson, S. (2000). Intracellular Cyanobacterial Symbionts in the Marine Diatom Climacodium Frauenfeldianum (Bacillariophyceae). Journal of Phycology. **36**:540–544.

Cavalier-Smith, T. (2006). Cell evolution and Earth history: stasis and revolution. Philosophical transactions of the Royal Society of London. Series B, Biological sciences. **361**:969–1006.

Cavalier-Smith, T. and Lee, J.J. (1985). Protozoa as Hosts for Endosymbioses and the Conversion of Symbionts into Organelles , 2. The Journal of Eukaryotic Microbiology. **32**:376–379.

Colón-López, M.S., Sherman, D.M. and Sherman, L.A. (1997). Transcriptional and translational regulation of nitrogenase in light-dark- and continuous-light-grown cultures of the unicellular cyanobacterium Cyanothece sp. strain ATCC 51142. Journal of bacteriology. **179**:4319–27.

Dagan, T., Blekhman, R. and Graur, D. (2006). The "domino theory" of gene death: gradual and mass gene extinction events in three lineages of obligate symbiotic bacterial pathogens. Molecular biology and evolution. **23**:310–6.

Darwin, Charles (1859), On the origin of species by means of natural selection. London: Murray.

Delmotte, F., Rispe, C., Schaber, J., Silva, F.J. and Moya, A. (2006). Tempo and mode of early gene loss in endosymbiotic bacteria from insects. BMC evolutionary biology. **6**:56.

Delwiche, C.F. and Palmer, J.D. (1996). Rampant horizontal transfer and duplication of Rubisco genes in eubacteria and plastids. Molecular biology and evolution. **31**:873–882.

Droop, S.J.M., Sims, P.A., Mann, D.G. and Pankhurst, R.J. (1993). A taxonomic database and linked iconograph for diatoms. Hydrobiologia. **269-270**:503–508.

Dyall, S.D., Brown, M.T. and Johnson, P.J. (2004). Ancient invasions: from endosymbionts to organelles. Science (New York, N.Y.). **304**:253–7.

Embley, T.M., van der Giezen, M., Horner, D.S., Dyal, P.L., Bell, S. and Foster, P.G. (2003). Hydrogenosomes, mitochondria and early eukaryotic evolution. IUBMB life. **55**:387–95.

Falkowski, P.G., Katz, M.E., Knoll, A.H., Quigg, A., Raven, J.A., Schofield, O. and Taylor, F.J.R. (2004). The evolution of modern eukaryotic phytoplankton. Science (New York, N.Y.). **305**:354–60.

Gibbs, S.P. (1979). The route of entry of cytoplasmically synthesized proteins into chloroplasts of algae possessing chloroplasts. J Cell Sci. **35**:253–266.

Glass, J.I., Lefkowitz, E.J., Glass, J.S., Heiner, C.R., Chen, E.Y. and Cassell, G.H. (2000). The complete sequence of the mucosal pathogen Ureaplasma urealyticum. **407**:757–762.

Gnerre, S. et al. (2011). High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proceedings of the National Academy of Sciences of the United States of America. **108**:1513–8.

Gray, M.W., Burger, G. and Lang, B.F. (1999). Mitochondrial Evolution. Annual review of genetics. **283**:1476–1481.

Gray, M.W. and Doolitle, W.F. (1982). Has the Endosymbiont Hypothesis Been Proven? Microbiological reviews. **46**:1–42.

Gray, M.W., Lang, B.F. and Burger, G. (2004). Mitochondria of protists. Annual review of genetics. **38**:477–524.

Haselkorn, R. and Buikema, W.J. (1992). Nitrogen fixation in cyanobacteria. In *Biological Nitrogen Fixation*. New York: Chapman & Hall, Inc., pp. 166–190.

Howe, C.J., Barbrook, A.C. and Lockhart, P.J. (2000). Organelle genes – do they jump or are they pushed ? Science. **16**:3–4.

Huang, C.Y., Ayliffe, M.A. and Timmis, J.N. (2003). Direct measurement of the transfer rate of chloroplast DNA into the nucleus. Nature. **422**:72–6.

Janson, W.B.C. (1999). Host specificity in the Richelia diatom symbiosis revealed by hetR gene sequence analysis. Environmental Microbiology. **1**:431–438.

Karl, D., Michaels, A., Bergman, B., Capone, D., Carpenter, E., Letelier, R., Paerl, H., Sigman, D. and Stal, L. (2002). Dinitrogen fixation in the world ' s oceans. Biogeochemestry. **57**:47–98.

Karlberg, O., Canbäck, B., Kurland, C.G. and Andersson, S.G. (2000). The dual origin of the yeast mitochondrial proteome. Yeast (Chichester, England). **17**:170–87.

Keeling, P.J. (2004). Diversity and Evolutionary History of Plastids and their Hosts. American journal of botany. **91**:1481–1493.

Klipp, W. (2004). Genetics and Regulation of Nitrogen Fixation in Free-Living Bacteria, Springer.

Kneip, C., Lockhart, P.J., Voss, C. and Maier, U.G. (2007). Nitrogen fixation in eukaryotes--new models for symbiosis. BMC evolutionary biology. **7**:55.

Kneip, C., Voss, C., Lockhart, P.J. and Maier, U.G. (2008). The cyanobacterial endosymbiont of the unicellular algae Rhopalodia gibba shows reductive genome evolution. BMC evolutionary biology. **8**:30.

Koskiniemi, S., Sun, S., Berg, O.G. and Andersson, D.I. (2012). Selection-driven gene loss in bacteria. PLoS genetics. **8(6)**:e1002787.

Krylov, D.M., Wolf, Y.I., Rogozin, I.B. and Koonin, E. V (2003). Evolution Gene Loss , Protein Sequence Divergence , Gene Dispensability , Expression Level , and Interactivity Are Correlated in Eukaryotic Evolution. Genome Research. **13**:2229–2235.

Larkum, A.W.D., Lockhart, P.J. and Howe, C.J. (2007). Shopping for plastids. Trends in plant science. **12**:189–95.

Lauterborn, R. (1895). Paulinella chromatophora. In *Protozoenstudien*. pp. 59537–59544.

Lawrence, J.G., Hendrix, R.W. and Casjens, S. (2001). Where are the pseudogenes in bacterial genomes? Trends in microbiology. **9**:535–40.

Liberton, M., Austin, J., Berg, R. and Pakrasi, H. (2009). Three-Dimensional Arrangement of Thylakoid Membranes in Cyanothece sp. ATCC 51142, a Unicellular Cyanobacterium. Microscopy and Microanalysis. **15**:876.

Loman, N.J., Constantinidou, C., Chan, J.Z.M., Halachev, M., Sergeant, M., Penn, C.W., Robinson, E.R. and Pallen, M.J. (2012). High-throughput bacterial genome sequencing: an embarrassment of choice, a world of opportunity. Nature reviews. Microbiology. **10**:599–606.

Lucattini, R., Likic, V.A. and Lithgow, T. (2004). Bacterial proteins predisposed for targeting to mitochondria. Molecular biology and evolution. **21**:652–8.

Macasev, D., Newbigin, E., Whelan, J. and Lithgow, T. (2000). Proteins ? Components of the Import Apparatus Tom20 and Tom22 from Arabidopsis Differ from Their Fungal Counterparts 1. Society. **123**:811–816.

Margulis, L. (1970). Origin of eukaryotic cells., New Haven: Yale University Press.

Margulis, L. and Fester, R. (1991). Symbiosis as a Source of Evolutionary Innovation. Sepeciation and morphogenesis., Bellagio: Massachusetts Institute of Technology Press.

Martin, W. and Müller, M. (1998). The hydrogen hypothesis for the first eukaryote. Nature. **392**:37–41.

Martin, W., Rujan, T., Richly, E., Hansen, A., Cornelsen, S., Lins, T., Leister, D., Stoebe, B., Hasegawa, M. and Penny, D. (2002). Evolutionary analysis of Arabidopsis, cyanobacterial, and chloroplast genomes reveals plastid phylogeny and thousands of cyanobacterial genes in the nucleus. Proceedings of the National Academy of Sciences of the United States of America. **99**:12246–51.

Martin, W. and Russell, M.J. (2003). On the origins of cells: a hypothesis for the evolutionary transitions from abiotic geochemistry to chemoautotrophic prokaryotes, and from prokaryotes to nucleated cells. Philosophical transactions of the Royal Society of London. Series B, Biological sciences. **358**:59–83; discussion 83–5.

Mcfadden, G.I. (1999). Plastids and Protein Targeting. **46**:339–346.

Medlin, L.K., Kooistra, W.H.C.F., Potter, D., Saunders, G.W. and Andersen, R.A. (1997). Phylogenetic relationships of the "golden algae" (haptophytes, heterokont chromophytes) and their plastids. Plant Systematics and Evolution. **11**:187–219.

Mereschkowski, C. (1905). Über Natur und Ursprung der Chromatophoren im Pflanzenreiche. Biologisches Centralblatt. **25**:593–604.

Millen, R.S. et al. (2001). Many parallel losses of infA from chloroplast DNA during angiosperm evolution with multiple independent transfers to the nucleus. The Plant cell. **13**:645–58.

Moran, N.A. (2003). Tracing the evolution of gene loss in obligate bacterial symbionts. Current Opinion in Microbiology. **6**:512–518.

Moran, N.A. and Mira, A. (2001). The process of genome shrinkage in the obligate symbiont Buchnera aphidicola. Genome biology. **2**:RESEARCH0054.

Moran, N.A. and Wernegreen, J. (2000). Lifestyle evolution in symbiotic bacteria: insights from genomics. Trends in ecology & evolution (Personal edition). **15**:321–326.

Moran, N.A. (2002). Microbial minimalism: genome reduction in bacterial pathogens. Cell. **108**:583–6.

Moran, N.A. (2006). Symbiosis. Current Biology. **16**:R866–R871.

Moshkovitz, S. and Ehrlich, A. (1983). Siliceous microfossils in the Upper Cretaceous Mishash Formation, Central Negev, Israel. Cretaceous Research. **4**:173–194.

Muller, H.J. (1964). The Relation of Recombination to Mutational Advance. Mutation Research. **204**:732–732.

Murry, M. and Wolk, C.P. (1989). Evidence that the barrier to the penetration of oxygen into heterocysts depends upon two layers of the cell envelope. Archives of microbiology. **151**:469–474.

Nagarajan, N. and Pop, M. (2009). Parametric complexity of sequence assembly: theory and applications to next generation sequencing. Journal of computational biology : a journal of computational molecular cell biology. **16**:897–908.

Nilsson, A.I., Koskiniemi, S., Eriksson, S., Kugelberg, E., Hinton, J.C.D. and Andersson, D.I. (2005). Bacterial genome size reduction by experimental evolution. Proceedings of the National Academy of Sciences of the United States of America. **102**:12112–6.

Nowack, E.C.M. and Grossman, A.R. (2012). Trafficking of protein into the recently established photosynthetic organelles of Paulinella chromatophora. PNAS. **109**:5340–5345.

Ochman, H. (2005). Genomes on the shrink. Proceedings of the National Academy of Sciences of the United States of America. **102**:11959–60.

Olson, J.M. (2006). Photosynthesis in the Archean era. Photosynthesis research. **88**:109–117.

Palenik, B. (2002). The genomics of symbiosis: hosts keep the baby and the bath water. Proceedings of the National Academy of Sciences of the United States of America. **99**:11996–7.

Parkinson, J. and Gordon, R. (1999). Beyond micromachining: the potential of diatoms. Trends in biotechnology. **17**:190–6.

Patron, N.J., Waller, R.F. and Keeling, P.J. (2006). A tertiary plastid uses genes from two endosymbionts. Journal of molecular biology. **357**:1373–82.

Pevzner, P.A., Tang, H. and Waterman, M.S. (2001). An Eulerian path approach to DNA fragment assembly. Proceedings of the National Academy of Sciences of the United States of America. **98**:9748–53.

Pisciotta, J.M., Zou, Y. and Baskakov, I.V. (2010). Light-dependent electrogenic activity of cyanobacteria. PloS one. **5**:e10821.

Pizzo, P. and Pozzan, T. (2007). Mitochondria-endoplasmic reticulum choreography: structure and signaling dynamics. Trends in cell biology. **17**:511–7.

Prechtl, J. and Maier, U.G. (2001). Zoology meets Botany: establishing intracellular organelles by endosymbiosis. Zoology (Jena, Germany). **104**:284–9.

Prechtl, J., Kneip, C., Lockhart, P.J., Wenderoth, K. and Maier, U.G. (2004). Intracellular spheroid bodies of Rhopalodia gibba have nitrogen-fixing apparatus of cyanobacterial origin. Molecular biology and evolution. **21**:1477–81.

Rai, A.N., Söderbäck, E. and Bergman, B. (2000). Cyanobacterium-Plant Symbioses. New Phytologist. **147**:449–481.

Raymond, J., Siefert, J.L., Staples, C.R. and Blankenship, R.E. (2004). The natural history of nitrogen fixation. Molecular biology and evolution. **21**:541–54.

Reumann, S., Inoue, K. and Keegstra, K. (2005). Evolution of the general protein import pathway of plastids (Review). Molecular Membrane Biology. **22**:73–86.

Rocha, E. (2002). Is there a role for replication fork asymmetry in the distribution of genes in bacterial genomes? Trends in microbiology. **10**:393–5.

Rossier, M.F. (2006). T channels and steroid biosynthesis: in search of a link with mitochondria. Cell calcium. **40**:155–64.

Rothpletz, A. (1896). Über die Flysch-Fucoiden und einige andere fossile Algen, sowie über liassische diatomeenführende Hornschwämme. Zeitschrift der Deutschen Geologischen Gesellschaft. **48**:854–914.

Round, F.E., Crawford, R.M. and Mann, D.G. (1990). Diatoms: biology and morphology of the genera., Cambridge University Press.

Sachs, J. (1882). Vorlesungen über Pflanzen-Physiologie., Leipzig: Verlag Wilhelm Engelmann.

Salzberg, S.L., Phillippy, A.M., Zimin, A., Earl, D., Bradnam, K. and John, J.S. (2012). GAGE : A critical evaluation of genome assemblies and assembly algorithms methods. Genome res. **22**:557–567.

Sanger, F., Nicklen, S. and Coulson, A.R (1977). DNA sequencing with chain-terminating inhibitors. Proceedings of the National Academy of Sciences. **74**:5463-5467.

Sato, N., Terasawa, K., Miyajima, K. and Kabeya, Y. (2003). Organization, Developmental Dynamics, and Evolution of Plastid Nucleoids. International Review of Cytology. **232**:217–262.

Schimper, A. (1883). Über die Entwicklung der Chlorophyllkörner und Farbkörper. Botanische Zeitung. **41**:105–114.

Schneider, A. and Ebert, D. (2004). Covariation of mitochondrial genome size with gene lengths: evidence for gene length reduction during mitochondrial evolution. Journal of molecular evolution. **59**:90–6.

Sheppard, A.E., Ayliffe, M.A., Blatch, L., Day, A., Delaney, S.K., Khairul-Fahmy, N., Li, Y., Madesis, P., Pryor, A.J. and Timmis, J.N. (2008). Transfer of plastid DNA to the nucleus is elevated during male gametogenesis in tobacco. Journal of Plant Physiology. **148**:328-336.

Shigenobu, S., Watanabe, H., Hattori, M., Sakaki, Y. and Ishikawa, H. (2000). Genome sequence of the endocellular bacterial symbiont of aphids Buchnera sp. APS. Nature. **407**:81–6.

Silva, F.J., Latorre, A. and Moya, A. (2001). Genome size reduction through multiple events of gene disintegration in Buchnera APS. Trends in genetics : TIG. **17**:615–8.

Silva-Filho, M.C. (2003). One ticket for multiple destinations: dual targeting of proteins to distinct subcellular locations. Current Opinion in Plant Biology. **6**:589–595.

Stover, C.K. et al. (2000). Complete genome sequence of Pseudomonas aeruginosa PAO1, an opportunistic pathogen. Nature. **406**:959–64.

Theissen, U. and Martin, W. (2006). Correspondences The difference between organelles and Response to Theissen and Martin. Current Biology. **16**:1016–1017.

Thorsness, P.E. and Fox, T.D. (1990). Escape of DNA from mitochondria to the nucleus in Saccharomyces cerevisiae. Nature. **346**:376–379.

Timmis, J.N., Ayliffe, M.A., Huang, C.Y. and Martin, W. (2004). Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes. Nature reviews. Genetics. **5**:123–35.

Waller, R.F. and McFadden, G.I. (2005). The apicoplast: a review of the derived plastid of apicomplexan parasites. Current issues in molecular biology. **7**:57–79.

Wallin, I. (1923). The Mitochondria Problem. The American Naturalist. **57**:255–261.

Williams, D.M. and Kociolek, J.P. (2007). Pursuit of a natural classification of diatoms: History, monophyly and the rejection of paraphyletic taxa. European Journal of Phycology. **42**:313–319.

Wolk, C.P. and Ernst, A. (1994). Heterocyst metabolism and development. In D. Bryant, ed. *The molecular biology of cyanobacteria.* Boston: Kluwer Academic Publishers, pp. 769–823.

Zhang, X.-P. and Glaser, E. (2002). Interaction of plant mitochondrial and chloroplast signal peptides with the Hsp70 molecular chaperone. Trends in plant science. **7**:14–21.

2 The frequency of gene loss events during plastid evolution

2.1 Abstract

The generally accepted view of plastid evolution is that approximately 1.5 billion years ago oxygenic photosynthesis was made available to eukaryotes via an endosymbiotic relationship between a heterotrophic eukaryote and a photosynthetic cyanobacterium. Subsequently host and symbiont have merged into one organism and the symbiont has been transformed into a cellular organelle, the plastid. Since then plastid-carrying eukaryotes have diverged into a multitude of lineages, each of which independently carries on the process of plastid/host integration. As part of this organelle genes have been transferred into the nucleus and lost from the plastid genome. I have assembled a dataset comprising the protein coding genes from 198 fully sequenced plastid genomes, sampled across the different lineages of plastid-carrying eukaryotes. A subset of these protein sequences was used to infer gene phylogenies using a Bayesian phylogenetic inference method. To reduce the impact of non-homogeneity of evolutionary process on the reconstruction of relationships within lineages, the gene phylogenies were reconstructed independently for the different major lineages and all resulting gene trees were then assembled into a supertree. The presence or absence of each gene across taxa was then mapped to this phylogeny and the minimum number of independent gene transfers to the nucleus necessary to produce a given pattern of gene presence was calculated. The results show that the pattern of presence/absence of the average plastid gene was shaped by at least 5 independent gene transfers to the nucleus. By using Maximum Likelihood methods it can be calculated that the number of gene transfers that can be expected to have produced these patterns is actually about 14% higher than this estimate. The results of this work demonstrate clearly that the transfer of plastid genes to the nucleus is not a rare or uncommon event, as has been suggested, but a common process that has shaped the plastid genomes of every lineage presumably according to the necessities of its lifestyle.

2.2 Introduction

Photoautotrophy involves using energy from the sun to convert CO_2 molecules into sugars. It has been the foundation of the earth's ecosystem since at least 3.5 billion years ago (bya) (Schopf, 2011). The first autotrophs were thought to carry out anoxygenic phostosynthesis meaning that electron transfer does not involve the splitting of oxygen. Extant bacterial species with this system utilize bacteriochlorophyll to capture light energy (Burke et al., 1993). Oxygenic photosynthesisers, in contrast, use chlorophyll to capture light energy, capture electrons from water and release oxygen in a process that is thought to have led to the oxygenation of the earth's atmosphere. The amazing ability to harness the sun's energy was one of the most important biological innovations and changed the planet forever. Beginning with the great oxidation event about 2.45 bya it not only sustained most of the life on earth but also provided biologically available oxygen required for aerobic respiration and ultimately prepared the way for multi-cellular organisms (Schopf, 1999).

To this day photosynthesis remains an exclusively bacterial domain. Eukaryotes were only able to utilize it by forming an endosymbiotic relationship with one of the bacterial photosynthesisers about 1.56 bya (Yoon et al., 2004). Whether the process of plastid endosymbiosis occurred in a singular event, multiple events or as the gradual adaptation of two species to each other is not clear but, once established, this partnership gave rise to most extant photosynthetic eukaryotes (the only generally recognized exception being a much younger symbiosis between a cyanobacterium and some Rhizaria in the genus *Paulinella*) (Larkum et al., 2007; Rodríguez-Ezpeleta et al., 2005; Bhattacharya et al., 2007; Martin and Schnarrenberger, 1997). Involving a process operating over a billion years of co-evolution, the endosymbiont has since been transformed into the plastid: a cellular organelle that is the site of photosynthesis reactions and the synthesis of carbohydrates in all photosynthetic eukaryotes (Margulis and Fester, 1991). Due to chance and necessity the integration of organelle and host was shaped in each lineage independently and in its own unique way. The similarity in gene content among contemporary plastid genomes is likely the result of extensive convergent patterns of gene loss (Timmis et al., 2004). We nonetheless see lineage specific patterns in gene content of the plastid genome as well as the nuclear genome (Dagan et al., 2006).

Plastid-carrying eukaryotes have diverged into three major lineages since the establishment of the putative primary endosymbiotic event: the Glaucophyta, Rhodophyta and Viridiplantae. The Glaucophyta are a small group of unicellular freshwater algae comprising only 13 known species. Their plastids have a peptidoglycan layer similar to the cell wall of some cyanobacteria and contain the photosynthetic pigment chlorophyll a as well as phycobiliproteins (Pfanzagl et al., 1996). The Rhodophyta or red algae are a group of mostly marine algae with ~6000 species, most of them multicellular (http://www.algaebase.org). The group is distinguished among other characteristics by the lack of flagella and centrioles, the use of chlorophyll a and phycobiliproteins as pigments for photosynthesis (giving them their red colour) and the morphology of their plastids, which lack an external endoplasmic reticulum and contain unstacked thylakoids (Graham and Wilcox, 2000; Gabrielson et al., 1990). The Viridiplantae comprise the embryophytes (higher plants) and the paraphyletic green algae. The green algae are made up of ~8000 species of unicellular, colony forming and multicellular species, most of which occur in freshwater. This lineage has lost phycobiliproteins as accessory pigments but uses chlorophyll a and b, giving the chloroplasts a bright green colour. Green algae thylakoids show a stacked morphology and some green algae have flagella (Lewis and McCourt, 2004). Together these three lineages constitute the Archaeplastida, named to recognize that they carry primary plastids.

To add to the diversity of photosynthetic organisms, some protist lineages have laterally acquired plastids by engulfing a plastid carrying eukaryote.[1] This process is called secondary endosymbiosis if the endosymbiont belongs to the Archaeplastida. If such a secondary algae in turn enters into an endosymbiosis with another eukaryote then the resulting organism is referred to as a tertiary alga. Serial endosymbioses have been described as well, where an existing plastid is replaced by a new endosymbiosis with another eukaryotic alga (Dorrell and Howe, 2011). This horizontal transfer of plastids as well as the nuclear genes necessary for maintaining the plastid among eukaryotes complicates the task to reconstruct the phylogeny of photosynthetic

[1] The terms Plantae and Archaeplastida are often used interchangeably, but Plantae is also sometimes used as synonym for Viridiplantae. I am using the term Archaeplastida to refer to eukaryotes with primary plastids and the term Plantae to refer to all plastid-carrying eukaryotes, including secondary and tertiary algae.

eukaryotes, and it remains unclear even how many of these secondary and tertiary endosymbioses have occurred. It also makes nuclear genes unreliable as a basis for the reconstruction of the phylogeny of plastids.

The clear division of the Archaeplastida into three major lineages is well established but the internal branching order of these lineages remains an issue of debate. Different phylogenetic studies have given support for the basal position of each of the three lineages. The Glaucophyta have traditionally been thought to be ancestral based on the retention of cyanobacterial characteristics in their plastids and analyses of plastid genes (Helmchen et al., 1995; Martin et al., 1998). Which lineage is supported as the most basal by plastid sequences though has been found to strongly depend on the selection of sequences used and highly conserved plastid genes lend support to a basal position of the Viridiplantae instead. Phylogenetic analyses of nuclear genes on the other hand have suggested a basal position of the Rhodophyta (Rodríguez-Ezpeleta et al., 2005; Hackett et al., 2007; Burki et al., 2008). The support for Rhodophyta as ancestral lineage has then again been found to depend on the choice of outgroup (Deschamps and Moreira, 2008). Nuclear genes of cyanobacterial origin strongly support Glaucophyta as basal as do eukaryotic nuclear markers, but a larger dataset of nuclear genes originating from endosymbiotic gene transfer favour the Viridiplantae (Reyes-Prieto and Bhattacharya, 2007; Deschamps and Moreira, 2008). Thus, the only thing that can be concluded with certainty is that genes of different evolutionary origin carry conflicting phylogenetic signals. The position of the root remains elusive despite the availability of cyanobacterial species as potential outgroups.

It has been estimated that about 3000 genes were present in the genome of the original endosymbiont but today's plastids retain only strongly reduced genomes with less than 200 genes (Prechtl and Maier 2001, Delwiche and Palmer 1997). The study of relatively recent endosymbiotic relationships allows us to reason that most gene losses would have occurred early in the process of endosymbiosis, most likely before the divergence of the major lineages (see Section 1.1.2). Vertically transmitted endosymbionts usually show a high rate of gene loss and reductive genome evolution, most likely due to Muller's ratchet (see Section 1.1.2 for a detailed description of Muller's ratchet) (Moran and Mira, 2001; Delmotte et al., 2006). In the early stages of the endosymbiotic relationship the precursor of the plastid would have quickly lost all genes that were not essential for the symbiosis. These might have been lost entirely or through

endosymbiotic gene transfer (EGT), transported and integrated into the nuclear genome. In the latter case, it has been suggested that after acquiring a functional promoter, many of these genes were adopted for functions elsewhere in the host (Martin et al., 2002).

Ultimately, complete merger of the two organisms transformed the endosymbiont into an organelle. The defining feature of which is a protein import system that allows nuclear encoded proteins to be imported into, and to fulfil functions within, the organelle. Once this was established it would have facilitated even more genes being lost from the endosymbiont genome provided that they were transferred to the host genome or could be functionally replaced by host proteins.

Modern plastids still require between 2100 and 4800 different proteins to function and the vast majority of these organelle proteins are now encoded in the host's nucleus (Richly and Leister, 2004). This constitutes an extraordinary level of integration of the symbiont into the host organism that has allowed for increasing host control over the endosymbiont genome. There are other possible reasons to relocate genes to the nucleus. Muller's ratchet continues to act even after all non-essential genes have been lost. Plastid genomes show an elevated AT content, probably due to impaired DNA repair mechanisms. This might impose an amino acid composition bias on plastid-encoded genes that is detrimental to their function (Howe et al., 2000). If so, then increasing AT levels might lead to a form of selection that effectively "pushes" genes to the nucleus (Howe et al., 2000).

Experiments with selectable marker genes in mitochondria and chloroplasts have shown that gene transfers to the nucleus occur at comparatively high frequencies (Thorsness and Fox, 1990; Huang et al., 2003). These findings suggest that gene transfer to the nucleus is potentially an ongoing process and not just one that has occurred deep in evolutionary time.

Selection pressure to retain some genes in the organelle genome seems to exist as well. Between 20 and 200 genes are still encoded in modern plastids. Their presence requires the maintenance of the biochemical machinery necessary to replicate and transcribe the plastid genome and carry out protein synthesis within the organelle.

Why these genes are retained in the plastid is not clear and several possible explanations have been proposed.

i) One suggestion is that some proteins for some reason cannot be imported into the plastid. Hydrophobicity, an alpha-helix rich secondary structure, three dimensional structures and interactions with cofactors have all been proposed as possible reasons, however evidence is not strong (Allen, 2003).

ii) Most plastid resident genes encode structural proteins involved in photosynthesis and the ribosomal machinery (Prechtl and Maier, 2001). One suggestion for their retention is that it might be easier to assemble ribosomes at the place of function, because ribosomal subunits are assembled in *statu nascendi* explaining why every plastid contains its own set of rRNA genes (Prechtl and Maier, 2001). In the mitochondria of many protists on the other hand all ribosomal proteins are encoded in the cell nucleus (Gray et al., 2001). Thus the retention of rRNA genes in the organelle is either not mandatory or the selection pressure is different for mitochondria and plastids.

iii) Genes that code for proteins in the core of the photosynthesis machinery must be tightly regulated to maintain the redox balance within the plastid and prevent the formation of reactive oxygen species. This might require the assembly of the major protein complexes to be immediately regulated according to the redox status of the plastid. This might be achieved by retaining at least one subunit of each protein complex in the plastid genome so that it can act as the limiting factor for the assembly of the whole protein complex. This hypothesis was proposed by Allen (1992) as co-location for redox regulation (CORR). In the plastid, gene expression is indeed redox-modulated by the plastoquinone pool (Race et al., 1999; Allen, 2003).

An alternative perspective is that there is in fact no reason for retention of the plastid genome. However, to replace the plastid-encoded copy, the new nuclear gene needs to acquire an appropriate promoter and plastid import sequence. If this is a rare event, it may simply be that the jump has not occurred yet for all genes, but will eventually do so, given time (Palmer, 1997; Herrmann, 1997).

Gene loss as a phylogenetic marker?

Plastids, like endosymbionts, appear to be isolated by their hosts from lateral gene transfer with the plastids of other species or bacteria. If this interpretation is correct, then the patterns of gene presence and absence observed in plastid genomes today must therefore be the result of gene loss only. Further, if it is assumed that successful gene transfers to the nucleus are rare events, then the presence or absence of plastid genes might be used to test alternative phylogenetic hypotheses (for example Nozaki et al., 2003; Ong et al., 2010). However, if convergent patterns of gene loss are common, then this is not possible. Martin et al. examined this issue in 2002 analysing all 16 completely sequenced plastid genomes available at the time. This original study suggested extensive gene loss during plastid evolution, suggesting that parallel losses outnumbered unique losses more than ten to one. The authors concluded from this that patterns of gene presence or absence in plastid genome are abundant with homoplasies and that Dollo parsimony analyses based on the presence or absence of plastid genes were therefore very likely to give incorrect results concerning phylogenetic relationships.

Since then the number of plastid genomes sequenced each year has steadily increased, spurred by advances in sequencing technology (Gao et al., 2010). By early 2011 more than 200 plastid genomes had been fully sequenced and the number is likely to reach 300 by the end of 2012. Given the vastly increased amount of data we can now hope to refine much of what we know about the history of gene loss during plastid evolution. In this project I provide estimates of the extent of transfer of genes to the nucleus in all major lineages of algae and plants and estimate the prevalence of gene transfers that have shaped modern plastid genomes.

2.3 Methods

2.3.1 Gathering and processing of data

Sequences and annotations were sourced from 'The Reference Sequence Collection' (RefSeq) provided by 'The National Center for Biotechnology Information' (NCBI).
The files plastid.1.protein.gpff and plastid.1.protein.faa as provided within RefSeq Release 45 (14/01/2011) were downloaded from the NCBI ftp server (ftp://ftp.ncbi.nlm.nih.gov/refseq/). These files contain every protein as annotated in every fully sequenced plastid genome available at the time in GenBank flat file format and fasta format respectively. This release comprises 17881 protein accessions with a total length of 4856424 amino acids and sequences form 201 different taxa. Of these accessions 15228 (85%) were provisional record and not curated.

A BioPerl script (MatrixMaker.pl, see Appendix A) was used to process the protein collection. The proteins provided in plastid.1.protein.gpff were sorted into groups of orthologs according to the four-letter gene name in the FEATURES section of the flat file. Each group of orthologous sequences was then written out to a text file in fasta format, ready for subsequent alignment. The script also produced a tab-delimited matrix with a column for each taxon and a row for each gene name encountered in the input. The accession number of a protein was used as entry in the matrix to mark the presence of the protein-coding gene in the plastid genome of the respective taxon.

Whenever two or more copies of a gene were present in a plastid genome only one of them was retained in the dataset. In case of identical copies one was discarded at random. In cases where the copies mainly differed in length, the shorter one was presumed truncated and discarded. If two apparently functional but non-identical copies were present, gene trees were estimated in PhyML with both copies and a selection of orthologous sequences (Criscuolo, 2011). The copy producing the shorter branch was kept in the dataset and the other copy/copies discarded. In no case did the tree topology differ between copies.

Hypothetical open reading frames whose products show no similarity to known proteins were at this stage excluded from the dataset. The columns of the data matrix were manually sorted into taxonomic groups and the groups labelled in accordance with the NCBI taxonomy database (Sayers et al., 2009).

The protein sequences derived from the annotated genome of the plastid of *Colocasia esculenta* were manually added to the dataset (Ahmed et al., 2013).

2.3.2 Data verification

Several rounds of data verification were carried out to ensure the completeness and correctness of the dataset. A visual inspection of the data matrix was used to detect entries that were missing or misplaced due to typos and formatting errors in the source files.

Genes are sometimes known under several aliases and GenBank records are not updated if the gene name is changed or assigned for the first time to a protein with previously unknown function or expression status. For this reason a local BLASTp search was performed to identify redundant gene names. A random sequence out of every set of orthologs was searched against the complete set of plastid proteins as provided by RefSeq. The results were then compared with the data matrix in search for proteins that showed high similarity to all entries in two or more rows of the matrix. Each of these cases was assessed individually based on the blast scores and e-values, length and similarity among the known orthologs of the protein. Similar proteins were assumed to be orthologous rather than paralogous. This conclusion is based on the assumption that the strong reductive pressure acting on plastid genomes makes gene duplication events, followed by divergent evolution of the two copies highly unlikely. The classification of gene names as aliases was in most cases backed up by information available in the NCBI Gene database. A similar BLASTp search was conducted with all ORF sequences to identify homologues to known plastid genes among them.

The nucleotide sequences of fully sequenced plastid genomes as provided by NCBI for download from its ftp server in RefSeq Release 49 (09/09/11) were used as subjects of local tBLASTn searches to verify annotations and identify unannotated genes based on their conserved coding sequences. For the purpose of this study the gene was assumed

to be present in a genome if the conserved sequence coincided with an open reading frame.

In cases where the BLAST results did not agree with annotations or suggested a missing annotation, the genomic region in question was translated with EXPASY-Translate (http://web.expasy.org/translate/) to identify start and stop codons. The presence of a conserved region within an open reading frame was then confirmed via online BLASTn and BLASTp searches on the NCBI BLAST server.

Local BLAST searches were performed on the Massey AWC servers using the BLAST+ command line applications (Altschul et al., 1990). BLAST reports were visualised with the Epos BLASTviewer (Griebel et al., 2008).

Following these computational steps used to verify data quality, the gene presence/absence matrix and the fasta files containing the ortholog protein sequences were updated manually.

The conifer species *Picea sitchens, Pinus geradia, Pinus krempfii* and *Pinus contorta* were excluded from the dataset because their plastid genome sequences were found to be incomplete.

2.3.3 Protein alignments

Orthologous protein sequences were aligned using the multiple sequence alignment (MSA) algorithm implemented in Dialign-TX (Subramanian et al., 2008). Dialign was chosen because unlike most MSA algorithms it does not make the implicit assumption of global homology, but produces multiple local alignments by identifying local pairwise similarities, or fragment alignments. This strategy does not penalize long indels and is therefore especially suitable for long sequences with low overall identity as can be expected to be the case for plastid proteins that are anciently diverged (Subramanian et al., 2008; Do and Katoh, 2008).

Alignments were visually inspected and edited in Seaview to identify and correct misalignments and subsequently trimmed using Gblocks_0.91b to remove poorly

aligned region of the alignment (Gouy et al., 2010; Castresana, 2000). Several sets of settings were tested and the default settings were chosen as the most suitable.

2.3.4 Phylogenetic Analyses

All non-photosynthetic taxa were excluded to simplify the analyses because they cannot be used in the subsequent gene loss analyses. In photosynthetic species a gene loss from the plastid genome can be assumed to equate to a gene transfer to the nucleus with subsequent re-import of the protein into the plastid. However, if photosynthetic function is lost the genes involved are usually lost from the organism as a whole. The latter is a fundamentally different process of gene loss and cannot be compared to the former (see Section 2.5.4).

2.3.4.1 Gene trees

Single-gene trees were reconstructed with PhyloBayes and PhyML for orthologous proteins that produced alignments with a length of more then 200 amino acids after trimming with Gblocks. Shorter alignments were not used to avoid problems with model fitting. Proteins that are known to give misleading signals due to shifts in site-specific evolutionary rates (heterotachy), like ribosomal proteins and RNA polymerase subunits, were not used (Wu et al., 2011; Lockhart et al., 2006).

In cases where a protein was present in several taxa of the green as well as the red algae, the dataset was split with a perl script and the phylogenies for both lineages were calculated separately to avoid systematic error due to different modes of sequence evolution and heterotachy in these two major lineages. The Glaucophyte *Cyanophora paradoxa*, if present, was included in both subsets to serve as an outgroup.

Phylogenetic relationships were reconstructed using a heuristic Maximum Likelihood method implemented in the PhyML software package and a Bayesian Inference Method implemented in PhyloBayes. All but two Angiosperm taxa (*Nymphea alba* and *Piper cenocladum*) were excluded from the phylogenetic analysis to reduce the complexity of the analyses.

For all PhyML analyses the following settings other than default settings were used: The substitution model cpREV was chosen as it is an empirical substitution model for plastid proteins after it was identified by Modeltest as the best fitting model for the selection of alignments that were tested (Posada, 2008).

The proportion of invariable sites as well as the gamma distribution parameter were estimated from the data. The support for internal branches was calculated with a non-parametric bootstrap analysis (100 replicates). The following parameters were used for all PhyloBayes analyses additional to default settings: The substitution rates across sites were modelled as a discrete gamma distribution with 4 categories (-dgam 4), while the prior on branch length was set to be the product of independent and identically distributed gamma distributions (-lgam). The non-parametric CAT model was chosen for the profile mixture in combination with the general time reversible matrix for the exchange rates (-cat –gtr).

For each alignment set two chains were run in parallel. The chains were allowed to run until convergence was reached. If it was noticed that the chains for a dataset ceased to move towards convergence, more chains were started for the respective protein alignment and different combinations of chains tested for convergence. Those chains that were found to not converge with the others were presumed trapped in a local maximum and were excluded from the analyses. The chains were tested for convergence following the recommendations specified in the PhyloBayes 3.3 manual.

2.3.4.2 Super tree assembly

Matrix Representation Parsimony (MRP) was used to merge the majority rule posterior consensus trees produced by Phylobayes. The software package Mesquite was used to edit the gene trees and produce MRP matrices from them (Maddison and Maddison, 2011). The MRP matrices were then used to calculate super trees with the maximum parsimony algorithm Pars from the Phylip package (Felsenstein, 2005).

Only nodes with a high posterior probability were used in super tree construction. Nodes below the cut-off value were collapsed prior to producing the MRP matix. Several cut-off values were tested and it was determined that the most appropriate cut-

off values were different for the red algae and green algae and for the two major green algae lineages, the Chlorophyta and Streptophyta.

The final super tree was then produced by manually merging the respective sections of the optimal super trees for the different lineages. At this point the Angiosperm taxa were added back to the phylogeny. In order to do so the Angiosperm phylogeny as published in Soltis et al., 2011 was used as guide to manually build a phylogeny for the Angiosperm taxa in the dataset.

At this point it was established that 59 of the angiosperm taxa had the same gene content as their closest relatives in the dataset. These taxa were omitted as non-informative from subsequent gene loss analyses, and from diagrams and figures to make them more readable.

2.3.5 Mapping of gene loss events

Given the equivocal evidence concerning the position of the root of the Archaeplastida, estimations of gene loss events were conducted for the three possible branching orders for the Glaucophyta, Rhodophyta and Viridiplantae.

2.3.5.1 Estimating the minimum number of independent gene loss events

The minimum number of independent gene loss events necessary to produce a certain presence/absence pattern for a gene among the taxa of a given phylogeny was calculated using Dollo parsimony. The parsimony calculations were conducted in MacClade 4.08a OS X (Maddison and Maddison, 2005).

To infer the minimum number of gene loss events each gene was coded as a character that is either present or absent in a taxon. All genes were assumed to have been present at the root. The character transformation type was defined as 'irreversible', only allowing changes in character state from present to absent, reflecting the assumption that due to genetic isolation of plastids within their lineage no gene gain has occurred and that the patterns of gene presence in extant plastid genomes are purely the result of gene losses.

2.3.5.2 Calculating the most likely gene loss frequency using Maximum Likelihood

A method proposed by Mike Steel (unpublished) was also used to calculate the most likely number of independent gene loss events using maximum likelihood. An algorithm for this was implemented into an R script written by Tim White (see Appendix A).

The script calculates the probability for any possible pattern of gene loss events that could have produced the given character state pattern at the tips of the given phylogeny. From the results the probabilities for that gene having been lost any particular possible number of times are calculated, ranging from the maximum number of gene loss events (where a gene was lost along each external branch leading to a tip where the gene is absent) to the minimum number of possible gene loss events, as calculated by the modified Dollo Parsimony method described above. The script outputs a maximum likelihood value that is the loss probability that maximises the total probability of the data over all possible reconstructions. For a detailed description of the calculations and their computational implementation see the documentation of the script in Appendix F.

2.4 Results

2.4.1 Quality of RefSeq data base entries

Eighty-five percent of the protein records retrieved from RefSeq for this project were provisional and were consequently found to contain errors of inadvertency, i.e. spelling mistakes and typos. In some cases the gene name had been appended to indicate different copies. As a consequence many proteins were missing from the output after an initial run of the MatrixMaker.pl script. These types of mistake were picked up via visual inspection of the presence/absence matrix. They became apparent in empty columns, and rows with only a single entry. Gene names were missing in the Features section of all plastid protein records for the following species: *Babesia bovis, Bambusa oldhamii, Micromonas sp., Micromonas pusilla,* and *Theileria parva*. The *Babesia bovis* records suffer from consistent misspellings. All gene names that are supposed to contain capital Is are instead written with lower case ls and in case of Rpl6 the lower case l was replaced with a 1 (Rp16). *Alveolata sp.* and *Helicosporidium sp*. were at first missing from the dataset because the species names were spelled differently across records. Rt in *Rhodomonas salina*, Urf-1 in *Vaucheria litorea*, and I-Cvul in *Chlorella vulgaris* are in fact hypothetical proteins and, like ORFs, were removed from the dataset. These mistakes were corrected after checking the original GenBank records. After merging genes with redundant names the list of genes was reduced from 307 to 274. Aliases for plastid proteins are listed in Appendix B.

A BLASTp search with all proteins that weren't assigned a gene name identified several proteins that had been annotated as ORFs but whose similarity to other genes strongly suggested orthology. It also identified many proteins of known identity for which the gene name had not been listed in the features section but elsewhere in the record. A table listing all protein sequences that were identified as orthologs of known plastid proteins in this way and added to the dataset can be found in Appendix C. Conversely, errors were also found where proteins had been incorrectly assigned a designation. For example a protein annotated as Ycf7 in *Porphyra purpurea* was found to be a hypothetical protein derived from an ORF7 and was removed from the dataset.

A local tBLASTn search conducted to recover not annotated but highly conserved gene sequences uncovered mistakes in the annotations for thirty-eight species, affecting fifty-one different sets of orthologs. Forty-two previously not annotated protein coding genes were added to the dataset and another twenty four hypothetical proteins were identified as likely functional orthologs by this BLAST search. Eleven proteins were removed from the dataset because they showed very low sequence similarity or were truncated relative to their supposed orthologs. Rpl12 in *Alveolata sp*. was found to be Rps12. The changes made to the dataset following this step are detailed in Appendix D.

2.4.2 The gene presence/absence matrix

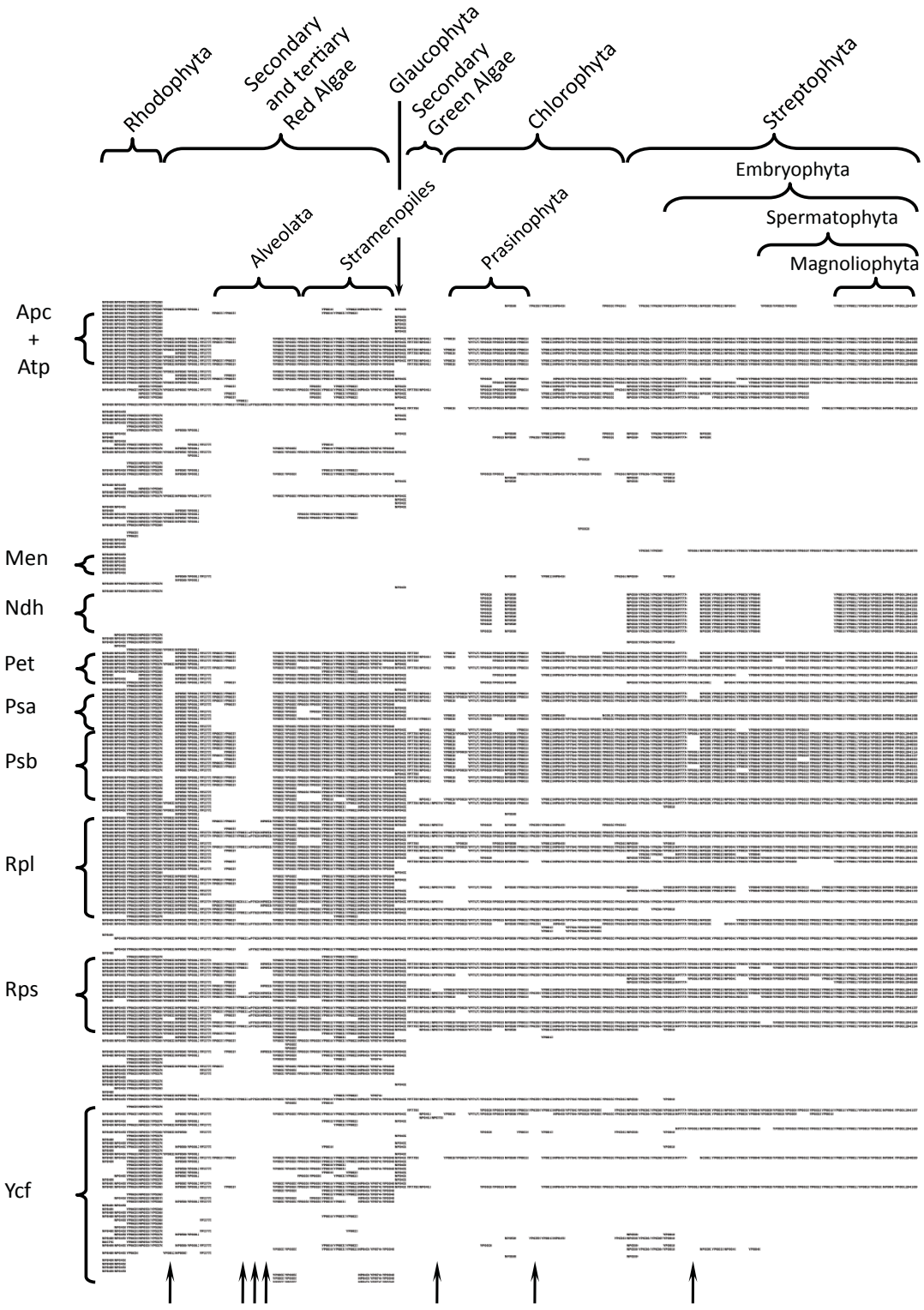
The final version of the gene presence/absence matrix includes the complete plastid proteomes of 193 species. It comprises a total of 15,588 protein sequences in 264 sets of orthologous proteins. These data are shown in Appendix A.

The dataset includes representatives of the three major lineages of the Archaeplastida, though one of them, the Glaucocystophyta, is only represented by a single species, *Cyanophora paradoxa*. The red lineage is represented by five Rhodophyta and twenty-one algae whose plastids stem from secondary or tertiary endosymbioses.

The gene content of the sequenced plastid genomes varies widely in a lineage specific way. Figure 2-1 shows an overview of the gene presence/absence matrix. Some taxa have been omitted from the most over-represented lineages to aid visualisation of lineage specific differences. Genes representing subunits of the major plastid protein complexes or functional groups of genes have been indicated as well as the rows

Figure 2-1 Schematic of the gene presence/absence matrix comprising 66 taxa that represent all major taxonomic groups in the dataset. (following page)

Parasitic species are indicated by arrows at the bottom of the figure. The rows are sorted alphabetically by gene name. The columns are sorted by taxonomic groups. Apc: Phycobillisome proteins, Atp: ATP Synthase subunits, Men: biosynthesis enzymes, Ndh: NADH-dehydrogenase, Pet: Cytochrome proteins, Psa: Photosystem I subunits, Psb: Photosystem II subunits, Rpl: Large subunit ribosomal proteins, Rps: Small subunit ribosomal proteins, Ycf: Proteins of unknown function


Rhodophyta
Secondary and tertiary Red Algae
Glaucophyta
Secondary Green Algae
Chlorophyta
Streptophyta
Embryophyta
Spermatophyta
Magnoliophyta
Alveolata
Stramenopiles
Prasinophyta
Apc + Atp
Men
Ndh
Pet
Psa
Psb
Rpl
Rps
Ycf

representing ycf proteins. The latter is not a functional group but comprises proteins of unknown function. Similarity in gene content is a conspicuous feature within lineages, as are the marked differences between lineages. Of all lineages, the Rhodophyta retain the most protein coding genes in their plastid genome, with between 198 in *Porphyra* and 188 in *Cyanidium caldarium*. Secondary and tertiary red algae have retained only around 130 genes in their plastids with the Alveolata containing the lowest number of genes. *Chromera velia* retains only 56 protein coding genes in its plastid genome. Compared to the Rhodophyta, the secondary red algae have lost genes that execute a variety of functions in the plastids metabolism and gene regulation. Those associated with major plastid encoded protein complexes tend to have been retained.

The green lineage has transferred more genes to the nucleus than other major lineages. The chloroplasts of primary green algae typically encode between 60 and 90 genes. Some Prasinophytes retain even fewer genes in their chloroplasts with *Micromonas pusilla CCMP1545* constituting a noteworthy extreme: its plastid genome retains only 27 protein coding genes, the lowest number in any photosynthetic eukaryote sequenced to date.

A distinctive pattern is apparent in terms of which genes have been lost from or retained in chloroplasts of the green lineage. This is the case despite the occasional presence/absence of many individual genes. The green lineage has lost the cyanobacterial phycobillisome proteins and translocated many other genes to the nucleus. Its members mostly only retain genes in the chloroplast that are coding for some of the subunits of these major protein complexes: ATP synthase, photochlorophyllide reductase (involved in chlorophyll biosynthesis), cytochrome, photosystems I and II, large and small ribosomal subunit, and RNA polymerase. The subunits of NADH dehydrogenase on the other hand are only encoded in the plastid genomes of most Streptophyta and some Prasinophytes, but not in the Glaucophyta or the red lineage. The secondary green algae in this dataset keep less than 60 protein coding genes in their plastids and are, like the Angiosperms and some Prasinophytes, missing the photochlorophyllide reductase genes.

Cyanophora paradoxa, the only example of a Glaucophyte, retains 129 protein-coding genes in its plastid genome. Its gene content is intermediate between the Rhodophyta

and the green lineage but shows a marked difference in gene retention to the secondary red algae.

2.4.3 The plastid phylogeny

The sequences of plastid encoded proteins were used to infer a phylogeny for the taxa in the dataset to serve as a basis for the reconstruction of gene loss events during the evolutionary history of plastids.

2.4.3.1 Protein Alignments

The total length of aligned concatenated protein sequence available for phylogenetic analyses in this project comprised 16892 amino acid positions. From these data 47 protein alignments were constructed containing subsets of taxa and genes. In 11 alignments, where homologs occurred in members of green as well as the red algae, alignments were calculated separately for both lineages to avoid reconstructions being impacted by lineage specific systematic biases that differed in these two major lineages. Thus, 58 separate protein alignments were available for subsequent phylogenetic reconstructions. In total 15144 amino acid positions were available to infer the phylogeny of the red lineage while 6762 positions were used for the green lineage. Of these, 5014 amino acid positions were shared by all lineages.

2.4.3.2 Gene trees

Fifty eight gene trees each were inferred from the above mentioned protein alignments using Bayesian and Maximum likelihood methods respectively to serve as input for the construction of a supertree of all taxa..

The individual gene trees are provided in Appendix A.


Glaucophyta
Cyanophora paradoxa
Rhodophyta
Cyanidium caldarium
Cyanidioschyzon merolae
Gracilaria tenuistipitata
Porphyra purpurea
Porphyra yezoensis
Secondary Red Algae
Emiliana huxleyi
Guillardia theta
Rhodomonas salina
Aureoumbra lagunensis
Aureococcus anophagefferens
Heterosigma akashiwo
Ectocarpus siliculosus
Vaucheria litorea
Phaeodactylum tricornutum
Durinskia baltica
Kryptoperidinium foliaceum
Odontella sinensis
Thalassiosira oceanica
Thalassiosira pseudonana
Chlorophyta
Nephroselmis olivacea
Pycnococcus provasolii
Euglena gracilis
Pyramimonas parkeae
Monomastix sp. OKE1
Ostreococcus tauri
Micromonas pusilla
Micromonas sp. RCC299
Chlamydomonas reinhardtii
Scenedesmus obliquus
Oedogonium cardiacum
Floydiella terrestris
Stigeoclonium helveticum
Bigelowiella natans
Bryopsis hypnoides
Leptosira terrestris
Pseudendoclonium akinetum
Oltmannsiellopsis viridis
Chlorella vulgaris
Parachlorella kessleri
Streptophyta
Mesostigma viride
Chlorokybus atmophyticus
Chaetosphaeridium globosum
Chara vulgaris
Staurastrum punctulatum
Zygnema circumcarinatum
Marchantia polymorpha
Physcomitrella patens
Syntrichia ruralis
Anthoceros formosae
Tracheophyta
Huperzia lucidula
Isoetes flaccida
Selaginella moellendorffii
Equisetum arvense
Angiopteris evecta
Psilotum nudum
Alsophila spinulosa
Pteridium aquilinum
Cheilanthes lindheimeri
Adiantum capillusveneris
Cycas taitungensis
Ephedra equisetina
Gnetum parvifolium
Welwitschia mirabilis
Cryptomeria japonica
Keteleeria davidiana
Cedrus deodara
Cathaya argyrophylla
Pinus thunbergii
Angiosperms
Nymphaea alba
Acorus americanus
Colocasia esculenta
Dioscorea elephantipes
Phalaenopsis aphrodite
Oncidium Gower Ramsey
Phoenix dactylifera
Coix lacryma-jobi
Zea mays
Saccharum officinarum
Oryza sativa var. Japonica
Oryza sativa var. Indica
Brachypodium distachyon
Agrostis stolonifera
Festuca arundinacea
Ranunculus macranthus
Spinacia oleracea
Panax ginseng
Trachelium caeruleum
Lactuca sativa
Parthenium argentatum
Helianthus annuus
Jasminum nudiflorum
Olea europaea
Nicotiana tomentosiformis
Ipomoea purpurea
Cuscuta gronovii
Cuscuta reflexa
Vitis vinifera
Citrus sinensis
Gossypium hirsutum
Carica papaya
Lobularia maritima
Oenothera parviflora
Eucalyptus grandis
Eucalyptus globulus
Pelagonium x hortorum
Monsonia speciosa
Geranium palmatum
Erodium texanum
Populus alba
Jatropha curcas
Manihot esculenta
Morus indica
Castanea mollissima
Glycine max
Phaseolus vulgaris
Lotus japonicus
Cicer arietinum
Pisum sativum
Lathyrus sativus
Trifolium subterraneum
Medicago truncatula

Figure 2-2 Supertree of 124 Plantae species (previous page)

Supertree constructed from 58 individual gene trees. The 124 taxa shown here represent the full diversity of photosynthetic species in the dataset in terms of gene distribution. Non-photosynthetic species were not included and monophyletic clades that share the same gene content are represented by only one species. The asterisks indicate secondary green algae.

2.4.3.3 Supertree

A single most parsimonious supertree was calculated from the gene trees inferred with Bayesian methods. For splits within the Streptophyta a posterior probability cut-off of 0.85 was used while the rest of the tree, comprising the red lineage and the green lineage of the Chlorophyta was assembled with a higher cut-off value of 0.9. It is shown in Figure 2.2 and includes the relationships among angiosperms as inferred by Soltis et al. (2011). Overall the relationships are similar to those previously reported, as discussed in Section 2.5.3.2.

2.4.4 The prevalence of plastid gene transfers to the nucleus

Using the principles of Dollo parsimony the minimum number of gene losses for each gene were reconstructed on the supertree. The numbers of losses across genes that were inferred for each branch of the crown groups are shown in Figure 2-3.

Because Dollo Parsimony can only produce the most conservative estimates, a Maximum Likelihood based approach suggested by Prof. Mike Steel was used to calculate an approximation of the effective number of gene loss events suggested by the data. This method calculated the number of gene losses about 14% higher than the conservative estimates produced with Dollo Parsimony. The results yielded by both methods are summarised in Table 2-1. The estimates for each individual gene in the dataset are listed in Appendix A.

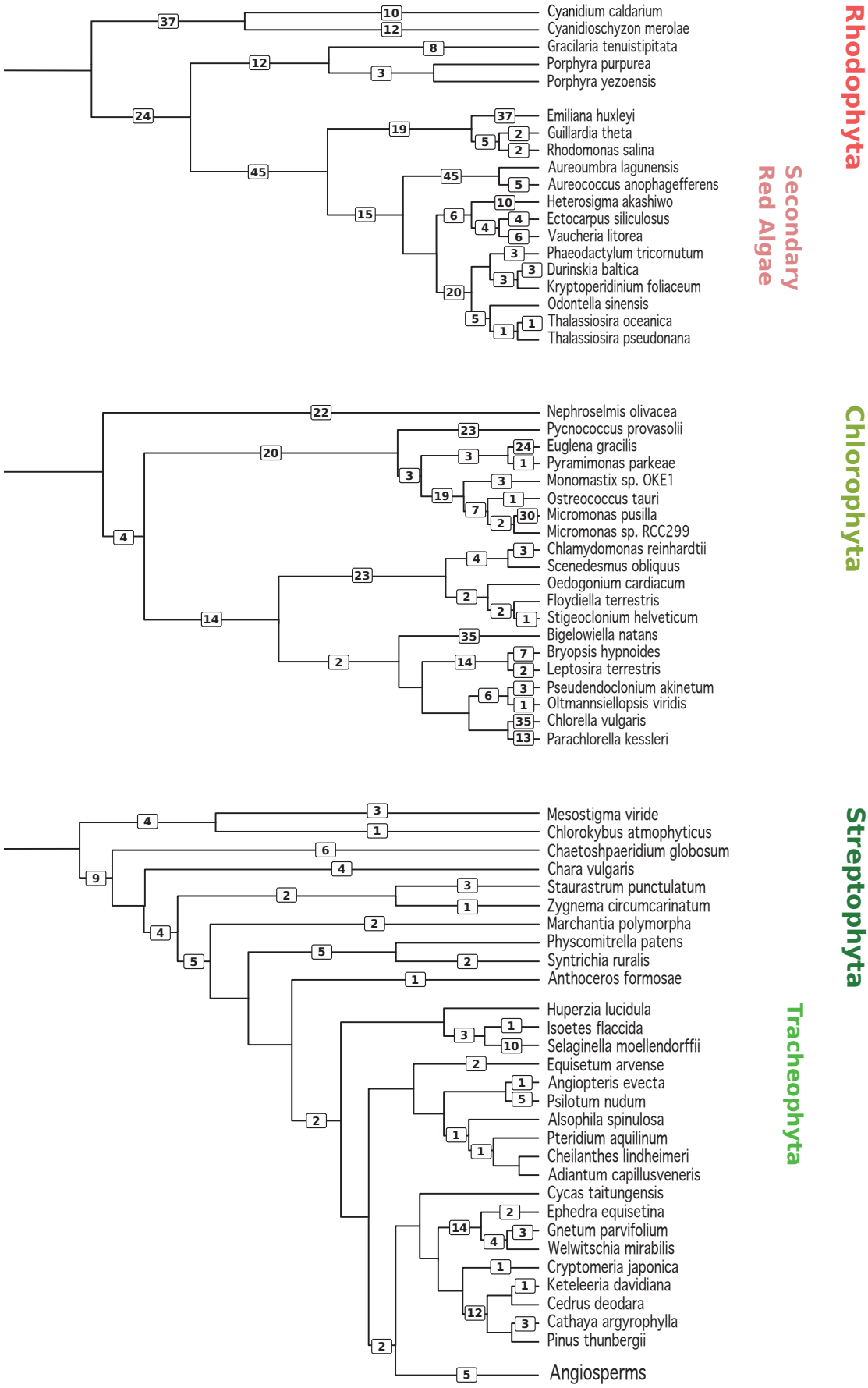
Table 2-1 Summary of results of Gene Loss reconstructions

Basic metrics summarising the Dollo Parsimony and Maximum Likelihood reconstructions of gene loss events from the plastid genome. The analysis was conducted for each of the three possible root positions, denoted by the name of the most basal lineage for a given root position. The values for "ML value in % Dollo" indicate the number of gene loss events predicted by Maximum Likelyhood as a percentage of the number inferred by Dollo Parsimony.

Basal lineage	**Glaucophyta**		**Rhodophyta**		**Viridiplantae**	
	Dollo	**ML**	**Dollo**	**ML**	**Dollo**	**ML**
Sum of losses for all genes	1243	1396.6	1271	1448.1	1370	1558.3
ML value in % Dollo	NA	112.4%	NA	113.9%	NA	113.7%
Average losses/gene	4.8	5.4	4.9	5.6	5.3	6.0
Parallel to unique losses	36:1	40:1	63:1	71:1	68:1	77:1

The results for the different positions of the root differ but nonetheless all suggest gene loss being a common event during plastid evolution. A basal placement of the Glaucophyta produced the smallest number of required gene loss events (1243) with 1397 being the most likely number. This equates to an average of 4.8 (Dollo) or 5.4 (ML) gene loss events to have shaped the evolution of the average plastid encoded gene in the dataset for this scenario. The highest number of gene loss events is inferred with a basal placement of the green lineage. This scenario requires at least 1370 independent gene losses with 1558 gene losses being most likely. The average number of losses per gene in this case is 5.3 (Dollo) and 6.0 (ML) respectively.

Another way of interpreting these data is to calculate the ratio of unique to parallel gene losses. Unique gene losses are defined as the cases where a gene was lost only once over the entire phylogeny. All other gene loss events are considered parallel gene losses as in these cases a gene was lost two or more times independently. In the initial study of 16 plastid genomes by Martin et al. in 2002 a ratio of >10:1 was calculated. The numbers were markedly higher with this larger dataset, ranging from > 36:1 (ML) for a basal placement of the Glaucophyta to > 77:1 (ML) for Rhodophyta as the most basal lineage.


Rhodophyta
Secondary Red Algae
Cyanidium caldarium
Cyanidioschyzon merolae
Gracilaria tenuistipitata
Porphyra purpurea
Porphyra yezoensis
Emiliana huxleyi
Guillardia theta
Rhodomonas salina
Aureoumbra lagunensis
Aureococcus anophagefferens
Heterosigma akashiwo
Ectocarpus siliculosus
Vaucheria litorea
Phaeodactylum tricornutum
Durinskia baltica
Kryptoperidinium foliaceum
Odontella sinensis
Thalassiosira oceanica
Thalassiosira pseudonana
Chlorophyta
Nephroselmis olivacea
Pycnococcus provasolii
Euglena gracilis
Pyramimonas parkeae
Monomastix sp. OKE1
Ostreococcus tauri
Micromonas pusilla
Micromonas sp. RCC299
Chlamydomonas reinhardtii
Scenedesmus obliquus
Oedogonium cardiacum
Floydiella terrestris
Stigeoclonium helveticum
Bigelowiella natans
Bryopsis hypnoides
Leptosira terrestris
Pseudendoclonium akinetum
Oltmannsiellopsis viridis
Chlorella vulgaris
Parachlorella kessleri
Streptophyta
Tracheophyta
Mesostigma viride
Chlorokybus atmophyticus
Chaetoshpaeridium globosum
Chara vulgaris
Staurastrum punctulatum
Zygnema circumcarinatum
Marchantia polymorpha
Physcomitrella patens
Syntrichia ruralis
Anthoceros formosae
Huperzia lucidula
Isoetes flaccida
Selaginella moellendorffii
Equisetum arvense
Angiopteris evecta
Psilotum nudum
Alsophila spinulosa
Pteridium aquilinum
Cheilanthes lindheimeri
Adiantum capillusveneris
Cycas taitungensis
Ephedra equisetina
Gnetum parvifolium
Welwitschia mirabilis
Cryptomeria japonica
Keteleeria davidiana
Cedrus deodara
Cathaya argyrophylla
Pinus thunbergii
Angiosperms

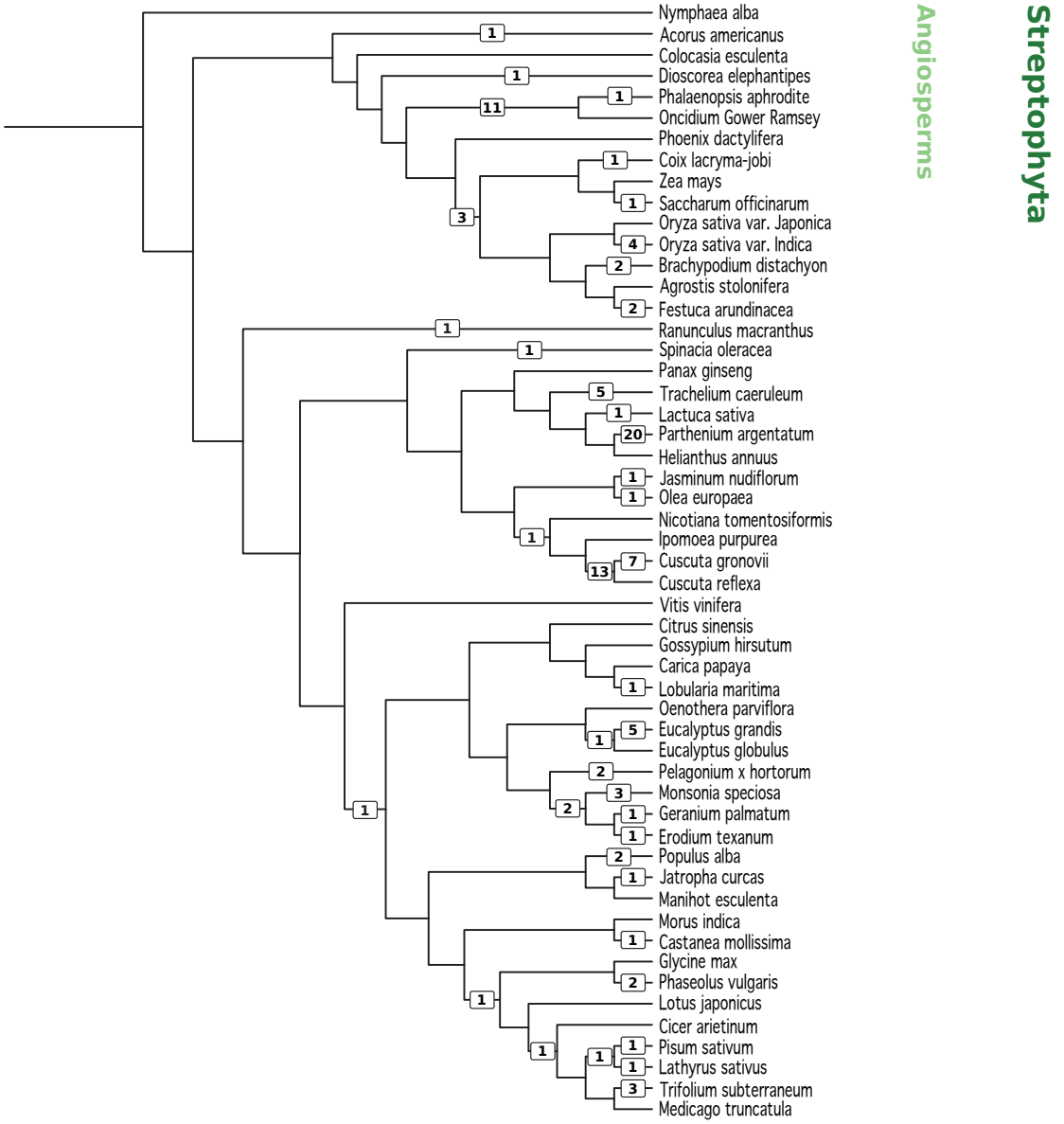


Figure 2-3 Gene losses in crown groups (this and previous page)

Shown are cladograms of the major lineages of the plantae. The branch labels indicate the number of genes that were inferred to have been lost along that branch using Dollo Parsimony. In case of internal branches it is therefore possible that the gene losses did not occur on that branch but instead multiple times along the descendent branches.

To put the 'unique' gene losses further into perspective: across the three different variants of the phylogeny, 32 different genes could have been lost in unique events. Of these 32, only twelve are missing in a wide range of taxa. Nine are absent from the entire green lineage (*atpG, dnaK, groeL, psbX, rbcS, rpl34, rps5, sufB, ysf39*) and three are missing in both the entire green and red lineages (*groeS, hemA, nadA*). The other 20 uniquely lost genes are missing only from single taxa in this dataset. *Micromonas pusilla* alone has lost 14 genes (*atpA, atpB, atpH, petG, psbB, psbC, psbD, psbE, psbH, psbK, psbN, psbT, rpl36, rps4*) that are present in all other taxa in this study, with the exception of two genes (*rpl36, atpB*), which are absent from the other Micromonad plastid genome in the dataset as well. The other unique gene losses are confined to *Saccharum officinarum* (*psbA*), *Chlorella vulgaris* (*psbF*), *Parthenium argentatum* (*rps3, rpl2*), and *Cathaya argyrophylla* (*rps7*).

2.4.5 Lineage specific gene loss and retention

The major lineages of plastid carrying eukaryotes show lineage specific gene loss and retention as shown by the presence/absence matrix in Section 2.4.2. Using Dollo Parsimony the minimal gene content of the plastid of the last common ancestor of a lineage can easily be reconstructed. Following the assumptions of Dollo Parsimony that a gene that is absent in all members of a clade has been lost prior to the clade's last common ancestor, it can also easily be calculated how many genes are likely to have been lost during evolution leading up to that clade. The results are summarised in Figure 2.4. The trees are schematics of the three phylogenies that were used to map gene loss events. The tree showing the Glaucophyta as most basal lineage is fully labelled. Each black number indicates the minimum number of genes that must have been present in the last common ancestor of the lineage represented by that branch. A change in the position of the root only changes the number of genes assigned to the internal branch that represents the common ancestor of the two sister lineages. The values for all other branches remain unchanged and were for that reason omitted from the schematics of the other phylogenies that represent a most basal placement of the Rhodophyta and Viridiplantae respectively. The number of likely losses of genes prior to the diversification of a lineage is indicated in a similar fashion as negative numbers in red. The change in root position does in this case affect the number of possible gene losses at the base of the major lineages as well as the internal branch. Figure 2.3 shows in detail where gene losses were reconstructed within the crown groups.

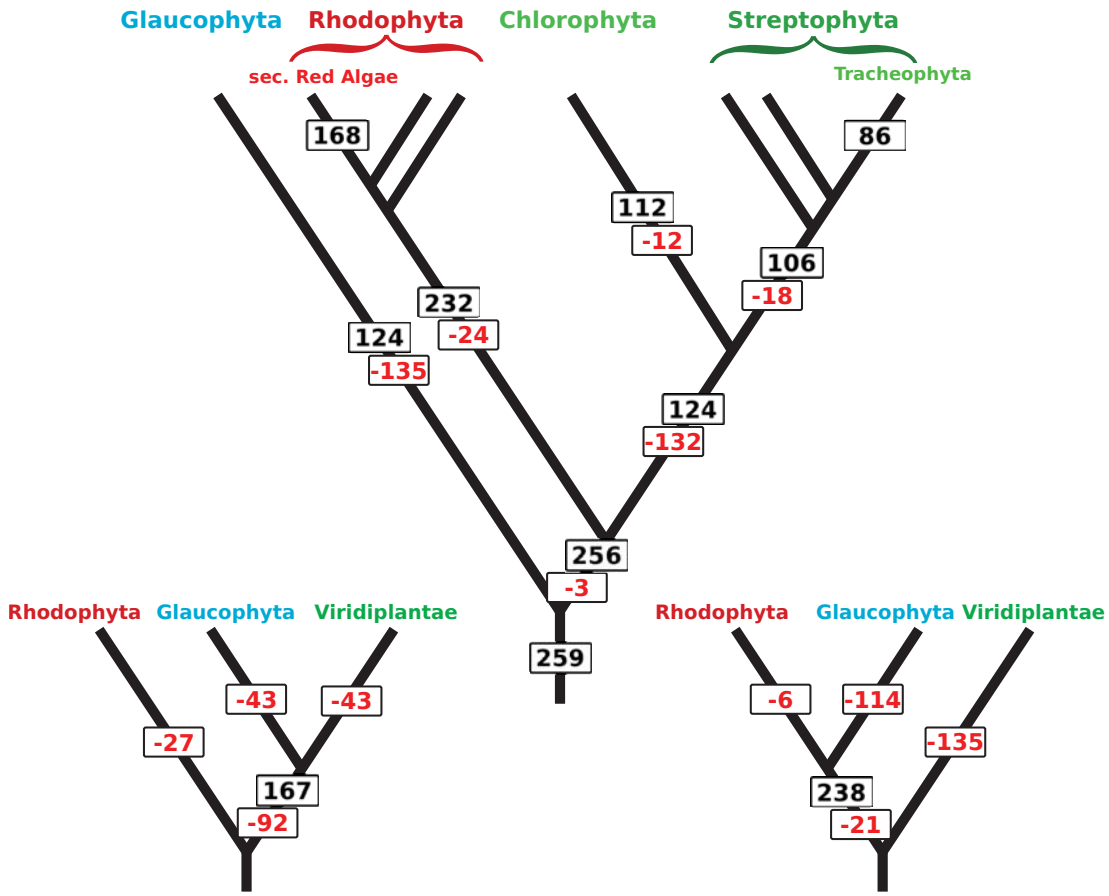


Figure 2-4 Summary Trees for the three possible root positions.

Shown are schematics of the plantae phylogeny with the three possible root positions. Numbers in black indicate the minimum number of genes that must have been present in the last common ancestor of the respective lineage or clade. The negative numbers in red indicate the number of genes that are likely to have been lost on that branch as they are not present in any of the taxa within that clade.

The numbers reflect what has been written about differential gene retention between the major lineages in Section 2.4.2. Of the 259 known plastid coded genes only very few would have been lost in the common ancestor of green and red algae (see the -3 on internal branch). The intermediate gene content of the Glaucophyta when compared to the red and green lineage is reflected in the markedly smaller numbers of genes retained on the internal branch of the small trees.

2.5 Discussion

2.5.1 Quality of RefSeq data base entries

This project is based on the groundwork of others who produced, annotated and submitted complete sequences of plastid genomes. Assuring the quality and completeness of the sequences and annotations prior to estimates being made in the present work was important to reliably estimate gene losses. The efforts made here in this respect have highlighted the patchy quality of GenBank records.

The GenBank flat file format and the sequence submission process seem to encourage the relegation of information into the comments section rather than the designated position in the GenBank record. This resulted in sporadically missing protein entries for some taxa. Gaps and invalid characters in the sequences are relatively easy to pick up, because they cause error messages from the programs used to work on them. Yet several taxa had to be removed from the study because of long gaps in their genomic sequences. In another case a protein sequence was found to contain a J, which is not a valid amino acid abbreviation. Within the scope of this project it was not possible to evaluate the quality of the plastid genome sequences apart from obvious shortcomings like these, but this should not be of much concern. Even though the problem of genome assemblies can have its pitfalls, the assembly of the relatively small plastid genomes with their well-known and conserved structure should not be very challenging. This fact gives some confidence that no genes should be missing from this analysis due to missing genomic sequence.

The quality of the annotations proved to be the far bigger issue, not only because of the number of mistakes found in GenBank records but because a gene that has either been missed or falsely annotated as present in a single genome can have a substantial effect on the calculations of gene loss frequency during plastid evolution. While most plastid genomes in this study were accurately annotated, there were also several examples of incomplete or incorrect annotations. Judging by the mistakes I found, one reason for missing annotations seems to be the common practise of simply copying the

annotations of the most closely related sequenced plastid genome. If the reference genome is missing a gene, then this gene is often not annotated in the new genome even if it is present there. This is a common problem for angiosperm species, an evolutionary young group that is the subject of most plastid sequencing projects due to its economical importance. The assumption that gene content will be highly similar between closely related angiosperms, has led to a lack of sufficient scrutiny when comparing genomes.

The rate at which new plastid genomes are sequenced has been steadily increasing, especially after the advent of high-throughput sequencing technology (Gao et al., 2010). As technological advancements make the production of new sequence ever faster and cheaper, the focus of work seems to be increasingly placed on output instead of careful analysis of sequences. A telling example is the plastid genome of *Oryza sativa var. Indica*, the strain of domesticated rice that originated in India as opposed to *O. sativa var. Japonica* from Japan. In the Indian domesticated rice, two genes (*CemA* and *PsbJ*) are missing and one (*Rps12*) is truncated compared to the Japanese strain. However, the GenBank record is missing annotations for a further fifteen protein coding genes that were missed or omitted despite the fact that with the *Oryza sativa* var. Japonica chloroplast genome a very close and well annotated reference has been available since 1989 (Hiratsuka et al., 1989; see Appendix D). The red alga *Porphyra yezoensis* is another example. Of the 198 proteins encoded in its plastid genome, thirty-four had been recognised as open reading frames but not identified, and one had been missed entirely.

Not only does the annotation process sometimes not get the attention it requires, but annotations produced *in silico* are most of the time not confirmed *in vivo.* The need for follow up work to confirm predicted annotations is exemplified in the questionable pseudo gene statuses of *Ycf15* and *Ycf68* in many Angiosperm species. In many taxa the coding sequences of these two genes are disrupted by premature stop codons. *Ycf15* and *Ycf68* are for this reason both annotated as pseudo genes in more than 100 taxa each. Yet the fact that these coding regions are conserved across angiosperms is conspicuous and indicates that these genes might in fact be expressed. Several authors have noted this situation but the expression status of these genes has not yet been systematically investigated (Sato et al., 1999; Schmitz-Linneweber et al., 2001; Steane, 2005; Raubeson et al., 2007; Daniell et al., 2008; Tangphatsornruang et al., 2011). A

possible explanation is that RNA editing might be used to remove these premature stop codons from the transcripts.
The case of *InfA* looks very similar at first but this gene has been shown to have lost its functionality several times independently in different angiosperm lineages. Analyses of nuclear expression data have shown that a copy has been independently transferred to the nucleus in several lineages as well. This seems to be a very interesting case of ongoing parallel gene transfer to the nucleus in several Angiosperm lineages (Millen et al., 2001). It also makes the expression status of the *InfA* genes that have been annotated in plastid genomes somewhat uncertain. Detailed expression studies would be necessary to establish the functionality of these genes. Given the absence of such information *InfA, Ycf15* and *Ycf68* were excluded from the analyses.

Another problem that GenBank users are facing is inconsistency of gene names. Existing records are usually not updated if a new gene name is assigned. This presents an obvious problem for records of Open Reading Frames (ORFs). Following convention a predicted open reading frame whose expression status and function is unclear, is named ORF followed by a number that indicates the number of amino acid residues that it may be coding for. This type of designation does by no means indicate orthology and is thus almost uninformative. The respective GenBank record is usually not updated, even if the function of the respective gene is subsequently characterised and a meaningful name assigned.

NCBI is trying to alleviate these problems by creating and maintaining records for every gene in every fully sequenced genome in its Gene database. Records in this database are updated when listed information changes and they list other gene names that are in use in the category 'Other Aliases'. This database has been a valuable resource in error checking my dataset. Other than that it does nothing to alleviate the problems for projects that, like this one, are sourcing their information from the original GenBank records.

2.5.2 The presence/absence matrix

With 163 out of 193, most species in the dataset belong to the green lineage. This is due to the economic importance of many Embryophyta. Consequently, only 25 taxa represent the vast diversity of green algae while the other 138 are Embryophyta. The

Angiosperms, or flowering plants, are the most strongly represented group with 113 fully sequenced chloroplast genomes. So far only three chloroplasts of secondary green algae from two different lineages have been sequenced: *Bigelowiella natans* represents the Chlorarachniophyta and *Euglena gracilis* and *Euglena longa* the Euglenoids.

A distinction must be made between the autotroph species that depend on photosynthesis and species that have adopted a parasitic life style and lost their ability to photosynthesise. Non-photosynthetic parasites are characterised by a tendency to lose genes from the plastid genome because they are not anymore required for survival. As a consequence the plastid genomes of parasitic taxa are much poorer in genes than those of autotroph relatives. The parasitic species included in Figure 2-1 clearly stand out from their neighbours as almost empty looking columns. The Apicomplexa have driven this process to the extreme with *Babesia bovis* encoding only sixteen proteins in its apicoplast genome, most of them ribosomal proteins. The plastid of the Cryptomonad *Cryptomonas reinhardtii* still encodes for 79 proteins, mainly ribosomal proteins and the ATP synthase subunits. All in all ten of the 193 species in the dataset are parasitic and have lost the ability to photosynthesise. The four representatives of the genus *Cuscuta* in this dataset have adopted a parasitic lifestyle but have so far retained photosynthetic function (Revill et al., 2005; McNeal et al., 2007).

2.5.3 The plastid phylogeny

2.5.3.1 Gene trees

Only 71 of the 193 taxa in the dataset were included in the phylogenetic analyses. Fitting complex models using Bayesian or Likelihood selection criteria is computationally demanding especially for a large number of taxa. This makes it necessary to restrict analyses to smaller subsets of taxa. For this reason the Angiosperms were omitted from the phylogenetic reconstructions. As detailed in Section 2.4.2 the Angiosperms constitute the vast majority of taxa (113 out of 193) in the dataset and are the best-studied lineage of all Plantae, making them the only numerous group in the dataset whose taxonomy is relatively well supported and for which a consensus of phylogenetic relationships has emerged (Soltis et al., 2011). Furthermore all non-photosynthetic taxa were excluded because they cannot be used

in the subsequent gene loss analyses. In photosynthetic species a gene loss from the plastid genome can be assumed to equate to a gene transfer to the nucleus with subsequent re-import of the protein into the plastid (see Section 2.5.4). However, where photosynthetic function is lost the genes involved are usually lost from the organism as a whole. Since the latter involves a different process it cannot be compared to the former.

Two different methods, Maximum Likelihood and Bayesian inference were used to estimate the individual gene trees. The gene trees that were reconstructed using maximum likelihood showed the effects of long branch attraction (LBA). This was especially obvious for the two Chromerida species, *Chromera velia* and *Alveolata sp*., which produced markedly longer branches than the other taxa. The topology of most phylogenies changed after the Chromerida were removed, consistent with heterotachy problems sufficient to impact on phylogenetic reconstructions (Lockhart et al., 2006; Lockhart et al., 1996). The two Chromerida taxa were for this reason removed from the analyses.

Using Bayesian inference the individual source trees were reconstructed under the CAT-GTR model for exchange rates and profile mixture. Empirically CAT-GTR has been found to be more robust than other models when dealing with LBA and is considered better suited to modelling sequence heterogeneity (PhyloBayes Manual; Wu et al., 2011). It is an infinite mixture model whose components differ by their equilibrium frequencies, but otherwise share the same set of relative exchange rates, which have been inferred from the data. The phylogenies produced by Bayesian inference were generally not as well resolved as those inferred with Maximum Likelihood but where they were, their nodes were better supported. The phylogenies produced by the two methods did in most instances not notably differ from each other in their topologies.

2.5.3.2 Supertree

A bifurcating phylogeny is required to reconstruct historic losses of protein coding genes. However, reconstructing the taxonomic and evolutionary relationships among the Plantae is in itself a difficult problem that still is an area of ongoing research. Traditional taxonomic classifications in use today are based on centuries of work on morphological characters. Often these have produced “synapomorphies” (shared derived character states) which are not supported by the results of molecular analyses.

The latter have in many cases prompted taxonomic revisions (Turmel et al., 2009). The findings of molecular studies have sometimes been conflicting as well and it remains unclear, which types of molecular data and which methods of analysis are least affected by stochastic or systematic error. The literature alone does not provide a consensus phylogenetic framework sufficient for analyses of gene loss. For this reason, phylogenetic estimates from protein sequence analyses were made in the present study in conjunction with reference to earlier findings reported in the literature.

The distribution of protein coding genes in the plastid genomes of photosynthetic eukaryotes is varied as highlighted by Figure 2-1. The single protein phylogenies that were constructed from the proteins in the dataset consequently comprised different subsets of taxa, depending on presence and absence of particular homologues. When a gene is absent in one or more species, this introduces gaps into the data of phylogenetic analyses. There are two general approaches to the problem of missing phylogenetic data, i) the use of a Supermatrix and ii) the construction of a Supertree or Supernetwork. A Supermatrix is a concatenation of all alignments that are part of the analysis. Such a matrix indicates missing data where a gene is absent from a taxon. There have been concerns that Supermatrix methods are susceptible to substitution model misspecification, though several comparative studies have suggested that the Supermatix approach can be effective, producing reasonable phylogenies and phylogenetic estimates (Philippe et al., 2004).

The Supermatrix method is generally recommended when genes can be expected to all have same evolutionary history and when phylogenetic reconstruction is not affected by lateral gene transfer (LGT) (Yang and Rannala, 2012). However, the approach can be impacted by differences in evolutionary dynamics among genes and between lineages. Though LGT does generally not occur in plastids (see Section 2.5.3), model misspecification due to heterotachy (differences in estimated rates of evolution in different lineages: Lockhart and Steel, 2005) is a significant problem especially in anciently diverged lineages, as it can induce long-branch attraction (LBA) effects which produce an artificial regrouping of the fast-evolving sequences (Felsenstein, 1978; Lockhart et al., 2006).

In long concatenated alignments, partitioning can be used to account for subsets of the data/genes that have evolved under different stationary models of sequence evolution

but it cannot account for lineage specific variation due to heterotachy. Different processes can potentially cause heterotachy, and its true nature has been poorly characterised (Lockhart et al., 2006). Heterotachy might be caused by the speeding up or slowing down of amino acid/nucleotide substitution rate in different lineages, and/or because of a change in the nature of the process of substitution between lineages. The later might include lineage specific changes in the equilibrium frequencies of amino acids or changes in the proportion of variable sites resulting from changes in structural/functional constraints.

Lockhart et al. (2006) have suggested that it is the proportion of variable sites that differs among orthologs in anciently diverged evolutionary lineages. These authors observed that the subunits of RNA polymerase (*Rpo*) produce gene trees incongruent with other plastid genes, and have presented analyses that indicate this is a consequence of an increased proportion of variable sites in the green algal lineage. The same phenomenon has been observed in eukaryotic translation factors that produce different tree topologies concerning the ciliates. Changes in protein-protein interactions have been suggested as one possible cause for this (Moreira et al., 2002). Lineage specific variations in evolutionary rates have been demonstrated for some seed plant lineages, in particular grasses and conifers (Zhong et al., 2011; Wu et al., 2011). Further, Wu et al. (2011) have observed that the functional class of proteins appears associated with whether or not plastid genes exhibit low or high heterotachy. Chloroplast proteins that fall into low heterotachous categories in conifers have functions related to photosynthesis, while the high heterotachous categories are associated with gene expression or other functions (Leebens-Mack et al., 2005; Wu et al. 2011). Since the sequences for gene expression can dominate the overall length of concatenated plastid datasets, Wu et al. (2011) have cautioned that the dominant feature of heterotachy in their seed plant Supermatrix could potentially mislead phylogenetic estimates. This point of concern has also been raised by others (e.g. Phillips et al., 2004), and in particular Jansen et al. (2007) have suggested that lineage-specific rate accelerations may be a general feature of plastid genome evolution. For these reasons the Supermatrix approach was not used in the present study.

An alternative approach to cope with missing genes in phylogenetic datasets is to construct a Supertree from source trees with varying taxa compositions. This approach was adopted here to obtain a phylogenetic framework for gene loss estimates, because

phylogenetic estimates could be made more independently for highly diverged evolutionary lineages and then these estimates combined. This was done to help reduce the impact of misleading signal, systematic error and tree building artefacts. It also accommodated missing data in the phylogenetic analyses. The main problem posed by this approach is in merging potentially conflicting tree topologies. The presently most commonly used and also most accurate method for building supertrees is MRP (Matrix Representation Parsimony) (Bininda-Emonds, 2004a; Bininda-Emonds, 2004b; Buerki et al., 2010). MRP converts the topological information of the source trees into a character matrix. The character matrix is subsequently used to reconstruct a phylogeny using maximum parsimony. Thus, the resulting supertree is a cladogram that merges the topologies of the source trees but does not retain any information on branch length. The resulting cladogram is essentially a majority consensus of all splits in the source trees regardless of their support. It is a weakness of MRP and all other current methods for Supertree assembly, that they ignore uncertainties in the source trees.

The single protein phylogenies calculated with PhyML and PhyloBayes include a substantial proportion of nodes with low bootstrap support or low posterior probability respectively. To ensure the accuracy of the supertree these nodes were excluded and only nodes with a bootstrap support or posterior probability above a certain cut-off value were coded into a MRP matrix. The most suitable cut-off values were determined empirically. A higher cut-off value reduces conflict introduced by weakly supported splits while a lower cut-off value includes more information. For the phylogenies inferred with Bayesian methods in Phylobayes several different values between 0.7 and 0.95 were tested to determine the most acceptable trade-off between sufficient information and lowest possible conflict. In practice this is the cut-off value that results in the least number of most parsimonious trees, ideally only one. As it turned out the most suitable cut-off value is not the same across all lineages. This is hardly surprising, as the different lineages of photosynthetic eukaryotes are known to follow different modes of sequence evolution at different rates (Lockhart et al., 2006). The branch representing the Streptophyta was assembled from nodes with a posterior probability higher than 0.85, while the rest of the tree, comprising the red lineage and the green lineage of the Chlorophyta, was assembled with a higher cut-off value of 0.9.

In the case of maximum likelihood trees produced by PhyML, nodes with a bootstrap support of at least 70% produced a single most parsimonious supertree. Bootstrap

values above 70% have in some cases been accepted to indicate reliable grouping (Baldauf, 2003). For convenience this level was adopted even though it has been shown that trees with 100% bootstrap support can be misleading if substitution models are miss-specified (Lockhart et al., 1994; Lockhart et al., 1996). For the green lineage the ML supertree showed the same topology as the Bayesian Supertree but was not fully resolved. The taxa of the red lineage on the other hand appeared jumbled as compared with the Bayesian Supertree and less congruent with phylogenies of red algae and Chromalveolata in the literature (Yoon et al., 2010). Phylogenetic reconstructions using CAT-GTR models have been reported to be more robust to some forms of LBA, and the higher level of concordance with other phylogenies reported in the literature led to a decision to base further analyses on the Bayesian Supertree rather than the inferred ML Supertrees.

The evolutionary history of most algae lineages, especially the deep branching ones, remains uncertain and much more in depth work is needed before we can have a clear picture of the relationships of even the major lineages. Phylogenetic studies based on molecular data often reveal that the existing morphology-based taxonomies are not sound and misled by synapomorphies (Turmel, Gagnon, et al., 2009; Turmel, Otis, et al., 2009). For this reason the topology of the supertree was accepted unchanged for most lineages and used in the subsequent analyses of the frequency of gene losses during plastid evolution even in cases where it does not agree with the NCBI taxonomy or other sources in literature. The evolution of the Embryophyta on the other hand is quite well studied and complex derived morphological and molecular characteristics greatly facilitate the reconstruction of the major lines of descent. For the Embryophyta clade the topology of the final supertree was for this reason constrained to the generally accepted consensus. This resulted in only one modification of the topology that was not directly supported by the source trees: the positions of the ferns and club-mosses were swapped to place the club-mosses basal to the vascular plants, a clade formed by the two sister groups, the seed plants (Spermatophyta) and ferns (Moniliformopses).

The topology of the supertree generally showed good agreement with the currently accepted algae taxonomy. The topology of the primary red algae in the supertree matches our current understanding of the Rhodophyta phylogeny (Yoon et al., 2010; Hagopian et al., 2004). The Florideophyceae *Gracilaria* is sister to the two

Bangiophyceae of the genus *Porphyra*, while the two Cyanidiophytina *Cyanidium* and *Cyanidioschyzon* form the most basal clade of the Rhodophytes in this analysis.

The secondary and tertiary red algae form a monophyletic clade within the red algae. This is in agreement with the Chromalveolata hypothesis (Harper et al., 2005; T Cavalier-Smith, 1999) but might well be an artefact of insufficient taxon sampling among the primary red algae. The different and diverse lineages of the secondary red algae group as expected based on published work. The two Dinoflagellates, *Durinskia* and *Kryptoperidinium*, form a clade within the Diatoms. Dinoflagellates are a lineage of Alveolata, which have replaced their plastid with one derived from a diatom (Imanian et al., 2010). A deviation of this phylogeny from accepted taxonomy is that the Pelagophytae *Aureoumbra* and *Aureococcus,* instead of the Diatoms, form the most basal clade of the Heterokonts (Janouskovec et al., 2010; Riisberg et al., 2009).

The two major lineages of the green alga, the Chlorophyta and Streptophyta, both form monophyletic groups. The Chlorophyta are commonly divided into four classes based on morphological features (Mattox and Stewart, 1984): the Prasinophytes form a paraphyletic assemblage basal to the core Chlorophyta, which comprise the Ulvophyceae, Trebouxiophyceae and Chlorophyceae (Friedl and Rybalka, 2012; Turmel, Gagnon, et al., 2009). The Prasinophyte *Nephroselmis* represents the most basal lineages of Chlorophyta (Turmel, Gagnon, et al., 2009). The other Prasinophytes in this study form a monophyletic clade, sister to the core Chlorophytes, rather than a paraphyletic assemblage, but they do group among each other in accordance with the findings of Turmel et al. (Turmel et al., 2008). *Pycnococcus* nominally belongs to the same lineage as *Nephroselmis*, the Pycnococcaceae, but a phylogenetic study using nucleotide sequences has found it to rather group with other Prasinophytes as it does here (Turmel, M.,-C., Gagnon, et al., 2009). In this phylogeny the Chlorophyceae form a sister clade to the Ulvophyceae and Trebouxiophyceae while other studies have found the Trebouxiophyceae to be ancestral in respect to the other two groups (Marin and Melkonian, 2010; Turmel et al., 2008). However, the placement of the major lineages of Chlophyceae is in full agreement with the results of Turmel et al. (2008). Furthermore, the Ulvophyceae and Trebouxiophyceae are not recovered as independent groups in this analysis but form a mixed clade with a basal lineage formed by one species of each group, the Ulvophycea *Bryopsis* and the Trebouxiophycea *Leptosira*. This finding is not unprecedented. A phylogenetic study using plastid encoded protein sequences

conducted by Zuccarello et al. in 2009 did not support the monophyly of the Trebouxiphycea either.

This analysis gives support to the hypothesis that both secondary green algae have acquired their plastids from Chlorophyta, albeit independently. The placement of the Euglenids close to the Prasinophyte *Pyramimonas* is supported by previous studies based on plastid sequences and gene order (Turmel et al., 2009). The plastid of the Chlorarachniophyte *Bigelowiella natans* is derived from a lineage ancestral to *Chlorella* and *Pseudendoclonium* as has been found by previous studies, even though its exact placement within the core Chlorophytes remains unclear (Rogers et al., 2007; Turmel, Gagnon, et al., 2009).

The other major lineage of green algae, the Streptophyta has been studied more comprehensively than the Chlorophyta but many aspects of its phylogeny are still under discussion. Generally the topology of the supertree is in agreement with the results of other studies as reviewed and summarised by Qiu (2008). The position of *Mesostigma viridae* has long been a point of discussion as several studies produced conflicting results. Based on morphology it had been classified as a Prasinophyte, while molecular plastid data tended to place it basal to all green algae (Turmel, Gagnon, et al., 2009; Robbens et al., 2007; Rogers et al., 2007). This is reflected in the topology of many of the source trees used to construct the supertree. After identification of systematic bias in the plastid data *Mesostigma* is now thought to represent the most basal lineage of Streptophyta together with *Chlorokybus atmophyticus* (Qiu, 2008; Rodríguez-Ezpeleta et al., 2007). This is as well supported by phylogenies based on combined plastid, mitochondrial and nuclear genes (Robbens et al., 2007).

Another point of discussion is the question whether the Charophyceae are sister to the Embryophyta or basal to the Zygnematales and Embryophyta. The former was suggested by a study on nuclear data by Karol et al. in 2001 and is supported by mitochondrial data. The removal of sites with greatest character state variation on the other hand supports Charophyceae branching basal to the Zygnematales. The same is supported by chloroplast sequences and structural features (Turmel et al., 2007; Turmel et al., 2006). The correct position of the Charophyceae remains equivocal but the chloroplast signal generally favours a basal placement as it does in this analysis (Qiu, 2008).

The topology of the basal land plants in my supertree agrees with the findings of Qui et al. (2007), which are based on a combination of chloroplast, mitochondrial and nuclear genes. The only exception is the placement of the ferns and club-mosses, which in my original supertree had swapped positions as compared to the accepted taxonomy. This was the only case in which the topology of my supertree was modified for the subsequent gene loss analyses as the position of the club-mosses as basal to ferns is well established based on a variety of different molecular and morphological traits. (Schneider et al., 2009; Banks et al., 2011).

The question whether Bryophytes are mono- or paraphyletic has long been under discussion and is yet to be settled (Qiu, 2008). Recent molecular data gives strong support to them being paraphyletic, while several studies based on plastid data support monophyly (Nishiyama et al., 2004; Goremykin and Hellwig, 2005; Shanker et al., 2011). In his analysis they appear paraphyletic even though plastid proteins were used.

The long branched Gnetales are notoriously hard to place and it is still not established which gymnosperms their closest relatives are. Many studies place them as sisters of the Pinaceae and some as sister to conifers (Chaw et al., 2000; Chaw and Zharkikh, 1997; Burleigh and Mathews, 2004; Finet et al., 2010; Qui et al., 2007). The latter is supported by this phylogeny.

Over all it can be concluded that disagreements were generally confined to groups that are known to be affected by unusual modes of sequence evolution and whose taxonomy remains more or less unresolved for this reason. It is quite likely that aspects of the phylogeny used to map the loss of genes from the plastid genome are incorrect for at least some of these lineages but as will be discussed in the following section, this should not have a marked impact on the main conclusions of this study.

2.5.4 The prevalence of plastid gene transfers to the nucleus

Two methods were used to reconstruct gene loss during plastid evolution. The first, Dollo parsimony, was first described by Farris in 1977 and is based on the assumption that a complex character does not evolve more than once but can be lost independently in different lineages. The method was modified to instead assume that a character

(plastid gene) cannot be gained but was present in the common ancestor. Thus, the pattern of presence or absence of a gene in the sequenced taxa must have been produced by loss of genes only. Based on these assumptions the minimum number of gene loss events necessary to produce a certain pattern can easily be calculated. If a gene is present in a taxon then it must have been present in all common ancestors of that taxon with other taxa. If a gene is missing from all members of a clade, then it can be inferred that at least one gene loss event prior to the last common ancestor of that clade was necessary to produce that result. It is possible that the gene was instead lost two or more times independently later in the evolution of the clade, but the data does not provide any information that would allow us to distinguish between these scenarios. Dollo Parsimony can for that reason only produce the most conservative estimates of gene loss events.

The reconstruction of gene loss events requires the reconstruction of a last common ancestor of all taxa and therefore a rooted phylogeny. The position of the root of the Archaeplastida is still a matter of debate as detailed in the introduction and the inclusion of cyanobacterial sequences as outgroup is likely to exacerbate the problems posed by long branches and lineage specific modes of sequence evolution discussed in Section 2.5.3. Instead gene loss events were inferred for each of the three possible branching orders of the major lineages. The numbers of gene loss events do vary for the different root positions as detailed in Table 2-1 but they do in all cases support the conclusion that gene loss events are not as rare as previously assumed.

Figure 2-3 shows the number of gene loss events that were inferred for the crown groups of the phylogeny. One of the most conspicuous aspects highlighted by this figure is how the relative propensity to lose genes varies between lineages. This holds true for the major lineages as a whole as well as for relatively closely related younger lineages. The Rhodophyta, being the gene richest of the major lineages, generally show the highest numbers of gene losses per branch but the variance is equally high. The Haptophyte *Emiliana huxleyi* for example has lost 37 genes as compared to its last common ancestor with the Cryptomonads. The two Cryptomonad taxa on the other hand have both only lost seven genes. The Chlorophyta show a comparably high variance given that the group as a whole has retained fewer genes than the Rhodophyta. Here especially the deep lineages and the two secondary green algae *Euglena gracilis* and *Bigelowiella natans* stand out with high numbers of gene losses but

especially the Micromonads demonstrate the vast differences in gene content that can be possible between relatively closely related species. The Streptophyta as a group retain a similar number of genes in their plastid as compared to the Chlorophyta but tend to be more stable in their gene content. Relatively old lineages like the Zygnematales have relinquished only five or less genes since their divergence while the markedly younger Gymnosperms have lost between none and 21 genes as compared to their last common ancestor.

My reconstruction of gene losses during plastid evolution was based on several fundamental assumptions. These assumptions are based on our current understanding of Symbiogenesis but also reflect my attempt to produce conservative estimates taking into account different sources of uncertainty. The reasoning behind these assumptions is discussed in detail below.

- ***All dispensable genes were lost from the plastome before diversification***.

The present study could only consider plastid genes present in at least some extant species. The relative small set of genes present in plastids today suggests that many gene loss events in primary plastid endosymbiogenesis are likely to have occurred prior to the last common ancestor of plastids. The study can only provide insight into later stages of reductive genome evolution following ancient endosymbiosis. Our best model for plastid establishment is the amoebae genus *Paulinella,* which acquired a photosynthetic endosymbiont about 60 Mya (Nowack et al., 2008). This endosymbiont is the only known organelle of recent origin, supported by the fact that EGT to nucleus and protein import into the endosymbiont have been demonstrated (Reyes-Prieto et al., 2010). Its organelle genome is about 1 Mbp in size. This is about a third of the size of the genomes of closely related Cyanobacteria but still almost an order of magnitude bigger than most plastid genomes, indicating that the process of symbiogenesis is not yet complete. The organelle genomes of two species of the genus have been sequenced and comparisons have found differential gene transfers into the nucleus (for example *psaI* and *psaK*) and established that different genes had been lost from the host/symbiont system in the two *Paulinella* species (Reyes-Prieto et al., 2010; Nowack et al., 2011). This shows that the process of purging genes that are not necessary for the symbiotic relationship (total losses from the system, not EGT) is not finished even after the establishment of a protein import system. This type of gene loss is not the subject of this study but the finding indicates that the evolution of protein import can happen

surprisingly quickly, likely facilitated by the presence of another protein import system (mitochondrial). The existence of secondary and tertiary algae likewise shows that protein import systems are not as difficult to establish as sometimes postulated (Cavalier-Smith and Lee, 1985; Theissen and Martin, 2006; Bodył and Moszczyński, 2006; Patron et al., 2006). The findings in *Paulinella* challenge the assumption that the loss of expendable genes and the EGT of most transferable genes were completed when the lineage diversified. This makes it likely that the number of independent gene losses is considerably higher than estimated here lending even more weight to the main conclusion.

- ***Genes cannot be acquired by the plastome***

Another assumption of this study is the complete genetic isolation of plastid genomes. The possibility of horizontal gene transfer (HGT) was ruled out, even though this has been suggested to be otherwise. In one study, Rice and Palmer (2006) surveyed 204 genes across 42 plastid genomes for signs of HGT. They suggest that *rbcS* and *rpl36* in Cryptophytes and Haptophytes, might have been directly transferred to red lineages from a bacterium via homolog recombination. However, other explanations have also been provided for unexpected phylogenies concerning nuclear encoded plastid genes (e.g. Barbrook et al., 1998). Thus the evidence for later gene transfer remains weak, and was not considered further in the present analyses.

- ***Monophyletic origin of the Archaeplastida***

Even though the idea of a single common ancestor of all plastids (with the exception of *Paulinella*) is now widely favoured, it is not unequivocally supported by the data and the possibility that the origin of photosynthetic eukaryotes was more complex cannot be ruled out. The assumption of monophyly of the Plantae was made to provide a framework for the reconstruction of gene loss events but violation of this assumption is also not expected to significantly affect the main conclusions of this study. While the total numbers of gene loss events calculated here depend on this assumption, gene losses are not confined to the basal branches but are extremely common within and throughout the crown groups (see figures 2-3 and 2-4). Thus the inference that gene losses are common events during plastid evolution is robust even if major lineages originated from independent or more complex endosymbiosis scenarios.

- ***A gene lost from the plastid genome remains functional in the organism***

Parasitic species were excluded from my reconstructions of gene loss events because the marked reduction of their plastid genomes has also been accompanied by loss of photosynthetic function. Thus very significant differences in evolutionary constraint are likely to concern photosynthetic and such non-photosynthetic lineages. EGT constitutes the continuation of the functional integration of the former endosymbiont into the host organism; it is the main symptom of symbiogenesis. But is it fair to assume that all genes that have become non-functional or have disappeared from the plastid genome have indeed been transferred to the nucleus? It can only up to a point because loss of function is a phenomenon that is not limited to parasites and though 1.5 billion years of co-evolution have long purged dispensable genes from the plastid genome, the aforementioned Domino Theory of gene loss remains theoretically valid (Dagan et al., 2006). It also needs to be taken into consideration that host proteins haven functionally replaced many cyanobacterial proteins in the plastid. Millen (2001) lists the following three examples for plastid gene substitutions: *Rpl23* has been functionally replaced by a nuclear gene in spinach (Bubunenko et al., 1994; K Yamaguchi and Subramanian, 2000); the plastid *accD* locus has become a pseudogene in grasses after functional replacement by a similar cytosolic gene (Konishi et al., 1996); and in an ancestor of angiosperms and gymnosperms the plastid *rpl21* was substituted by a mitochondrial version (Martin et al., 1990).

Thus, the import of a protein into the plastid is not always preceded by the translocation of a plastid gene to the nucleus, yet it is another milestone in host/symbiont integration. The type of gene loss from the plastid genome that is the main focus of this study must therefore not be understood to simply be EGT but to also include the replacement of a plastid gene with one of different origin that is encoded outside the plastid genome.

Empirical data on the extent of EGT is somewhat sparse because for most taxa in this dataset the nuclear genome has not been sequenced. A missing gene can potentially be detected via PCR, hybridisation or the sequencing of expressed sequences tags, if the nuclear genome is not fully sequenced, but negative results are not meaningful because of a possible lack of sensitivity of these methods. In the few cases where nuclear genomic data is available, searches are usually conducted only for genes that are absent in the plastid of the species in question but present in the plastids of the closest

available relatives. In these cases the gene is generally either detected in the nucleus or the species is missing the ability that was conveyed by it. Examples include *infA*, which was lost from Angiosperm lineages several times independently and has been identified in the nucleus of at least 4 of these species (Millen et al., 2001), and the *ndh* genes which are plastid encoded in all Angiosperms except *Phalaenopsis*, where nuclear transcripts of the genes have been detected (Chang et al., 2006). For *rpl32* it has been shown in poplars and rice that the gene has in fact been transferred to the nucleus (Ueda et al., 2007). All genomes of the Prasinophyte *Ostreococcus tauri* have been sequenced and its reduced plastid genome is one of the smallest known (Robbens et al., 2007). Of seven genes (*rpl21, rpl22, rpl33, rps15, rps16, odpB*, and *ndhJ*) that were reported to have been lost at the base of the Chlorophyta lineage by Grzebyk et al. (2003), five were detected in the nuclear genome of *O. tauri*. Other genes missing from its plastid genome include *chlB, chlI, chlL* and *chlN*, all of which play a role in chlorophyll synthesis in dark. Only chlI was found in the nuclear genome while the others seem to be missing, consistent with the fact that *O. tauri* is unable to produce chlorophyll in dark (Robbens et al., 2007; Derelle et al., 2006). Seven of another 16 genes missing from the *O. tauri* plastid genome have been detected in the nucleus. Whether the complete loss of the other nine genes is related to a loss of function is not known.

The assumption that imported proteins replace all lost genes that were considered in this study is thus the weakest of all and the only one that results in a likely over-estimation of gene transfers. It was made out of necessity because not enough data is available to verify the presence of protein substitutes. However, the loss of a plastid gene would almost certainly result in a loss of function, something that is usually strongly selected against. It should therefore be relatively rare.

- ***The phylogeny reflects the true evolutionary history of plastids***

The last and probably most dubious assumption is that the phylogeny used to infer gene losses is in fact a fair representation of plastid evolutionary history. Even though a fully resolved phylogeny is absolutely crucial for the reconstruction of gene losses, the details of its topology are not. Given the pervasiveness of gene loss on all levels of the phylogeny as highlighted by Figure 2-3, swapping some branches would not signigicantly affect the order of magnitude of the resulting numbers. It should be noted

at this point that given the great number of sequenced plastid genomes, even a Maximum Parsimony tree based on the distribution of plastid genes requires a high number of gene loss events. A Dollo Parsimony analysis based on gene absence in the taxa of my dataset produces a multitude of equally parsimonious trees that are at least 744 steps long (not shown). This equates to roughly half as many gene losses as inferred by my analyses but is still a considerably high number.

2.5.5 Effects of uneven lineage sampling

Grouping the columns of the matrix according to the NCBI taxonomy highlighted the very uneven sampling of complete plastid genomes across the Archaeplastida and secondary algae. The selection of taxa in this study reflects the many research interests that motivate the sequencing of plastid genomes. Most species included here are crop plants, pests, parasites or of other economical or medical importance. Others are extremophiles or major players in the global ecosystem. Interest in the phylogenetic history and diversity of photosynthetic eukaryotes has contributed to the sequencing of the plastid genomes of many of the known main lineages though many lineages are not yet represented. There are at best rough estimates on how many species the different major lineages comprise and the taxonomic structure within them is work in progress. It is to be expected that many species remain yet undiscovered, many lineages yet unrecognized. The number of extant and extinct species is not an indicator of evolutionary diversity and it is hardly possible to adequately represent the Glaucophyta with around ten extant species when comparing them to the several hundred thousands of Streptophyta. This makes it hard, maybe impossible, to tell exactly how biased and uneven the sampling is and in conjunction with the huge variance in the propensity to lose or retain genes across lineages makes any attempt to extrapolate the findings of this study over all photosynthetic eukaryotes impractical, if not impossible. The only prediction that can be made with certainty is that the number of gene loss events that are needed to explain the patterns of gene absence across plastids can only increase as more plastid genomes are sequenced. The ratio of unique to multiple parallel gene losses is equally bound to increase. A greater sample size can only corroborate the fact that gene loss is prevalent during plastid genome evolution.

However, more sequencing will help produce a more complete and refined picture of the necessities behind the retention of a gene in the plastome. One of the most

immediate benefits of a broader sampling would be the detection of more cases where genes that are strongly conserved across all lineages are found to be missing from single, relatively recent lineages as is the case for 20 of the "unique" gene losses listed in Section 2.4.4. Only eight genes (*psaA, psaB, rpl14, rpl16, rps11, rps18, rps19, rps8*) have not been lost from any of the taxa in this analysis. This can be taken as an indication that the presence of these eight genes in the plastid genome is indispensable for photosynthetic function. However, the same conclusion would have been drawn for the 20 genes that are missing in only one species had that species by chance not been sequenced yet. These losses of genes, otherwise very strongly conserved as plastid encoded, force us to re-evaluate our ideas on how indispensible any specific gene is and can be for plastid function.

2.6 Conclusion and outlook

It was the aim of this project to systematically investigate how frequently gene losses from the plastome have occurred during plastid evolution. To do so the gene contents of 198 fully sequenced plastid genomes were surveyed and collected into a data matrix, revealing lineage specific difference in gene content. The sequences of suitable plastid encoded proteins were then used to infer gene phylogenies. To reduce the impact of non-homogeneity of evolutionary processes on the reconstruction of relationships within lineages, the gene phylogenies were reconstructed independently for the different major lineages and the resulting gene trees were then assembled into a Supertree. The Supertree showed good agreement with current understanding of the Plantae evolutionary history. The minimum number of independent gene transfers to the nucleus necessary to produce the presence/absence patterns for 259 genes were calculated using Dollo Parsimony and Maximum Likelihood methods. This was done for the three different root positions that are currently supported by data. The results show that the pattern of presence/absence of the average plastid gene was shaped by at least 5 independent gene transfers to the nucleus. The number of parallel gene losses was found to vastly outnumber that of unique losses regardless of the position of the root. Maximum Likelihood estimates suggested that the number of gene losses was in fact approximately 14% higher.

The inference that gene loss is as prevalent as my analyses show leads to the conclusion that genes are encoded in the plastid not because chance has kept them there but because their presence is maintained by specific selection pressures. Without this selection pressure genes appear to be transferred relatively quickly, as the example of *InfA* suggests. This gene has been transferred to the nucleus and become dysfunctional in the plastome at least four times independently in the Angiosperms alone (Millen et al., 2001). The rate at which genes are lost from the plastid genome can show strong lineage specific differences as well. This indicates a constant stabilising selection pressure in lineage with few gene transfers while in some other lineages changes in selections pressures result in a high rate of gene transfer from the plastid genome. From this it follows that the gene content of any plastid genome fits the organism's needs as much as possible with a mechanism that is not time reversible (i.e.

once lost, a gene cannot be recovered should it become useful again). By analog and with respect to what has been suggested in the Domino Theory of Gene Loss (Dagan et al., 2006, see general Introduction) the loss of a gene would result in a shift in constraints acting on the remaining genes. It is likely that this mechanism was involved in the evolution of the lineage specific differential gene content that is such a conspicuous feature of the gene presence/absence matrix as exemplified in Figure 2-1. The loss of a certain set of genes from the plastid genome of one lineage would have increased the selection pressure to retain other genes in the plastid genome even though the same genes were transferred the nucleus in other lineages. Conversely the pressure to retain other genes would have decreased while the same genes remained essential for plastid function in other lineages.

The nature of the selection pressures acting on the gene content of plastid genomes remains elusive. Of the different mechanisms that have been suggested to underlie the maintenance of organellar genomes, the CORR hypothesis is most consistent with the conclusions that can be drawn from our results. It requires the retention of some subunits of a protein complex as limiting factors during assembly, but generally does not place much importance on which subunit is retained. Whether CoRR is indeed the mechanism behind the retention of genes in the plastid genome or maybe one of several mechanisms requires further investigation.

Bibliography

Ahmed, I., Biggs, P.J., Matthews, P.J., Collins, L.J., Hendy, M.D. and Lockhart, P.J. (2013). Mutational dynamics of Aroid chloroplast genomes. Genome biology and evolution. **in press**:

Allen, J.F. (1992). How does protein phosphorylation regulate photosynthesis ? Trends in Biochemical Sciences. **17**:12–17.

Allen, J.F. (2003). The function of genomes in bioenergetic organelles. Philosophical transactions of the Royal Society of London. Series B, Biological sciences. **358**:19–37; discussion 37–8.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990). Basic local alignment search tool. Journal of Molecular Biology. **215**:403–410.

Baldauf, S.L. (2003). Phylogeny for the faint of heart: a tutorial. Trends in genetics : TIG. **19**:345–51.

Banks, J.A. et al. (2011). The Selaginella genome identifies genetic changes associated with the evolution of vascular plants. Science (New York, N.Y.). **332**:960–3.

Barbrook, A.C., Lockhart, P.J. and Howe, C.J. (1998). Phylogenetic analysis of plastid origins based on SecA sequences. Current Genetics. **34**:336–341.

Bhattacharya, D., Archibald, J.M., Weber, A.P.M. and Reyes-Prieto, A. (2007). How do endosymbionts become organelles? Understanding early events in plastid evolution. BioEssays : news and reviews in molecular, cellular and developmental biology. **29**:1239–46.

Bininda-Emonds, O.R.P. (2004a). Phylogenetic supertrees: combining information to reveal the tree of life. Vol. 4. O. R. P. Bininda-Emonds, ed., Springer Verlag.

Bininda-Emonds, O.R.P. (2004b). The evolution of supertrees. Trends in ecology & evolution. **19**:315–22.

Bodył, A. and Moszczyński, K. (2006). Did the peridinin plastid evolve through tertiary endosymbiosis? A hypothesis. European Journal of Phycology. **41**:435–448.

Bubunenko, M.G., Schmidt, J. and Subramanian, A.R. (1994). Protein Substitution in Chloroplast Ribosome Evolution - A Eukaryotic Cytosolic Protein has Replaced its Organelle Homologue (L23) in Spinach. Journal of molecular biology. **240**:28–41.

Buerki, S., Forest, F., Salamin, N. and Alvarez, N. (2010). Comparative Performance of Supertree Algorithms in Large Datasets Using the Soapberry Family (Sapindaceae) as a Case Study. Systematic biology. **60**:32–44.

Burke, D.H., Hearst, J.E. and Sidow, A. (1993). Early evolution of photosynthesis: clues from nitrogenase and chlorophyll iron proteins. Proceedings of the National Academy of Sciences of the United States of America. **90**:7134–8.

Burki, F., Shalchian-Tabrizi, K. and Pawlowski, J. (2008). Phylogenomics reveals a new "megagroup" including most photosynthetic eukaryotes. Biology letters. **4**:366–9.

Burleigh, J.G. and Mathews, S. (2004). Phylogenetic signal in nucleotide data from seed plants: implications for resolving the seed plant tree of life. American journal of botany. **91**:1599–613.

Castresana, J. (2000). Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Molecular biology and evolution. **17**:540–52.

Cavalier-Smith, T. (1999). Principles of protein and lipid targeting in secondary symbiogenesis: euglenoid, dinoflagellate, and sporozoan plastid origins and the eukaryote family tree. The Journal of eukaryotic microbiology. **46**:347–66.

Cavalier-Smith, T. and Lee, J.J. (1985). Protozoa as Hosts for Endosymbioses and the Conversion of Symbionts into Organelles , 2. The Journal of Eukaryotic Microbiology. **32**:376–379.

Chang, C. et al. (2006). The chloroplast genome of Phalaenopsis aphrodite (Orchidaceae): comparative analysis of evolutionary rate with that of grasses and its phylogenetic implications. Molecular biology and evolution. **23**:279–91.

Chaw, S. and Zharkikh, A. (1997). Molecular Phylogeny of Extant Gymnosperms Analysis of Nuclear 18s rRNA Sequences and Seed Plant Evolution : Molecular biology and evolution. **14**:56–68.

Chaw, S.M., Parkinson, C.L., Cheng, Y., Vincent, T.M. and Palmer, J.D. (2000). Seed plant phylogeny inferred from all three plant genomes: monophyly of extant gymnosperms and origin of Gnetales from conifers. Proceedings of the National Academy of Sciences of the United States of America. **97**:4086–91.

Criscuolo, A. (2011). morePhyML: improving the phylogenetic tree space exploration with PhyML 3. Molecular Phylogenetics and Evolution. **61**:944–948.

Dagan, T., Blekhman, R. and Graur, D. (2006). The "domino theory" of gene death: gradual and mass gene extinction events in three lineages of obligate symbiotic bacterial pathogens. Molecular biology and evolution. **23**:310–6.

Daniell, H., Wurdack, K.J., Kanagaraj, A., Lee, S.B., Saski, C. and Jansen, R.K. (2008). The complete nucleotide sequence of the cassava (Manihot esculenta) chloroplast genome and the evolution of atpF in Malpighiales: RNA editing and multiple losses of a group II intron. Theoretical and Applied Genetics. **116**:723-737.

Delmotte, F., Rispe, C., Schaber, J., Silva, F.J. and Moya, A. (2006). Tempo and mode of early gene loss in endosymbiotic bacteria from insects. BMC evolutionary biology. **6**:56.

Derelle, E. et al. (2006). Genome analysis of the smallest free-living eukaryote Ostreococcus tauri unveils many unique features. Proceedings of the National Academy of Sciences of the United States of America. **103**:11647–52.

Deschamps, P. and Moreira, D. (2008). Signal Conflicts in the Phylogeny of the Primary Photosynthetic Eukaryotes. Molecular Biology. **26**:2745-2753.

Do, C.B. and Katoh, K. (2008). Protein Multiple Sequence Alignment J. D. Thompson, M. Ueffing, and C. Schaeffer-Reiss, eds. Methods in Molecular Biology. **484**:379–413.

Dorrell, R.G. and Howe, C.J. (2011). What makes a chloroplast? Reconstructing the establishment of photosynthetic symbioses. Journal of Cell Science. **125**:1865-1875.

Farris, J.S. (1977). Phylogenetic Analysis Under Dollo's Law. Systematic biology. **26**:77–88.

Felsenstein, J. (1978). Cases in which Parsimony or Compatibility Methods will be Positively Misleading. Systematic biology. **27**:401–410.

Felsenstein, J. (2005). Phylip (Phylogeny Inference Package) version 3.6.

Finet, C., Timme, R.E., Delwiche, C.F. and Marlétaz, F. (2010). Multigene Phylogeny of the Green Lineage Reveals the Origin and Diversification of Land Plants. Current biology : CB.2217–2222.

Friedl, T. and Rybalka, N. (2012). Systematics of the Green Algae: A Brief Introduction to the Current Status U. Lüttge, W. Beyschlag, B. Büdel, and D. Francis, eds. Progess in Botany. **73**:259–280.

Gabrielson, P.W., Garbary, D.J., Sommerfeld, M.R., Townsend, R.A. and Tyler, P.L. (1990) Phylum Rhodophyta, In: Margulis, L., Corliss, J.O., Melkonian, M. and Chapman, D.J (eds.). Handbook of Protoctista: The Structure, Cultivation, Habitats and Life Histories of the Eukayotic Microorgan- isms and Their Descendants Exclusive of Animals, Plants and Fungi. Jones & Bartlett, Boston, MA, pp. 914.

Gao, L., Su, Y.-J. and Wang, T. (2010). Plastid genome sequencing, comparative genomics, and phylogenomics: Current status and prospects. Journal of Systematics and Evolution. **48**:77–93.

Goremykin, V.V. and Hellwig, F.H. (2005). Evidence for the most basal split in land plants dividing bryophyte and tracheophyte lineages. Plant Systematics and Evolution. **254**:93–103.

Gould, S.B., Waller, R.F. and McFadden, G.I. (2008). Plastid evolution. Annual review of plant biology. **59**:491–517.

Gouy, M., Guindon, S. and Gascuel, O. (2010). SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. Molecular biology and evolution. **27**:221–4.

Graham, L.D. and Wilcox, L.W. (2000). Algae. Prentice-Hall, Upper Saddle River, NJ

Gray, M.W., Burger, G. and Lang, B. (2001). Minireview The origin and early evolution of mitochondria. Genome Biology. **2**:1–5.

Griebel, T., Brinkmeyer, M. and Böcker, S. (2008). EPoS: a modular software framework for phylogenetic analysis. Bioinformatics (Oxford, England). **24**:2399–400.

Grzebyk, D., Schofield, O. and Falkowski, P.G. (2003). Minireview: The Mesozoic Radiation of Eukaryotic Algae : **267**:259–267.

Hackett, J.D., Yoon, H.S., Li, S., Reyes-Prieto, A., Rümmele, S.E. and Bhattacharya, D. (2007). Phylogenomic analysis supports the monophyly of cryptophytes and haptophytes and the association of rhizaria with chromalveolates. Molecular biology and evolution. **24**:1702–13.

Hagopian, J.C., Reis, M., Kitajima, J.P., Bhattacharya, D. and Oliveira, M.C. de (2004). Comparative analysis of the complete plastid genome sequence of the red alga Gracilaria tenuistipitata var. liui provides insights into the evolution of rhodoplasts and their relationship to other plastids. Journal of molecular evolution. **59**:464–77.

Harper, J.T., Waanders, E. and Keeling, P.J. (2005). On the monophyly of chromalveolates using a six-protein phylogeny of eukaryotes. International journal of systematic and evolutionary microbiology. **55**:487–96.

Helmchen, T.A., Bhattacharya, D. and Melkonian, M. (1995). Analyses of ribosomal RNA sequences from glaucocystophyte cyanelles provide new insights into the evolutionary relationships of plastids. Journal of molecular evolution. **41**:203–10.

Herrmann, R.G. (1997). Eukaryotism and Symbiosis H. E. A. Schenk, ed., Springer Verlag.

Hiratsuka, J. et al. (1989). The complete sequence of the rice (Oryza sativa) chloroplast genome: intermolecular recombination between distinct tRNA genes accounts for a major plastid DNA inversion during the evolution of the cereals. Mol. Gen. Genet. **217**:185–194.

Howe, C.J., Barbrook, A.C. and Lockhart, P.J. (2000). Organelle genes – do they jump or are they pushed ? Science. **16**:3–4.

Huang, C.Y., Ayliffe, M.A. and Timmis, J.N. (2003). Direct measurement of the transfer rate of chloroplast DNA into the nucleus. Nature. **422**:72–6.

Imanian, B., Pombert, J.F. and Keeling, P.J. (2010). The complete plastid genomes of the two 'dinotoms' Durinskia baltica and Kryptoperidinium foliaceum. PLoS ONE. **5**:e10711.

Janouskovec, J., Horák, A., Oborník, M., Lukes, J. and Keeling, P.J. (2010). A common red algal origin of the apicomplexan, dinoflagellate, and heterokont plastids. Proceedings of the National Academy of Sciences of the United States of America. **107**:10949–54.

Jansen, R.K. et al. (2007). Analysis of 81 genes from 64 plastid genomes resolves relationships in angiosperms and identifies genome-scale evolutionary patterns. Proceedings of the National Academy of Sciences of the United States of America. **104**:19369–74.

Karol, K.G., McCourt, R.M., Cimino, M.T. and Delwiche, C.F. (2001). The closest living relatives of land plants. Science (New York, N.Y.). **294**:2351–3.

Konishi, T., Shinohara, K., Yamada, K and Sasaki, Y. (1996). Acetyl-CoA carboxylase in higher plants: most plants other than gramineae have both the prokaryotic and the eukaryotic forms of this enzyme. Plant & cell physiology. **37**:117–22.

Larkum, A.W.D., Lockhart, P.J. and Howe, C.J. (2007). Shopping for plastids. Trends in plant science. **12**:189–95.

Leebens-Mack, J., Raubeson, L.A., Cui, L., Kuehl, J.V., Fourcade, M.H., Chumley, T.W., Boore, J.L., Jansen, R.K. and Depamphilis, C.W. (2005). Identifying the basal angiosperm node in chloroplast genome phylogenies: sampling one's way out of the Felsenstein zone. Molecular biology and evolution. **22**:1948–63.

Lewis, L.A. and McCourt, R.M. (2004). Green algae and the origin of land plants. American Journal of Botany. **91**:1535-1556.

Lockhart, P.J. and Steel, M. (2005). A Tale of Two Processes. Systematic Biology. **54**:948–951.

Lockhart, P.J, Larkum, A.W.D., Steel, M.A., Waddell, P.J. and Penny, D. (1996). Evolution of chlorophyll and bacteriochlorophyll: the problem of invariant sites in sequence analysis. Proceedings of the National Academy of Sciences of the United States of America. **93**:1930–4.

Lockhart, P.J, Novis, P., Milligan, B.G., Riden, J., Rambaut, A. and Larkum, T. (2006). Heterotachy and tree building: a case study with plastids and eubacteria. Molecular biology and evolution. **23**:40–5.

Lockhart, P.J, Steel, M.A., Hendy, M.D. and Penny, D. (1994). Recovering evolutionary trees under a more realistic model of sequence evolution. Molecular biology and evolution. **11**:605–12.

Maddison, D.R. and Maddison, W.P. (2005). MacClade 4: Analysis of phylogeny and character evolution. Version 4.08a.

Maddison, W.P. and Maddison, D.R. (2011). Mesquite: a modular system for evolutionary analysis. Version 2.75.

Margulis, L. and Fester, R. (1991). Symbiosis as a Source of Evolutionary Innovation. Sepeciation and morphogenesis., Bellagio: Massachusetts Institute of Technology Press.

Marin, B. and Melkonian, M. (2010). Molecular phylogeny and classification of the Mamiellophyceae class. nov. (Chlorophyta) based on sequence comparisons of the nuclear- and plastid-encoded rRNA operons. Protist. **161**:304–36.

Martin, W, Lagrange, T., Li, Y.F., Bisanz-Seyer, C. and Mache, R. (1990). Hypothesis for the evolutionary origin of the chloroplast ribosomal protein L21 of spinach. Current genetics. **18**:553–6.

Martin, W and Schnarrenberger, C. (1997). The evolution of the Calvin cycle from prokaryotic to eukaryotic chromosomes: a case study of functional redundancy in ancient pathways through endosymbiosis. Current genetics. **32**:1–18.

Martin, W., Rujan, T., Richly, E., Hansen, A., Cornelsen, S., Lins, T., Leister, D., Stoebe, B., Hasegawa, M. and Penny, D. (2002). Evolutionary analysis of Arabidopsis, cyanobacterial, and chloroplast genomes reveals plastid phylogeny and thousands of cyanobacterial genes in the nucleus. Proceedings of the National Academy of Sciences of the United States of America. **99**:12246–51.

Martin, W., Stöbe, B., Goremykin, V., Hansmann, S., Hasegawa, M. and Kowallik, K.V. (1998). Gene transfer to the nucleus and the evolution of chloroplasts. Nature. **393**:162–165.

Mattox, K.R. and Stewart, K.D. (1984). Classification of the green algae. A concept based on comparative cytology. In D. E. G. Irvine and D. M. John, eds. *Systematics of the green algae.* London Orlando: Academic Press, pp. 29–72.

McNeal, J.R., Arumugunathan, K., Kuehl, J.V., Boore, J.L. and Depamphilis, C.W. (2007). Systematics and plastid genome evolution of the cryptically photosynthetic parasitic plant genus Cuscuta (Convolvulaceae). BMC biology. **5**:55.

Millen, R.S. et al. (2001). Many parallel losses of infA from chloroplast DNA during angiosperm evolution with multiple independent transfers to the nucleus. The Plant cell. **13**:645–58.

Moran, N.A. and Mira, A. (2001). The process of genome shrinkage in the obligate symbiont Buchnera aphidicola. Genome biology. **2**:RESEARCH0054.

Moreira, D., Kervestin, S., Jean-Jean, O. and Philippe, H. (2002). Evolution of eukaryotic translation elongation and termination factors: variations of evolutionary rate and genetic code deviations. Molecular biology and evolution. **19**:189–200.

Nishiyama, T. et al. (2004). Chloroplast phylogeny indicates that bryophytes are monophyletic. Molecular biology and evolution. **21**:1813–9.

Nowack, E.C.M., Melkonian, M. and Glöckner, G. (2008). Chromatophore genome sequence of Paulinella sheds light on acquisition of photosynthesis by eukaryotes. Current biology : CB. **18**:410–8.

Nowack, E.C.M., Vogel, H., Groth, M., Grossman, A.R., Melkonian, M. and Glöckner, G. (2011). Endosymbiotic gene transfer and transcriptional regulation of transferred genes in Paulinella chromatophora. Molecular biology and evolution. **28**:407–22.

Nozaki, H., Ohta, N., Matsuzaki, M., Misumi, O. and Kuroiwa, T. (2003). Phylogeny of plastids based on cladistic analysis of gene loss inferred from complete plastid genome sequences. Journal of molecular evolution. **57**:377–82.

Ong, H.C. et al. (2010). Analyses of the Complete Chloroplast Genome Sequences of two Members of the Pelagophyceae : *Aureococcus anophagefference* CCMP1984. **615**:602–615.

Palmer, J.D. (1997). Organelle genomes: going, going, gone. Science. **275**:790.

Patron, N.J., Waller, R.F. and Keeling, P.J. (2006). A tertiary plastid uses genes from two endosymbionts. Journal of molecular biology. **357**:1373–82.

Pfanzagl, B., Zenker, A., Pittenauer, E., Allmaier, G., Martinez- Torrecuadrada, J., Schmid, E.R., De Pedro, M.A. and Loffelhardt, W. (1996). Primary structure of cyanelle peptidoglycan of Cyanophora paradoxa: a prokaryotic cell wall as part of an organelle envelope. Journal of Bacteriology. **178**:332–339.

Philippe, H., Snell, E.A., Bapteste, E., Lopez, P., Holland, P.W.H. and Casane, D. (2004). Phylogenomics of eukaryotes: impact of missing data on large alignments. Molecular biology and evolution. **21**:1740–52.

Phillips, M.J., Delsuc, F. and Penny, D. (2004). Genome-scale phylogeny and the detection of systematic biases. Molecular biology and evolution. **21**:1455–8.

Posada, D. (2008). ModelTest: Phylogenetic Model Averaging. Molecular biology and evolution. **25**:1253–1256.

Prechtl, J. and Maier, U.G. (2001). Zoology meets Botany: establishing intracellular organelles by endosymbiosis. Zoology (Jena, Germany). **104**:284–9.

Qiu, Y. (2008). Phylogeny and evolution of charophytic algae and land plants. **46**:287–306.

Qui, Y.-L. et al. (2007). A Nonflowering Land Plant Phylogeny Inferred from Nucleotide Sequences of Seven Chloroplast , Mitochondrial , and Nuclear Genes Reviewed. International Journal of Plant Sciences. **168**:691–708.

Race, H.L., Herrmann, R.G. and Martin, W. (1999). Why have organelles retained genomes? TIG. **15**:364.

Raubeson, L.A., Peery, R., Chumley, T.W., Dziubek, C., Fourcade, H.M., Boore, J.L. and Jansen, R.K. (2007). Comparative chloroplast genomics: analyses including newsequences fromthe angiosperms Nuphar advena and Ranunculusmacranthus. BMC Genomics. **8**:174.

Revill, M.J.W., Stanley, S. and Hibberd, J.M. (2005). Plastid genome structure and loss of photosynthetic ability in the parasitic genus Cuscuta. Journal of experimental botany. **56**:2477–86.

Reyes-Prieto, A. and Bhattacharya, D. (2007). Phylogeny of nuclear-encoded plastid-targeted proteins supports an early divergence of glaucophytes within Plantae. Molecular biology and evolution. **24**:2358–61.

Reyes-Prieto, A., Yoon, H.S., Moustafa, A., Yang, E.C., Andersen, R.A., Boo, S.M., Nakayama, T., Ishida, K. and Bhattacharya, D. (2010). Differential gene retention in plastids of common recent origin. Molecular biology and evolution. **27**:1530–7.

Rice, D.W. and 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 biology. **4**:31.

Richly, E. and Leister, D. (2004). NUMTs in sequenced eukaryotic genomes. Molecular biology and evolution. **21**:1081–4.

Riisberg, I., Orr, R.J.S., Kluge, R., Shalchian-Tabrizi, K., Bowers, H.A., Patil, V., Edvardsen, B. and Jakobsen, K.S. (2009). Seven gene phylogeny of heterokonts. Protist. **160**:191–204.

Robbens, S., Derelle, E., Ferraz, C., Wuyts, J., Moreau, H. and Van de Peer, Y. (2007). The complete chloroplast and mitochondrial DNA sequence of Ostreococcus tauri: organelle genomes of the smallest eukaryote are examples of compaction. Molecular biology and evolution. **24**:956–68.

Rodríguez-Ezpeleta, N., Brinkmann, H., Burey, S.C., Roure, B., Burger, G., Löffelhardt, W., Bohnert, H.J., Philippe, H. and Lang, B.F. (2005). Monophyly of primary photosynthetic eukaryotes: green plants, red algae, and glaucophytes. Current biology : CB. **15**:1325–30.

Rodríguez-Ezpeleta, N., Philippe, H., Brinkmann, H., Becker, B. and Melkonian, M. (2007). Phylogenetic analyses of nuclear, mitochondrial, and plastid multigene datasets support the placement of Mesostigma in the Streptophyta. Molecular biology and evolution. **24**:723–31.

Rogers, M.B., Gilson, P.R., Su, V., McFadden, G.I. and 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.

Sato, S., Nakamura, Y., Kaneko, T., Asamizu, E. and Tabata, S. (1999). Complete structure of the chloroplast genome of Arabidopsis thaliana. DNA Research. **6**:283–290.

Sayers, E. et al. (2009). Database resources of the National Center for Biotechnology Information. Nucleic Acids Research. **D5**:15.

Schmitz-Linneweber, C., Maier, R.M., Alcaraz, J.P., Cottet, A., Herrmann, R.G. and Mache, R. (2001). The plastid chromosome of spinach (Spinacia oleracea): complete nucleotide sequence and gene organization. Plant Molecular Biology **45**:307–315.

Schneider, H., Smith, A.R. and Pryer, K.M. (2009). Is Morphology Really at Odds with Molecules in Estimating Fern Phylogeny ? Is Morphology Really at Odds with Molecules in Estimating Fern Phylogeny ? **34**:455–475.

Schopf, J.W. (1999). Deep divisions in the Tree of Life--what does the fossil record reveal? The Biological bulletin. **196**:351–3; discussion 354–5.

Schopf, J.W. (2011). The paleobiological record of photosynthesis. Photosynthesis research. **107**:87–101.

Shanker, A., Sharma, V. and Daniell, H. (2011). Phylogenomic evidence of bryophytes' monophyly using complete and incomplete datasets from chloroplast proteomes. Journal of Plant Biochemistry and Biotechnology. **20**:288–292.

Soltis, D.E. et al. (2011). Angiosperm phylogeny: 17 genes, 640 taxa. American journal of botany. **98**:704–30.

Steane, D.A. (2005). Complete nucleotide sequence of the chloroplast genome from the Tasmanian blue gum, Eucalyptus globulus (Myrtaceae). DNA Research. **12**:215–220.

Subramanian, A.R., Kaufmann, M. and Morgenstern, B. (2008). DIALIGN-TX: greedy and progressive approaches for segment-based multiple sequence alignment. Algorithms for molecular biology : AMB. **3**:6.

Tangphatsornruang, S., Uthaipaisanwong, P., Sangsrakru, D., Chanprasert, J., Yoocha, T., Jomchai, N. and Tragoonrung, S. (2011). Characterization of the complete chloroplast genome of Hevea brasiliensis reveals genome rearrangement, RNA editing sites and phylogenetic relationships. Gene. **475**:104–12.

Theissen, U. and Martin, W. (2006). Correspondences The difference between organelles and Response to Theissen and Martin. Current Biology. **16**:1016–1017.

Thorsness, P.E. and Fox, T.D. (1990). Escape of DNA from mitochondria to the nucleus in Saccharomyces cerevisiae. Nature. **346**:376–379.

Timmis, J.N., Ayliffe, M.A., Huang, C.Y. and Martin, W. (2004). Endosymbiotic gene transfer: organelle genomes forge eukaryotic chromosomes. Nature reviews. Genetics. **5**:123–35.

Turmel, M., Brouard, J., Gagnon, C., Otis, C. and Lemieux, C. (2008). Deep Division in the Chlorophyceae (Chlorophyta) Revealed by Chloroplast Phylogenomic Analyses. Journal of Phycology. **44**:739–750.

Turmel, M., Gagnon, M.-C., O'Kelly, C.J., Otis, C. and 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**:631–48.

Turmel, M., Otis, C. and Lemieux, C. (2006). The chloroplast genome sequence of Chara vulgaris sheds new light into the closest green algal relatives of land plants. Molecular biology and evolution. **23**:1324–38.

Turmel, M., Otis, C. and Lemieux, C. (2009). The chloroplast genomes of the green algae Pedinomonas minor, Parachlorella kessleri, and Oocystis solitaria reveal a shared

ancestry between the Pedinomonadales and Chlorellales. Molecular biology and evolution. **26**:2317–31.

Turmel, M., Pombert, J.-F., Charlebois, P., Otis, C. and Lemieux, C. (2007). The Green Algal Ancestry of Land Plants as Revealed by the Chloroplast Genome. International Journal of Plant Sciences. **168**:679–689.

Ueda, M., Fujimoto, M., Arimura, S., Murata, J., Tsutsumi, N. and Kadowaki, K. (2007). Loss of the rpl32 gene from the chloroplast genome and subsequent acquisition of a preexisting transit peptide within the nuclear gene in Populus. Gene. **402**:51–6.

Wu, C.-S., Wang, Y.-N., Hsu, C.-Y., Lin, C.-P. and Chaw, S.-M. (2011). Loss of different inverted repeat copies from the chloroplast genomes of Pinaceae and cupressophytes and influence of heterotachy on the evaluation of gymnosperm phylogeny. Genome biology and evolution. **3**:1284–95.

Yamaguchi, K. and Subramanian, A.R. (2000). The plastid ribosomal proteins. Identification of all the proteins in the 50 S subunit of an organelle ribosome (chloroplast). The Journal of biological chemistry. **275**:28466–82.

Yang, Z. and Rannala, B. (2012). Molecular phylogenetics: principles and practice. Nature reviews. Genetics. **13**:303–14.

Yoon, H.S., Hackett, J.D., Ciniglia, C., Pinto, G. and Bhattacharya, D. (2004). A molecular timeline for the origin of photosynthetic eukaryotes. Molecular biology and evolution. **21**:809–18.

Yoon, H.S., Zuccarello, G.C. and Bhattacharya, D. (2010). Evolutionary History and Taxonomy of Red Algae. In *Red Algae in the Genomic Age*. pp. 25–42.

Zhong, B., Deusch, O., Goremykin, V.V., Penny, D., Biggs, P.J., Atherton, R.A., Nikiforova, S.V. and Lockhart, P.J. (2011). Systematic error in seed plant phylogenomics. Genome biology and evolution. **3**:1340–8.

Zuccarello, G. C., Price, N., Verbruggen, H. and Leliaert, F. (2009). Analysis of a Plastid Multigene Dataset and the Phylogenetic Position of the Marine Macroalga Caulerpa Filiformis (Chlorophyta). Journal of Phycology. **45**:1206–1212.

3 Sequencing the genome of the spheroid body of *Rhopalodia gibba*

3.1 Abstract

The diatom *Rhopalodia gibba* harbours a nitrogen-fixing endosymbiont of cyanobacterial origin, the spheroid body. Limited sequencing work on its genome has indicated extensive gene losses while the genes involved in nitrogen fixation seem to be highly conserved. This, and the observation of major physiological changes that integrate the endosymbiont into the host system, led to the suggestion that the spheroid body might in fact be in the process of becoming an organelle. To further investigate this endosymbiosis, a fosmid library for the spheroid body's genome was sequenced using Illumina sequencing technology. The aim was to determine the suitability of the new sequencing technology for this kind of sample, to compare the fidelity of the sequence produced by it with that of sequences produced with the traditional Sanger sequencing method and to obtain novel spheroid body genomic data.

A pooled sample of 165 fosmid inserts was sequenced. An iterative assembly strategy was used to produce 2.87 Mbp of assembled sequence. The high quality of the assembled data was confirmed using multiple assessment methods. BLAST similarity searches in conjunction with analyses in MEGAN were successfully used to segregate putative spheroid body genomic sequences and diatom sequences from bacterial contaminations and to taxonomically classify these contaminants. The data produced in this project showed that fosmid libraries can be successfully sequenced using high throughput sequencing, though *de novo* assembly in the presence of bacterial contaminants is challenging. The sequences determined provide a genome resource and contribution towards further study of the diatom-spheroid body endosymbiotic relationship.

3.2 Introduction

This project was undertaken as part of an international cooperation with the cell biology research group at the Philips Universität, Marburg. It investigates genome evolution in the endosymbiotic relationship between the diatom *Rhopalodia gibba* and its cyanobacterial endosymbiont, the so-called spheroid body.

Symbiotic relationships have always played an essential role in eukaryote evolution, most prominently with the symbiogenesis of mitochondria and plastids. These milestones in the evolution of eukaryotic life lie too deep in time to allow a detailed reconstruction of the processes leading to organelle formation based on modern organelles, and instead modern symbioses are studied to shed light on these ancient events. The very close phylogenetic relationship between the intracellular spheroid bodies and an extant and fully sequenced cyanobacterium makes the *R. gibba* symbiosis an excellent model system.

3.2.1 *Rhopalodia gibba* and its spheroid body

Rhopalodia gibba is a pennate diatom of the family Rhopalodiaceae, first described by O. F. Müller in 1895 in freshwater samples from East Africa and Germany (Jahn, 1996). It is a motile pennate diatom that occurs adhered to the surface of water plants and soil in freshwater habitats worldwide (Geitler, 1977). Several ellipsoidal bodies are usually present in the central part of the cell in addition to the nucleus. They are enclosed by a double membrane and contain thylakoids, characteristic membrane structures that are the site of photosynthesis in cyanobacteria and plastids. These "spheroid bodies" were first described by Pfitzer in 1869 and can be found in all species of the genera *Epithemia* and *Rhopalodia* (Pfitzer, 1869).

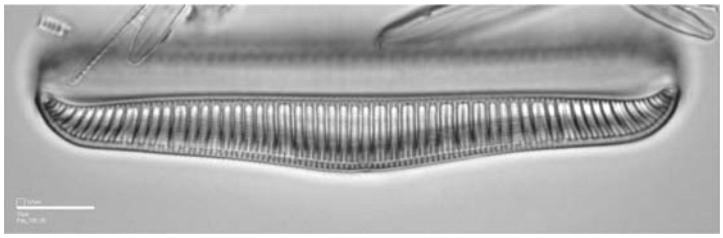


Figure 3-1 *Rhopalodia gibba*

Valve view under brightfield light microscopy (Droop, Sims et al. 1993)

Floener and Bothe (1980) later demonstrated via an acetylene reduction test that *Rhopalodia gibba* has the capacity to fix nitrogen. The bio-fixation of nitrogen is a feature known only in prokaryotes. Given this and the morphological features of the spheroid body and the fact that it contained DNA, they suggested that it might in fact be an endosymbiotic cyanobacterium. Phylogenetic analyses of 16S rRNA and *nifD* genes have since demonstrated that the spheroid body is indeed a close relative of the cyanobacterial strain *Cyanothece sp.* ATCC 51142 (Prechtl et al., 2004). In phylogenetic reconstructions of both genes, the branch lengths separating free-living cyanobacteria and the cell inclusions of *R. gibba* are very short, indicating that the origins of these symbioses are relatively recent. This is in contrast to the situation for plastids and extant cyanobacteria, which have an ancient phylogenetic relationship (Kneip et al., 2007).

3.2.2 The cyanobacterium *Cyanothece sp*. ATCC 51142

The members of the genus *Cyanothece* are marine, unicellular cyanobacteria that are able to fix nitogen under aerobic conditions (Reddy et al., 1993). The 5.4 Mbp genome of *Cyanothece sp*. ATCC 51142 (from now on referred to as *Cyanothece* 51142) was the first unicellular diazotrophic cyanobacterium to be completely sequenced (Welsh et al., 2008). The strain provides an experimental system to study the relationship between nitrogen fixation and oxygenic photosynthesis in unicellular cyanobacteria. In *Cyanothece* 51142 these antagonistic processes are accommodated by temporal separation (Colón-López et al., 1997). Cyanobacteria remain the only prokaryotes known to have a circadian lifestyle (Kondo et al., 1997; Woelfle et al., 2004).

Morphological analyses, genome and expression studies on *Cyanothece* 51142 have revealed this temporal separation is achieved by major metabolic changes under strict genetic control. Photosynthetic activity and related gene expression peaks during the light phase. The resulting carbohydrates are stored in the form of large glycogen granules between the thylakoid membranes. With the beginning of the dark phase the photosynthetic machinery is rearranged to facilitate ATP synthesis during nitrogen fixation (Schneegurt et al., 1994; Schneegurt et al., 1997). The expression of the nitrogenase enzyme complex peaks in the first half of the dark phase (Wittenberg et al., 1972; Hill et al., 1996). The degradation of the stored glycogen through glycolysis and

respiration provides the necessary ATP for energy intensive N_2 fixation (Colón-López et al., 1997; Stöckel et al., 2008).

A tight control of gene expression, the ability to adjust the cellular machinery, as well as the ability to synthesize, store, degrade, and use large quantities of storage products are all required to allow for the fixation of nitrogen in the same cellular compartment in which photosynthesis takes place. *Cyanothece* 51142 evolved these features out of the need to reconcile these two processes. With a symbiotic host that can alleviate the endosymbiont from the need to photosynthesise, such complex and costly measures would become obsolete and the genes responsible lost during reductive genome evolution.

3.2.3 The spheroid body of *Rhopalodia gibba*

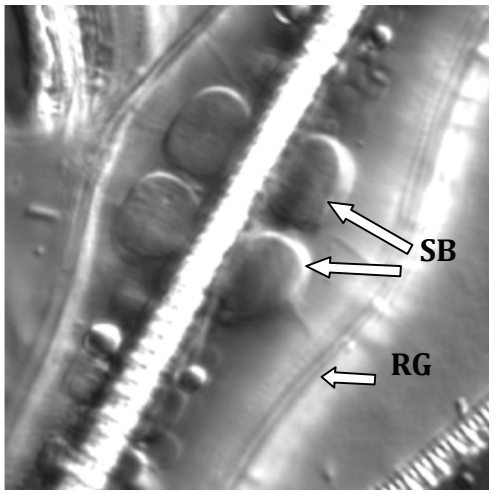


Figure 3-2 The spheroid body of *Rhopalodia gibba*

SB: spheroid body; RG *Rhopalodia gibba* frustule (Kneip 2002)

In *Cyanothece sp.* nitrogen fixation is temporally separated from the photosynthetic light reactions but in the spheroid body of *R. gibba* nitrogen fixation is a strictly light dependent process, where photosynthesis and nitrogen fixation are now spatially separated. This reversed photoperiodicity is likely the result of adaptations to the endosymbiotic lifestyle. Spheroid bodies (SB) are not photosynthetically active as they lack photosynthetic pigments, as well as certain photosynthesis related genes (Prechtl et al., 2004). Photosynthesis is the main source of energy for cyanobacteria. Its loss equals a loss of autonomy for the bacterium, turning this relationship into an obligate

symbiosis, as the endosymbiont is now dependent on the import of energy-rich molecules from the host for housekeeping and nitrogen fixation. Consistent with a permanent nature of the endosymbiosis, the spheroid bodies were found to be unable to grow outside the diatom host (Kneip, 2004).

A comparison of the sequenced regions of *R. gibba's* spheroid body genome with cyanobacterial genomes revealed reductive genome evolution in the endosymbiont (Kneip et al., 2008). Gene losses by deletion or replacement with AT rich non-coding DNA were observed, in addition to accumulation of deleterious mutations in genes for cell wall biosynthesis and transposase controlled processes (Kneip et al., 2008).

Taken together, this evidence suggests that the spheroid bodies are vertically transmitted permanent endosymbionts in the transition from a 'true' free-living cyanobacterium to a nitrogen-fixing eukaryotic organelle (Prechtl et al., 2004).

To establish how far it has progressed on this path, three main questions need to be answered: (i) What is the extent of gene loss in the symbiont's genome? (ii) Have any genes been transferred into the diatom's nuclear genome? (iii) Has a protein import system been established into the endosymbiont?

To answer the first question and help answer the other two it is necessary to fully sequence the genome of the endosymbiont. This is a challenging task, not because of the expected size of the endosymbiont's genome but because of difficulties in isolating it. *Rhopalodia gibba* like many other diatoms and algae live in obligate association with bacteria in its extracellular matrix (Kawai et al., 2005; Lorenz et al., 2005). These bacteria will inevitably contaminate any extraction of endosymbiont DNA (Uwe Maier, pers. com.). To separate endosymbiont DNA from contaminations, a fosmid library of the genome was constructed by Kneip (2004) and candidate endosymbiont sequences identified via Sanger sequencing of the insert ends. Even though only a fraction of the fosmid library was screened this way, 165 inserts of 30 to 40 kbp length were identified as sequencing targets. It was decided to sequence these inserts with the traditional Sanger sequencing technology as well as the newly introduced Illumina Genome Analyser using Solexa sequencing chemistry. The goal was to establish whether the new technology was suitable for the task of sequencing a fosmid library and to compare the fidelity of the two sequencing methods.

3.3 Methods

3.3.1 Illumina sequencing

This project used genetic material from *R. gibba* cultures cultivated at the Phillips Universität in Marburg. The workgroup under Prof. Uwe Maier kindly provided 165 extractions of fosmids with about 40kb long inserts, isolated from a fosmid library of spheroid body DNA. The fosmid library was constructed by Christoph Kneip (2004), using the CopyControlTM Fosmid Library Production Kit (Epicentre). The fosmids were chosen from the library after Sanger sequencing of one end of the insert suggested that they were not derived from bacterial contaminants. This inference was based on the GC composition of the sequenced fragments.

Illumina sequencing was carried out by the Massey University Genome Service in Palmerston North, New Zealand.

To prepare the samples for Illumina sequencing the DNA concentration of each of the 165 fosmid extractions was measured with a Bioanalyzer NanoDrop spectrometer. The DNA extractions were then pooled at similar concentrations into two samples. The first sample comprised 155 fosmids at a total amount of 37.5 µg of DNA in 500µL. The second sample comprised the remaining ten fosmids with a total amount of 5 µg in 50 µL. The two samples were indexed and both run on the same lane of the flow cell in a 75 bp paired-end sequencing run

3.3.2 Assembly

Reads with both indices were combined for assemblies. The reads were quality trimmed with the DynamicTrim script from the SolexaQA toolkit (Cox et al., 2010) because test assemblies showed an improved performance on trimmed reads. Reads were assembled using the software package Velvet 0.7 (Zerbino and Birney, 2008).

The perl script Velvet3parameterTest.pl (see Appendix A) was used to make systematic parameter sweeps of the assemblies and select the most suitable combinations of the minimum coverage cutoff, k-mer length and expected coverage parameters.

Assemblies used all reads that passed the quality filters. The Sanger sequences of fosmid insert ends supplied by the Prof. Uwe Maier workgroup served as scaffolds in some assemblies. In a repetitive assembly strategy, contigs produced by previous assemblies were used as scaffolds for subsequent assemblies with changed parameters.

3.3.2.1 Workflow

Following each step the reads were mapped to the contigs in BWA to remove misassemblies and verify that the resulting sequences were still in agreement with the reads.

The contigs of two assemblies with k-mers of 23 and 55 respectively were combined into lager contigs by aligning them with the Sequencher sequence analysis software version 4.9 (Gene Codes Corporation). The resulting contigs were mapped in BWA and reads that didn't map were used in another assembly. The results were added to the Sequencher project. Based on the results of local BLASTn searches, CG content was used to identify bacterial contaminations in the resulting contigs. Another local BLASTn search against all *Cyanothece* sequences available in GenBank was conducted to identify potential spheroid body sequences.

A final assembly was carried out with an intermediate k-mer value of 41. In this assembly, sequences produced by the workflow above that showed significant similarity to *Cyanothece* sequences were used for scaffolding (i. e. to guide the assembly). The assembly was carried out on a set of reads from which reads that mapped contaminant sequences had been removed. The contigs produced by this assembly were used as query sequences in a BLASTx search against the NCBI nr (non redundant proteins) database to identify the taxonomic classification of the source organism. The BLAST results were analysed in MEGAN (Huson et al., 2007).

3.3.3 Contig assembly and sequence comparison

Contigs produced by different assemblies were combined with the long read sequence assembly feature in the Sequencher sequence analysis software version 4.9 (Gene Codes Corporation). The alignments within all contig clusters were hand edited and then used to produce consensus sequences.

Sequences produced with Sanger shotgun sequencing of the same templates were kindly supplied by the workgroup of Prof. Uwe Maier. These sequences amounted to 1.9 Mbp in 663 contigs. These sequences were aligned to the Illumina sequencing results using the same sequence assembly feature in Sequencher to produce a spheroid body genome assembly from all available data.

The sequence alignment algorithm Exonerate version 2.2 was used to produce sequence identity statistics (Slater and Birney, 2005). The Sanger reads were assembled into contigs in Sequencher where possible and used as query sequences in an alignment against the consensus sequences produced from the combined short read assemblies. Exonerates 'roll your own' output options were used to produce statistics on alignment length and sequence identity.

3.3.4 Mapping and editing

The quality filtered reads were mapped to the consensus sequences of contig alignments produced in Sequencher using the Bowtie and the BWA algorithms (Langmead et al., 2009; Li and Durbin, 2009). The mapping results in SAM format were processed with a perl script (SAMtrimmer.pl, see Appendix A) to remove entries for reads that did not map. The resulting SAM files were then converted into sorted bam files using Samtools 0.1.18. Mapping results were visualised in Tablet (Milne et al., 2010). The contigs were then hand edited in Sequencher according to mapping results to remove misassemblies, and close or shorten gaps.

3.3.5 Identification of bacterial contaminations

BLAST searches were performed on the NCBI BLAST website and locally on the AWC servers at Massey University (Altschul et al., 1990). The code for local BLAST as well as preformatted GenBank databases were obtained from the NCBI BLAST website. Custom BLAST databases were built from sequences obtained from GenBank.
BLAST searches were facilitated by the use of the Blast2GO software (Conesa et al., 2005). BLAST results were analysed with MEGAN (Huson et al., 2007).

CG contents of contig sequences were analysed in Artemis and Tablet (Rutherford et al., 2000; Milne et al., 2010).

3.4 Results

3.4.1 Illumina sequencing

Illumina sequencing of the indexed and pooled fosmid inserts produced 1620.47 Mbp of data in 10 803 145 pairs of 75 bp reads. The indices indicate that of these, 10 538 587 read pairs comprising 1580.79 Mbp were derived from 155 pooled fosmid inserts. The pooled sample of 10 fosmid inserts produced 39.68 Mbp of sequence in 264 546 read pairs.

After trimming low quality sequence from the ends of the reads with the DynamicTrim tool a total of 1370.73 Mbp of sequence data remained.

3.4.2 Assembly

The quality trimmed paired-end reads were assembled in Velvet following the workflow described in Section 3.3.2.1. The parameters used and the basic statistics of the assemblies are listed in Table 3-1. The different assemblies are identified by letters for convenience. The assemblies A and B were performed on the complete set of quality trimmed paired-reads. Those reads that were not used in A or B were separately assembled in assembly C. The contigs produced by assemblies A, B and C were combined by alignment in Sequencher. Potential spheroid body sequences and sequences derived from contaminants were identified based on BLAST results and GC content (for details see Section 3.4.4). For the final assemblies D and E reads that mapped to contaminant sequences were removed from the read set and putative spheroid body sequences were used as scaffolds to guide the assemblies.

The statistics given in Table 3-1 show that combining assemblies with different k-mer values and iterative assemblies that use contigs produced by previous assemblies do indeed produce more assembled sequence. Though the n50 value is less than half that in the initial assemblies A and B, the maximum contig length and total amount of assembled sequence were significantly increased in D and E.

Table 3-1 Parameters and statistics of Velvet assemblies

The top half of the table details the parameters used in five Velvet assemblies labelled A to E. The bottom half of the table gives the basic assembly statistics for the resulting sequences. Exp_cov: expected coverage, cov_cutoff: coverage cutoff, nodes: number of nodes in the assembly graph, n50: contigs longer than this value amount to 50% of assembled sequence, max: maximum contig length, total: total assembled sequence in contigs longer than 100 bps.

	A	B	C	D	E
k-mer	55	23	55	41	41
exp_cov	150	150	20	15	100
cov_cutoff	20	30	10	7	7
nodes	361	2 030	435	2 327	1 672
n50 (in bps)	38 233	33 245	1 137	14 604	17 028
max (in bps)	86 846	88 313	7 074	104 714	104 714
total (in bps)	2 182 205	3 386 120	120 225	4 247 334	4 257 268

The contigs produced by the final assemblies D and E were again combined in Sequencher and contigs longer than 500 bps checked for misassemblies. Overall 604 contigs longer than 500 bps were assembled, comprising a total of 4.241 Mbp of sequence.

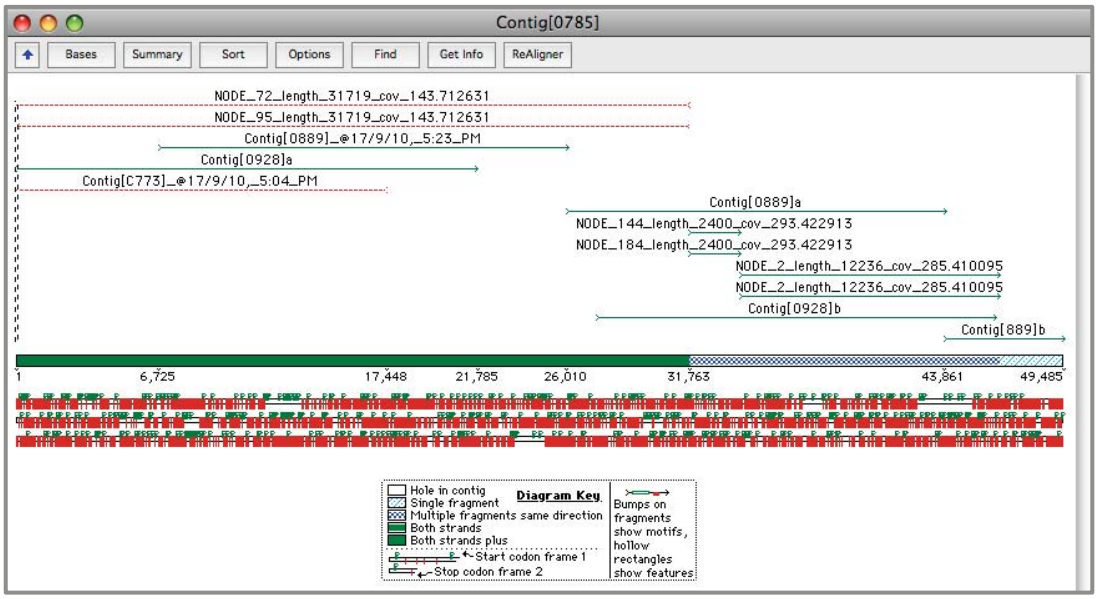


Figure 3-3 Example of a long sequence assembly for contigs from different Velvet assemblies

Screenshot of an assembly by alignment in Sequencher. Contigs labelled NODE and Contig respectively were produced in different assemblies.

Figure 3-3 shows an alignment of contigs in Sequencher. The sequences labelled as NODE or Contig respectively were assembled with different k-mer values. The example illustrates that sequence regions that assemble well under one k-mer value produce fragmented assemblies under other values.

3.4.3 Identification of misassemblies by mapping

Mappings were used extensively to ensure the quality of the assembled sequences and to verify "supercontigs" that had been produced based on short overlaps in Sequencher alignments of contigs. The following screenshots of mappings show examples of the kind of assembly errors that can be identified and corrected based on mappings. Figure 3-4 shows a misassembled repeat region where the 6 bp repeat unit was inferred to be present eighteen instead of nine times in the assembly. The repeat region is in this case bridged completely by reads but was nonetheless misassembled because the assembly used a k-mer shorter than the repeat region and incorrect coverage calculations.

Figure 3-4 Reads mapped to a misassembled repeat region, visualised in Tablet

Reads aligning to a 6 bp repeat region in a contig are displayed in packed mode. Each column represents nucleotides that were aligned to a position in the reference sequence. The nucleotide base is indicated by colour. Nucleotides that do not match the reference sequence are indicated in lighter colours. The mapping reveals that this repeat region was assembled incorrectly with too many repeats. The white ovals indicate base pairs at the ends of reads that extend beyond the repeat region and the correct position where these bases should align.

Figure 3-5 shows another misassembly. The sequences on both sides were joined based on paired-end positional information but the gap was filled with incorrectly assembled sequence that does not align well to any reads and doesn't match the read overhang from properly aligned reads to the right.

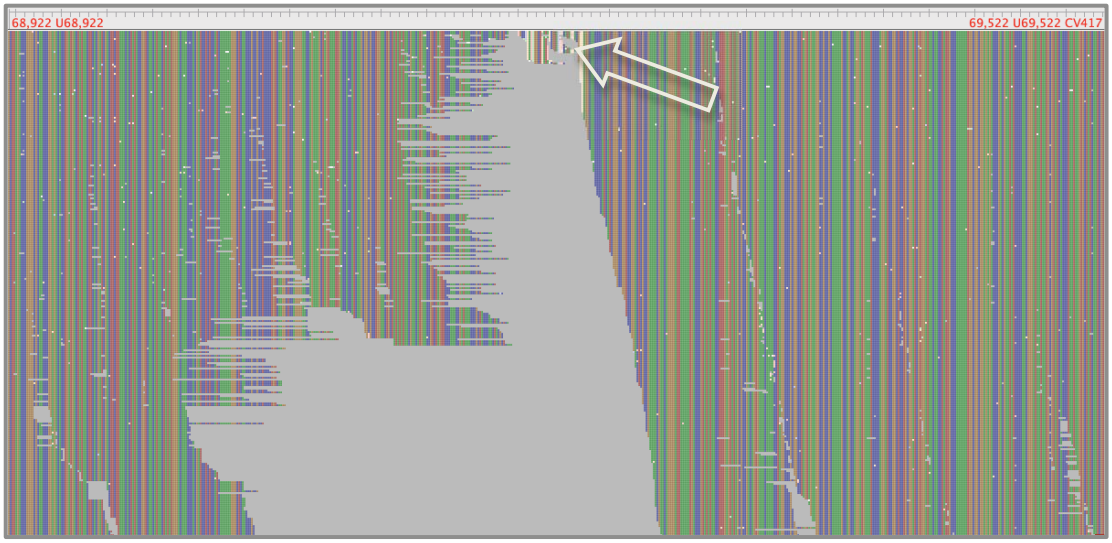


Figure 3-5 Example of a misassembly

Reads that align to the reference sequence are shown in packed mode in Tablet. Nucleotides that do not match the reference sequence are indicated in lighter colours. The arrow points out a short region where no reads align correctly, flanked by regions of high coverage.

Another common type of misassembly were single erroneous bases that were not supported by the reads. Velvet's scaffolding function introduces N-stretches into contig sequences to preserve paired-end positional information. A small number of reads usually align on the boundaries of N-stretches, with part of the read continuing into the unassembled region. In most cases these reads show strong sequence agreement and their consensus sequence was used to resolve the sequence at the edges of these N-stretches.

3.4.4 Assembly quality

3.4.4.1 Test for chimeric sequences

For the purpose of quality control of the assembly, a subset of ten fosmid extractions had been indexed differently to the remaining 155 fosmid preparations. To test if the

reads derived from different template sequences had been mixed up during the assembly, the subset of reads was mapped against two samples of contigs produced by the final assembly. The first comprised a random selection of 70 contigs ranging in size from 500 bp to 2 kb, the second comprised 32 fully assembled fosmid inserts or several overlapping fosmid inserts that were identified as putative spheroid body sequences based on their similarity to Cyanothece sequences.

Figure 3-6 Example of a coverage peak of reads derived from a subsample of fosmid inserts

The coverage along the entire reference sequence is indicated in blue at the top. The red box indicates the sequence region that is displayed in detail at the bottom of the figure. The middle panel shows the coverage for the sequence region below. Reads that align to the reference sequence are shown in packed mode in the bottom part of the figure. Mismatches between reads and reference sequence are indicated in lighter colours.

Most of the random sample of 70 contigs only aligned to few reads, scattered along the full length of the reference sequence. These isolated reads were most likely produced by sequencing errors in conjunction with chance sequence similarity. For three contigs short regions of relatively high coverage were observed. An example is shown in Figure 3-6. These high coverage regions showed consistent mismatches between reads and reference sequences in all cases. Reads in the subset only aligned to six of these 70 contigs with a significant coverage. The mappings showed an even coverage across the entire contig sequences.

The mapping of the read subset against the full-length fosmid inserts produced very similar results. Most of these 32 contigs only very infrequently aligned to single or few

reads as can be expected based on random sequence similarities. However, the reads did align to four of the contigs across their entire length and with an even coverage that was sufficiently high for assembly, suggesting that these contigs had been assembled from these reads while the other 28 contigs had been assembled without input from

Table 3-2 Statistics of BWA mapping of indexed subset of reads to completely assembled fosmid inserts

The average coverage was calculated assuming an average read length of 70 bps.

Contig ID	Contig length in bps	Number of Reads	Average Coverage
2	35544	455	0.90
5	81274	1005	0.87
8	67314	904	0.94
9	38132	32115	58.95
12	37259	695	1.31
15	59935	671	0.78
21	56789	657	0.81
23	37285	401	0.75
25	42113	494	0.82
30	57006	705	0.87
39	39505	17542	31.08
49	41628	490	0.82
50	38135	43073	79.06
51	36152	406	0.79
2503	38328	432	0.79
2810	38127	1021	1.87
2847	150963	1799	0.83
2850	35211	828	1.65
2883	104680	1284	0.86
2888	91490	1001	0.77
2891	93015	8196	6.17
2894	84971	1079	0.89
2895	75793	852	0.79
2906	44794	560	0.88
2910	36595	411	0.79
2911	35979	473	0.92
2912	101401	1193	0.82
2926	61931	667	0.75
2927	37098	500	0.94
2957	102924	1357	0.92
2980	75791	953	0.88
3000	48117	599	0.87

this subset of reads. The average coverage for the four contigs that were assembled from these reads ranged from 15 to 79, illustrating the differences in coverage that resulted from the experimental setup. Table 3-2 lists the average coverage for contig

2891 as 6.17 instead of 15. The reason for this is that the reads only mapped to a 38484 bps long region in the middle of the contig (see Figure 3-7). The table lists the average coverage along the entire contig length while the actual coverage within that fosmid insert size region was close to 15.

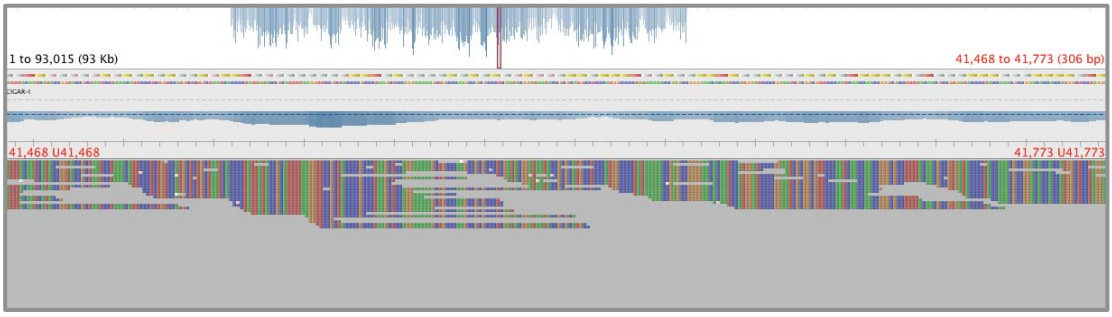


Figure 3-7 Mapping of the indexed subset of reads to Contig 2891

Screenshot of a mapping visualised in Tablet. The coverage across the contig is indicated in blue at the top, indicating that reads align only to a clearly delimited 40kb long region in the middle of the contig. The bottom sections shows a detail of the mapped reads in packed view. Miss-matches are indicated in white.

3.4.4.2 Comparison to Sanger sequences

Sanger sequences were aligned to contigs produced by short read assemblies as described in Section 3.3.2.1 to compare the quality of the sequences. Table 3-3 lists the average sequence identities of alignments within the given length categories. Alignments longer than 1000 bps were generally produced from Sanger sequences that were assembled into contigs, while the majority of shorter alignments were produced from single Sanger sequences. Alignments shorter than 500 bps were excluded. Only the best alignment of a Sanger sequence against any of the contigs assembled from short reads was included.

Table 3-3 Average sequence identity between Sanger sequences and short read assemblies by alignment length.

Alignment length category	Number of alignments	Average sequence identity
> 5000 bps	50	99.97%
1500 – 4999 bps	140	99.88%
1000 – 1499 bps	104	99.25%
500 – 999 bps	230	98.55%

3.4.5 Identification of putative spheroid body sequences and contaminants

Preliminary local BLASTn searches suggested that sequences that showed strongest similarities to *Cyanothece* sequences were consistently GC-poor with a GC content of only ~35%. Sequences characterised by a markedly higher GC content of ~60% consistently showed strong similarities to other types of bacteria. Following the initial three assemblies, contigs with a GC content of 55% or higher were excluded from subsequent analyses. The final assembly was in this way split up in 98 contigs of GC-rich sequences, comprising more than 1.82 Mbp, and 506 contigs of GC-poor sequence, comprising 2.87 Mbp.

A local BLASTx search against the nr protein database confirmed that the GC-rich sequences were indeed most likely to be derived from bacterial contaminants. The analysis of these BLAST results in MEGAN showed that the taxonomic composition of the BLAST hits was very similar for GC-rich sequences produced at the different stages of the assembly workflow. The results of one of these MEGAN analyses are shown in Figure 3-8. They were produced from the results of a BLASTx search that used the 98 GC-rich sequences longer than 500 bps produced in the final assembly and comprising a total of 1.37 Mbp as query sequences (Appendix A).

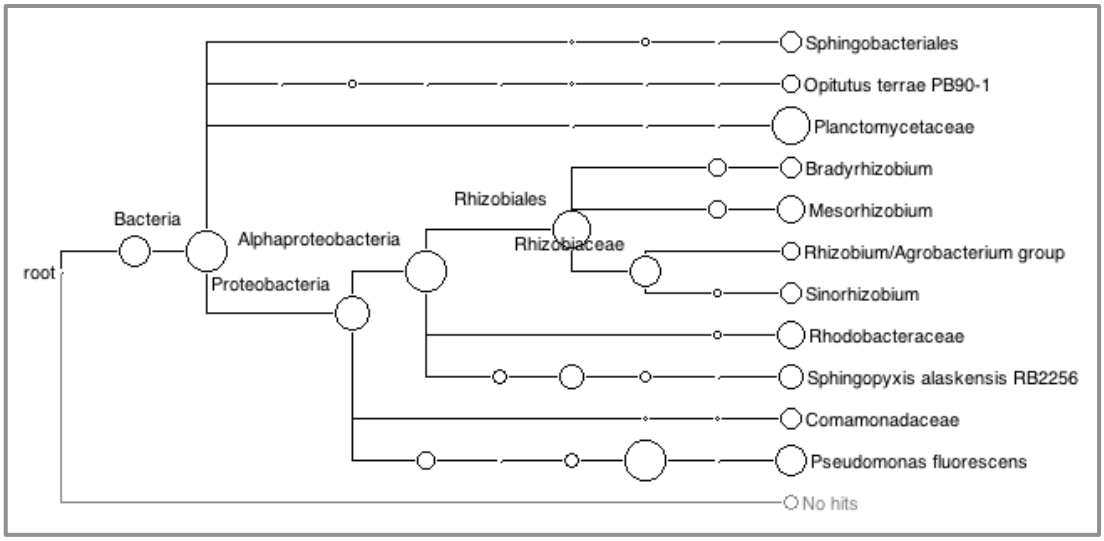


Figure 3-8 MEGAN visualisation of BLAST results for CG-rich contaminant sequences

The number of reads assigned to a node is indicated by the diameter of the circle.

A similar BLAST search was conducted on the remaining 506 AT-rich contigs longer than 500 bps as query sequences in order to identify potential spheroid body sequences. Figure 3-9 shows an overview of the results as visualised in MEGAN. Based on the results of this analysis the contigs were split into sets according to their taxonomic classification. Fifty-eight contigs were assigned to the Cyanobacteria and *Cyanothece* in particular. These sequences were assumed to be derived from the endosymbiont and comprised a total of 1.824 Mbp of sequence. The 36 contigs assigned to Bacillariophyta are presumed to be *R. gibba* genomic or organelle genome sequences and comprise 0.183 Mbp of sequence. This analysis also identified sequences from bacterial contaminants. These sequences show a very similar taxonomic spread to those of high GC content as is evident from Figures 3-8 and 3-9. Sequences that were assigned neither to diatoms nor cyanobacteria comprised 0.862 Mbp in 410 contigs. Fasta files with the contig sequences for each taxonomic set as well as the GC-rich sequences are provided in Appendix A.

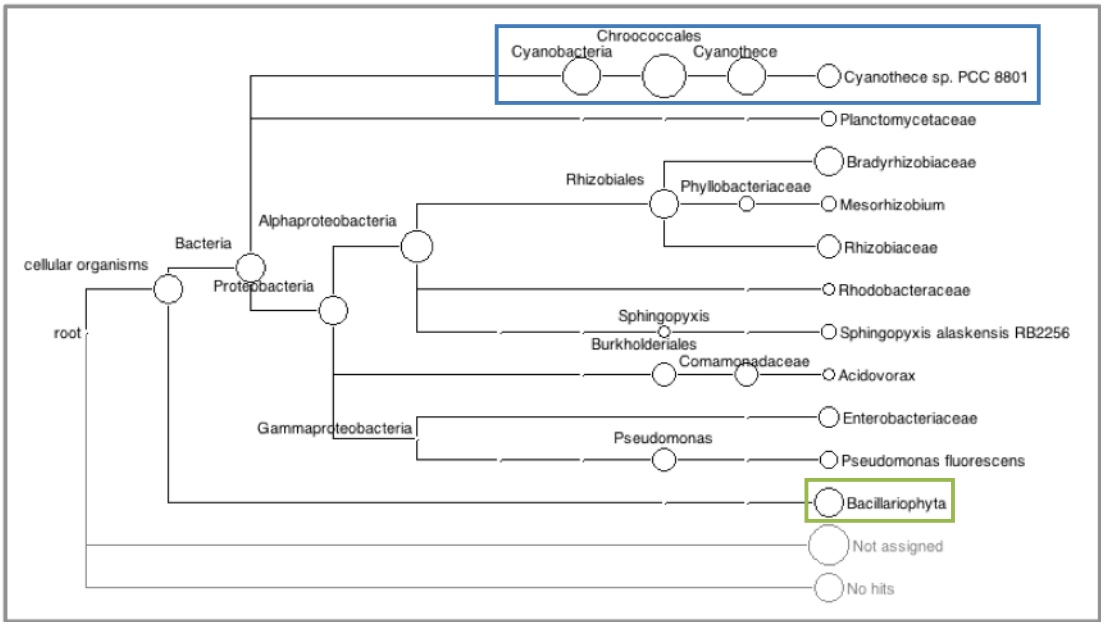


Figure 3-9 MEGAN visualisation of BLAST results for AT-rich contigs

The number of reads assigned to a node is indicated by the diameter of the circle. A green and blue box highlights Bacillariophyta and Cyanobacteria clades respectively.

The putative spheroid body sequences identified above were combined in Sequencher with the 663 contigs of Sanger sequences provided by our collaborators in Marburg. The Sanger sequences comprised a total of 1.9 Mbp of sequence. The combined data yielded 2.25 Mbp of putative spheroid body genome sequence, indicating that the

Illumina sequencing produced at least 0.35 Mbp of novel sequence information for the spheroid body genome.

3.5 Discussion

The sequencing of the genome of the spheroid body of *R. gibba* requires *de novo* sequence assembly. Even though the fully sequenced genome of a relatively closely related bacterium, *Cyanothece sp*. ATCC 51142, was available, the evolutionary distance from the spheroid body genome prevented its use as a reference sequence for short read mapping or to guide assemblies. Preliminary sequencing results indicated that extensive genome changes and gene losses have taken place in the spheroid body. As is to be expected, the changes in evolutionary dynamics acting on endosymbiont genomes (see Section 1.1.2 for more detail) also result in an accelerated mutation rate and lower GC contents. This reduces the suitability of the *Cyanothece* genome as reference even more, because alignments of short reads are more strongly affected by a high frequency of nucleotide substitutions than those of longer reads. However, my analyses of BLAST results in MEGAN suggested that sequence similarity to *Cyanothece* can be used to identify potential spheroid body sequences among longer sequences.

3.5.1 Sequence assembly

The velvet assembly algorithm requires the specification of several parameters that greatly affect the assembly. Three of these parameters, the minimum coverage cutoff, k-mer length and expected coverage, can only be determined by trial and error. This process is complicated by the fact that these parameters influence each other (Zerbino and Birney, 2008). In part this is due to the fact, that the relative coverage available for the assembly is influenced by the length of the k-mers used as described by the following formula:

$C_k = C \times (L - k\text{-mer} + 1) / L$

C_k : k-mer coverage

C: coverage

L: read length

k-mer: k-mer length

Preliminary assemblies were carried out with a range of different values for k-mer length, minimum coverage cutoff and expected coverage. The assembly statistics showed that while changes in parameter values had marked effects on the assembly outcome, no single assembly produced satisfying results (results not shown). Analyses of the contigs identified strong differences in read coverage as the likely cause for the high fragmentation and relatively little total assembled sequence in these assemblies.

High throughput methods of sequencing (HTS) produce a large number of relatively short reads. With no reference sequence to guide the assembly, as is the case for *de novo* sequencing, the reconstruction of the original sequence from these short reads presents one of the most challenging computational task in all of biology. It becomes even harder if the sequencing depth across the template is uneven. However, the problem of uneven coverage is exacerbated in the presence of mixed template DNA as is the case here, or metagenomic samples. The abundance of the different species in the sample is quite uneven to start with but conserved sequences that are shared between different species cause perceived variations in coverage even within genomes.

The standard approach to analysing HTS metagenomics samples has therefore been to forgo assembly and directly BLAST the short reads against databases. Software like MEGAN (Huson et al., 2007) was designed for the analysis of these BLAST results but the main drawback of this approach is the computing power and time required to perform BLAST searches with millions of reads.

The Short Read Assemblers that are most commonly used including Velvet (Zerbino and Birney, 2008), Abyss (Simpson et al., 2009) or SOAPdenovo (Li et al., 2010) use the de Bruijn graph approach. See Section 1.4 for details on de Bruijn graph assemblers. This approach very effectively processes the huge amount of information produced by HTS but it has some inherent problems that more often then not result in incomplete and fragmented assemblies. Peng et al. (2012) discuss this in more detail. Their conclusions can be summarized as follows:

Sequencing errors that result in incorrect k-mers complicate the graph. Short read assemblers were developed to sequence single genomes and assume an even coverage. When this is the case, incorrect k-mers, which should be less numerous then correct ones, can be identified and excluded from the assembly based on their lower coverage.

However, even in pure samples this assumption doesn't hold. The read coverage usually fluctuates due to amplification bias. High GC content and the propensity of some sequence regions to form secondary structures is thought to be the underlying cause. The problem is significantly exacerbated in mixed samples with varying template abundances. Regions with a high read coverage inevitably also produce a higher number of erroneous reads than regions of low coverage. If coverage is used to discriminate between correct and erroneous reads then the presence of a relatively high number of erroneous k-mers in high coverage regions can push up the cut-off value to an extent that it exceeds the total k-mer coverage available for low-coverage regions in the sample. The assumption that correct k-mers always outnumber incorrect k-mers is invalid in these cases. As a result these low coverage regions are not assembled because the k-mers available for them are excluded from the assembly.

The choice of the appropriate value for k is another problem. Large values for k can result in failure to assemble some regions because of missing k-mers, especially in low coverage regions. This would break up the assembly. Small values for k on the other hand introduce more branches to the graph, especially in repeat regions and in the presence of erroneous reads. This can make the graph complicated to a degree that it becomes un-resolvable. If a repeat region coincides with one of low coverage then no value of k may exist with which that section of the graph can be resolved.

The development of assemblers like IDBA-UD (Peng et al., 2012) that are better prepared to deal with these issues has begun and will hopefully significantly improve the usefulness of HTS for metagenomics, transcriptomics and any other kind of sample with contaminations or uneven template abundance.

Both of the issues discussed above – uneven coverage and presence of contaminations - affect the sample used in this project. The pooling of fosmid DNA extractions inevitably introduced variations in template abundance and therefore read coverage. This problem is exacerbated by overlap between fosmid inserts. Figure 3-10 shows an example for a contig produced from overlapping forsmid inserts and the marked decline in coverage at the end of the overlap region. The fact that some fosmid inserts originated from bacterial contaminants introduced a higher level of complexity to the assembly, as they are likely to show different sequence characteristics (GC content among others) and inevitably contribute to a more fragmented assembly. It was for

that reason that it was attempted to exclude reads derived from contaminant sequences from the final assemblies to make the assembly problem computationally tractable.

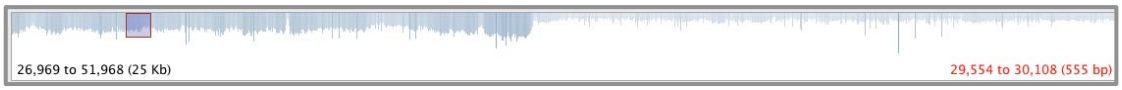


Figure 3-10 Tablet screenshot of the read coverage of a 52 kb long contig that was assembled from overlapping sequencing templates.

Because the k-mer length influences the relative coverage, different k-mer values are best suited for different template coverage. For this reason the combination of the results of several assemblies with different k-mers by sequence alignment is likely to produce more and overall longer contigs than any single assembly. Iterative assemblies make use of Velvet's ability to use long sequences for scaffolding. By using contigs produced by an assembly as an scaffold in subsequent assemblies, the computational task can be simplified. Both strategies were combined in this project. Contigs produced in initial assemblies with k-mers of 23 and 55 were used as scaffolds in the final assemblies with an intermediate k-mer value of 41 (see Table 3-1). While the initial assemblies A and B used a very high coverage cutoff of 20 and 30 respectively, a much lower value of seven was used for the subsequent assemblies to capture sequences with relatively low coverage. The final assemblies D and E only differ from each other in the very different values for the expected coverage. This strategy was devised to overcome issues caused by extreme differences in coverage that occur when fosmid inserts in the sample overlap. This approach did indeed produce more contigs that reached or exceeded the average fosmid insert size of 35 Kb than any of the standard assemblies. It also gave a greater maximum contig length as detailed in Table 3-1. The n50 value though was smaller in the final assemblies in comparison to the initial assemblies A and B. The n50 is defined so that contigs longer than the n50 value amount to 50% of the assembled sequence. This metric is a commonly used indicator for how fragmented an assembly is (Baker, 2012). However, in this case the lower n50 values of the final assemblies were not the result of a break up of long contigs. Instead the fact that many of the long contigs have been extended compared to the initial assemblies did not suffice to compensate for the effect on the n50 value of a higher number of short contigs in the assembly.

Overall 4.24 Mbp of sequence in contigs longer than 500 bps were produced in the final assembly. At an average size of 35 kbs the 165 fosmid inserts in the sample amount to approximately 5.78 Mbp of sequence, less the regions of overlap between them. This suggests that despite the problems posed by the characteristics of the sample, the assembly recovered the majority of the template, even though many of the fosmid inserts remain fragmented.

The size of the spheroid body genome remains unknown due to experimental difficulties caused by bacterial contaminants. Based on the analyses of a sequenced fragment of the SB genome Kneip (2004) concluded that, while a significant proportion of the genes present in the free-living ancestor of the endosymbiont had been inactivated and their coding regions eroded, they have not yet been deleted from the genome. Kneip suggested for that reason that the SB genome is not yet significantly smaller than the 5.4 Mbp genome of *Cyanothece* 51142. This work produced 1.823 Mbp of putative spheroid body sequence. When combined with the Sanger sequences provided by our collaborators in Marburg a total of 2.25 Mbp of putative SB sequence in 141 contigs were successfully assembled. Figure 3-11 shows an example of an alignment between the two sequence types. The example shows that while the short read assemblies produced in this project did not add much new sequence, they produced longer continuous sequences and a less fragmented assembly. Despite this many gaps remain. In combination with the Sanger sequencing data 32 contigs of putative spheroid body sequence were produced that represent at least 52 fully sequenced fosmid inserts, some of them overlapping and forming longer continuous sequences. This was determined based on the presence of vector sequence at one or in most cases both ends of the contigs, and their length. This indicates that the fragmented state of this assembly is due to the fact that the fosmid inserts sequenced so far were not sufficient to cover the entire SB genome, corroborating Kneip's conjecture.

Combining the results of different assemblies into longer continuous sequences was helped and at the same time hindered by Velvet's scaffolding function. Scaffolding introduces N-stretches that preserve the positional information of paired reads if the assembly breaks up within a sequence region that is bridged by read pairs. The length of N-stretches inserted by Velvet is based on the expected paired-end insert length (see Section 1.3 for details on paired-end sequencing). Empirically the insert length has a high variance and consequently the length of N-stretches is only a rough approximation

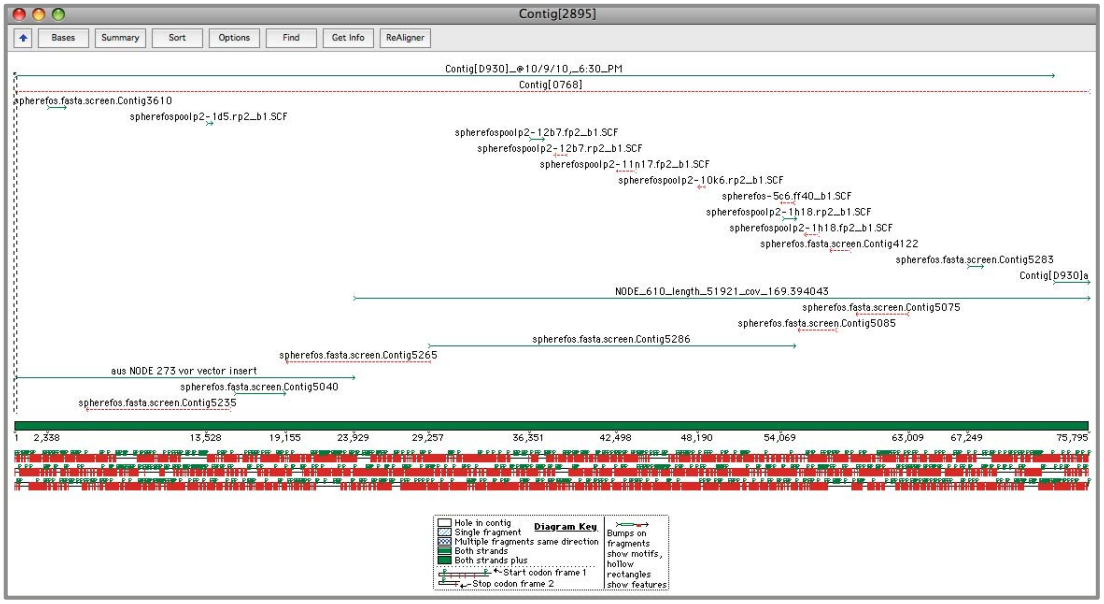


Figure 3-11 Alignment of short read assemblies and Sanger sequences

Shown is a screenshot of an alignment done in Sequencher. The sequences labelled as 'Contig[D930]', 'Contig[D930]a' and 'Contig[0768]' were produced in the final assemblies D and F. Sequences labelled as 'NODE' were assembled from short reads in preliminary test assemblies. All other sequences are Sanger sequences, some of them assembled into contigs.

of the length of missing sequence. In alignments of scaffolded contigs with contigs in which that region has been fully assembled, this produces the same effects as insertions or deletions. The sequence alignment algorithm implemented in Sequencher 4.9 was not designed to handle large insertion or deletions and sequence regions following an N-stretch were usually miss-aligned. The alignments and subsequent mappings also revealed that the rates of assembly errors increased steeply close to the ends of contigs. This has to be expected, as the same characteristics of the data that cause an assembly to break up prematurely are also likely to cause misassemblies nearby the break-off point. Therefore, consensus sequences could be improved substantially by allowing sequences from the main body of a contig to overrule conflicting sequences at the ends of another, overlapping contig. For this reason each alignment had to be hand edited before a consensus sequence was produced. The reads were then mapped to these consensus sequences to confirm the edits and identify further misassemblies (see Figures 3-4 and 3-5 for examples) and then edited a second time according to the results of the mappings. Spikes in coverage with characteristic mismatches for example revealed identical or even non-identical repeats that had been collapsed. In most cases

misassemblies manifested themselves simply in single erroneous base pairs. In many cases, sequences could be extended as well based on read overhangs.

3.5.2 Assembly quality and comparison to Sanger sequences

Indexing was used to test the quality of the assembly with regards to chimeric contigs. The reads derived from a subset of 10 fosmids in the sample were flagged with a different index than the remaining reads and mapped against contigs produced by the assemblies. Chimeric contigs that were misassembled from templates that were indexed differently can be expected to comprise regions that map to this subset of reads as well as regions that do not align with any of these reads. Figure 3-6 shows an example for a coverage peak in the mapping of a contig that over most of its length does not align with these reads. This could be interpreted as an indicator for a chimeric sequence, but the fact that the reads show consistent mis-matches with the reference sequence indicate that they are derived from a template that shows a high similarity but is not identical to the reference sequence. This finding is reassuring with regards to the assembly quality as it indicates that despite high sequence similarity these reads did not partake in the assembly of this region. Very similar observations were made in all cases of coverage peaks along reference sequences that generally did not align to the read subset.

The only contig that showed regions of coverage bordering on regions without coverage, as would be characteristic of a chimeric sequence is Contig 2891 (see Figure 3-7). This could indeed be a chimeric contig, misassembled from three different templates. The fact that the length of 38 kb of the high coverage region corresponds to the expected size of a fosmid insert (30 – 40 kb) on the other hand suggests that this 93 kb sequence was assembled correctly from overlapping fosmid inserts.

In all instances in which reads mapped to a contig sequence in a non-random way, they did so with an even coverage that was consistently high enough to suggest that the sequences were in fact assembled from these reads. Overall the results of these mappings are encouraging in that they suggest that misassemblies in the form of chimeric sequences are not common in this data, though it cannot be ruled out that some may have been produced.

The quality of the assembly was also assessed by comparison with Sanger sequencing results. These sequences were produced by shotgun sequencing of the same templates. They comprise 1.9 Mbp of sequence data in 663 contigs, many of which are single Sanger reads that failed to assemble into longer contigs, due to insufficient coverage. These sequences were aligned to the contigs produced from Illumina reads in this project to assess the agreement between sequences produced by the two methods. Sanger sequences remain the gold standard for sequencing in terms of error rate (Kircher and Kelso, 2010). The error rate is very low, with a sequencing error occurring on average every 10,000-100,000 nucleotides (Ewing and Green, 1998). However, the error rate increases steeply towards the end of long Sanger reads, among other reasons due to reduced separation by the electrophoresis that affects the accuracy of base calls (Ewing et al., 1998). The ends of Sanger reads are for this reason usually trimmed back, albeit often not enough to remove all sequencing errors. This became apparent when Sanger sequences and the contigs produced in this project were used to produce a spheroid body genome assembly from all available data. Alignments usually show a steep increase in mis-matches towards the end of a contig, regardless of the sequencing method used. In case of Sanger sequences the aforementioned base calling issues are to blame, while in short read assemblies properties of the data cause misassemblies that eventually result in a breaking off of the assembly.

The effect of increased error rates towards the ends of contigs can be seen in the statistics given in Table 3-3. The alignments between longer contigs show consistently sequence identities of over 99% across the alignment. The proportion of mismatches increases with decreasing alignment lengths, from an average of 0.03 % for alignments and therefore contigs that are longer than 5 kb, to 0.12% in alignments between 1.5 and 5 kb in length. The average sequence identity calculated for alignments that are between 0.5 and 1 kb long is significantly lower at 98.55%. The vast majority of alignments used to calculate this value involved a single Sanger read aligned to a longer short-read contig. The increased proportion of mis-matches is very likely caused by the relatively greater effect of sequencing errors at the end of the Sanger reads on short alignments. The longest alignments involve sequences that are not much affected by increased sequencing error rates towards the ends and are likely to give a good approximation of the concordance of sequences produced by the two methods. Overall this comparison indicates that the assembly of Illumina reads produced high quality sequence data. The observed disagreement between the Illumina assemblies and long

Sanger contigs of 0.03% lies in the same order of magnitude as the empirical error rate of Sanger sequencing.

3.5.3 Contaminant sequences

Bacterial contaminants cannot as yet be removed from the culture of *R. gibba* cells and comprehensive attempts to eliminate them from DNA extractions have so far failed. It was therefore necessary to remove them with a bioinformatic approach post sequencing. The fosmid library was produced before the introduction of high throughput sequencing. It provided the possibility to enrich the sequencing template for spheroid body sequence by screening the inserts for GC-content by Sanger sequencing one of their ends, but this strategy was not entirely successful as the presence of contaminant sequences in the assembly results demonstrates.

Potential contaminant sequences in the fosmid library were identified based on two factors: BLAST results and GC content. Empirically putative SB sequences are GC-poor with GC-contents of only approximately 35%, while many of the contaminant sequences had a GC-content of approximately 60%. Of the total sequence produced in the final assembly 32% were GC-rich contaminant sequence. The validity of GC-content as marker for at least a proportion of contaminant sequences was backed up by BLAST results that never showed significant similarities of the GC-rich contigs to Cyanobacteria. The same BLAST analyses, when applied to the GC-poor sequences, on the other hand found significant similarities to Cyanobacteria and *Cyanothece* sp. in particular in 63.5% of sequences and significant similarities to available diatom sequences in 6 %. This amounts to 43% of the total assembled sequence being putative endosymbiont genomic sequence and 4% putative *R. gibba* sequence. Twenty percent of the sequence produced in the final assembly was GC-poor, but assigned by MEGAN to non-cyanobacterial lineages similar to those represented by the GC-rich sequences or not assigned at all. It is possible that at least some of these sequences are spheroid body sequences but the results show nonetheless that bacterial sequences constituted a very significant proportion of fosmid inserts, despite attempts to exclude GC-rich fosmid inserts from the sample.

A very similar approach for the segregation of sequences in mixed samples has been used by Kumar & Blaxter (2012) to simultaneously sequence the genomes of nematode

worms and their endosymbionts with HTS. They have developed a computational pipeline to separate metazoan and bacterial sequences *in silico* by using sequencing coverage information, base composition and sequence similarity searches. Similar to the strategy developed in this project, Kumar and Blaxter also recommend reassemblies following the separation of sequences from different organisms produced in preliminary assemblies.

The analyses of BLAST results in MEGAN suggest that the bacteria associated with *R. gibba* belong to the Proteobacteria, predominantly the α-Proteobacteria. They also suggest that *R. gibba* cells in the culture from which the fosmid library was prepared are associated with a diverse community of bacteria from a variety of bacterial lineages (see Figures 3-8 and 3-9). The characterisation and study of these bacteria was outside the scope of this project but has the potential to produce valuable insights into the micro-communities that *R. gibba* and diatoms in general form with extracellular bacteria. It is commonly observed in cultures of diatoms and studies of aquatic ecosystems that these algae form potentially obligate relationships with bacteria but nothing is known yet about the exact nature of these relationships (Grossart et al., 2005; Schäfer et al., 2002; Bruckner et al., 2008).

3.6 Conclusion and outlook

The inserts of 165 fosmids were sequenced using Illumina sequencing technology with the aim to sequence the genome of the endosymbiotic spheroid body of the diatom *R. gibba*. The fosmid library had been produced to exclude bacterial contaminations and enrich the sample for spheroid body sequence but also contributed to uneven read coverage in the assembly. An iterative assembly strategy that combined the results of assemblies using different k-mer values was developed to overcome assembly problems caused by uneven coverage and was used successfully to extend long contigs as compared to single assembly strategies.

BLAST searches in conjunction with MEGAN analyses were used to successfully identify putative SB and diatom sequences in the assemblies based on their similarity to available cyanobacterial and diatom sequences and to segregate them from bacterial contaminations. Several strategies were used to assess the quality of the assembled sequences. Mappings of reads were used to uncover and correct misassemblies. Mappings with an indexed subset of reads did not detect any chimeric contigs. Alignments with sequences produced with Sanger sequencing showed that the rate of sequence errors in the sequences assembled from short reads is comparable with the high quality observed in Sanger sequencing. All these points indicate that the assembled sequences are of excellent quality.

As a result of these efforts a total 2.25 Mbp of putative spheroid body genomic sequence were assembled. The long contigs produced from short-reads in this project were successfully used to significantly reduce fragmentation in the previously available Sanger shotgun sequencing data. Even though, the genomic sequence is not yet complete due to insufficient coverage in the available fosmid inserts.

The annotation of the sequence data and further sequencing efforts to complete the genomic assembly will be undertaken under the responsibility of the cell biology work group at the Universität Marburg. In principal the same approach could be used to continue sequencing work on the spheroid body. The fosmid library allows for subsets of the template to be indexed, thereby breaking up the assembly problem into several

smaller ones but it also handicaps the assembly by compounding uneven coverage, though modern assemblers, some of which have been developed with transcriptomics in mind, should be more apt to work with this type of data than the tools used in this work.

The construction of fosmid libraries is costly and time intensive. For this reason sequencing of genomic extractions has become the standard approach when using HTS. Sequencing of a complete genomic extraction would be the most suitable strategy to finish the sequence of the SB genome and enable the full assessment of gene loss from the endosymbiont genome. HTS can achieve sufficient coverage to sequence both endosymbiont and host simultaneously, as well as bacterial contaminants (Kumar and Blaxter, 2012). This will provide complete gene catalogues for both the host and symbiont and allow the systematic study of metabolic pathways contributed by each organism.

Bibliography

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990). Basic local alignment search tool. Journal of molecul. **215**:403–410.

Baker, M. (2012). De novo genome assembly: what every biologist should know. Nature Methods. **9**:333–337.

Bruckner, C.G., Bahulikar, R., Rahalkar, M., Schink, B. and Kroth, P.G. (2008). Bacteria associated with benthic diatoms from Lake Constance: phylogeny and influences on diatom growth and secretion of extracellular polymeric substances. Applied and environmental microbiology. **74**:7740–9.

Colón-López, M.S., Sherman, D.M. and Sherman, L.A. (1997). Transcriptional and translational regulation of nitrogenase in light-dark- and continuous-light-grown cultures of the unicellular cyanobacterium Cyanothece sp. strain ATCC 51142. Journal of bacteriology. **179**:4319–27.

Conesa, A., Götz, S., Garcia-Gomez, J.M., Terol, J., Talon, M. and Robles, M. (2005). Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics. **21**:3674–3676.

Cox, M.P., Peterson, D.A. and Biggs, P.J. (2010). SolexaQA: At-a-glance quality assessment of Illumina second-generation sequencing data. BMC bioinformatics. **11**:485.

Droop, S.J.M., Sims, P.A., Mann, D.G. and Pankhurst, R.J. (1993). A taxonomic database and linked iconograph for diatoms. Hydrobiologia. **269-270**:503–508.

Ewing, B. and Green, P. (1998). Base-Calling of Automated Sequencer Traces Using Phred . II . Error Probabilities. Genome research. **8**:186–194.

Ewing, B., Hillier, L., Wendl, M.C. and Green, P. (1998). Base-Calling of Automated Sequencer Traces Using Phred . I . Accuracy Assessment. Genome research. **8**:175–185.

Floener, L. and Bothe, H. (1980). Nitrogen fixation in Rhopalodia gibba, a diatom containing blue-greenish inclusions symbiotically. In W. Schwemmler and H. E. A. Schenk, eds. *Endocytobiology: Endosymbiosis and Cell Biology, a Synthesis of Recent Research, Vol. 1*. Berlin: Walter de Gruyter, pp. 541–552.

Geitler, L. (1977). Life history of the Epithemiaceae Epithemia, Rhopalodia and Denticula (Diatomophyceae) and their presumable symbiotic spheroid bodies. Journal of Plant Systematics and Evolution. **128**:259–275.

Grossart, H.-P., Levold, F., Allgaier, M., Simon, M. and Brinkhoff, T. (2005). Marine diatom species harbour distinct bacterial communities. Environmental microbiology. **7**:860–73.

Hill, D.R., Belbin, T.J., Thorsteinsson, M. V, Bassam, D., Brass, S., Ernst, A., Boger, P., Paerl, H., Mulligan, M.E. and Potts, M. (1996). GlbN (cyanoglobin) is a peripheral membrane protein that is restricted to certain Nostoc spp. Journal of Bacteriology. **178**:6587–6598.

Huson, D.H., Auch, A.F., Qi, J. and Schuster, S.C. (2007). MEGAN analysis of metagenomic data. Genome research. **17**:377–86.

Jahn, R. (1996). The historical East African freshwater algae collection at the Botanical Museum Berlin-Dahlem (B). Willdenowia.333–340.

Kawai, H., Motomura, T. and Okuda, K. (2005). Isolation and purification techniques for macroalgae. In R. A. Anderssen, ed. *Algal Culturing Techniques*. Elsevier Academic Press, p. 141.

Kircher, M. and Kelso, J. (2010). High-throughput DNA sequencing--concepts and limitations. BioEssays : news and reviews in molecular, cellular and developmental biology. **32**:524–36.

Kneip, C. (2004). Sphaeroidkoerper der Diatomee Rhopalodia gibba – Obligate Endosymbionten zur molekularen Stickstofffixierung.

Kneip, C., Lockhart, P.J., Voss, C. and Maier, U.G. (2007). Nitrogen fixation in eukaryotes--new models for symbiosis. BMC evolutionary biology. **7**:55.

Kneip, C., Voss, C., Lockhart, P.J. and Maier, U.G. (2008). The cyanobacterial endosymbiont of the unicellular algae Rhopalodia gibba shows reductive genome evolution. BMC evolutionary biology. **8**:30.

Kondo, T., Mori, T., Lebedeva, N. V, Aoki, S., Ishiura, M. and Golden, S.S. (1997). Circadian rhythms in rapidly dividing cyanobacteria. Science. **275**:224–227.

Kumar, S. and Blaxter, M.L. (2012). Simultaneous genome sequencing of symbionts and their hosts. Symbiosis (Philadelphia, Pa.). **55**:119–126.

Langmead, B., Trapnell, C., Pop, M. and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome biology. **10**:R25.

Li, H. and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. **25**:1754–60.

Li, Y., Hu, Y., Bolund, L. and Wang, J. (2010). State of the art de novo assembly of human genomes from massively parallel sequencing data. Human genomics. **4**:271–277.

Lorenz, M., Friedl, T. and Day, J.G. (2005). Perpetual maintenance of actively metabolizing microalgal cultures. In R. A. Anderssen, ed. *Algal Culturing Techniques*. Elsevier Academic Press, p. 155.

Milne, I., Bayer, M., Cardle, L., Shaw, P., Stephen, G., Wright, F. and Marshall, D. (2010). Tablet--next generation sequence assembly visualization. Bioinformatics (Oxford, England). **26**:401–2.

Peng, Y., Leung, H.C.M., Yiu, S.M. and Chin, F.Y.L. (2012). IDBA-UD: A de Novo Assembler for Single-Cell and Metagenomic Sequencing Data with Highly Uneven Depth. Bioinformatics. **28**:1420–1428.

Pfitzer, E. (1869). Über den Bau und die Zellteilung der Diatomeen. Botanische Zeitung. **27**:774–776.

Prechtl, J., Kneip, C., Lockhart, P.J., Wenderoth, K. and Maier, U.G. (2004). Intracellular spheroid bodies of Rhopalodia gibba have nitrogen-fixing apparatus of cyanobacterial origin. Molecular biology and evolution. **21**:1477–81.

Reddy, K.J., Haskell, J.B., Sherman, D M and Sherman, L.A. (1993). Unicellular, aerobic nitrogen-fixing cyanobacteria of the genus Cyanothece. Journal of bacteriology. **175**:1284–92.

Rutherford, K., Parkhill, J., Crook, J., Horsnell, T., Rice, P., Rajandream, M.A. and Barrell, B. (2000). Artemis: sequence visualization and annotation. Bioinformatics (Oxford, England). **16**:944–945.

Schneegurt, M., Sherman, D. and Sherman, L. (1997). Composition of the carbohydrate granules of the cyanobacterium, Cyanothece sp. strain ATCC 51142. Archives of microbiology. **167**:89–98.

Schneegurt, M.A., Sherman, Debra M, Nayar, S. and Sherman, L.A. (1994). formation and dinitrogen fixation in the Oscillating Behavior of Carbohydrate Granule Formation and Dinitrogen Fixation in the Cyanobacterium Cyanothece sp . Strain ATCC 51142.

Schäfer, H., Abbas, B., Witte, H. and Muyzer, G. (2002). Genetic diversity of “satellite” bacteria present in cultures of marine diatoms. FEMS microbiology ecology. **42**:25–35.

Simpson, J.T., Wong, K., Jackman, S.D., Schein, J.E., Jones, S.J.M. and Birol, I. (2009). ABySS: a parallel assembler for short read sequence data. Genome research. **19**:1117–23.

Slater, G.S.C. and Birney, E. (2005). Automated generation of heuristics for biological sequence comparison. BMC bioinformatics. **6**:31.

Stöckel, J., Welsh, E.A, Liberton, M., Kunnvakkam, R., Aurora, R. and Pakrasi, H.B. (2008). Global transcriptomic analysis of Cyanothece 51142 reveals robust diurnal oscillation of central metabolic processes. Proceedings of the National Academy of Sciences of the United States of America. **105**:6156–61.

Welsh, E. a et al. (2008). The genome of Cyanothece 51142, a unicellular diazotrophic cyanobacterium important in the marine nitrogen cycle. Proceedings of the National Academy of Sciences of the United States of America. **105**:15094–9.

Wittenberg, J.B., Appleby, C.A. and Wittenberg, B.A. (1972). The kinetics of the reactions of leghemoglobin with oxygen and carbon monoxide. Journal of Biological Chemistry. **247**:527–531.

Woelfle, M.A., Ouyang, Y., Phanvijhitsiri, K. and Johnson, C.H. (2004). The Adaptive Value of Circadian Clocks : An Experimental Assessment in Cyanobacteria. **14**:1481–1486.

Zerbino, D.R. and Birney, E. (2008). Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome research. **18**:821–9.

4 High-throughput sequencing of an environmental sample of the diatom *Epithemia sorex*

4.1 Abstract

Recent work has shown that the nitrogen fixing endosymbionts (spheroid bodies) in diatoms of the order Rhopalodiales originate from a single endosymbiotic event followed by diversification of the lineage (Nakayama et al., 2011). To date, only the genome of one of these cyanobacterial endosymbionts, the spheroid body of *Rhopalodia gibba,* has been partially sequenced. A comparison with an endosymbiont genome from another genus in the Rhopalodiales is likely to provide valuable insights into the early stages of an endosymbiosis founded on nitrogen fixation that has the potential to lead to organelle evolution. My aim was to sequence the spheroid body genome of a local sample of *Epithemia sorex*. In the absence of a pure culture this was attempted through high throughput sequencing of an environmental sample. However, bioinformatic analyses of the resulting sequence data were not able to detect diatom sequences from a genomic preparation of the environmental sample. Although a diversity of sequences were determined, these were not represented in public sequence databases. To investigate the nature of organisms in the environmental sample genomic preparation, targeted Sanger sequencing of 16S and 18S rDNA sequences was then undertaken. Analyses of these data revealed that the majority of the DNA in the sample originated from Naididae worms, nematodes and ciliates. These results highlight not only the limitations of high throughput data for work on environmental samples but it also suggested that the relative effectiveness of DNA extractions protocols on different organisms could impact assessment of taxonomic composition of metagenomic samples. The absence of *E. sorex* sequences in the sample investigated here is presumed to result from absorption of *E. sorex* DNA to the diatom's silicate shell in the presence of RNAlater. Further work needs to be undertaken to investigate the relative efficiency of different RNAlater/DNA extraction protocols for diatom environmental samples and *E. sorex* in particular.

4.2 Introduction

Several diatom species, like *Rhizosolenia* species or *Climacodium frauenfeldianum*, maintain facultative symbiotic relationships with cyanobacteria to take advantage of their ability to fix molecular nitrogen. The only known obligate symbiosis occurs in diatoms of the order Rhopalodiales, whose cyanobacterial endosymbionts are inherited trough the sexual cycle of the hosts (Prechtl et al., 2004; Geitler, 1977). The Rhopalodiaceae are the only family in the order Rhopalodiales. The family comprises three extant genera - *Rhopalodia, Epithemia* and *Protokeelia* - and the fossil genus *Yoshidaia* (Round et al., 1990). The genera *Rhopalodia* and *Epithemia* are known to harbour cyanobacterial endosymbionts referred to as spheroid bodies (SBs) (DeYoe et al., 1992; Geitler, 1977; Kies, 1992). If these endosymbionts are or were present in *Protokeelia* or *Yoshidaia* is not known (Nakayama et al., 2011).

The analysis of both the hosts' 18S and spheroid bodies' 16S ribosomal DNA sequences in three rhopalodiacean diatom species - *Epithemia turgida, Epithemia sorex* and *Rhopalodia gibba* - indicates a single origin for their spheroid bodies. This means that the SBs were acquired in one initial endosymbiosis with a common ancestor of *Rhopalodia* and *Epithemia* and have since undergone speciation along with their hosts (Nakayama et al., 2011).

The same analysis places the Surirellales as the group closest to the Rhopalodiales. This is in agreement with the fact that both groups share a derived trait, silicon bridges (fibrulae) to stabilise their raphe (a slit along the apical axis characteristic for the Bacillariophyceae). The fossil record suggests that Surirellacean diatoms first appeared approximately 12 million years ago (Mya) during the middle Miocene (Sims et al., 2006; Hajós, 1986). If so, then the symbiosis that gave rise to the rhopalodian SBs is younger than 12 million years.

Since then, each endosymbiont containing lineage has continued integration of the endosymbiont into the host organism independently. The predicted progression of reductive genome evolution is likely to follow a path similar to that described in section 1.2. In short, relaxed or even discontinued purifying selection is expected to lead to the accumulation of mutations and eventual deletions from the genome (Nilsson,

Koskiniemi et al. 2005). This will follow as consequence of "Muller's ratchet" and genetic drift (Muller, 1964; Kneip et al., 2008) and has the potential to lead to a more or less random selection of genes (and co-dependent genes : (Dagan et al., 2006)) loosing their functionality.

The SB's function within the symbiotic relationship is thought to involve the fixation of molecular nitrogen. Several lines of evidence corroborate this hypothesis. *Rhopalodia gibba* has been shown to be capable of fixing nitrogen and the genome of its spheroid bodies encodes the highly conserved *nif* operon, comprising the genes coding for the proteins involved in nitrogen fixation (Kneip et al., 2008). Further, the alpha-subunit of dinitrogenase (NifD) has been immunolocalized to the spheroid body of *R. gibba* (Prechtl et al., 2004). An expected physiological response has also been inferred, consistent with a role in nitrogen fixation. That is, when provided with sufficient levels of phosphorus the number of endosymbionts per cell in *R. gibba* and *Epithemia turgida* have been observed to increased under low nitrogen conditions (DeYoe et al., 1992).

The genome of *Mycobacterium leprae* is likely to be at a similar stage of reductive evolution. It is unclear when the ancestor of *M. leprae* adopted an intracellular live style, but it has been estimated that massive gene inactivation occurred in its genome within the last 20 million years (Gómez-Valero et al., 2007). Given that the rhopalodian SBs are not likely to be older than 12 million years and have a longer generation time than *M. leprae*, it is likely that they are still at the 'chance driven stage' of massive gene loss. It is not clear at this point if spheroid body genes have been successfully transferred to the diatom nucleus, or if these symbionts-host systems have developed a protein import system for the endosymbiont. However, the example of the amoeba *Paulinella* shows that this crucial step for converting an endosymbiont into an organelle does not require a significant amount of evolutionary time. *Paulinella* carries a cyanobacterial photosynthetic endosymbiont, the chromatophore, whose genome has been estimated to have shrunken to about 30% of its original size. More than 30 functional endosymbiont derived genes have been identified in the nuclear genome and of these, three have been shown to be imported back into the chromatophore. Thus the chromatophores of *Paulinella* fulfil all criteria of being a genuine organelle, yet the endosymbiont that it derives from was acquired only about 60 mya (Nowack and Grossman, 2012).

The selection pressure to maintain a gene depends strongly on what other genes are still available to the system (Dagan et al., 2006). With this in mind the comparison of the different lineages of rhopalodian endosymbionts and their hosts can be expected to be very informative with respect to nitrogen fixation. It would allow us to identify aspects of symbiogenesis that evolve in concordance in different lineages, which might suggest convergent processes. Such study would allow more insight to be gained into details and mechanisms of gene loss and the course that reductive evolution of a nitrogen-fixing genome typically takes. Generalities might be concluded as to the necessities underlying this complex process of symbiogenesis.

To date only very limited sequence information is available for rhopalodian diatoms. Across the Rhopalodiaceae only 30 single gene entries for four different nuclear genes are available in GenBank and two genes of the SB in *R. gibba*. The sequencing of a fosmid library of potential *R. gibba* SBs sequences described in Chapter 3 has made a major part of one of the rhopalodian SB's genomes available. This first reference sequence represents a stepping stone to provide a more in depth understanding of the whole genus. Building on this groundwork a plan was developed to sequence another diatom from the family using high throughput sequencing technology and to isolate and assemble the spheroid body sequences using the *R. gibba* sequences as reference. The species *Epithemia sorex* was chosen as a representative of the only other genus known to harbour spheroid bodies. Reasons for choosing this specific species were its use in the analysis of SB phylogeny by Nakayama et al. (2011) and its reported occurrence in New Zealand.

Other than the one used for the *Rhopalodia gibba* project, no cultures of a rhopalodian diatom are currently available. Given that even diatoms in mono-culture are inseparably associated with satellite bacteria in their extracellular matrix, mostly Proteobacteria and Bacteroidetes, it stands to reason that there might be little advantage in developing a culture compared to characterising an environmental sample with *E. sorex* as the dominant species (Schäfer et al., 2002; Bruckner et al., 2008). Thus we sought to sequence an environmental sample with the goal of isolating and assembling a SB genome from it as proof of principle for this approach. High throughput sequencing technology is still fairly new and its capabilities and limits especially with environmental sampling are still being tested. At the outset of this project it was not clear if the methods available were capable of recovering *E. sorex* nuclear and SB sequences from a metagenomic sample.

4.3 Methods

4.3.1 Study species

Epithemia is a cosmopolitan genus of freshwater diatoms that is commonly found in a wide range of habitats including lakes, ponds, streams and swamps in climatic conditions from tropical to temperate and alpine. *Epithemia* species prefer alkaline freshwater but are tolerant of a wide range of salt concentrations and pH. The members of the genus have an epiphytic and epipelic life style where solitary cells are found closely appressed to the substratum, usually aquatic plants or filamentous algae, and thus less likely to become buried beneath sediment (Round et al., 1990). The species *Epithemia sorex* was first described by Kützing in 1844. It is easily distinguished from other diatoms and other species of the genus by its strong resemblance to a bicorne hat in valve view and the biarcuate raphe forming a distinct 'V' (see Figure 4-1).

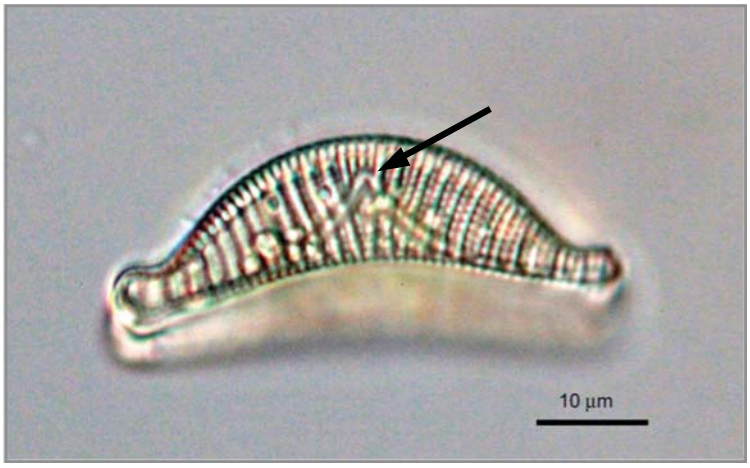


Figure 4-1 Silicate shell of *Epithemia sorex* in valve view
The arrow indicates the point of the V-shaped raphe.

4.3.2 Sample Selection

Samples were collected over two days in November 2010 at different locations in Canterbury, south of Christchurch, and in the Southern Alps south of Otira. Sampling sites were chosen based on suitability and previous sightings of *E. sorex* and are shown in Figures 4-5 to 4-7. Algae were scraped off shallow rocks and water plants and examined under a field microscope to determine the presence of *Epithemia* cells at the site. *Epithemia* cells were found only in samples taken at sites 2_10 and 2_15, marked in green in Figure 4-7. Site 2_10 is at the southern shore of Lake Sarah where epiphytic algae were scraped off stalks of reed. Microscopic examination showed a mixture of a wide variety of unicellular algae with sporadic sightings of *E. sorex* cells. Site 2_15 is a ditch at the side of State Highway 73, ca. 1.5 km south of Arthur's Pass (Figure 4-2).

Figure 4-2 *Epithemia sorex* sampling site

New Zealand Traverse Mercator Projection: Northing 5243095.0,

Easting 1483355.0

GPS / Google Earth:

42°57'34.35"S, 171°34'25.11"E

Altitude: 710 m

Microscopic examinations showed *E. sorex* cells covering aquatic plants and filamentous green algae at very high density (see Figure 4-3). Other than *Epithemia* only very few cells of other unicellular algae species were observed. Based on visual observation it was estimated that *E. sorex* cells constituted more than 95% of unicellular algae biomass in the sample.

Figure 4-3 ***Epithemia* cells *in situ* covering the stem of an aquatic plant**

Portions of aquatic plant matter were transferred into falcon tubes filled with RNAlater® Solution (Ambion) and shaken out to detach the diatoms from their substrate. Once separated the diatoms precipitated rapidly due to their heavy silicate shells. The plant matter was then removed and the procedure repeated to enrich the diatoms in the sample. The sediment was examined under the microscope to confirm that it was mainly composed of *Epithemia* cells (see Figure 4-4). At this stage the diatoms were identified as *Epithemia sorex* based on the characteristically arched raphe (see Figure 4-1). The samples were stored in RNAlater on ice until the DNA was extracted.

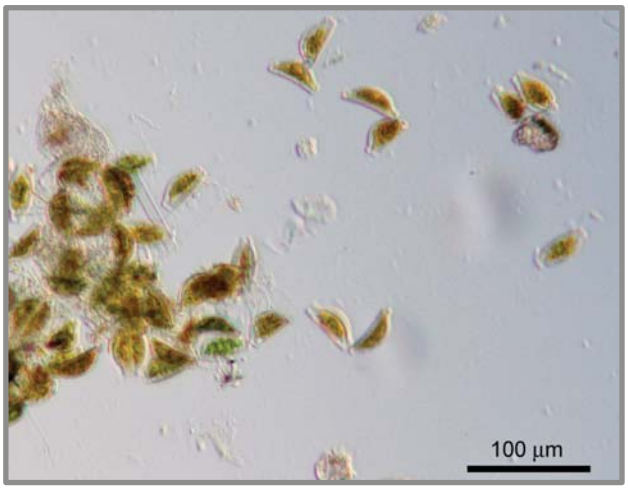


Figure 4-4 **Environmental sample after enrichment for diatom cells**

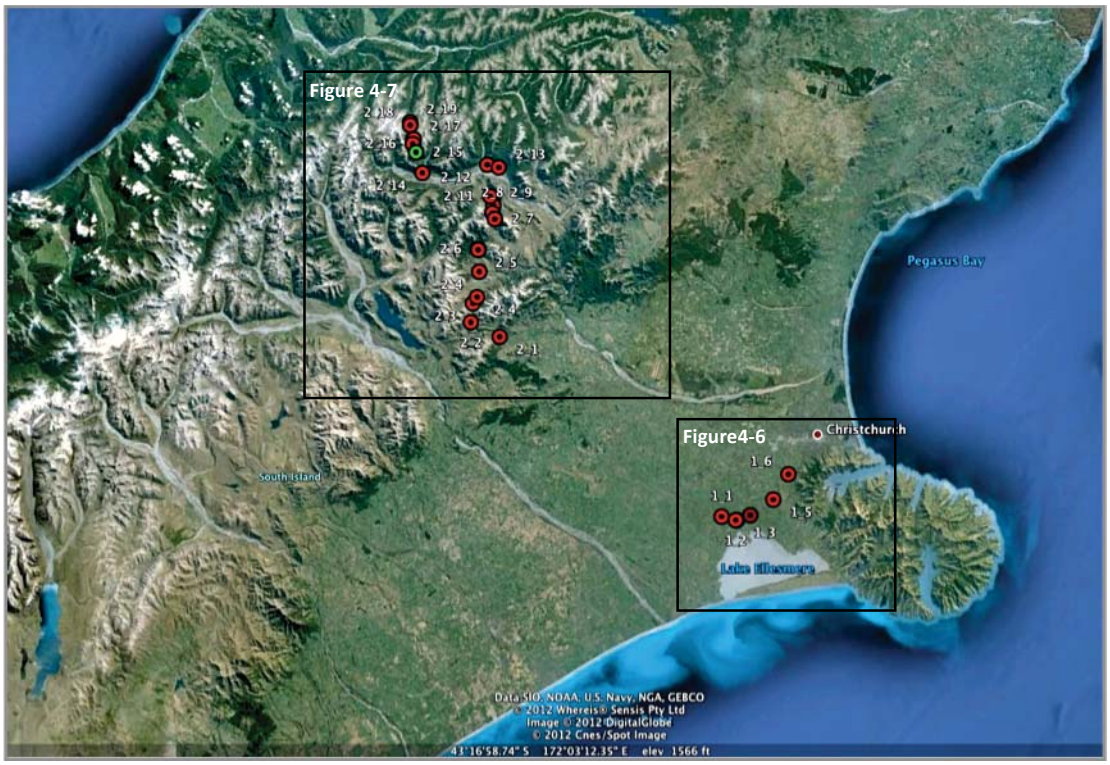


Figure 4-5 Overview of sampling sites on the south island of New Zealand.

The black boxes indicate areas that have been magnified in Figures 4-6 and 4-7. The dots mark the positions of sampling sites. *E. sorex* was successfully sampled from the site indicated in green.

Figure 4-6 Sampling sites south of Christchurch

Sampling sites are indicated by red markers.

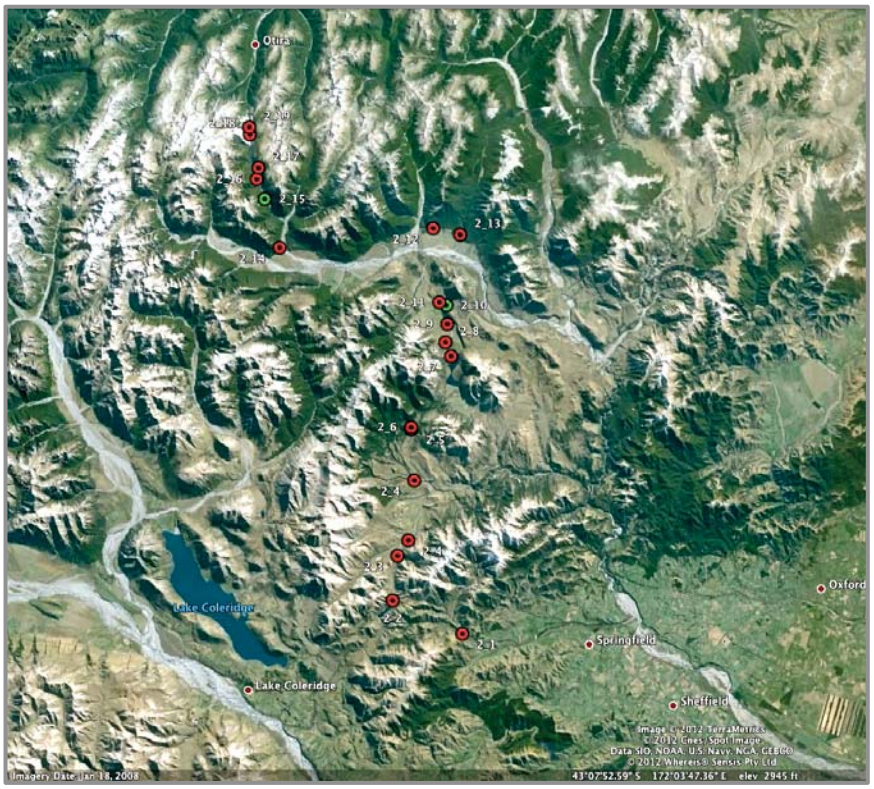


Figure 4-7 Sampling sites around Arthur's Pass

The markers indicate sites that were sampled. Sites where *E. sorex* was found are indicated in green.

4.3.3 DNA extraction and Illumina sequencing

The mostly diatomaceous sediment was washed twice in RNAlater. Aliquots of the sediment of about 200 - 300 µL volume were transferred to 1.5 mL Eppendorf tubes. Fast setting sediments (i.e. sand) were removed at this stage. The aliquots were washed one more time in RNAlater, removing the supernatant as thoroughly as possible. At this stage the composition of the samples was checked again under the microscope for composition and signs of degradation. Aliquots that were not used immediately were stored at -80ºC.

DNA extractions were initially performed on two aliquots using CTAB (cetyltriammonium bromide) and Phenol/Chloroform for lysis and the removal of proteins and contaminants, and Quiagen DNeasy columns for DNA recovery. One mL pre-heated heat extraction buffer (0.1M Tris-HCl, pH8.0; 2% CTAB; 1.5M NaCl; 0.02M EDTA) each was added to each aliquot and the samples were incubated at 65ºC for 30 min. One of the samples was mechanically disrupted with a pre-heated pestle during incubation while in the other this step was forgone. The tubes were centrifuged (18,000xg) for 5 min at room temperature. The aqueous phases were then transferred to new tubes containing 750 µL 49:1 chloroform:isoamyl alcohol mixture and inverted several times to mix. Centrifugation, transfer of the aqueous phase and mixing with chloroform/isoamyl alcohol were repeated once more and the tubes then centrifuged (18,000xg) for one minute at room temperature. The aqueous phase was then transferred into a new tube and the DNA extraction was continued following the Qiagen DNeasy protocol from step 13 onwards as described in the DNeasy Plant Handbook (07/2006, Mini Protocol: Purification of Total DNA from Plant Tissue). Agarose gel electrophoresis and the Nanodrop were used to determine DNA concentration and quality. The Qiagen RepliG mini Kit was used to amplify 5 µL of the higher concentrated DNA extraction following the standard protocol (Qiagen). Of the resulting 10 µg of DNA about 7 µg was submitted for sequencing on the Illumina GA II at the Massey University Genome Service in Palmerston North, New Zealand.

Following technical difficulties with the Illumina GAII at the sequencing provider during the first sequencing run, a second run was arranged. Tentative mapping of the single-end reads produced by this run showed no reads mapping to the available *E. sorex* sequences, casting doubt on the suitability of the extraction method. For this reason another DNA extraction protocol was developed for the second run based on the

Roche High pure PCR template preparation kit taking into consideration that diatom DNA is most commonly extracted with an initial proteinase K digestion step (Iwatani et al., 2005; Ravin et al., 2010; Scala et al., 2002; Jacobs et al., 1992).

This extraction procedure was as follows: Approximately 200 µL of diatom cells were mixed with 400 µL of lysis buffer and 80 µL proteinase K (as provided by kit). The mixture was then incubated for 3 hours at 55ºC. During incubation the sample was shaken every 30 minutes. The incubation was stopped when a microscopic examination of the sediment confirmed that the frustules were empty and the cells had in fact lysed. Four hundred µL of binding buffer and 200 µL isopropanol were added and the tube inverted several times to mix. The mixture was centrifuged (13,000xg) for 5 min and the supernatant then loaded onto two Roche High Pure Filter tubes. Both tubes were centrifuged (8,000xg) for one minute and the wash steps carried out as described in section 2.8 of the kit manual. The DNA was eluted from both columns with the same 100 µL of 70ºC warm elution buffer.

Electrophoresis on a 1% agarose gel confirmed that the extraction yielded high molecular weight DNA at a high concentration. A RepliG amplification was therefore omitted to avoid introducing replication bias.

Illumina sequencing was carried out again by the Massey University Genome Service in Palmerston North, New Zealand. Following the recommendations of the sequencing service a sample of 7 µg genomic DNA was again submitted for 100 bp paired-end sequencing on one lane on the Illumina GA II.

4.3.4 Mapping

The reads obtained from the sequencing runs were mapped to reference sequences using Bowtie2 (Langmead and Salzberg, 2012). Different parameter combinations were used depending on how close the available reference sequences were expected to be to the sample. The mapping results in SAM format were processed with a perl script (SAMtrimmer.pl, see Appendix A) to remove entries for reads that did not map. The resulting SAM files were then converted into sorted bam files using Samtools 0.1.18 (Li et al., 2009). Mapping results were visualised in Tablet (Milne et al., 2010).

The reads were mapped against three sets of reference sequence to identify reads originating from the genome of the diatom's endosymbiont: i) the putative genomic sequences of the spheroid body of *R. gibba* (see chapter 3), ii) *R. gibba* spheroid body *Nif* operon and 16S rRNA sequence (Prechtl et al., 2004) and iii) all sequences available at the time on GenBank for the cyanobacterial genus *Cyanothece*. The mapping parameters as detailed with the results in Table 4-2 were chosen to allow for relatively frequent base substitutions between template and reads to accommodate the high mutation rate commonly observed in endosymbionts.

To identify reads originating from the diatom's nuclear genome all reads were mapped against three sets of diatom sequences sourced from GeneBank. The reads were mapped against i) 18S rRNA sequences that represent the whole taxonomic breadth of diatoms, including rhopalodian diatoms, and some representatives of aquatic plants; ii) all sequences available for *Epithemia sorex* at the time; iii) all sequences available for rhopalodian diatoms at the time.

4.3.5 Assemblies

The reads were assembled using Abyss 1.3.3 (Simpson et al., 2009). Tentative assemblies of subsets of the read data were performed on a local server (awc3, Massey University, Turitea). These tentative assemblies were carried out in single end mode (i.e. ignoring positional information for paired reads) to reduce the computational requirements. Based on the results the full assemblies were performed in paired-end mode on the complete sequence data as produced by both Illumina GA II sequencing runs. The full assemblies had very high memory requirements due to the high complexity of environmental data and were run on Pinnacle, a high memory server located at Massey University, Albany. The assembly results were exported in form of several fasta files, containing the assembled contigs and scaffolds that join several contigs based on paired-end information.

Three paired-end assemblies, with the addition of available single read data, were performed testing two different k-mer values and using reads that had been trimmed with the DynamicTrim script from the SolexaQA toolkit or untrimmed reads (Cox et al., 2010). No other parameter combinations could be tested due to the high requirements in computing time and memory of these assemblies.

A perl script (AbyssContigSifter.pl, see Appendix A) was used to filter out short contigs and assess the level of fragmentation of the assembly results.

4.3.6 BLAST analyses

BLAST (Altschul et al., 1990) searches were performed on the NCBI BLAST website and locally on the awc servers. The code for local BLAST as well as preformatted GenBank databases were obtained from the NCBI BLAST website. Custom BLAST databases were built from sequences obtained from GenBank.

BLAST searches were performed on contigs longer than 500bps produced by the Abyss paired-end assembly of untrimmed reads with a k-mer value of 25. BLAST reports were visualised in MEGAN4 (Huson et al., 2011).

4.3.7 Analyses of eukaryotic signature proteins (ESPs)

Local tBLASTn searches were performed on contigs longer then 500bps to identify eukaryotic signature proteins based on sequence similarity. The ESP sequences of 274 proteins from *Giardia lamblia* and of 8000 proteins from *Homo sapiens* were used as queries in two separate searches. *Homo* and *Giardia* sequences were kindly provided by Jian Han (Han, 2012).

The tBLASTn search results were visualised with EPoS 0.9.1 (Griebel et al., 2008). Contigs that were likely to contain conserved ESP sequences were identified based on bit score, e-value and query coverage of the hits. The respective sequences were first verified via read mapping and then used as queries for BLASTx searches against the non-redundant protein database (nr) on the NCBI blast severs. The BLAST results for each query sequence were searched for representatives of the different main branches of the eukaryotic tree and an as diverse as possible sample of homolog protein sequences was retrieved. In each case sequences from the two fully sequences Diatoms – *Thalassiosira pseudonana* and *Phaeodactylum tricornutum* – were included in the data set. Each contig sequence was translated using the ExPASy translation tool (http://web.expasy.org/translate/,(Gasteiger, 2003)). Based on the BLASTx results the correct frame and sequence section were chosen and added to the fasta file of homolog protein sequences for subsequent alignment.

Protein alignments were performed using ClustalW. The alignments were edited by hand to minimise misalignments and taxa that did not align well were removed. Neighbour joining was used to build phylogenetic trees from the alignments. Protein alignment, alignment editing and tree building steps were all performed in Geneious Pro 5.6 (Drummond et al., 2012).

4.3.8 Sequencing of 16S and 18S rDNA genes

Universal primers were based on those developed by Medlin et al. (1988) and Giovannoni (1991) and used to amplify 18S and 16S rDNA sequences respectively. Primer sequences and expected product sizes are listed in Table 4-1. The template was the same whole DNA extraction that was used in the second Illumina sequencing run. The primer sequences are given in Table 4-1.

Table 4-1 Primer sequences and expected product sizes

Primer ID	Primer sequence	Product size
18S F	AAC CTG GTT GAT CCT GCC AGT	1.8kb
18S R	TGA TCC TTC TGC AGG TTC ACC TAC	
F 27 (16S F)	AGA GTT TGA TCC TGG CTC AG	1.5kb
1492 (16S R)	TAC GGY TAC CTT GTT ACG ACG AC	

The polymerase chain reactions (PCR) were conducted using Roche FastStart Taq DNA polymerase as recommended by the manufacturer in 20 µL reactions. The temperature profile was as follows: 94ºC for 5 min; 35 cycles of 94ºC for 50 sec, 52ºC for 50 sec, 72ºC for 90 sec; 72ºC for 5 min. The PCR reactions produced products of the expected size. The products were purified by gel extraction using the Zymoclean™ GelDNA Recovery Kit.

The PCR products were then cloned with the Invitrogen TOPO-Cloning KIT and cells. Ten colonies with 16S rDNA sequences and thirty with 18S rDNA sequences were picked for colony PCRs. Colony PCR reactions were performed under the same conditions as described above. Colony PCR products were submitted for Sanger sequencing at the Massey University Genome Service in Palmerston North, New Zealand. Sequences were edited and aligned in the Sequencher sequence analysis software (Gene Codes Corporation).

4.4 Results

4.4.1 DNA extraction and sequencing

Agarose gel electrophoresis (see Figure 4-8) and Nanodrop measurements showed that the mechanically disrupted sample yielded substantially less DNA (2 ng/µL versus 5 ng/µL), however both extractions produced high molecular weight DNA with a 260:280 ratio of about 2.1.

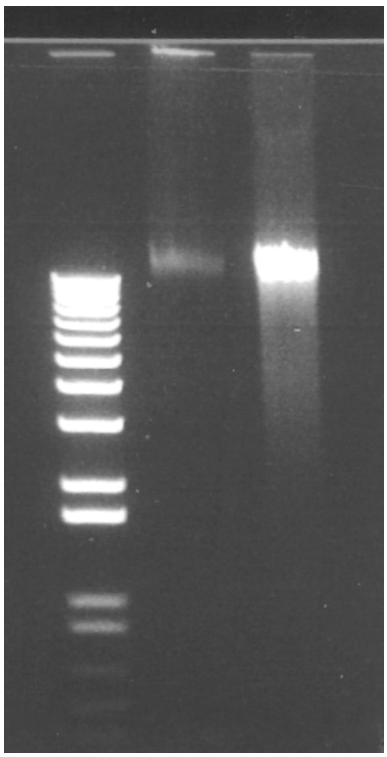


Figure 4-8 Environmental sample whole DNA extraction (CTAP extraction protocol)

Electrophoresis was carried out in 1% Agarose gel. Left lane: 1kb+ ladder; Middle lane: 1 µL of environmental DNA extraction; Right lane: 10 µL of environmental DNA extraction. This DNA extraction was used for paired-end Illumina sequencing.

The first sequencing run was aborted after sequencing the first 100bp end due to technical difficulties that caused unacceptably high error rates. This run did however yield about 8.4 million 100bp reads of good quality. See Appendix A for Solexa QA assessment of sequence qualities.

The second sequencing run was completed successfully and yielded 33.15 million 2 x 100bp reads.

4.4.2 Mapping

The Illumina reads each of length 100 bps were mapped against reference sequences sourced from GenBank to gauge the taxa composition and coverage. The settings used and basic statistics of the mappings are summarized in Table 4-2. Relatively relaxed settings that allowed for mismatches within the alignment seed were used for reference sequences that were expected to be divergent from the sequencing template. Stricter settings were used on reference sequences that could be expected to be taxonomically close and/or well conserved (for example the *E. sorex* 18S rDNA sequence).

Very few reads aligned to genomic sequences from the *R. gibba* endosymbiont. On average only one read aligned per 225 or 171 bps of reference sequence, depending on mapping parameters. The distribution of reads along the template was uneven. At no section of the reference sequence did reads map with a significant coverage that would indicate a conserved region, but rather aligned to low-complexity regions.

Mapping the reads against all sequences available for the cyanobacterial genus *Cyanothece* produced very similar results (see Table 4-2). More reads aligned to these sequences, proportional to the greater overall length of the reference sequences, but all alignments appeared to be due to random sequence similarities.

Figure 4-9 shows a detail of the mapping of the paired-end data against the 16S rDNA sequence of the SB of *R. gibba*. The overview pane along the top of the screenshot shows how uneven the distribution of reads is along the reference sequence. The detailed visualisation of the reads mapping a region with high coverage shows a high frequency of mismatches, demonstrating that this is a very non-specific mapping of reads to a reference sequence due to the relaxed parameters used. More stringent parameters failed to align a significant number of reads. The *nif* operon can be expected to be highly conserved among all rhopalodian spheroid bodies due to its crucial importance for the symbiotic relationship. However, the mappings to the sequence of the *R. gibba nif* operon did produce only very infrequent alignments of reads, mostly to low-complexity regions. Over all the mappings did not provide any indications that DNA of a rhopalodian endosymbiont had been sequenced.

Reference sequence	Ref seq. Length	L	N	Preset	Read pairs aligning concordantly	Read pairs aligning disconcordantly	Single reads aligning
R. gibba SB contigs	1925279	17	0	n/a	2860	226	2360
"	"	20	1	n/a	3292	292	4052
SB nif genes	51475	20	1	n/a	193	9	52
CyanotheceAll	27391697	20	1	n/a	23536	500	57770
NCBI Rhopalodiales1	40071	17	1	n/a	7097	170	5732
NCBI Rhopalodiales2	"	-	-	Very sensitive	6324	130	5692
NCBI Rhopalodiales3	"	20	0	Sensitive	4836	111	2868
NCBI Rhopalodiales4	"	30	1	n/a	5899	116	4782
NCBI Rhopalodiales5	"	30	0	n/a	3846	60	2168
Diatom/plant 18S	41638	17	0	Very sensitive	2660	15	2232
NCBI E. sorex1	4158	30	0	n/a	266	23	662
NCBI E. sorex2	"	70	1	n/a	1756	11	1152

Table 4-2 Bowtie2 parameters and statistics for each mapping.

L: Seed length; N: maximum number of mismatches allowed within seed; all other settings are the same as in the 'Sensitive' preset unless the preset 'Very sensitive' was used; D: maximum number of consecutive seed extension attempts, i: sets the interval at which alignments seeds are extracted along a read, R: maximum number of times a read is 're-seeded'.

Default settings for Presets:

Very-sensitive : -D 20 -R 3 -N 0 -L 20 -i S,1,0.50

Sensitive : -D 15 -R 2 -N 0 -L 22 -i S,1,1.15

Figure 4-9 Mapping of paired-end reads against *R. gibba* 16S rRNA sequence

The coverage across the whole reference sequence is indicated in blue at the top. Reads mapping to a subsection of the reference sequenced are displayed in packed mode at the bottom of the figure. Mismatching bases are indicated in a lighter colour.

Mapping the reads to all available *E. sorex* sequences as well as the collection of all sequences available for rhopalodian diatoms failed to identify any reads from a diatom source, even though a broad range of settings was tested (see Table 4-2), ranging from

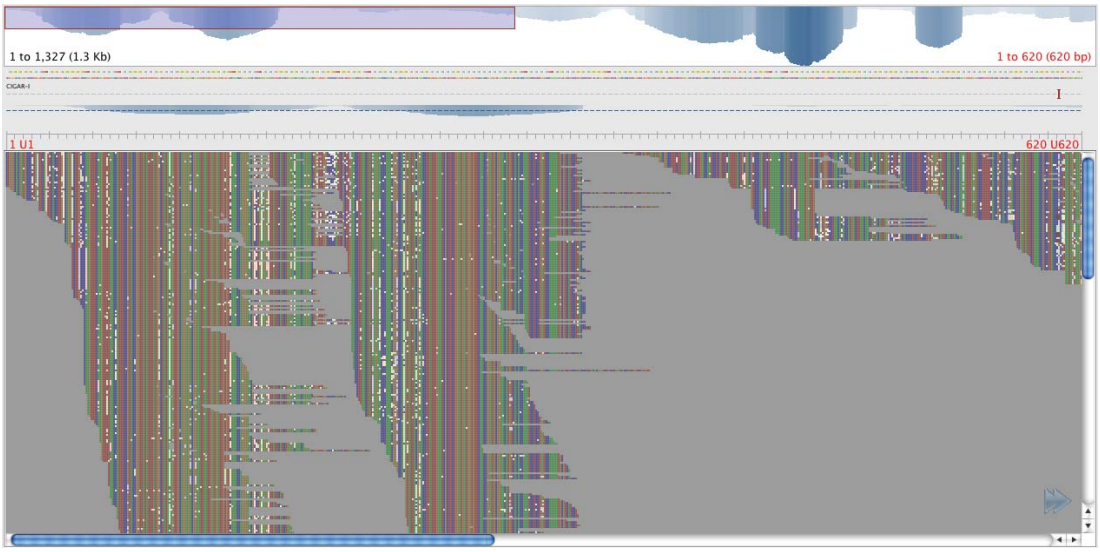


Figure 4-10 Mapping of paired-end reads against the *E. sorex* 18S rRNA sequence

The coverage across the whole reference sequence is indicated in blue at the top. Reads mapping to a subsection of the reference sequenced are displayed in packed mode at the bottom of the figure. Mismatching bases are indicated in a lighter colour.

high specificity to identify *E. sorex* reads to a high sensitivity to identify reads from distantly related diatom species.

The 18S rDNA sequence from *E. sorex* produced results very similar to the 16S rDNA mapping described above. A section of the mapping is shown in Figure 4-10. Diatom specific genes like those for a silicic acid transporter only produced very infrequent alignments. Diatom chloroplast genes showed very unspecific mappings due to the high level of conservation of the protein domains across lineages but with much lower overall coverage compared to the 16S and 18S rDNA sequences. Figure 4-11 shows a screenshot of the mapping for the chloroplast gene *psbC* from *Rhopalodia contorta* as an example.

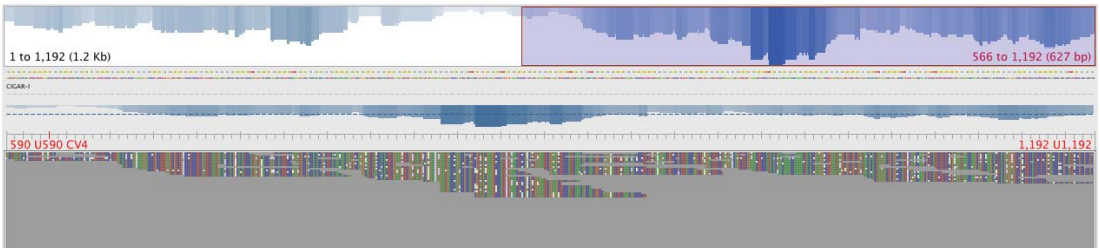


Figure 4-11 Mapping of paired-end reads against the sequence of the gene *PsbC* from *R. contorta* (chloroplast)

The coverage across the whole reference sequence is indicated in blue at the top. Reads mapping to a subsection of the reference sequenced are displayed in packed mode at the bottom of the figure. Mismatching bases are indicated in a lighter colour.

The paired-end reads were also aligned to a selection of twenty-five different 18S rDNA sequences of diatoms and several aquatic plants. None of the trialled 18S sequences produced alignments with a low number of mismatches that would have been indicative of a close taxonomic relationship to the template source.

4.4.3 Assemblies

Two assemblies of the untrimmed reads were performed with k-mer values (see Section 1.4) of twenty and twenty-five respectively. A k-mer value of twenty was found to produce a more fragmented assembly with a lower N50 value and markedly less assembled sequence. A third assembly was then performed with a k-mer of twenty-five, using reads that had been trimmed according to base call qualities using the

DynamicTrim algorithm on default settings. The use of quality-trimmed reads had a negative effect on the assembly performance. This assembly resulted in less assembled sequence and shorter contigs than the assembly of identical settings using the raw reads. Table 4-3 lists the basic statistics for the three assemblies. Figure 4-12 shows the contig size distribution in absolute numbers. The shorter k-mer produced a more fragmented assembly with no contig longer than 1366 bps while the longer k-mer of twenty-five produced more and longer contigs. The use of trimmed reads affected predominantly the longer contigs and markedly reduced the number of contigs in the largest size categories. All available statistics show the best results for the assembly were produced using a k-mer of twenty-five and the untrimmed reads. All subsequent analyses were performed on contigs from this assembly.

Table 4-3 Summary of the assembly statistics

The three different assemblies are indicated by the kmer value used and whether the reads used were trimmed with DynamicTrim or not. There were the only differences between assemblies. All values are given for the complete scaffolds, i. e. the contigs might contain N-stretches that were inferred based on paired-end positional information.

n:100: number of contigs longer than 100 bps; n:N50: number of contigs longer than N50 threshold; n50: contigs longer than this value amount to 50% of assembled sequence.

	k = 25 trimmed	**k = 20 untrimmed**	**k = 25 untrimmed**
Number of reads used	6889371	48.1e6	9128603
n:100	1503923	1044785	1838326
n:N50	361599	381891	402325
Median contig length	144 bps	125 bps	139 bps
Mean contig length	207 bps	144 bps	211 bps
n50	231 bps	140 bps	237 bps
Maximum contig length	15675 bps	1366 bps	25128 bps
Total assembled sequence	311.8e6 bps	151.1e6 bps	388.9e6 bps

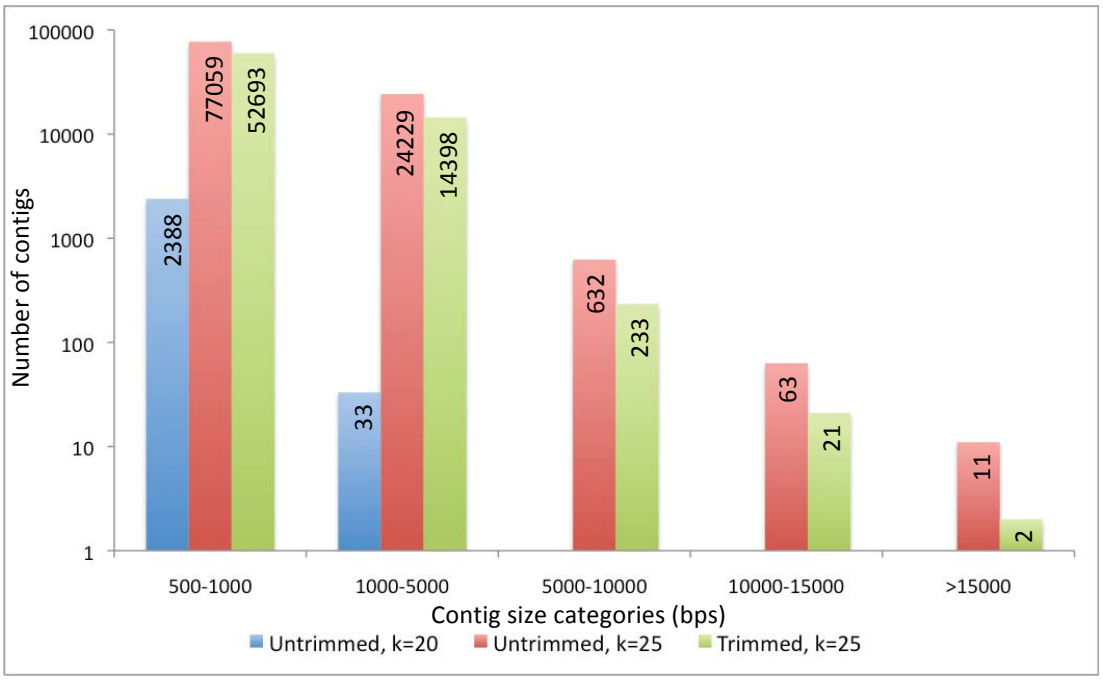


Figure 4-12 Contig size distribution of assembly results

The number of contigs within size categories are visualised as columns for the three different assemblies. The assemblies are identified by their kmer value and template type. The y-axis is logarithmic and shows the number of contigs in each category, the exact value is given as label for each column; categories on the x-axis are for contig length in bps.

The average coverage and standard deviation were calculated for the contigs in the different size categories. The results are given in Table 4-4 and show that the larger contigs generally have a higher coverage. This indicates that they originate from a more abundant DNA source in the sample, while templates that are less abundant yield lower coverage and generally produce more fragmented assemblies.

Table 4-4 Average coverage of contigs assembled from untrimmed reads using a k-mer of 25

Contig size category	Average coverage	Standard deviation of coverage
> 500bps	6.833080542	5.30179171
> 5 000bps	10.89139618	4.42462813
> 10 000bps	9.438021874	1.128461644

4.4.4 BLAST results

Based on the assumption that the longest contigs were produced by the most abundant template in the sample, the eleven contigs longer than 15kb were searched with BLASTn against the complete GenBank nucleotide collection (nt, as of May 2012). They did not produce any significant hits.

This search was then extended to all contigs longer than 500bps in a local BLASTn search against the complete nt database. The results are shown in Figure 4-13 as visualised in MEGAN.

The contigs were assigned to clades across the entire taxonomic breadth of the GenBank nucleotide collection, but as indicated by the green arrow no contigs were specifically assigned to diatoms. A more sensitive tBLASTx search with the same sequences could not be conducted because it exceeded the available computational resources.

Subsequent BLAST searches were limited to collections of sequences closer to the target organism. A BLASTn search against all diatom sequences deposited in GenBank at the time (as of May 2012) only resulted in fourteen contigs being assigned to the diatom lineage. Ten of these were assigned based on very short BLAST hits with scores below one hundred, while the remaining four contigs hit the highly conserved rRNA sequences and chloroplast genes. All these similarities are thus not significant. A more sensitive BLASTx search against all diatom protein sequences assigned 42 000 of the 102 000 contigs to a diatom species or clade (see Figure 4-14). However, even this result is likely to be insignificant given that the distribution of assignments indicates that the matches are strongly biased by the database composition with most hits assigned to *P. tricornutum* and *T. pseudonana*, the only two diatoms that have been fully sequenced.

Neither a BLASTn nor a tBLASTx search against the Sequences of the *Rhopalodia gibba* spheroid body (see Chapter 3) resulted in any significant hits (not shown).

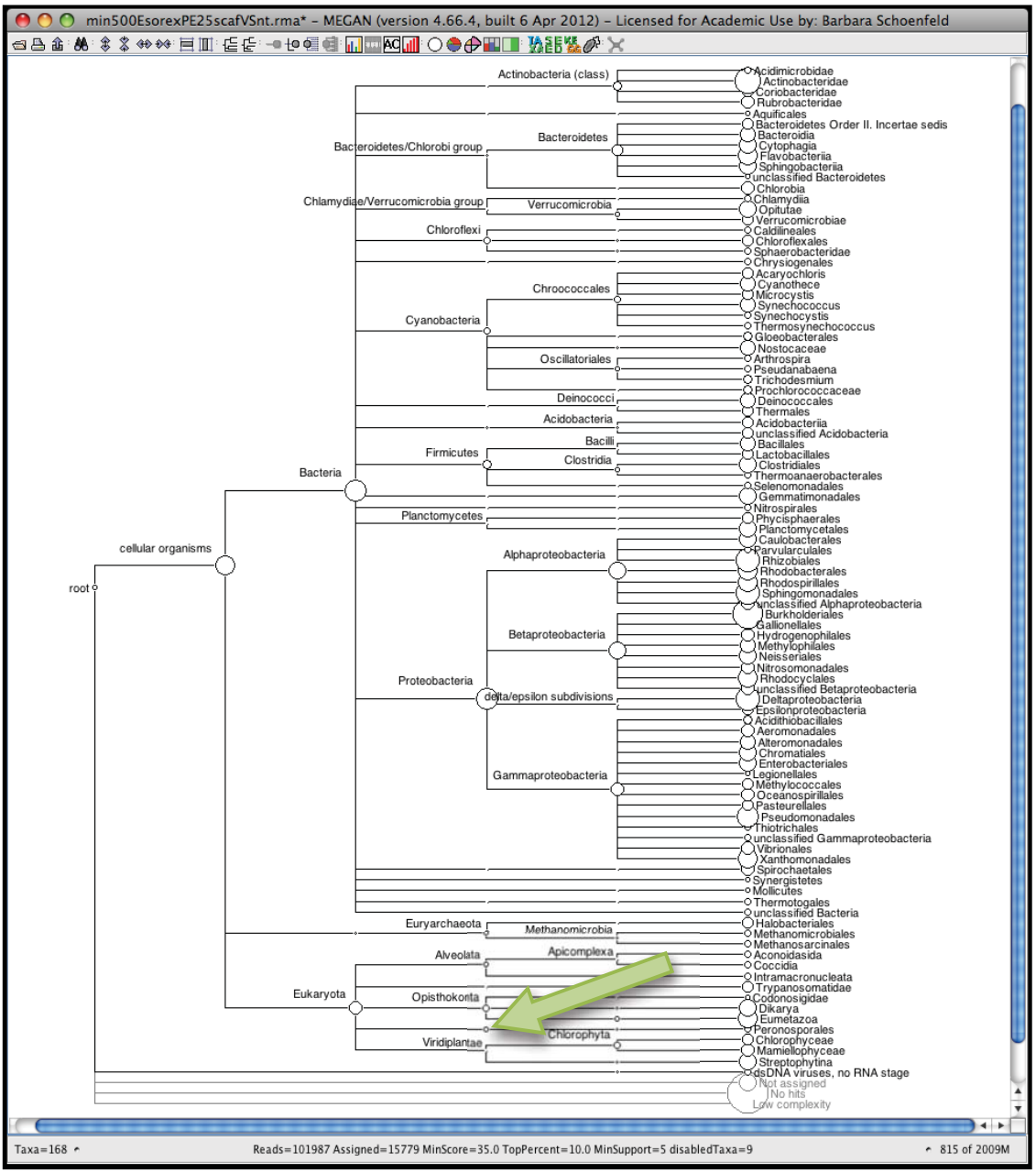


Figure 4-13 Results of a BLASTn search against the nt database visualised in MEGAN

The diagram shows the parts of the NCBI taxonomy to which sequences were assigned. The number of reads assigned to a node are indicated by the diameter of the circles. The green arrow marks where the Diatom lineage would branch off in this tree had any sequences been assigned to that lineage.

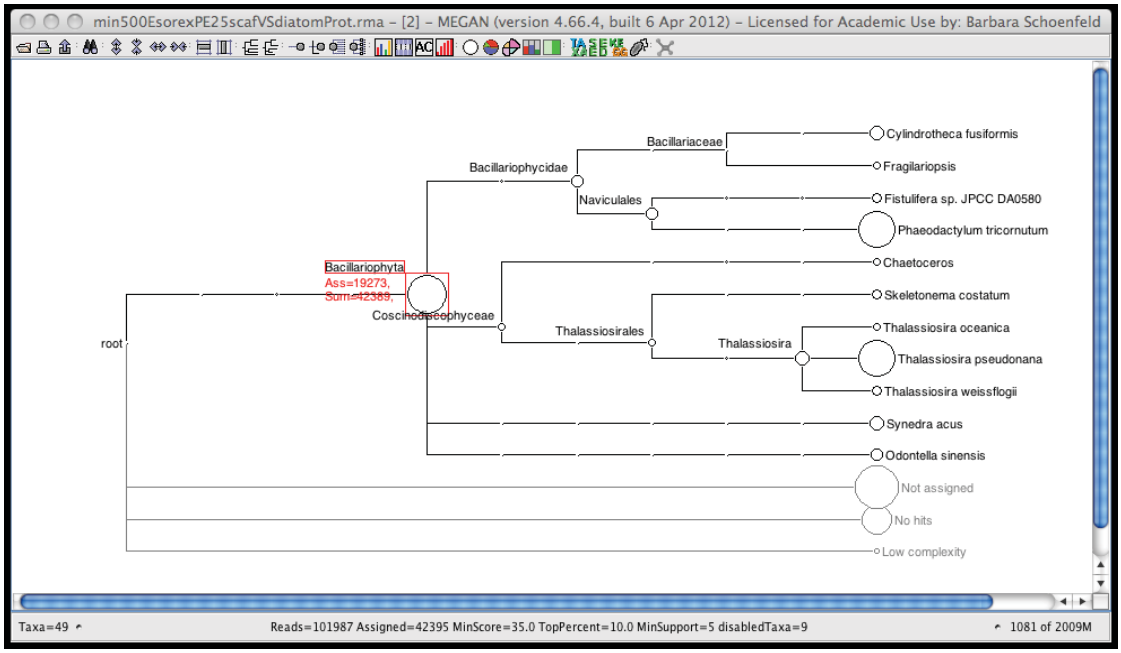


Figure 4-14 Results of a BLASTx search against all diatom sequences in GenBank visualised in MEGAN

The diagram shows the parts of the NCBI taxonomy to which sequences were assigned. The number of reads assigned to a node are indicated by the diameter of the circles.

4.4.5 Eukaryotic signature proteins

Eukaryotic signature proteins are highly conserved and orthologs are likely to be present in eukaryotic genomes. This makes them relatively easy to detect in an assembly even if close reference sequences for the sequenced organisms are not available. Once identified the sequences of the ESPs can be used to test if the source organism a diatom based on the similarity to known diatom ESPs.

Local tBLASTn searches of eukaryotic signature proteins (ESPs) from *Giardia lamblia* and *Homo sapiens* against contigs longer than 500 bps found significant sequence similarities to nine contigs. Two of these contigs matched the same ESPs. Mapping the reads against the ESP containing contigs did not reveal any misassemblies. The contigs generally mapped without mismatches, apart from the random sequencing error, and with an even, albeit low coverage. Table 4-5 lists the average coverage for each ESP containing contig.

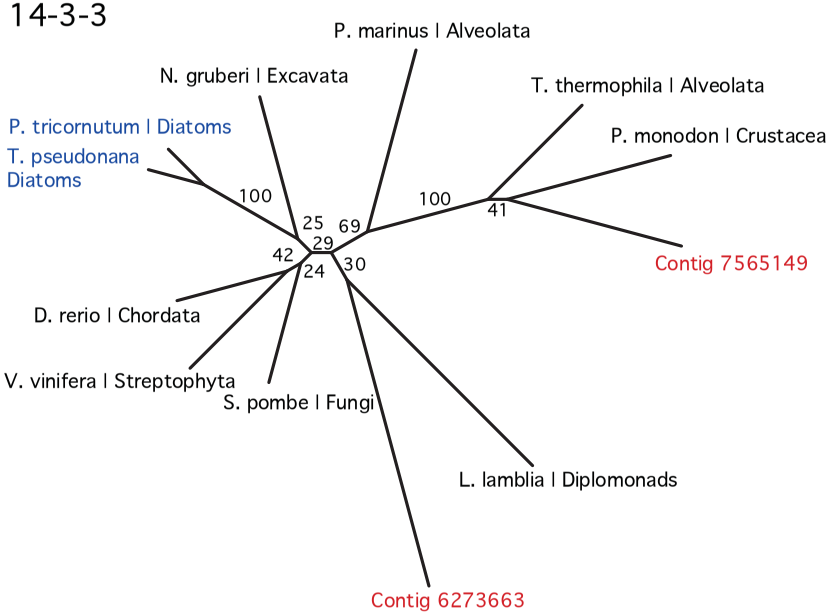
Each of these ESP-matched contigs was used as query sequence for BLASTx searches of the GenBank protein database for homologous proteins. A taxonomically as diverse as possible set of homologous sequences was then selected from the search results for each contig, downloaded and aligned with the translated contig sequence. The Table given in Appendix E lists all sequences used in these ESP alignments with a functional description of the protein and taxonomic information for the source organism as provided by GenBank.

Table 4-5 ESP containing contigs and their average coverage

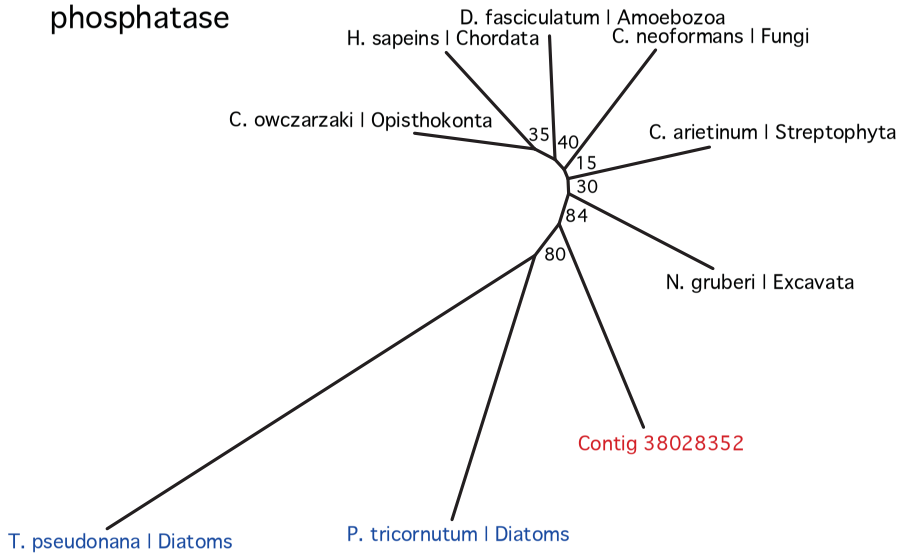
Contig name	Average coverage
>11211941 778 3234	4.300531915
>7565149 586 2165	3.866071429
>266513 695 2841	4.246636771
>3802835 1434 5838	4.146306818
>11454230 968 11548	12.25902335
>6273663 598 2864	5.006993007
>7697359 977 3829	4.026288118
>11428777 783 2920	3.857331572
>8191597 1058 4476	4.337209302

The alignments were then used to build Neighbour Joining trees. Figure 4-15 shows four of these phylogenies with their bootstrap support. Each phylogeny includes two diatom sequences, from *T. pseudonana* and *P. tricornutum* respectively. These two reliably group together in all nine phylogenies with high bootstrap support. The positions of the contig sequences are incongruent between the different gene trees, as are the positions of the representatives of the major eukaryotic lineages relative to each other. The bootstrap support for internal branches is generally low. However, the contig sequences do not ever group with the diatom sequences. In summary while this analysis did not conclusively identify to which eukaryotic lineage the ESPs detected in the environmental sample belong, it can nevertheless be concluded that they are not derived from *E. sorex* or other diatoms.

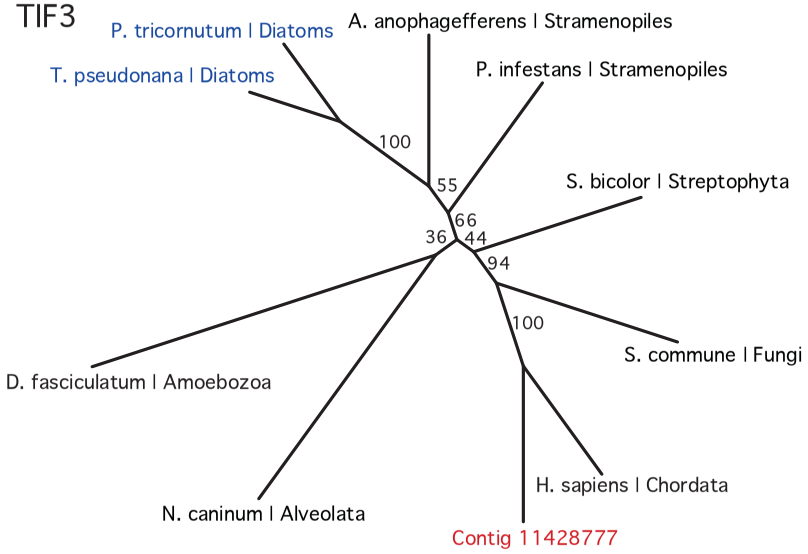
14-3-3



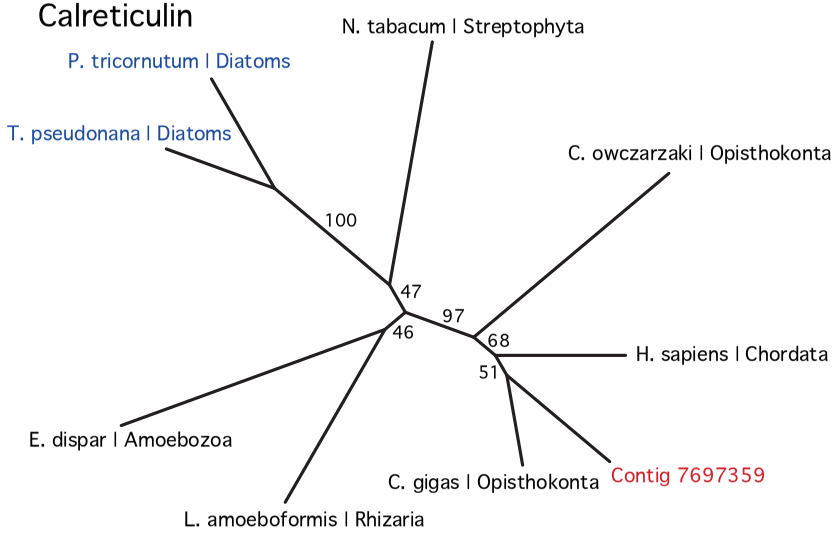
Ser/Thr protein phosphatase



TIF3



Calreticulin



4.4.6 16S and 18S rDNA sequencing

Universal primers were used to amplify 16S and 18S rDNA sequences in the environmental DNA extraction. The PCR products were cloned and nine clones of 16S rDNA sequences and twenty-nine clones of 18S sequences were Sanger sequenced. The sequences were then used in a remote BLASTn search of the complete GenBank nucleotide database (nt) for similar sequences. The results are shown in Table 4-6. The table lists the percentage of the most notable taxonomic groups among the 100 best BLAST hits and also the range of sequence identity between the query sequence and the 100 most similar sequences in the search database. No identical match was found for any of the query sequences. All 16S rDNA sequences most closely match uncultured and unidentified bacteria. For two of them the bacterial lineage can be narrowed down to Proteobacteria, and for another two to Verrucomicrobia because a proportion of the target sequences have been classified to that degree. Based on the BLAST results only one of the 16S sequences, 16S-4 could potentially originate from *E. sorex* as it shows strong similarity to nine diatom chloroplast 16S sequences in the database.

The twenty-nine clones of 18S rDNA produced nine different 18S sequences. Five of these sequences, representing twenty-two of the clones, show a strong similarity to 18S rDNA sequences from members of a family of oligochaete worms, the Naididae. Two sequences show strongest similarity to Ciliates and one to Nematodes. Only one of the twenty-nine clones produced a sequence that is likely to stem from a Diatom, though not one of the group Bacillariophyceae but Fragilariaceae.

Figure 4-15 Four examples of ESP gene trees (previous page)

Shown are four unrooted phylogenies produced by Neighbour Joining of ESP sequences. Sequences translated from assembly contigs are indicated in red. The two diatom sequences included in each of the phylogenies are indicated in blue. The labels give the bootstrap support for internal branches.

Table 4-6 Summary of BLAST results for 16S and 18S rDNA sequences

Each row summarises the taxonomic classification and level of sequence identity in percent for the hundred most similar sequences in the GenBank nt database for each unique 16S and 18S rDNA sequence. The sequence identifiers listed in the first column specify the clone or clones that yielded the respective sequence.

Sequence identifier	Taxonomic distribution of 100 best hits	Range of seq. identity
16S-1	100% uncultured soil bacteria	99-96%
16S-3	100% uncultured bacteria	98-84%
16S-4	9% diatom (chloroplast), 77% uncultured bacteria, 14% uncultured organisms	99-96%
16S-5	100% uncultured bacteria, of these 39% Verrucomicrobia	97-85%
16S-6	41% uncultured bacteria, 59% uncultured organism	99-97%
16S-7	100% uncultured bacteria	98-89%
16S-8	100% uncultured bacteria, of these 28% Verrucomicrobia	98-91%
16S-9	100% uncultured bacteria, of these 31% Proteobacteria	97-89%
16S-10	100% uncultured bacteria, of these 35% Proteobacteria	98-96%
18S-1, 2, 3, 5, 7, 8, 9, 10	98% Naididae (type of segmented worms), 2% uncultured Metazoa	99-97%
18S-4	85% Ciliates, 15% uncultured Eukaryotes	97-89%
18S-6	88% Diatoms (75% Fragilariaceae), 12% uncultured Eukaryotes	99-96%
18S-11, 12, 13, 19, 20, 21, 22, 23, 25, 26, 28	95% Naididae (type of segmented worms), 3% uncultured Metazoa, 2% uncultured Eukaryotes	99-97%
18S-14, 17	73% Ciliates, 20% uncultured Eukaryota	98-86%
18S-15	96% Naididae (type of segmented worms), 2% uncultured Metazoa, 2% uncultured Eukaryotes	98-96%
18S-16	96% Naididae (type of segmented worms), 2% uncultured Metazoa, 2% uncultured Eukaryotes	97-95%
18S-18, 24, 27	94% Nematodes, 3% uncultured Eukaryota	94-83%
18S-30	95% Naididae (type of segmented worms), 3% uncultured Metazoa, 2% uncultured Eukaryotes	93-91%

4.5 Discussion

4.5.1 Sample Quality

The main prerequisite for this project was to find a sampling site where *Epithemia sorex* occurred with such a high abundance that it would constitute the dominant species in an environmental sample. Despite the global distribution and wide range of tolerated environmental parameters and habitat types, *E. sorex* is not a commonly occurring member of epiphytic and epipelic communities in New Zealand. It had in the past only been sighted very sporadically within the sample area (Phil Novis, personal communication). Given the relatively low expectations for discovering a suitable natural population of *E. sorex* the observed high ratio of *Epithemia*-like cells to contaminants in the environmental sample collected near Arthur's pass led to high expectations for the success of this project.

Other observations were also encouraging: the occurrence of *E. sorex* at this sample site was accompanied by very low levels of sand and mineral contamination; further the epipelic life style of *E. sorex* made the removal of other algae and organisms that were not attached to the same substrate relatively easy.

The diatom samples were stored and transported in RNAlater Solution to protect the DNA for sequencing. As the name indicates RNAlater Solution was developed to prevent RNA degradation in tissue samples to allow RNA based downstream applications even after prolonged storage in adverse conditions. It does so, among other things, by denaturing proteins, and it can be expected that such disruption equally benefits the stability of DNA within a sample. Thus RNAlater has routinely been used for DNA extractions as well as for RNA extractions (see www.invitrogen.com). RNAlater was initially developed for animal tissue samples and no empirical studies on its efficiency preserving diatom samples appear to have been undertaken. Nevertheless, after approximately four hours of incubation of the *E. sorex* sample in RNAlater, a change in colour became evident as shown in Figure 4-16. Figure 4-16a shows the natural brown colour of *E. sorex* cells stored in water. The cells stored in RNAlater shown in Figure 4-16b turned yellow-green. This colour change is most likely due to the denaturation of the proteins in the antennae complexes in the chloroplast. This suggests that the RNAlater Solution was successful to relatively quickly permeate the cells and denature intracellular proteins thereby protecting the DNA from degradation. However, it can be

speculated that denaturing intracellular proteins also resulted in an exposure of the frustules' silicate surfaces. DNA extraction from diatoms is known to be problematic due to the propensity of diatom silicate shells to bind DNA (Carter and Milton, 1993). This effect was observed in the present study during the first DNA extraction, where a mechanical disruption of the diatom cells resulted in a greater silicate surface area and much lower DNA yields. In the current project, a DNA extraction very similar to the one used by Nakayama et al. (2011) was used. These authors had previously used this method to successfully extract DNA from an *E. sorex* culture to sequence rDNA genes. The only significant difference between the protocol of Nakayama et al. (2011) and that used here was the use of RNAlater to preserve the sample.

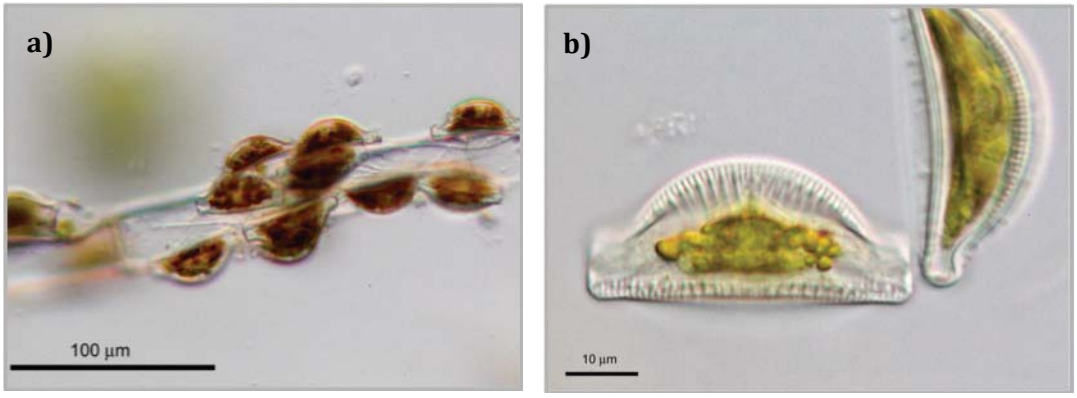


Figure 4-16 *Epithemia sorex* cells *in situ* and after incubation in RNAlater.
a) *E. sorex* cells *in situ*; **b)** *E. sorex* cells after four hours incubation in RNAlater

In the present study, microscopic surveys of the diatom frustules were made after the lysis step, and this examination showed that the destruction of the frustules was not necessary to release the cell contents. All frustules appeared empty under the microscope with neither pigments nor any cellular structures visible. Further, the extraction protocol produced high molecular mass DNA at a relatively high concentration. At the time these observations were encouraging in that the sampling and storage condition, and extraction protocol were optimal for the project and the template likely to be very suitable for high-throughput sequencing. The two sequencing runs produced good quality reads at a yield high enough to provide enough coverage for the assembly of the genomes of a diatom as well as its organelles and endosymbiont. The fully sequenced diatom genomes available on GenBank are smaller than 30 Mbp. Based on this size, the number of reads produced from this sample was expected to have been sufficient to achieve an average coverage of nearly 250, assuming no contaminants were present. Even in the presence of a limited number of

contaminants we expected that this would be sufficient to assemble at least partial genomes for *E. sorex*.

4.5.2 Sequence quality and assembly

Tentative assemblies of a subset of reads in single end mode resulted in little and highly fragmented assembled sequence at extremely low coverage due to the high complexity of the data. The final assemblies were therefore performed in paired-end mode on a high memory server. This highlighted the need to include as much information as possible into the assembly process. All available quality-trimmed reads were used in the assembly including the single end reads from the first, aborted sequencing run. Given the low coverage values calculated in the initial test assemblies, very small k-mer values were chosen, as longer k-mers result in lower k-mer coverage. The first assemblies established a k-mer of twenty-five to be more suitable for the data than a shorter k-mer of twenty. A third assembly was then conducted with a k-mer of twenty-five on untrimmed reads to test if the inclusion of all available sequence information would improve the assembly despite introducing more sequencing errors. This was indeed the case, and the likely explanation is corroborated by the low read coverage of most contigs in the assembly. The statistics given in Table 4-4 show that the average coverage of the contigs produced with the entire available data ranged from below two to not much higher than fifteen. For short read assemblies these are very low values given that whole genome assemblies are commonly based on data with 50 to 100 times coverage (Salzberg et al., 2012).

The fact that the untrimmed reads produced a "better" assembly in terms of contig length and total assembled sequence than the quality filtered reads shows that in this case quality filtering obstructed the assembly more by removing information than it aided it by removal or erroneous bases. This reflects that due to its metagenomic origin the data used here was of high complexity and low coverage. Both factors cause the removal of even small amounts of information to have severe effects on the success of the assembly, because the algorithms commonly used on short read data are designed to follow the basic assumption that the template is present in a homogenous and relatively high coverage. For a discussion of the effects of low and uneven coverage on assemblies see Section 3.5.1.

Given the lack of reference sequences to verify the assembly other than by mapping of reads it is difficult to assess whether the chosen assembly is in fact the most accurate

and correct. Assembly statistics are often insufficient to appropriately describe an assembly. They can only describe the amount and length of sequences produced but cannot give any information on the propensity of misassemblies. The most commonly used statistic, the N50 is in this case quite meaningless as its value can be severely skewed by an abundance of very small contigs as can be expected in an environmental sample containing traces of DNA from a multitude of organisms (Baker, 2012). The eleven longest contigs, which were also among the highest in average coverage (see Table 4-4), as well as those contigs that were identified as containing putative ESP sequences (see Table 4-5) were for this reason validated by mapping of the raw reads back onto them. None of these mappings uncovered obvious misassemblies but showed relatively even read coverage, both indicators for good assembly quality.

4.5.3 Mapping

Mapping is the most common and basic approach to assess and analyse high-throughput sequencing data. Mapping tools were developed to align short reads to long reference sequences. It is a fast and easy method to detect sequence changes compared to a close reference sequence and to detect misassemblies. Bowtie2 was chosen for this project because it supports gapped, local and paired-end alignment modes, and is therefore well suited to map reads to references from different species that can be expected to align to the reads with frequent mismatches. The reads were not trimmed but instead the local mode was used for mapping. The local alignment mode of Bowtie2 is able to ignore mismatches at the end of reads. Increased sequencing error rates or adaptor sequences are therefore less likely to impair its performance. It was thus not necessary to remove nucleotides with a low quality value as assigned by the sequencing pipeline, which would inevitably have resulted in the loss of genuine sequence information.

A wide range of mapping parameters was tested to improve the specificity of the mappings against 16S and 18S rDNA sequences but the results shown in Figures 4-9 and 4-10 could not be improved upon as settings that allowed less mismatches or required longer seed lengths resulted in no alignments.

Mapping settings were at first chosen to be very sensitive, with short seed length and mismatches allowed in the seed region to enable reads with several mismatches to align. This produced alignments of a relatively high number of reads, but these

alignments were non-specific with a high frequency of mismatches. The settings were then systematically changed to improve the specificity of the alignments and filter out less similar reads. At the end of this process reads only aligned with very low coverage but still showed multiple mismatches. This lead to the conclusion that the targeted diatom sequences were not represented in the sequencing results.

Given that the reads that mapped the 16S and 18S rDNA sequences were likely to themselves originate from 16S or 18S sequences, an attempt was made to identify the organisms they originated from. This was made difficult because the reads clearly represented a mixture of sequence variants with at least four clearly discernible patterns of mismatching bases. The tools available to generate consensus sequences from mapped reads, like mpileup from the SAMtoolkit (Li et al., 2009), are not designed to handle a mixture of reads that requires the reconstruction of several consensus sequences in parallel. Instead inconsistent mismatches are interpreted as sequencing errors. Attempts to distinguish the different consensus sequences by hand failed because low coverage regions were not bridged by paired reads. This meant that it was not possible to identify which consensus sequence fragments on both sides of the gaps came from the same source. BLAST searches with the resulting sequences were inconclusive.

Thus, all that could be achieved with this method was to show that no discernible reads from the target organisms were present in the sequencing results. None of the mappings was able to detect reads that aligned well with few mismatches to the available reference sequences. All attempts to identify the origin of sequences that were mapping to conserved sequences failed. It was concluded that while it might be possible to use a mapping approach in this way, to identify conserved sequences and produce a consensus sequence from the reads, this would only be practical from a pure culture, but not as yet from an environmental sample.

4.5.4 BLAST analyses with MEGAN

The standard bioinformatics tool to assess sequencing results is BLAST. The BLAST algorithms are tools to perform similarity searches of large sequence databases and the standard method of identifying sequencing results based on similarity to known sequences. BLAST is commonly used in metagenomics to assign reads to lineages or species based on sequence similarities. Given the amount of data produced by high-

throughput sequencing technologies this is very computing intensive. The computational resources required can be reduced by first assembling the reads into longer contigs. This removes redundant reads from the analyses but can blur estimates of the abundance of the source organisms.

The identification of reads in metagenomic samples is in general dependant on the availability of reference sequences in the databases. Due to the statistics involved the results, especially for short query sequences, are generally biased by the subject database composition. The problem is even more significant if MEGAN is used for the analysis of BLAST results. This became very apparent in this case where in the absence of genuine reference sequences the BLAST searches triggered MEGAN to assign reads to the taxa that were most strongly represented in the respective subject databases. MEGAN assigns every read that shows significant similarity to one or more sequences in the BLAST database to the lowest common ancestor of the matching sequences according to the NCBI taxonomy as shown in Figure 4-17a). MEGAN accepts by default BLAST hits with very low bit score[1] because it was designed for the analysis of reads as short as 35 bps. It disregards e-values[2] that could otherwise be used to correct for data base size for the same reason. This means that in the absence of genuine matches, reads can easily be assigned to clades or species based on chance similarities or short stretches of highly conserved sequence as shown in Figure 4-17b).

The analysis of the results of a BLASTx search of the contigs against all available diatom protein sequences in MEGAN is an informative example. The results were strongly biased by the database composition with most contigs assigned to *P. tricornutum* and *T. pseudonana*, the only two diatoms that have been fully sequenced. *T. pseudonanan* is a marine diatom and could not have been present in the sample. A change of the settings to increase the bit score threshold resulted in a dramatic reduction in the number of contigs that were assigned to any taxonomic level.

[1]The bit score is the sum of scores for each matching base or amino acid in a BLAST hit. It increases with quality as well as alignment length.

[2] The e-value is the probability that by chance there is another alignment with a similarity greater than the given score. It is dependent on the size of the searched database and properties of the query sequence.

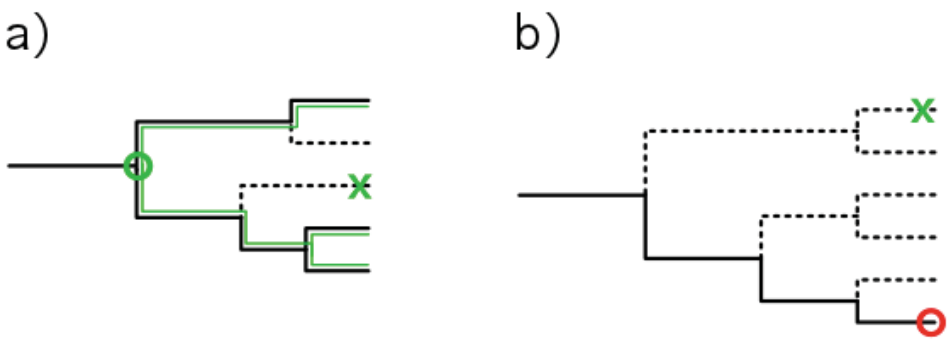


Figure 4-17 Working principle of the Last Common Ancestor (LCA) algorithm used in MEGAN and the effects of poor database representation

Solid lines represent taxa for which a blast search has produced hits to homologous sequences. The dashed lines represent taxa that are not represented in the database with sequences homologous to the querry. The circles indicate the nodes or tips to which LCA assigns the read while the crosses marks the actual source of the read within the taxonomy. **a)** The source organism is not represented in the database but several related taxa produce BLAST hits. LCA resolves the conflict by assigning the read to the node that represents the most recent common ancestor of all taxa that produce significant BLAST hits. The result gives a correct indication as to what taxonomic group the source organism belongs to. **b)** In this case the taxonomic group of the source organism is very poorly represented in the database. Only distantly related taxa produce significant hits and are used to assign the read to the latest common ancestor of those taxa. The result suggests that the source organism belongs to a different taxonomic group than is actually the case. Note that this effect tends to miss-assign reads disproportionally to those taxa that have been fully sequenced as they usually provide the only homologue sequences for the majority of the genome. Given the lack of fully sequenced genomes for many groups of prokaryotes and protozoa this can result in dramatic miss assignments.

A survey of random samples of assigned reads revealed that all of them seem to be either i) very short but highly conserved sequences, ii) long sequences that show relatively weak similarity to the query sequence or iii) rDNA sequences that are relatively long and highly conserved across organisms. In all these cases the fact that the homolog sequence was available for only some of the species in the database led to an erroneous assignment of the read to a clade, even though almost no genuine diatom sequences were represented in the reads. Genomic sequences are scarce for most species except those that have become the subject of a whole-genome sequencing project. The taxonomic sampling for 16S and 18S sequences, which are common taxonomic markers, is much more extensive. Contigs containing these types of sequence would have been invaluable to identify species that contributed to this DNA sample. Mapping the reads against rDNA showed that the sequencing data did indeed contain a mixture of 16S and 18S sequences from different sources. This is very likely the reason why the *de novo* assembly did not produce any of these ribosomal RNA

sequences. The differences between these very similar sequences would have been interpreted as sequencing errors by the assembly algorithm and prevented an assembly. This is also a problem commonly observed for assembly in the presence of heterozygous alleles, paralogue genes and other cases of sequence variants (Baker, 2012).

This effect was observed in the evaluation of all BLAST searches done with the contigs from this sample. Contigs were assigned to the entire taxonomic breadth represented in the search subject database but almost no contigs were assigned to the taxonomic groups of the target organism, diatoms or cyanobacteria. The fact that this is an environmental sample means that diverse evolutionary lineages are expected to be present. However, the fact that little to no contigs are assigned to diatoms and cyanobacteria lead to the conclusion that the DNA extracted from the environmental sample comprised little if any diatom DNA. Instead the sequences that were sequenced from the environmental sample are not yet represented in GenBank and thus were not identified with BLAST and MEGAN analyses.

This conclusion is corroborated by the fact that the eleven longest contigs did not produce significant BLAST hits either when used to search the entire GenBank nucleotide database. These long contigs have a higher than average coverage and can be assumed to originate from the most abundant organisms in the DNA extraction. They are also unlikely to be assembly artefacts based on the way the raw reads map to them. Like the contigs surveyed in the MEGAN analyses they only produced non-specific BLAST hits indicating that the genome of their source organism or anything closely related has not yet been submitted to GenBank.

4.5.5 Analysis of Eukaryotic Signature Proteins

At this point the standard tools for the analysis of high-throughput sequencing data – mapping and BLAST – both indicated that the intended target organism had not been sequenced. The observation that the environmental sample had yielded a copious amount of high quality genomic DNA which in turn produced a high number of good quality short reads after Illumina sequencing (see Section 4.4.1) still posed the question as to what organism had been sequenced. The classic approach to identify the taxonomic origin of a sequence sample is to amplify, clone and sequence 16S and 18S sequences from it. This requires intensive lab work and makes an alternative

bioinformatic approach desirable. Eukaryotic signature proteins are not only specific to eukaryotes, but each protein is highly conserved across most lineages of eukaryotes. This makes them potentially suitable taxonomic markers to identify related sequences in the GenBank databases. The work of J. Han on ESPs in *Giardia lamblia* and *Homo sapiens* suggests that ESPs can be identified based on similarity in BLAST searches, and despite their strong functional constraint, these sequences retain enough phylogenetic information to be useful as taxonomic markers (Han, 2012). Mappings of 18S sequences showed that a high proportion of the reads did indeed originate from Eukaryotes. A tBLASTn search with known ESPs from *G. lamblia* and *H. sapiens* did find putative homolog sequences in nine of the contigs. As already established by comprehensive BLAST searches of the assembly contigs, none of them showed strong similarities to published nucleotide sequences. A more sensitive BLASTx search did however detect potential homolog protein sequences across a wide range of eukaryotic lineages for each of these contigs. A selection of putative homolog proteins was downloaded from GeneBank for each contig and was used to reconstruct Neighbour Joining trees (see Figure 4-15 and Appendix A). The putative homologs were selected to represent as many of the major eukaryotic lineages as possible.

Given the ancient divergence of the major eukaryotic lineages, sequence similarities indicate the importance of these proteins for eukaryote organisms. This similarity is presumably due to high functional constraint, which has preserved the identity of these proteins despite processes like genetic drift, lineage specific evolution, metabolic innovations and changing environments. However, while this conservation assists identification of ancient ESP homologs it can also limit the inference of evolutionary relationships as there can be relatively few sites indicating phylogenetic relationships among ESP homologues. For this reason, the expectation was that the gene trees constructed from sets of very deeply divergent ESP homologs would be star-like with little support for the internal branches, no matter which tree-building method was used. The Neighbour Joining trees did indeed show low support for most internal branches and the grouping of representatives of the different major eukaryotic lineages was incongruent across the ESP phylogenies. However, all gene trees did include a much younger divergence: that between the two diatoms *T. pseudonana* and *P. tricornutum*. The genomes of both diatoms have been fully sequenced and contain putative homologs for all ESP that were detected in the sample. Sequences of both diatoms were included in each phylogeny to test if relatively distantly related diatom sequences would group together. This was always found to be the case, suggesting that

for close evolutionary relationships, expected phylogenetic relationships were recovered in tree building analyses. Based on this observation it is a reasonable assumption that if the ESPs assembled from the environmental sample were from a diatom then they should also group with the other diatom sequences in the phylogeny. This was not found to be the case for any of the ESP gene trees. It was evident from the assembled ESP sequences that the contig sequences were strongly diverged from the two diatom sequences, which, in turn, showed strong similarities to each other. In summary, this final attempt to identify the source of DNAs in the environmental sample using a bioinformatics approach was unsuccessful. Thus, a further empirical approach utilizing Sanger sequencing was investigated next.

4.5.6 rDNA sequencing

The DNA sequences of the ribosomal RNAs 16S and 18S are among the most commonly used taxonomic markers for prokaryotes and eukaryotes respectively. They are available in the public sequence databases for a much wider and detailed taxonomic range than most other types of sequence. Previous bioinformatic work has already demonstrated that our DNA extraction contained a mixture of rDNA sequences, but mapping as well as assembly failed to disentangle the mix of sequence variants. This was only achieved by targeted PCR amplification and subsequent cloning.

Nine clones of 16S rDNA sequences and twenty-nine of 18S rDNA were sequenced. The assumption was that the chance of the different sequences in the sample to be picked and sequenced should be proportional to their abundance. The number of clones is not sufficient for a statistical significant evaluation of the DNA sample composition but it can give an idea of the type of organism that produced the highest yields of DNA in the extraction.

Nine 16S sequences determined for the environmental sample all differed from each other, indicating a high bacterial diversity as should be expected in an environmental sample. None of them were identical to sequences already present in the GenBank nucleotide database, or showed strong similarities to sequences of described species. Instead the closest matches were with uncultured bacteria. Though the level of sequence identity was generally very high, the level of taxonomic information given for the most similar sequences was rather low. It ranged from uncultured bacteria, over uncultured soil bacteria, to a general taxonomical classification as Proteobacteria or

Verrucomicrobia. This lack of taxonomic information for the GenBank sequences made any more precise classification of the cloned sequences impossible. However, one of the 16S sequences stood out in that it showed similarity to chloroplast sequences of diatom species of the group Fragilariaceae, but it should be noted that over its whole length the sequence shows only 98% sequence identity to diatom sequences while it is 99% identical to some uncultured bacteria.

Twenty-nine clones of 18S rDNA sequences from the environmental sample yielded only nine different sequences. One of these is likely to be of diatom origin, more specifically a Fragilariaceae. This appears consistent with the results for the 16S rDNA sequences, one of which showed similarity to Fragilariaceae chloroplast sequences. The Fragilariaceae are a lineage of araphid, pennate diatoms. They are not close to either of the two fully sequenced diatom species or the raphid diatoms that include the Rhopalodiales and *E. sorex.* Compared to these groups they are rather underrepresented in the database. A compositional bias of the database favouring their identification is therefore unlikely to be the reason why the only diatom sequences that were identified in the sample matched a lineage other than *E.sorex*, rhopalodian diatoms or raphid diatoms in general. This might hint at the diatom in the sample not being the raphid diatom *E. sorex*, but rather an araphid species. However, given the clearly visible derived feature of an arched raphe and the characteristic shape of the frustule (see Section 4.3.1) a misclassification of the sample is very unlikely. This notion is compounded by the fact that Fragilariaceae, albeit pennate in shape, do not look very similar to rhopalodian diatoms (see Figure 4-18). The species of the genus *Synedra* showed the highest sequence similarity to the putative diatom sequence. They are non-motile, but commonly occur singular or in colonies as epiphytes on aquatic plants to which they attach by a pad of mucilage at the base. They are widespread and common in freshwater habitats. This means that they are likely to co-occur with *E. sorex* and that some could have contaminated the sample. It is conceivable that morphological factors result in very different efficiencies of the DNA extraction in the two types of diatoms. It is for example possible that in *E. sorex,* more silicon dioxide groups are exposed on the inner surface of the frustule or that in other species specific differences in the polymers that cover the silicate structures could affect the proportion of DNA that is released during the DNA extraction. Convergent evolution of frustule shape may be an explanation but it is more likely that other factors caused the DNA extractions to have very different efficiencies in the two types of diatom.

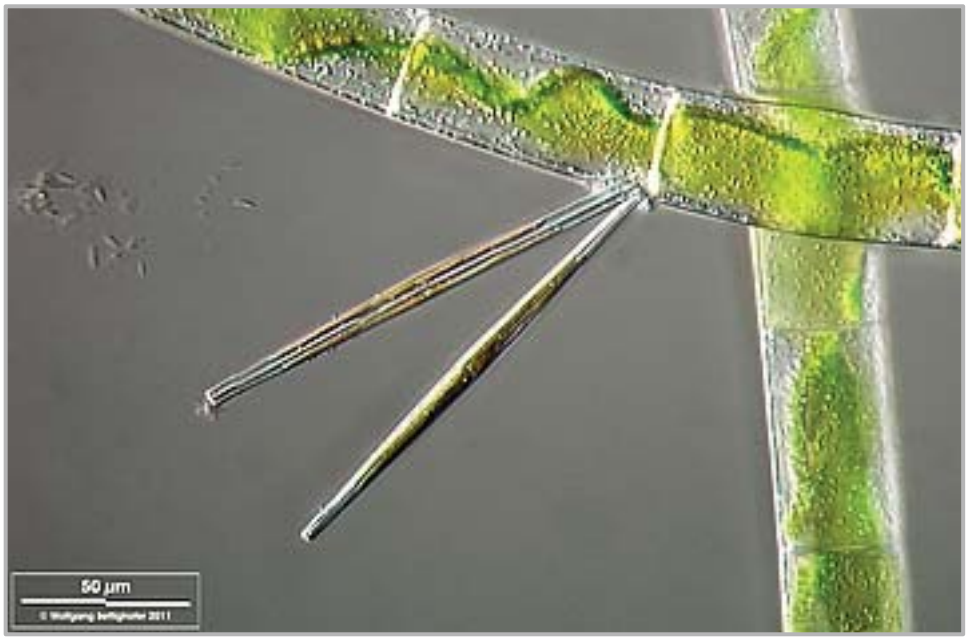


Figure 4-18 The araphid, pennate diatom *Fragilaria ulna*.

Shown are two specimens of the pennate, araphid diatom Fragilaria ulna. Both of them are attached to the algae Mougeotia by a pad of mucilage at one tip.
Image under Creative Commons License V 3.0 (CC BY-NC-SA). Image copyright by Wolfgang Bettighofer

As with the 16S rDNA sequences, none of the 18S ones had a perfect match in the GenBank database. Instead a variety of very similar sequences was found that strongly indicated the taxonomic group of the sequence's source organism. The majority of 18S sequences show a clear similarity to Naididae, a family of citellate oligochaete worms. Of the nine different 18S sequences, five show strong similarities to Naididae sequences. These five sequences represent 22 of the 29 clones. The remaining sequences show strong similarities to Nematodes (one sequence in three clones) and Ciliates (two sequences in two clones). The fact that none of the source organisms is represented in GenBank explains why the BLAST searches did not return any confident identifications. Only very little genomic data is available for the Naididae, the 71 species recognised in the NCBI taxonomy have less than 500 nucleotide records, mostly for rDNA sequences. The nematoda are one of the most diverse animal phyla but only the model system *Caenorhabditis elegans* has been sequenced in depth. The situation is even more dire for the ciliates. Only one of them has been sequenced to date (*Tetrahymena thermophila*). If the majority of sequences in the assembly originated from organisms of these taxonomic groups, it is not surprising that none of the BLAST searches found similar sequences.

Naididae are important elements of many marine and freshwater benthic communities (Verdonschot, 2006). Nematodes are ubiquitous in most described ecosystems and ciliates are common in almost all aquatic habitats. It was therefore not surprising to detect sequences from any of these eukaryotic lineages in an environmental freshwater sample. What is surprising is that these organisms dominate the genomic DNA that was extracted from a sample whose biomass was, based on microscopic evaluation, made up of more than 90% diatoms, and that did not contain any visible animals (i.e. worms). I can only speculate about the reasons. The silicate of diatom frustules is known to adsorb DNA (Carter and Milton, 1993), a phenomenon that was directly observed while testing different DNA extraction protocols on this sample. Crushing the diatom frustules resulted in markedly reduced DNA yields, most likely because DNA was bound to the freshly exposed silicate surfaces at the break lines. Microscopic examination of the sediments after DNA extractions showed that the diatom cells did in fact lyse, but it is possible that the diatom DNA was never released from the frustule but instead bound to it. The DNA of metazoa and other eukaryotes in the sample on the other hand would have been more easily extracted, especially because DNA extraction kits are generally optimised for animal DNA extraction. Worm eggs for example or larvae would have been very difficult to spot under the microscope but would contain copious amounts of DNA, accounting for the high yields of the extraction.

4.6 Conclusions

This project attempted to characterise the genome of *E. sorex* and its endosymbiotic spheroid body from an environmental sample using Illumina sequencing. A site with an extremely high density of diatom cells was sampled and morphological examination gave confidence in their identification as *E. sorex*. High quality high molecular mass DNA was isolated from the sample with a protocol similar to one previously used for diatom DNA extraction. However an exhaustive bioinformatic analysis failed to detect *E. sorex* or spheroid body sequences in Illumina sequencing reads produced from this genomic preparation. This disappointing result might be best explained by a failure to extract the DNA from RNAlater-preserved *E. sorex* tissue. To establish the root cause DNA extraction protocols would have to be repeated on freshly collected samples, which were not stored in RNAlater. Overall our endeavours to analyse the sequencing results have highlighted the dependence of bioinformatic methods on available reference sequences and the extent to which sample preparation and DNA extraction methods can bias metagenomic studies.

Bibliography

Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990). Basic local alignment search tool. Journal of Molecular Biology. **215**:403–410.

Baker, M. (2012). De novo genome assembly: what every biologist should know. Nature Methods. **9**:333–337.

Bruckner, C.G., Bahulikar, R., Rahalkar, M., Schink, B. and Kroth, P.G. (2008). Bacteria associated with benthic diatoms from Lake Constance: phylogeny and influences on diatom growth and secretion of extracellular polymeric substances. Applied and environmental microbiology. **74**:7740–9.

Carter, M.J. and Milton, I.D. (1993). An inexpensive and simple method for DNA purifications on silica particles. Nucleic acids research. **21**:1044.

Cox, M.P., Peterson, D.A. and Biggs, P.J. (2010). SolexaQA: At-a-glance quality assessment of Illumina second-generation sequencing data. BMC bioinformatics. **11**:485.

Dagan, T., Blekhman, R. and Graur, D. (2006). The "domino theory" of gene death: gradual and mass gene extinction events in three lineages of obligate symbiotic bacterial pathogens. Molecular biology and evolution. **23**:310–6.

DeYoe, H.R., Lowe, R.L. and Marks, J.C. (1992). Effects of Nitrogen and Phosphorus on the Endosymbiont Load of Rhopalodia gibba and Epithemia turgida (Bacillariophyceae). Journal of Phycology. **28**:773–777.

Drummond, A. et al. (2012). Geneious v5.6, Available from http://www.geneious.com.

Gasteiger, E. (2003). ExPASy: the proteomics server for in-depth protein knowledge and analysis. Nucleic Acids Research. **31**:3784–3788.

Geitler, L. (1977). Life history of the Epithemiaceae Epithemia, Rhopalodia and Denticula (Diatomophyceae) and their presumable symbiotic spheroid bodies. Journal of Plant Systematics and Evolution. **128**:259–275.

Giovannoni, S.J. (1991). The Polymerase Chain Reaction. In E. Stackerbrandt and M. Goodfellow, eds. *Nucleic Acid Techniques in Bacterial Systematics*. Chichester: John Wiley and Sons, pp. 177–203.

Griebel, T., Brinkmeyer, M. and Böcker, S. (2008). EPoS: a modular software framework for phylogenetic analysis. Bioinformatics (Oxford, England). **24**:2399–400.

Gómez-Valero, L., Rocha, E.P.C., Latorre, A. and Silva, F.J. (2007). Reconstructing the ancestor of Mycobacterium leprae: the dynamics of gene loss and genome reduction. Genome research. **17**:1178–85.

Hajós, M.A. (1986). Magyarországi miocén diatomás képződmények rétegtana: Stratigraphy of Hungary's miocene diatomaceous earth deposits., Institutum Geologicum Hungaricum.

Han, J. (2012). Eukaryotic Signature Proteins: Guides to pathogenic eukaryotic parasites. Massey University, Palmerston North.

Huson, D.H., Mitra, S., Ruscheweyh, H.-J., Weber, N. and Schuster, S.C. (2011). Integrative analysis of environmental sequences using MEGAN4. Genome research. **21**:1552–60.

Iwatani, N., Murakami, S. and Suzuki, Y. (2005). A sequencing protocol of some DNA regions in nuclear , chloroplastic and mitochondrial genomes with an individual colony of Thalassiosira nordenskioeldii Cleve (Bacillariophyceae). 35–45.

Jacobs, J.D., Ludwig, J.R., Hildebrand, M., Kukel, a, Feng, T.Y., Ord, R.W. and Volcani, B.E. (1992). Characterization of two circular plasmids from the marine diatom Cylindrotheca fusiformis: plasmids hybridize to chloroplast and nuclear DNA. Molecular & general genetics : MGG. **233**:302–10.

Kies, L. (1992). Glaucocystophyceae and other protists harbouring prokaryotic endocytobionts. In *Algae and Symbiosis. Plants, Animals, Fungi, Viruses, Interactions Explored*. Bristol: Biopress Ltd., pp. 353–377.

Kneip, C., Voss, C., Lockhart, P.J. and Maier, U.G. (2008). The cyanobacterial endosymbiont of the unicellular algae Rhopalodia gibba shows reductive genome evolution. BMC evolutionary biology. **8**:30.

Kützing, F.T. (1844). Die kieselschaligen Bacillarien oder Diatomeen. (Mit 30 vom Verfasser gravirten Tafeln)., W. Köhne.

Langmead, B. and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat Meth. **9**:357–359.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Juan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R. and Subgroup, 1000 G.P.D.P. (2009). The Sequence alignment/map (SAM) format and SAMtools. Bioinformatics. **25**:2078–2079.

Medlin, L.K., Elwood, H.J., Stickel, S. and Sogin, M.L. (1988). The characterisation of enzymatically amplified eukaryotic 16S-like rRNA-coding regions. Gene. **71**:491–499.

Milne, I., Bayer, M., Cardle, L., Shaw, P., Stephen, G., Wright, F. and Marshall, D. (2010). Tablet--next generation sequence assembly visualization. Bioinformatics (Oxford, England). **26**:401–2.

Muller, H.J. (1964). The Relation of Recombination to Mutational Advance. Mutation Research. **204**:732–732.

Nakayama, Takuro, Ikegami, Y., Nakayama, Takeshi, Ishida, K.-I., Inagaki, Y. and Inouye, I. (2011). Spheroid bodies in rhopalodiacean diatoms were derived from a single endosymbiotic cyanobacterium. Journal of plant research. **124**:93–7.

Nowack, E.C.M. and Grossman, A.R. (2012). Trafficking of protein into the recently established photosynthetic organelles of Paulinella chromatophora. PNAS. **109**:5340–5345.

Prechtl, J., Kneip, C., Lockhart, P.J., Wenderoth, K. and Maier, U.-G. (2004). Intracellular spheroid bodies of Rhopalodia gibba have nitrogen-fixing apparatus of cyanobacterial origin. Molecular biology and evolution. **21**:1477–81.

Ravin, N.V., Galachyants, Y.P., Mardanov, A.V., Beletsky, A. V, Petrova, D.P., Sherbakova, T.A., Zakharova, Y.R., Likhoshway, Y.V., Skryabin, K.G. and Grachev, M.A. (2010). Complete sequence of the mitochondrial genome of a diatom alga Synedra acus and comparative analysis of diatom mitochondrial genomes. Current genetics. **56**:215–23.

Round, F.E., Crawford, R.M. and Mann, D.G. (1990). Diatoms: biology and morphology of the genera., Cambridge University Press.

Salzberg, S.L., Phillippy, A.M., Zimin, A., Earl, D., Bradnam, K. and John, J.S. (2012). GAGE : A critical evaluation of genome assemblies and assembly algorithms methods. Genome res. **22**:557–567.

Scala, S., Carels, N., Falciatore, A., Chiusano, M.L. and Bowler, C. (2002). Genome Properties of the Diatom Phaeodactylum tricornutum. Plant Physiology. **129**:993–1002.

Schäfer, H., Abbas, B., Witte, H. and Muyzer, G. (2002). Genetic diversity of "satellite" bacteria present in cultures of marine diatoms. FEMS microbiology ecology. **42**:25–35.

Simpson, J.T., Wong, K., Jackman, S.D., Schein, J.E., Jones, S.J.M. and Birol, I. (2009). ABySS: a parallel assembler for short read sequence data. Genome research. **19**:1117–23.

Sims, P.A., Mann, D.G. and Medlin, L.K. (2006). Evolution of the diatoms: insights from fossil, biological and molecular data. Phycologia. **45**:361–402.

Verdonschot, P.F.M. (2006). Beyond masses and blooms: the indicative value of oligochaetes. Hydrobiologia. **564**:127–142.

5 Conclusions and future work

Bacterial endosymbionts present themselves in two very different guises: Those which only relatively recently took up an endosymbiotic lifestyle seem to be nothing more than deranged bacteria of a very narrowly defined usefulness to their hosts, and their genomes seem to have entered a path to oblivion. The descendants of the two ancient endosymbionts recognized today on the other hand have become fully integrated parts of eukaryotic cells that, apart from few exceptions, are indispensable components of the eukaryotic cell's physiology, taking part in many aspects of its metabolism. It is fascinating to think that both guises have evolved by going down the same path of endosymbiogenesis, differing only in the distance that they have travelled it.

The focus of this work was to advance our understanding of endosymbiotic relationships by computational means. The advancements in sequencing technology in recent years have produced a wealth of sequence information and opened up new possibilities for the study of symbiotic systems. In the first part of this work I have made use of the available plastid genome data to study the frequency of gene losses during the evolution of an ancient endosymbiosis. This was done across an unprecedented taxonomic breadth and was to my knowledge the first study comprehensive enough to give a meaningful answer to the basic question: How frequently do gene losses occur in established organelle genomes? The answer is that gene losses are not as rare as was generally presumed, and more importantly, they can occur over short evolutionary time. This implies that the retention of genes in the plastid genome is the result of specific evolutionary pressures and rebuts assumptions that gene losses are rare and chance driven, adding valuable insight to our understanding of organelle evolution.

For this analysis a comprehensive dataset of protein coding plastid genes was compiled and curated. This consolidation of plastid protein information highlighted that the available fully sequenced plastid genomes to date still only inadequately represent the taxonomic breadth of the Plantae. More taxa need to be sequenced to fill the gaps, especially for the Glaucophyta, which are only represented by a single species, and the diversity of the red algae, which is only represented by five species. This dataset is also a valuable resource for future work and can potentially be used to address a multitude

of questions, concerning the evolution of plastids, mechanisms of sequence evolution and the evolution, as well as genetic foundations, of eukaryote photosynthesis, to name but a few. Furthermore, the detailed analyses of lineage specific differences in gene content and gene loss frequency were outside the scope of this PhD and require future work. Correlations between the loss of certain genes from plastid genomes, and the retention or loss of others, as well as correlation with physiological characteristics of the taxa concerned, have the potential to reveal much of the mechanisms that encourage or prevent the transfer of genes into the nuclear genome.

The advent of short-read high-throughput sequencing technologies has not only steeply increased the available data but also opened up new possibilities for the sequencing of complex samples. The new technologies do not require specific primers, cell cultures or the construction of genetic libraries. They can be used on modest amounts of genetic data and are capable of producing the amount of sequence data required to paint a comprehensive genetic picture of environmental samples, as well as for example micro-communities that are formed by organisms in obligate relationships. All this makes HTS technologies potentially very useful for the sequencing of host-endosymbiont systems, especially when these cannot be isolated into pure samples, as is the case for many micro-algae. Yet micro-algae have formed some of the most interesting 'young' endosymbioses. The rhopalodian diatoms with their nitrogen fixing spheroid bodies are one of the most striking examples.

The Illumina sequencing technology had been newly released when this project began. The method's true capabilities and limitations were largely unknown and the tools available for the analysis of the data it produced were in the early stages of development, designed according to the theoretical properties of the data rather than the empirical. Both, the sequencing of a pooled fosmid library as well as that of an environmental sample with the aim to sequence a specific target organism had, to my knowledge, not been attempted before.

The work on a fosmid library that had been constructed from a spheroid body enriched extraction of *R. gibba* demonstrated the feasibility of this sequencing strategy and the high sequence quality that can be achieved with it. It also demonstrated how unfit early short read assemblers were to deal with data that violated their assumptions of an ideal sample. The recognition that real data rarely shows an even read coverage or

homogenous sequence composition and the desire to use HTS in the field of transcriptomics has prompted much work to improve the assembly tools' ability to deal with these issues. It is fair to assume that the task of assembling a fosmid library from Illumina reads would be much easier with current tools.

The attempt to perform targeted sequencing of a diatom and its endosymbiont from an environmental sample, on the other hand, was not successful but produced valuable insights into the potential biases that can affect commonly used metagenomic methods. Further work is required to determine which aspects of the handling of the sample in conjunction with the DNA extraction method prevented the extraction of DNA from the abundant diatom cells in the sample, but it is apparent that not all types of organisms were affected equally by these issues, as the metazoa that were present in the sample produced a high yield of high quality genomic DNA. It is a reminder that one must not assume that the DNA extracted from a mixed sample is a fair representation of the sample composition. The analyses of the sequences produced from this DNA extraction highlighted not only the dependence of the bioinformatic methods on the availability of well characterised reference sequences but also the susceptibility of the popular metagenomics software MEGAN to produce results that are strongly biased by the composition of the BLAST databases used. The method behind MEGAN is supposed to be conservative because it retains conflicting information in the data by assigning reads to nodes ancestral to all the taxa involved. However, my analyses demonstrate that MEGAN is nonetheless vulnerable to missing data. This is an important caveat given that, despite the staggering amount of sequence data that is available to date, this data nonetheless only represents a small sample of the genetic diversity of life.

Appendix A

List of files in electronic appendix

The names of files and folders are indicated in bold.

Chapter 2

Plastid_Gene_Matrix.xlsx

Excel spread sheet containing a data matrix representing the complete data set. Columns represent species, rows homologous proteins encoded in the plastid genome. Proteins that are present in a plastid genome are represented by their GenBank accession number. Taxonomic groupings are indicated in the top 11 rows. Non-photosynthetic species are indicated in red. The members of the genus *Cuscuta* are parasitic but have retained photosynthetic function to varying degrees. Rows are sorted alphabetically by gene name. Most protein sequences are represented on a single row. However, the situation for *RpoB* and *RpoC* which code for the β and β' subunits of the DNA directed RNA polymerase is more complex. The β subunit is encoded by a single gene, *RpoB*, in most taxa in this dataset. However, in the Trebouxiophyceae *Leptosira terrestris* and the Chlorophyceae *Stigeoclonium helveticum, Oedogonium cardiacum, Floydiella terrestris* and *Scenedesmus obliquus RpoB* is broken up into two separate genes, *RpoBa* and *RpoBb*. The β' subunit is encoded by a single gene, *RpoC* in only one species in this data set, the Rhodophyte *Cyanidioschyzon merolae.* In most taxa it is encoded by the genes *RpoC1* and *RpoC2*. In the Chlorophyceae *Chlamydomonas reinhardtii* the *RpoC1* gene is split yet another time into *RpoC1a* and *RpoC1b*. Thus the genes coding for this RNApolymerase are listed over seven rows in total. Genes with doubtful expression status are indicated in grey.

Gblocks_trimmed_alignments (Folder)

Alignments of each set of homologous proteins in the data set in phylip format. The alignments have been trimmed with the Gblocks algorithm on default settings.

PhyloBayes_Gene_trees (Folder)

Individual gene trees calculated by PhyloBayes. The results of the convergence checks for chains run for each gene are detailed in convergence_check_results.txt.

Gene_losses_per_gene.xlsx

Excel spread sheet that lists the results for the Dollo Parsimony and Maximum Likelihood calculations for each gene and each of the three possible root positions. The root position is indicated by the name of the most basal lineage.

Scripts (Folder)

MatrixMaker.pl: Script written to extract all proteins from the RefSeq collection of fully sequenced plastid genomes. The script uses the information in the GenBank flat files to build a data matrix (see Plastid_Gene_Matrix.xlsx) from the accession numbers, species and gene names and also writes out the sets of homolog sequences to separate fasta files.

GeneFilesSorter.pl: Script used to sort homolog sequence sets alphabetically by species name and to pull out genes with multiple copies.

Gene_loss.r: Script that calculates the expected number of gene loss events given a phylogeny and character information for the tips. Written by Tim White.

Gene_loss_feeder.pl: Script to create input files and feed them to the Gene_loss.r script.

Chapter 3

Velvet3parameterTest.pl

Script that performs Velvet assemblies for all possible combinations of a range of values for each of the three parameters k-mer length, expected coverage and coverage cutoff.

SolexaQA (Folder)

Output files of the SolexaQA quality assessment of the 75bp paired end reads produced by the Illumina GAII sequencer at Massey University Genome Service, Palmerston North. Reads were indexed with two different indices and analyses results are given for both sets of reads separately. The folder Index1 contains the quality assessment for reads derived from 155 pooled fosmid inserts, while the folder Index2 contains those for reads derived from 10 pooled fosmid inserts.

SBcompleteFosmids.fa
Completely assembled fosmid inserts, putative spheroid body sequences only. Assembled from short reads and Sanger sequences.

AllPutSBsequences.fa
Consensus sequences of an assembly of all putative spheroid body sequences, combining Sanger sequence data and short read assemblies.

FinalAssemblyContigs (folder)

PutativeSB.fa: Contigs that show strongest sequence similarity to Cyanobacteria.

GCrich.fa: GC-rich sequences.

Diatom.fa: Contigs that show strongest sequence similarity to diatoms.

GCpoorContaminations.fa: GC-poor sequences that showed no unambiguous similarity to either Cyanobacteria or diatom sequences. Presumably a mixture of bacterial contaminations and unidentified SB and *R. gibba* sequences.

Chapter 4

SolexaQA_1stRun (Folder)

Output files of the SolexaQA quality assessment of the 100bp single end reads produced by the Illumina GAII at Massey University Genome Service, Palmerston North.

SolexaQA_2ndRun (Folder)

Output files of the SolexaQA quality assessment of the 100bp paired end reads produced by the Illumina GAII at Massey University Genome Service, Palmerston North.

ESP_alignments (Folder)

The alignments of eukaryotic signature protein sequences used to infer the ESP phylogenies.

AbyssContigSifter.pl

Perl script used to assess Abyss assembly results by contig size and calculate the average coverage of contig sets.

SAMtrimmer.pl

Perl script used to remove reads that didn't map from SAM files to make the files more manageable.

Appendix B

List of gene name aliases

Ordered alphabetically by most commonly used gene name.

AcsF = Ycf59
CbbX = CfxQ
Ccs1 = Ycf44
CcsA = Ycf5
CysA = Ycf85 = MbpX
DesA = CrtR
Dfr = Ycf26 =Tsg1=TilS
FtrB = FtrC
HlpA = Hlp = HupA
NblA = Ycf18
NdhK = PsbG
OmpR = Ycf27 = Ycf29
PetG= PetE
PetM = Ycf31 = PetX
PsbW = Psb28
PsbY = Ycf32
RbcR = Ycf30
SufB = Ycf24
SufC = Ycf16
TatC = Ycf43
ThiS = Ycf40
TrmE = ThdF
Ycf17 = HliP = HemH
Ycf42 = Bas1
Ycf47 = SecG

Apendix C

List of plastid proteins that were assigned gene names based on protein sequence similarity

The entries are ordered alphabetically by species.

Species	Accession No.	Gene Name
Amborella trichopoda	NP_904117	PetL
	NP_904109	PsaI
	NP_904093	PsbM
Babesia bovis T2Bo	YP_002290875	Rps4
Chara vulgaris	YP_635738	MatK
Chlorella vulgaris	NP_045891	Ycf1
	NP_045861	Ycf47
	NP_045859	Ycf62
Chromera velia	YP_003795326	PsaC
Cyanidium caldarium	NP_045115	MoeB
Cyanophora paradoxa	NP_043273	CysA
Ectocarpus siliculosus	YP_003289187	Rpl36
Eimeria tenella	NP_852626	Rpl16
	NP_852627	Rpl17
	NP_852632	Rpl36
	NP_852622	Rpl4
	NP_852630	Rpl6
	NP_852647	RpoB
	NP_852645	RpoC2
	NP_852629	Rps8
Emiliania huxleyi	YP_277329	PsbX
	YP_277382	PsbY
Euglena gracilis	NP_041930	RpoA
Euglena longa	NP_074977	RpoA
Gracilaria tenuistipitata	YP_063623	Ycf42
Guillardia theta	NP_050704	Ycf80
Helicosporidium sp.	YP_635915	Ycf62
Leptosira terrestris	YP_001382139	Ycf47
Marchantia polymorpha	NP_039290	MatK
	NP_039317	PetL
	NP_039285	PsaM
	NP_039357	Ycf1
	NP_039292	Ycf2

	NP_039274	Ycf66
Micromonas pusilla	YP_002808515	Rpl2
Odontella sinensis	NP_043699	Ycf66
	NP_043612	Ycf88
	NP_043579	Ycf89
	NP_043683	Ycf90
Oryza sativa Japonica	NP_039406	PsaJ
Parachlorella kessleri	YP_003058278	Ycf62
Phalaenopsis aphrodite	YP_358596	PetL
	YP_358586	PsaI
	YP_358638	Rpl32
	YP_358599	Rpl33
Picea sitchensis	YP_002905069	AtpF
	YP_002905074	RpoC1
Pinus contorta	YP_002905131	AtpF
	YP_002905136	RpoC1
Pinus krempfii	YP_002905266	AtpF
	YP_002905271	RpoC1
Pinus thunbergii	NP_042487	Ycf1
	NP_042505	Ycf2
Populus trichocarpa	YP_001109565	Ycf1
Porphyra purpurea	NP_053974	DsbD
	NP_053945	MoeB
	NP_053952	NtcA
	NP_053831	PetL
	NP_053954	ThiS
	NP_053964	upp
	NP_053882	Ycf42
	NP_053891	Ycf45
	NP_053884	Ycf47
	NP_053802	Ycf52
	NP_053812	Ycf53
	NP_053814	ycf54
	NP_053814	Ycf54
	NP_053813	Ycf55
	NP_053818	Ycf56
	NP_053891	Ycf58
	NP_053970	Ycf60
	NP_053993	Ycf62
	NP_054003	Ycf63
Porphyra yezoensis	YP_537073	Dfr
	YP_537045	DsbD
	YP_537017	MoeB
	YP_537047	NblA
	YP_537023	NtcA
	YP_537024	OmpR
	YP_536902	PetL

	YP_536957	PetM
	YP_536887	PsbY
	YP_537025	ThiS
	YP_537035	upp
	YP_537036	Ycf17
	YP_537034	Ycf19
	YP_536896	Ycf20
	YP_537061	Ycf21
	YP_537053	Ycf22
	YP_537054	Ycf23
	YP_537011	Ycf33
	YP_536910	Ycf34
	YP_536955	Ycf36
	YP_537003	Ycf38
	YP_536954	Ycf42
	YP_536956	Ycf47
	YP_536873	Ycf52
	YP_536883	Ycf53
	YP_536885	ycf54
	YP_536885	Ycf54
	YP_536884	Ycf55
	YP_536889	Ycf56
	YP_537041	Ycf60
	YP_537064	Ycf62
	YP_537074	Ycf63
Rhodomonas salina	YP_001293517	Ycf80
Selaginella moellendorffii	YP_003097532	Ycf1
Staurastrum punctulatum	YP_636383	MatK
Toxoplasma gondii RH	NP_044550	Rpl16
	NP_044555	Rpl36
	NP_044546	Rpl4
	NP_044553	Rpl6
	NP_044569	RpoB
	NP_044568	RpoC1
	NP_044556	Rps11
	NP_044551	Rps17
	NP_044548	Rps19
	NP_044567	Rps2
	NP_044549	Rps3
	NP_044545	Rps4
	NP_044554	Rps5
	NP_044558	Rps7

Apendix D

List of corrections made to annotations based on sequences similarity and verification of ORF boundaries

The entries are ordered alphabetically by species name. For newly established ORFs the start and end positions are given for the GenBank record of the complete plastid sequence. The comments indicate weather a sequence was added or removed from the dataset or its information in the data matrix otherwise corrected.

Species	Accession No.	Name	Comment
Alveolata sp.	YP_003795444	Rpl12	Renamed Rps12
Anomochloa marantoidea	NC_014062 pos. 82319-82588	Rps19	Added
Anthoceros formosae	NP_777472	Ycf1	Added
Babesia bovis	NC_011395 pos.22463-22353	Rpl36	Added
	YP_002290866	Rps13	Removed
Bambusa oldhamii	YP_003029736	AtpF	Added
	ACF32450	MatK	Added
Bryopsis hypnoides	NC_013359 pos.47417-47310	PsaI	Added
	NC_013359 pos.73140-73179	PsaM	Added
Chlamydomonas	NP_958415	Ycf1	Added
Chlorella vulgaris	NP_045873	MinE	Removed
Cicer arietinum	NC_011163 pos.58466-59 89	Ycf4	Added
Cryptomeria japonica	AP009377 pos.24620-25216	ClpP	Added
Cryptomonas paramecium	NC_013703 pos.63491-63637	Rpl32	Added
Cucumis sativus	YP_247623	Rps12	Added
Cyanidioschyzon merolae	NP_848940	InfB	Removed
	NP_849009	Rps1	Removed
	NP_848961	Ycf55	Removed
Cyanidium caldarium	NP_045204	PetM	Added
	NC_001840 pos.122065-122184	PsbX	Added
	NP_045115	ChlN	Removed
Cycas taitungensis	NC_009618 pos.9848-9944	PsaM	Added
Dendrocalamus latiflorus	YP_003097634	Ycf1	Removed
Ephedra equisetina	NC_011954 pos.33246-33338	PsaM	Added
	NC_011954 pos.81025-81174	Rpl32	Added
Eucalyptus grandis	NC_014570 pos.61165-62307	AccD	Added
	NC_014570 pos.53530-54382	NdhK	Added
	NC_014570 pos.89366-88935	Rpl2	Added
Festuca arundinacea	NC_011713 pos.65892-66338	Rps18	Added
Glycine max	NC_007942 pos.58628-59221	Ycf4	Added
Gossypium	NC_007944 pos.88002-88454	Rpl22	Added
Lactuca sativa	NC_007578 pos.123027-128200	Ycf1	Added
Marchantia polymorpha	NC_001319 pos.5171-5257	PetN	Added
	NC_001319	Ycf3	Added
Medicago truncatula	NC_003119 pos.66202-65546	Ycf4	Added
Mesostigma viride	NP_038386	PetN	Added
Monomastix sp. OKE1	YP_002600998	PetN	Removed
Nephroselmis olivacea	NP_050844	PetN	Added

	NP_050877	Ycf2	Removed
Olea europaea	FN996943.2 pos.82200-81994	InfA	Added
Oryza sativa var. Indica	AY522329 pos.53253-51760	AtpB	Added
	NC_008155 pos.32700-34067	AtpF	Added
	NC_008155 pos. 110602..112690	NdhA	Added
	NC_008155 pos. 87596-85364	NdhB	Added
	NC_008155 pos.71997-72641	PetB	Added
	NC_008155 pos.63485-63577	PetL	Added
	NC_008155 pos.17574-17488	PetN	Added
	NC_008155 pos.57167-57274	PsaI	Added
	NC_008155 pos.64576-64686	PsaJ	Added
	NC_008155 pos.61771-61884	PsbL	Added
	NC_008155 pos.70442-70546	PsbT	Added
	NC_008155 pos.11876-12061	PsbZ	Added
	NC_008155 pos.66666-66310	Rpl20	Added
	NC_008155 pos.76081-75971	Rpl36	Added
	NC_008155 pos.43789-41814	Ycf3	Added
Porphyrea yezoensis	YP_537023	Ycf28	Added
	NC_007932 pos.85391-83695	Ycf45	Added
	YP_536916	Ycf58	Added
Parthenium argentatum	NC_013553 pos.26899-28155	AtpF	Added
	NC_013553 pos.117650-118144	NdhI	Added
	NC_013553 pos.10194-10280	PetN	Added
	NC_013553 pos.64742-64990	PsbE	Added
	NC_013553 pos.64474-64587	PsbL	Added
	YP_003330991	Rpl16	Added
Psilotum nudum	NC_003386 pos.73124-73219	PsbT	Added
Selaginella moellendorfii	NC_013086 pos.60524-60408	PsbF	Added
	YP_004021257 pos.122435-	PsbZ	Added
	NC_013086 pos.65011-65250	Rps18	Added
	NC_013086 pos.19015-19683	Rps2	Added
Solanum bulbocrastrum	YP_538884	InfA	Removed
Theileria parva	XP_762671	Rpl4	Removed
Theobroma cacao	NC_014676 pos.37834-38019	PsbZ	Added
	NC_014676 pos.88595-88999	Rpl22	Added
Trifolium subterraneum	NC_011828 pos.141910-141674	Rpl32	Added
	NC_011828 pos.4158-4340	Rps18	Added
	NC_011828 pos. 89791-84435	Ycf1	Added
	NC_011828 pos. 12611-11985	Ycf4	Added
Vaucheria litorea	YP_002327461	TatC	Added

Appendix E

List of sequences used in ESP phylogenies

Accession number	Description	Species	Taxonomic group
7565149 586 2165	Contig		
6273663 598 2864	Contig		
gi157434854	14-3-3 protein	Giardia lamblia ATCC 50803	Diplomonads
gi223998024	14-3-3-like protein	Thalassiosira pseudonana CCMP1335	Diatoms
gi219124680	14-3-3-like protein	Phaeodactylum tricornutum CCAP 1055/1	Diatoms
gi146162627	14-3-3 protein	Tetrahymena thermophila	Alveolata
gi330690205	14-3-3 protein	Penaeus monodon	Crustacea
gi19115079	14-3-3 protein Rad24	Schizosaccharomyces pombe	Fungi
gi225423491	14-3-3-like protein	Vitis vinifera	Streptophyta
gi47086819	14-3-3 protein epsilon	Danio rerio	Chordata
gi290986410	predicted protein	Naegleria gruberi	Excavata
gi294953313	14-3-3 protein, putative	Perkinsus marinus ATCC 50983	Alveolata

Accession number	Description	Species	Taxonomic group
7697359 977 3829	Contig		
spP27797	Calreticulin	Homo sapiens	Chordata
gi224003875	calreticulin-like protein	Thalassiosira pseudonana CCMP1335	Diatoms
gi219129933	calreticulin	Phaeodactylum tricornutum CCAP 1055/1	Diatoms
gi148717307	calreticulin	Crassostrea gigas	Opisthokonta
gi226427123	calreticulin	Lotharella amoeboformis	Rhizaria
gi320165251	calreticulin	Capsaspora owczarzaki ATCC 30864	Opisthokonta
gi167395304	Calreticulin precursor	Entamoeba dispar SAW760	Amoebozoa
gi732893	calretulin	Nicotiana tabacum	Streptophya
11211941 778 3234	Contig		
gi157433613	Axoneme central apparatus protein	Giardia lamblia ATCC 50803	Diplomonads
gi219124247	predicted protein	Phaeodactylum tricornutum CCAP 1055/1	Diatoms
gi224013122	importin alpha 1 subunit-like protein	Thalassiosira pseudonana CCMP1335	Diatoms
gi323448763	hypothetical protein, partial	Aureococcus anophagefferens	Stramenopiles
gi301120334	conserved hypothetical protein	Phytophthora infestans T30-4	Stramenopiles
gi255089465	hypothetical protein	Micromonas sp. RCC299	Chlorophyta
gi338721475	sperm-associated antigen 6-like, partial	Equus caballus	Chordata
gi302783224	hypothetical protein	Selaginella moellendorffii	Streptophyta
gi340052205	axoneme central apparatus protein	Trypanosoma vivax Y486	Euglenozoa

Accession number	Description	Species	Taxonomic group
11404237 721 2384	Contig		
spP62736ACTA	Actin, aortic smooth muscle	Homo sapiens	Chordata
gi224012529	actin-like protein	Thalassiosira pseudonana CCMP1335	Diatoms
gi219126085	predicted protein	Phaeodactylum tricornutum CCAP 1055/1	Diatoms
gi204022116	actin	Pyropia yezoensis	Rhodophyta
gi330803828	hypothetical protein	Dictyostelium purpureum	Amoebozoa
gi359744459	actin	Cryptocaryon irritans	Alveolata
gi1022821	actin	Naegleria fowleri	Excavata
gi157887507	actin I	Plasmodiophora brassicae	Rhizaria
gi353236678	probable actin	Piriformospora indica DSM 11827	Fungi
11428777 783 2920	Contig		
spQ14152EIF3A	Euk. translation initiation factor 3 subunit A	Homo sapiens	Chordata
gi223999087	Euk. translation initiation factor 3, subunit 10	Thalassiosira pseudonana CCMP1335	Diatom
gi219119962	predicted protein	Phaeodactylum tricornutum CCAP 1055/1	Diatom
gi302682306	hypothetical protein	Schizophyllum commune H4-8	Fungi
gi242052283	hypothetical protein	Sorghum bicolor	Streptophyta
gi301125229	Euk. translation initiation factor 3 subunit A	Phytophthora infestans T30-4	Stramenopiles
gi323456395	hypothetical protein, partial	Aureococcus anophagefferens	Stramenopiles
gi325116964	Euk. translation initiation factor 3 subunit 10	Neospora caninum Liverpool	Alveolata
gi328865094	PCI domain-containing protein	Dictyostelium fasciculatum	Amoebozoa

Accession number	Description	Species	Taxonomic group
9777398 652 2435	Contig		
spP61962DCAF7	DDB1- and CUL4-associated factor 7	Homo sapiens	Chordata
gi224008941	hypothetical protein THAPSDRAFT_42258	Thalassiosira pseudonana CCMP1335	Diatoms
gi219123489	predicted protein	Phaeodactylum tricornutum CCAP 1055/1	Diatoms
gi323454218	hypothetical protein AURANDRAFT_52888	Aureococcus anophagefferens	Stramenopiles
gi168057818	predicted protein	Physcomitrella patens subsp. patens	Streptophyta
gi268638131	WD40 repeat-containing protein	Dictyostelium discoideum AX4	Amoebozoa
gi221124606	similar to WD repeat-containing protein 68	Hydra magnipapillata	Cnidaria
gi302890020	predicted protein	Nectria haematococca mpVI 77-13-4	Fungi
266513 695 2841	Contig		
gi157434817	Ubiquitin	Giardia lamblia ATCC 50803	Diplomonads
gi219123892	ubiquitin extension protein	Phaeodactylum tricornutum CCAP 1055/1	Diatoms
gi224000942	predicted protein	Thalassiosira pseudonana CCMP1335	Diatoms
gi61741149	polyubiquitin	Fistulifera pelliculosa	Diatoms
gi167945	ubiquitin	Dictyostelium discoideum	Amoebozoa
gi384496533	polyubiquitin	Rhizopus oryzae RA 99-880	Fungi
gi118370596	Ubiquitin family protein	Tetrahymena thermophila	Alveolata
gi290984061	polyubiquitin	Naegleria gruberi	Excavata
gi323452163	polyubiquitin	Aureococcus anophagefferens	Stramenopiles
gi158753	ubiquitin	Drosophila melanogaster	Insecta
gi32400967	polyubiquitin	Griffithsia japonica	Rhodophyta

Accession number	Description	Species	Taxonomic group
3802835 1434 5838	Contig		
spP301532AAA_	Serine/threonine-protein phosphatase 2A	Homo sapiens	Chordata
gi219127721	predicted protein	Phaeodactylum tricornutum CCAP 1055/1	Diatoms
gi224007317	predicted protein	Thalassiosira pseudonana CCMP1335	Diatoms
gi3928142	protein phosphatase	Cicer arietinum	Streptophyta
gi58261744	hypothetical protein	Cryptococcus neoformans	Fungi
gi320164714	protein phosphatase 2	Capsaspora owczarzaki ATCC 30864	Opisthokonta
gi328877027	protein phosphatase 2A scaffold subunit	Dictyostelium fasciculatum	Amoebozoa
gi290999102	phosphoprotein phosphatase A	Naegleria gruberi	Excavata
8191597 1058 4476	Contig		
spO94906PRP6	Pre-mRNA-processing factor 6	Homo sapiens	Chordata
gi224002959	RNA splicing factor	Thalassiosira pseudonana CCMP1335	Diatoms
gi219118732	predicted protein	Phaeodactylum tricornutum CCAP 1055/1	Diatoms
gi298705024	conserved unknown protein	Ectocarpus siliculosus	Stramenopiles
gi301106837	pre-mRNA-processing factor, putative	Phytophthora infestans T30-4	Stramenopiles
gi328770819	hypothetical protein BATDEDRAFT_29905	Batrachochytrium dendrobatidis JAM81	Fungi
gi323448694	hypothetical protein AURANDRAFT_38946	Aureococcus anophagefferens	Stramenopiles
gi221123192	PREDICTED: similar to predicted protein	Hydra magnipapillata	Cnidaria
gi328866185	TPR repeat-containing protein	Dictyostelium fasciculatum	Amoebozoa

Accession number	Description	Species	Taxonomic group
11454230 968 11548	Contig		
spP61964WDR5	WD repeat-containing protein 5	Homo sapiens	Chordata
gi219110699	predicted protein	Phaeodactylum tricornutum CCAP 1055/1	Diatoms
gi224006458	platelet-activating factor	Thalassiosira pseudonana CCMP1335	Diatoms
gi281410775	HET-E	Podospora anserina	Fungi
gi323449228	hypothetical protein	Aureococcus anophagefferens	Stramenopiles
gi145529465	hypothetical protein	Paramecium tetraurelia strain d4-2	Alveolata
gi340504897	WD repeat protein	Ichthyophthirius multifiliis	Alveolata
gi156366072	predicted protein	Nematostella vectensis	Cnidaria
gi168018581	WD40 repeat protein	Physcomitrella patens subsp. patens	Streptophyta

Appendix F

Gene_loss.r Documentation

Written by Tim White

It is often reasonable to hypothesise that a particular gene was present in the common ancestor of some taxa, and over the course of evolution was lost (possibly multiple times, independently) somewhere in the lineages leading to the taxa in which it is absent today. Whenever monophyletic groups of taxa lack the gene, multiple gene loss patterns are possible. Given a tree relating a set of n taxa for which we have experimentally determined the presence or absence of the gene, the question then arises: how many independent loss events probably occurred?

A simple model for exploring this question consists of a rooted tree in which the root is labelled 1, and each taxon appears at a distinct leaf that is labelled 1 or 0 according to whether the gene is observed to be present or absent in that taxon. These labels are gene presence/absence states. A *reconstruction* is an assignment of presence/absence states to all internal nodes that entails no 0-1 edges. A *loss event* is a 1-0 edge in the tree. A *retention event* is a 1-1 edge. Loss events are defined to occur with probability p, retention events with probability 1 - p, and 0-0 edges with probability 1, with these probabilities all being conditional on the parent node having the necessary presence/absence state. (This model bears similarities to Dollo parsimony, but has the location of the single permitted 0-1 transition fixed at an imagined edge leading to the root, rather than being free to vary.) The probability of a given reconstruction that has k loss events and h retention events is therefore p^k * (1-p)^h.

Given a tree and a per-edge loss probability p, and assuming that every reconstruction on that tree is equally likely, the probability that a given dataset evolved on that tree can be calculated by summing the probabilities of every possible reconstruction on that tree. Furthermore the probability that the dataset arose as a result of exactly k loss events on that tree can be calculated by summing over only those possible reconstructions having that many loss events. This calculation can be done for each value of k, and the values of k

(there may in general be more than one) that produce a maximum probability are then the maximum likelihood solutions.

Restricting to a rooted binary tree enables convenient calculation of the probability of a particular reconstruction: as we now show, it depends only on the number of loss events. (If polytomies are allowed, a reconstruction having k loss events can have a variable number of retention events, with this number depending on the tree topology and the locations of the loss events.) The maximum number of loss events occurs uniquely in the reconstruction where each taxon lacking the gene experiences its own, separate loss event on the pendant edge leading to that taxon. In this case, every other edge in the tree is a retention event. There are 2n - 2 edges in a rooted binary tree on n leaves, so if there are n taxa in total, d of which lack the gene, the probability of this d-loss reconstruction is p^d * (1 - p)^(2n - 2 - d). Starting from here, every other reconstruction can be reached by repeating the following process some number of times: find an internal node in the tree that has loss events on both of its child edges, remove these 2 loss events, and change the retention event on the edge leading to the parent to a loss event. This transformation can be applied some (topology- and dataset-dependent) maximum number of times, eventually producing a unique minimum-loss tree with some number c of loss events. Each application of the transformation reduces the number of loss events and the number of retention events by 1 each, so a reconstruction with d - m losses must have probability p^(d - m) * (1 - p)^(2n - 2 - d - m). Rearranging, the probability that the data was produced by a given reconstruction that has k = d - m losses is p^k * (1 - p)^(2n - 2 - 2d + k). This holds for any c <= k <= d, irrespective of the tree topology or the locations of the loss events.

The above formula gives the probability of a particular reconstruction in which the locations of the k loss events are specified. In order to calculate the probability that the data was produced by any of the possible reconstructions that have exactly k loss events, we need to multiply by the number of distinct k-loss reconstructions, r(k).

This can be calculated recursively. Let r(k, t, x) be the number of k-loss reconstructions on the subtree rooted at any node t, assuming that the presence of the gene in t is indicated by x (x must be 1 or 0). Clearly r(k) = r(k, root, 1). First consider the case where the gene is missing from t (i.e. x is 0). Because the regaining of genes is forbidden, there is only one way in which r(k, t, 0) can be nonzero: when k = 0 and all taxa beneath t are missing the

gene. In this case there is exactly 1 possible reconstruction, in which all edges below t are 0-0; in all other cases having x = 0 there are no possible reconstructions:

r(k, t, 0) = 1 if k = 0 and all taxa beneath t are missing the gene, otherwise 0

The remainder concerns the case where x = 1.

For a node u having parent t, define the *extended subtree of u* as the subtree rooted at u plus the edge from t to u. A k-loss reconstruction on the subtree rooted at any node t must distribute all of its k losses between its two child extended subtrees somehow, with some number 0 <= i <= k of losses being allocated to the left extended subtree and the remaining k - i losses going to the right extended subtree. For every way that i losses can be assigned to edges in the left extended subtree, each of the ways that the remaining k - i losses can be assigned to edges in the right extended subtree produces a distinct reconstruction on the subtree rooted at t, so these paired reconstruction counts can be multiplied. Also, every reconstruction having i losses in the left extended subtree is distinct from every reconstruction having j != i losses in that extended subtree, so these k + 1 products can be summed. This establishes that no reconstructions are double-counted; and since every possible reconstruction having k losses must have some number 0 <= i <= k of losses in the left extended subtree, and thus must be counted in exactly one of these k + 1 terms, no reconstructions are omitted. Thus the number of distinct k-loss reconstructions on the subtree rooted at any internal node t, given that the gene is present in t, is

r(k, t, 1) = \sum{i=0}^k{s(i, t.left) * s(k - i, t.right)}

where s(k, u) is the number of k-loss reconstructions on the extended subtree of u, assuming that the gene is present in u's parent t. To calculate s(k, u), observe that two cases are possible: either a loss event occurs on the edge from u's parent t to u, or it does not. If it does, which can only happen if k >= 1, then one loss event has been expended and the gene is now absent from u, so the number of reconstructions for the extended subtree in this case is equal to r(k - 1, u, 0). If it does not, then the number of reconstructions for this extended subtree is equal to r(k, u, 1). The total number of reconstructions is the sum of the counts for these two mutually exclusive possibilities:

s(k, u) = r(k, u, 1) + w(k, u)

w(k, u) = r(k - 1, u, 0) if k >= 1, otherwise 0

Finally, we have the base case

r(k, t, 1) = 1 if k = 0 and t possesses the gene, otherwise 0

for any leaf node t.

A direct implementation of the above recursion would be excessively slow for large inputs, but this can be ameliorated by dynamic programming: whenever a value of r(k, t, x) is sought, a table in memory is first consulted to see whether the result has already been computed for that combination of parameter values. If so, the pre-computed result is returned immediately; if not, it is computed as usual, and stored in memory before the function returns. This memoisation technique reduces the total execution time of the algorithm to the number of *distinct* parameter value combinations that the function can be called with, multiplied by the time taken for each function call to execute. For a rooted binary tree on n taxa, the number of distinct parameter value combinations is upper-bounded by n * (2n - 1) * 2. Each function invocation performs O(n) internal work summing products of counts, so the overall time complexity for calculating all n + 1 counts -- namely, r(i) for 0 <= i <= n -- is O(n^3). The memory cost of storing the memoisation table is O(n^2), and this dominates the space complexity.

This algorithm is implemented in an R program, gene_loss.r, which uses the freely available R modules ape and phylobase. The program accepts as input a binary tree in Newick format, a 2-column table mapping taxon names to gene presence/absence values, and the probability p of gene loss on an edge, and outputs the maximum likelihood estimate of the number of loss events. Alternatively the loss probability p can be omitted, in which case the maximum likelihood value (the loss probability that maximises the total probability of the data over all possible reconstructions) will first be estimated using the R optimize() function. The program also produces a table showing the probability given the data for each number of loss events (calculated by dividing each r(i) value by the total probability of the data), as well as the minimum, maximum, and expected numbers of loss events. Running it with the option --help gives further details.