UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Università degli Studi di Padova

SCUOLA DI DOTTORATO DI RICERCA IN : Bioscienze e biotecnologie

INDIRIZZO: Biochimica e Biofisica

CICLO XXII

Going ultra deep to unravel the secret recipe of biofuel

Direttore della Scuola : Ch.mo Prof. Giuseppe Zanotti

Coordinatore d'indirizzo: Ch.mo Prof. Maria Catia Sorgato

Supervisore :Ch.mo Prof. Giorgio Mario Giacometti

Dottoranda : Elisa Corteggiani Carpinelli

"To this end, they procured a royal patent for erecting an academy of projectors in Lagado; and the humour prevailed so strongly among the people, that there is not a town of any consequence in the kingdom without such an academy. In these colleges the professors contrive new rules and methods of agriculture and building, and new instruments, and tools for all trades and manufactures; whereby, as they undertake, one man shall do the work of ten; a palace may be built in a week, of materials so durable as to last for ever without repairing. All the fruits of the earth shall come to maturity at whatever season we think fit to choose, and increase a hundred fold more than they do at present; with innumerable other happy proposals. The only inconvenience is, that none of these projects are yet brought to perfection; and in the mean time, the whole country lies miserably waste, the houses in ruins, and the people without food or clothes."

Jonathan Swift

GULLIVER'S TRAVELS INTO SEVERAL REMOTE NATIONS OF THE WORLD PART III. A VOYAGE TO LAPUTA, BALNIBARBI, LUGGNAGG, GLUBBDUBDRIB, AND JAPAN. CHAPTER IV

"The first man I saw was of a meagre aspect, with sooty hands and face, his hair and beard long, ragged, and singed in several places. His clothes, shirt, and skin, were all of the same colour. He has been eight years upon a project for extracting sunbeams out of cucumbers, which were to be put in phials hermetically sealed, and let out to warm the air in raw inclement summers. He told me, he did not doubt, that, in eight years more, he should be able to supply the governor's gardens with sunshine, at a reasonable rate: but he complained that his stock was low, and entreated me "to give him something as an encouragement to ingenuity, especially since this had been a very dear season for cucumbers." I made him a small present, for my lord had furnished me with money on purpose, because he knew their practice of begging from all who go to see them."

Jonathan Swift

GULLIVER'S TRAVELS INTO SEVERAL REMOTE NATIONS OF THE WORLD PART III. A VOYAGE TO LAPUTA, BALNIBARBI, LUGGNAGG, GLUBBDUBDRIB, AND JAPAN. CHAPTER IV

SCIENTIST ARE DREAMERS..

AND ENJOY CHALLENGES..

"Solar energy into fuel:

If a leaf can do it, we can do it"

James Barber

Abstract

Nonostante le microalghe rivestano una particolare importanza per l'ecologia, la biochimica e le biotecnologie, solo poche specie sono state sequenziate a tutt'oggi e per lo più nell'era del sequenziamento di tipo Sanger. Mentre i sequenziatori di nuova generazione hanno fornito straordinari mezzi al risequenziamento di genomi di singoli individue per cui il genoma della specie di riferimento era già disponibile, il sequenziamento di genomi eucariotici completamente nuovi, utilizzando sequenze corte, presenta ancora grandi difficoltà. Le principali difficoltà si riscontrano in fase di assemblaggio e sono principalmente dovute alle notevoli dimensioni dei genomi eucariotici e alla presenza di regioni ripetute. Le microalghe, nei loro genomi, grandi in genere dalle 30Mb alle 100Mb, contengono poche regioni a bassa complessità e costituiscono pertanto degli interessanti organismi sui quali sperimentare il sequenziamento *ex novo* utilizzando esclusivamente i sequenziatori di seconda generazione. In questo lavoro, descriviamo il sequenziamento e l'assemblaggio del genoma della microalga *Nannochloropsis gaditana*, ottenuto utilizzando le sequenze prodotte dal 454 della Roche e dal SOLiD. Il sequenziamento di 'mate-pairs' utilizzando il SOLiD ha prodotto una copertura di sequenza di circa 250 volte la grandezza del genoma, mentre il sequenziamento 454 è stato utilizzato per produrre una ulteriore copertura di 7 volte con sequenze di media lunghezza. Le sequenze sono state assemblate in una versione preliminare non del tutto finita di 32Mb, dove 18.7Mb sono state incluse in 167 grandi scaffolds. Il 50% degli scaffolds ottenuti è più grande di 50Kb, mentre un terzo del genoma è stato assemblato in circa 20 contigs. Il nostro lavoro conferma la previsione che le tecniche di nuova generazione sono adatte al sequenziamento del genoma di una microalga e consentono di ottenere risultati utili in un tempo più breve di quelli tradizionali e ad un minor costo. I dati ottenuti potranno esser utilizzati per analisi comparative del genoma e saranno anche un importante prerequisito per l'applicazione di tecniche ricombinati alla microalga di interesse biotecnologico.

Inoltre, durante questo progetto, sono state portate avanti anche diverse analisi del trascrittoma tramite sequenziamento, finalizzate alla produzione di una lista di geni utile all'annotazione del genoma e all'identificazione di geni differenzialmente espressi in condizioni di accumulo di lipidi. Vengono presentate, in questo lavoro, anche alcune innovazioni tecniche che hanno permesso di sfruttare al meglio il sequenziamento SOLiD per produrre un'accurata annotazione della struttura del trascrittoma e una più completa analisi dei geni differenzialmente espressi codificati negli organelli. I risultati dimostrano che una sola corsa SOLiD su campioni preparati in modo adeguato, è sufficiente per l'annotazione accurata di un genoma completamente nuovo e per l'individuazione di geni differenzialmente espressi in condizioni di interesse.

Abstract

Microalgae are of great importance in ecology, biochemistry, and biotechnology. Nevertheless, just a few genomes have been sequenced so far, most of them by Sanger sequencing. While next generation sequencing techniques have revolutionized genome resequencing, genetic mapping, or transcriptome and ChIP analyses, *de novo* assembly of eukaryotic genomes still presents significant hurdles, because of their large size and stretches of repetitive sequences. Microalgae contain fewer repetitive regions in their 30–100 Mb genomes than genomes of mammals or higher plants and thus are suitable candidates to test *de novo* genome assembly from short sequence reads. Here, we present a draft sequence of the *Nannochloropsis gaditana* genome that was obtained by a combination of SOLiD and Roche 454 sequencing. Mate-Pair SOLiD sequencing of genomic DNA to 250-fold coverage and an additional 7-fold coverage by single-end 454 sequencing resulted in 15 Gb of raw sequence data. Reads were assembled to a 32 Mb draft version (N50 of 50kb) with the a pipeline of tools evolved in our group. 167 scaffolds were produced accounting for 18.7Mb.

Our study supports the expectation that for typical microalgae, *de novo* assembly of genomes from short sequence reads alone is feasible, cheap and efficient; that a mixture of SOLiD and 454 sequencing substantially improves the assembly; and that the resulting data can be used for comparative studies and to provide a valuable framework to plan the application of recombinant techniques.

Furthermore a whole transcriptome analyses was carried out to identify, characterize and catalogue all the transcripts expressed, exploiting the great potential of RNA-Seq using the SOLiD platform to determine the correct gene annotation, the identification and characterization of the splicing patterns and to obtain the structure of genes, also defining—at single nucleotide resolution—the transcriptional boundaries of genes and the expressed Single Nucleotide Polymorphisms (SNPs). Important technical innovations were also introduced in this work that allowed a precise mapping of the transcription start to support a robust prediction of the regulatory region on the genome sequence.

RNA-Seq was also used to study the differential expression of transcripts in cultures able to accumulate substantially different amounts of lipids, in order to obtain insights on lipid metabolism of *N.gaditana*. The study was carried on using as input sequences for the SOLiD run both polyadenylated mRNA enriched fractions and ribo-depleted RNA samples that allowed to recovery also the plastidial mRNA.

Our results shows how data obtained from a single SOLiD run, applying specific *ad hoc* variation on the standard protocols, provide enough coverage to support a valuable annotation of a

completely new genome and to provide all the information necessary to underline the main features of pathways of interest if different biological samples are compared.

The *N. gaditana* sequencing project fulfilled the two important aims of assembling a draft of the genome sequence using sole next generation short reads and providing a careful genome annotation, with the goal of a better understanding of *N. gaditana* biology and the idea of improving its value as a model organism for biotechnological applications related to biofuel production.

Table of contents

1. Introduction

The arrival of next generation sequencing technologies has changed the way we think about genome research and is changing even more our perspectives in biological and biotechnological investigation. The ability to produce cheaply an enormous volume of data - up to billions of short reads per instrument run- expanded the realm of experimentation beyond any imagination, opening up the capability to investigate the biodiversity of natural communities, to spot rare transcripts, to speculate about genes' structure and sequence variations at a very fine level.

Over the past five years, since the potential for ultra deep sequencing has become a reality, a number of surveys was made to investigate biodiversity of different natural environments, showing us that metabolic modes of unicellular microorganisms are much more diverse than we expected and that the general map of the basic energy-carbon metabolism of a cell must be somehow integrated, revised, rebalanced.

In the very same period the growing interest of the society for renewable energy sources has encouraged the scientific community to further investigate the metabolic activity of a growing number of photosynthetic and non photosynthetic microorganism in order to exploit the available biomass for the production of fuels and electricity. An incredible amount of evidences about the extraordinary capacity of certain microorganisms for lipids or starch accumulation, hydrogen evolution or survival in impairing conditions has flooded the scientific literature stressing how little we know about the possible metabolic modes and about the control that allows the flux of energy through alternative metabolic pathways.

In other words, this is one of those exciting moments in science in which the capability to look at reality in major detail and the collection of a growing number of data requires an improvement in the interpretation in order to cope with the new evidences.

To cast a light on a corner of this scenario we considered one of the most popular photosynthetic microorganism able to accumulate triacylglycerols in nitrogen deprivation and we did ultra deep sequencing to obtain the genome and an accurate annotation of the transcriptome in order to unravel its metabolic profile. Furthermore we applied the RNA-Seq to characterize the differentially expressed genes responsible for changes in the metabolic flux in the different growth conditions.

1.1. Next generation sequencing technology: new adventures in biology

DNA sequence represents now a single format onto which a broad range of biological phenomena can be projected for high throughput data collection. A few main modification in the pipeline of the standard sequencing procedure, that has been used for almost 30 years, helped to make the leap and boost the production of sequences in past few years. The most important innovations were the abandon of the cloning procedure in bacteria, the substitution of capillary electrophoresis with imaging methods and the extraordinary capability for parallelization achieved with the new systems. Despite sharing these main characteristics, each of the different commercially available technology realized the goal through a different strategy.

The Sanger method

Since the early 1990s, DNA sequence production has almost exclusively been carried out with capillary-based, semi-automated implementations of the Sanger biochemistry. In high-throughput genome sequencing projects, DNA to be sequenced is obtained by one of two approaches: first, for shotgun de novo sequencing, randomly fragmented DNA is cloned into a high-copy-number plasmid, which is then used to transform *E.coli*; or second, for targeted resequencing, PCR amplification is carried out with primers that flank the target. The output of both approaches is an amplified template, either as many clonal copies of a single plasmid insert present within a spatially isolated bacterial colony that can be picked, or as many PCR amplicons present within a single reaction volume. The sequencing biochemistry takes place in a cycle sequencing reaction, in which cycles of template denaturation, primer annealing and primer extension are performed. The primer is complementary to known sequence immediately flanking the region of interest. Each round of primer extension is stochastically terminated by the incorporation of fluorescently labelled dideoxynucleotides (ddNTPs). In the resulting mixture of end-labelled extension products, the label on the terminating ddNTP of any given fragment corresponds to the nucleotide identity of its terminal position. Sequence is determined by high-resolution electrophoretic separation of the single-stranded, end-labelled extension products in a capillary- based polymer gel. Laser excitation of fluorescent labels as fragments of discreet lengths exit the capillary, coupled to four-colour detection of emission spectra, provides the readout that is represented in a Sanger sequencing trace. Software translates these traces into DNA sequence, while also generating error probabilities for each base-call. Simultaneous electrophoresis in 96 or 384 independent capillaries is the maximum parallelization that can be obtained.

Second-generation or cyclic-array DNA sequencing

A number of different alternative strategies for DNA sequencing have been evolved in the past few years, in particular various implementations of cyclic-array sequencing have been realized in successful commercial products: 454 sequencing (used in the 454 Genome Sequencers, Roche Applied Science; Basel); Solexa technology (used in the Illumina (San Diego) Genome Analyzer); and the SOLiD platform (Applied Biosystems; Foster City, CA, USA)).

The concept of cyclic-array sequencing can be summarized as the sequencing of a dense array of DNA features by iterative cycles of enzymatic manipulation and imaging-based data collection (Mitra et al. 2003). In shotgun sequencing with cyclic-array methods, common adaptors are ligated to fragmented genomic DNA, which is then subjected to one of several protocols that results in an array of millions of spatially immobilized PCR 'colonies'. Each 'colony' consists of many copies of a single shotgun library fragment. As all colonies are tethered to a planar array, a single microliter-scale reagent volume (e.g., for primer hybridization and then for enzymatic extension reactions) can be applied to manipulate all array features in parallel. Imaging-based detection of fluorescent labels incorporated with each extension is used to acquire sequencing data on all features in parallel. Successive iterations of enzymatic interrogation and imaging are used to build up a contiguous sequencing read for each array feature. Although the different commercial platforms are quite diverse in sequencing biochemistry as well as in how the array is generated, their work flows are conceptually similar. Library preparation is accomplished by random fragmentation of DNA, followed by in vitro ligation of common adaptor sequences. Alternative protocols can be used to generate jumping libraries of mate-paired tags with controllable distance distributions. The generation of clonally clustered amplicons to serve as sequencing features can be achieved by two main different approaches: the emulsion PCR (used in both SOLiD and 454) or bridge PCR (used by Illumina). What is common to these methods is that PCR amplicons derived from any given single library molecule end up spatially clustered, either to a single location on a planar substrate (bridge PCR), or to the surface of micron-scale beads, which can be recovered and arrayed (emulsion PCR). The sequencing process itself consists of alternating cycles of enzyme-driven biochemistry and imaging-based data acquisition. Serial extension of primed templates are carried out, where the enzyme driving the synthesis can be either a polymerase (454 and Illumina) or a ligase (SOLiD). Data are acquired by imaging of the full array at each cycle.

Global advantages of cyclic-array strategies, relative to Sanger sequencing, include the following:

- *in vitro* construction of a sequencing library, followed by in vitro clonal amplification to generate sequencing features, circumvents several bottlenecks that restrict the parallelism of conventional sequencing (that is, transformation of *E. coli* and colony picking);
- array-based sequencing enables a much higher degree of parallelism than conventional capillary-based sequencing. As the effective size of sequencing features can be on the order of 1 μm, hundreds of millions of sequencing reads can potentially be obtained in parallel by rastered imaging of a reasonably sized surface area;
- because array features are immobilized to a planar surface, they can be enzymatically manipulated by a single reagent volume.

The major disadvantage of second-generation DNA sequencing seems to be at the moment the read-length, that for all of the new platforms, is currently much shorter than that obtained using conventional Sanger sequencing. Raw accuracy is also an important issue since, on average, base-calls generated by the new platforms are still less accurate than base-calls generated by Sanger sequencing.

454 technology

Sample preparation

Fragments of DNA are ligated to adapters that facilitate their capture on beads (one fragment per bead). A water-in-oil emulsion containing PCR reagents and one bead per droplet is created to amplify each fragment individually in its droplet. After amplification, the emulsion is broken, DNA is denatured and the beads, containing one amplified DNA fragment each, are distributed into the wells of a fiber-optic slide.

Pyrosequencing

The wells are loaded with sequencing enzymes and primer (complementary to the adapter on the fragment ends), and then exposed to a flow of one unlabelled nucleotide at a time, allowing synthesis of the complementary strand of DNA to proceed. When a nucleotide is incorporated, pyrophosphate is released and converted to ATP, which fuels the luciferase-driven conversion of luciferin to oxyluciferin and light. As a result, the well lights up.

Throughput

The average read length of the 'GS FLX Titanium Series' is 400 nucleotides and it can produce up to >1 million high-quality reads per a 10 hour run, with an estimated accuracy of 99%.

SOLiD technology

Sample preparation

Fragments of DNA are ligated to adapters and amplified on beads by emulsion PCR. The DNA is denatured and the beads deposited onto a glass slide.

Sequencing by ligation

A sequencing primer is hybridized to the adapter and its 5' end is available for ligation to an oligonucleotide hybridizing to the adjacent sequence. A mixture of octamer oligonucleotides compete for ligation to the primer (the bases in fourth and fifth position on these oligos are encoded by one of four colour labels). After its colour has been recorded, the ligated oligonucleotide is cleaved between position 5 and 6, which removes the label, and the cycle of ligation-cleavage is repeated. In the first round, the process determines possible identities of bases in positions 4, 5, 9, 10, 14, 15, etc. The entire process is repeated, offset by one base by using a shorter sequencing primer, to determine positions 3, 4, 8, 9, 13, 14, etc., until the first base in the sequencing primer (position 0) is reached. Since the identity of this base is known, the colour is used to decode its neighbouring base at position 1, which in turn decodes the base at position 2, etc., until all sequence pairs are identified.

Throughput

The current maximum read length of the '5500xl SOLiD' is 75 nucleotides for fragments libraries, 75bp plus 35bp for paired-end sequencing and 60bp plus 60bp for mate-paired samples. The throughput per run using nanobeads can reach 300 Gb for fragments libraries or greater than 4.8 billion tags while running paired-end or mate-paired. A single run takes up to 7 days for mate-paired or paired-end libraries. The system accuracy is guaran-

teed to be 99.99%. Up to 96 different samples can be loaded on the same ran using the bar coding protocol.

Illumina technology

Sample preparation

Fragments of DNA are ligated to end adapters, denatured and bound at one end to a solid surface already coated with a dense layer of the adapters. Each single-stranded fragment is immobilized at one end, while its free end bends over and hybridizes to a complementary adapter on the surface, which initiates the synthesis of the complementary strand in the presence of amplification reagents. Multiple cycles of this solid-phase amplification followed by denaturation create clusters of ~1,000 copies of single-stranded DNA molecules distributed randomly on the surface.

Sequencing with reversible terminators

Synthesis reagents, added to the flow cell, consist of primers, DNA polymerase and four differently labelled, reversible terminator nucleotides. After incorporation of a nucleotide, which is identified by its colour, the 3' terminator on the base and the fluorophore are removed, and the cycle is repeated for a read length of nucleotides.

Throughput

The current maximum read length of the 'HiSeq 2000' is 100 nucleotides. The throughput per run can reach 200 Gb: up to one billion cluster passing filter and up to two billions paired-ends reads. A single run takes up to 8 days for mate-paired or paired-end libraries. Typically about 80 to 90% of the bases obtained in paired 100bp runs are high quality.

Single molecule real time sequencing

Several academic groups and companies are working on technologies for ultra-fast DNA sequencing that are substantially different from the current crop of available next-generation platforms. One approach is nanopore sequencing, in which nucleic acids are driven through a nanopore (either a biological membrane protein such as alpha-hemolysin or a synthetic pore). Fluctuations in DNA conductance through the pore, or, potentially, the detection of interactions of individual bases with the pore, are used to infer the nucleotide sequence. Although progress has been made in achieving early proof-of-concept demonstrations with such methods, major technical challenges remain along the path to a truly practical nanopore-based sequencing platform.

Another approach involves the real-time monitoring of DNA polymerase activity. Nucleotide incorporations can potentially be detected through FRET (fluorescence resonance energy transfer) interactions between a fluorophore-bearing polymerase and gamma phosphate-labelled nucleotides (visigen; Houston), or with zero-mode waveguides (Pacific Biosciences; Menlo Park, CA, USA), with which illumination can be restricted to a zeptoliter-scale volume around a surface tethered polymerase such that incorporation of nucleotides (with fluorescent labels on phosphate groups) can be observed with low background. Pacific Biosciences recently demonstrated substantial progress toward a working technology, including the potential for longer reads than Sanger sequencing (Chin et al. 2010).

Advantages and disadvantages of different approaches

Although second generation sequencing have revolutionized genomics and biology providing new incredible tools, Sanger sequencing is still used where in terms of costs, limitations and practical aspects of implementation, represents an advantage. The applications of conventional sequencing for small-scale projects in the kilobase to megabase range, will likely remain the technology of choice for the immediate future. This is a consequence of its greater 'granularity' (that is, the ability to efficiently operate at either small or large production scales) relative to the new technologies. Even so, it is clear that despite limitations relative to Sanger sequencing (e.g., in terms of read-length and accuracy), large-scale projects have quickly come to depend entirely on next-generation sequencing. As we make comparisons between traditional sequencing and next generation in terms of applications, there are also important differences among the second-generation platforms themselves that may result in advantages with respect to specific applications. Some applications (e.g., resequencing) may be more tolerant of short read-lengths than others (e.g., *de novo* assembly). For applications relying on tag counting (e.g., quantification of protein-DNA interactions), one would actually prefer a given amount of sequencing to be split into as many reads as possible (above some minimum length that allows placement to a reference). The overall accuracy as well as the specific error distributions of individual technologies (e.g., the rate of insertion-deletion versus substitution errors; the propensity for systematic consensus errors) may also be highly relevant. Mate-paired reads, useful in *de novo* assembly and for mapping structural variants, for example, are now available with all of the second generation platforms, but the extent to which the distance distribution with which the read pairs are separated can be controlled or varied may be an important factor. Finally, of course, the cost of sequencing varies greatly between the second-generation platforms.

In addition to reducing the per base cost of sequencing by several orders of magnitude, second generation instruments have fewer laboratory work to prepare the sequencing samples; instead, the principle challenge is represented by downstream data management. The diversity and rapid evolution of next generation sequencing technology is posing challenges for bioinformatics in areas including sequence quality scoring, alignment, assembly and data release. Moreover, as the research community is becoming more familiar with the potential of these platforms to interrogate biology, it is looking to ask new questions and run new types of experiments beyond the current bioinformatics.

No doubts that the complexity of analysis will rise markedly, but the opportunities for an immensely deeper understanding will be even greater.

What would we do if we could sequence everything?

Although we have not quite reached the point where cost and technology are no object, new sequencing techniques have not simply changed the landscape but have placed basic, clinical and translational research scientists into a new and unfamiliar world in which entirely different types of questions can be addressed.

Cataloguing sequences and their variation

The most direct and obvious result of enhanced sequencing capabilities has been the simple accumulation of much more sequence data, and hence, many sequences from different

species that allow more informative analyses of phylogeny and evolution. The rapid resequencing of multiple strains of the most popular model organisms provides a taste of things to come with the new techniques that are supplanting the earlier methods. With the much larger range of sequences becoming available, it will now be possible to study the effects of selective forces in evolution at the level of individuals rather than entire populations. The search for individual variation has been used to great effect in recent whole-genome studies for understanding disease associations in humans with dozens of novel genes implicated in various phenotypes. A special interest is represented by the study of SNPs associated to human diseases, but these are not the sole class of variation associated with important phenotypic consequences. From the early days of cytogenetics, structural genomic aberrations have been known to play a role in human disease. Until recently, only relatively large-scale rearrangements such as chromosomal aneuploidies and megabase-sized deletions, duplications, translocations, or inversions, detectable by traditional karyotyping techniques, were studied and used as diagnostic markers. With the advent of high-resolution genome-wide technologies and techniques, previously undetected submicroscopic structural variations have been shown to exist, both in the genomes of diseased individuals, as well as in the genomes of normal populations.

Dynamic DNA and metagenomics

The rapidly changing genomes of some viruses (such as HIV) can be analysed by deep resequencing of samples from multiple patients over time and with treatment.

High-throughput sequencing can also identify viruses, bacteria and other organisms present in a complex biological sample by identifying their genomic signatures. Traditionally, 16S ribosomal RNA or other highly conserved genes have been used to characterize the variety of organisms in a given sample to minimize the amount of sequencing required, but that approach limits the amount of information that can be derived from a sample. A deep characterization of the genomes represented in a mixture as well as how that composition changes over time can now be carried out and allows us to picture a complex system in detail. The unexpected diversity of populations, such as the mixture of syntrophic microbes that utilize methane from deep sea vents, is causing a rethinking of how those complex ecosystems work and evolve (Pernthaler et al. 2008). A variety of other ecosystems is now just beginning to be examined with millions of sequences used to characterize the tremendous genetic diversity present in the different environments. The tremendous potential for the use of this diversity as substrate for novel industrial enzymes and processes can only be speculated on at this point. The ability to directly sequence and quantify gene expression in those complex systems gave us also the possibility to speculate about the main biochemistry carried on in the system. Moreover, measuring changes in transcription might prove a sensitive technology for assessing the dynamic characteristics and the health of those ecosystems. In addition to extreme or interesting environments, attention is also being directed at the variety of ecosystems present on and within the human body and how they may affect health and diseases.

Epigenome

Increasingly, we are coming to recognize that the message encoded within the DNA sequence is regulated in a variety of complex ways, including modifications of the genome

itself. One example, alterations in the pattern of DNA methylation in human DNA, has been associated with the level of transcription, a variety of disease states and a host of other phenotypes. The most common variation in human DNA, DNA methylation at the 5'-position in cytosine, can be detected by sequencing of bisulfite-treated DNA, a method that allows methylated and unmethylated positions to be distinguished. The new technology allows a sufficient sequencing capacity for studying this kind of patterns at whole genome level. In addition to methylation patterns, it is possible to detect other types of structural variation as well. Numerous enzymes and chemical agents are differentially sensitive to various forms of DNA strain, structure, accessibility and other features. With sufficient sequencing capacity, the ability to see breakpoints induced by nucleases or modifications induced by chemical agents becomes practical at the whole genome el. Hypersensitive regions are frequently associated with regulatory and protein binding phenomena, so the ability to detect them at high resolution will provide further understanding of many processes.

Interactome

DNA sequence, structure and modifications are recognized by a wide variety of proteins like histones, which bind unspecifically to DNA and help to compact it, and transcription factors, which bind to specific sequence motifs and activate or repress transcription. Chromatin immunoprecipitation (ChIP) was developed to identify novel protein binding sites across the genome. This allows researchers to identify protein-binding sites within living cells and has been expanded to a genome wide assay with the improvement of microarray and sequencing technologies. The hybridization technologies have traditionally had the benefit of low cost, but have been limited by incomplete coverage of the genome, difficulties in distinguishing closely related sequences and poor resolution of sequence boundaries. As sequencing technologies achieve lower cost and higher throughput, ChIP-Seq approaches are yielding more sensitive and complete coverage of the entire genome, providing more accurate pictures of the complex regulatory landscape within the genome.

Transcriptome: more variants and greater precision for measuring RNA

The ultimate goal of understanding epigenomics, transcription factors and histone modifications is deducing how the genome responds to the complex collection of factors that influence cellular physiology. Measuring RNA expression has been accomplished by increasingly sophisticated tools of greater specificity and higher throughput, which have allowed the detection of increased numbers of RNA species in larger sample sets over a wider dynamic range. Although studies with the recently developed technologies of qPCR, DNA microarrays and serial analysis of gene expression (SAGE) have provided tremendous insight into biological processes, each suffers from its own biases and limitations. In contrast, RNA-Seq on NGS platforms has clear advantages over the existing approaches. First, unlike hybridization-based technologies, RNA-Seq is not limited to the detection of known transcripts, thus allowing the identification, characterization and quantification of new splice isoforms. In addition, it allows researchers to determine the correct gene annotation, also defining -at single nucleotide resolution- the transcriptional boundaries of genes and the expressed Single Nucleotide Polymorphisms (SNPs). Other advantages of RNA-Seq compared to microarrays are the low background signal, the absence of an up-

per limit for quantification and consequently, the larger dynamic range of expression levels over which transcripts can be detected. RNA-Seq data also show high levels of reproducibility for both technical and biological replicates. Moreover, compared to the most dynamic and accurate qPCR, RNA-Seq presents the advantages of extending the data collection at a transcriptome wide level and of sampling the important information carried by the non coding RNAs.

de novo sequencing of small eukaryotic genomes

While the capability to output an extraordinary number of very short reads has been a resource for the resequencing of already available genomes, providing enough coverage to support the robust assignment of sequence and structural variations, rather complicated seems to be the application of the new technology to boost the *de novo* sequencing of eukaryotic genomes. As a general situation the maximum length of a read, for any of the available sequencing method (including the Sanger system), is orders of magnitude smaller than a genome whose sequence is desired. Therefore, in any sequencing project, libraries representing the source DNA are randomly oversampled and the parent sequence is reconstructed by inferring mutual overlaps between the reads. This basic procedure can be eventually extended to entire genomes and the resulting whole genome shotgun (WGS) sequencing method has been the basis of all the most recent genome projects. While calculating the oversampling needed a critical number is reached where the percentage of genome covered rises very poorly in response to an enormous increase in reads number. Reached this point a percentage of bases will be found not sequenced or sequenced by a number of reads below the cut-off necessary to support a robust assembly. While considering small and relatively simple genomes, such those of viruses and prokaryotes, the high parallelization of second generation systems allows incredibly high average coverage and makes it possible to obtain the whole genome draft of a new species even in a single run. Short reads present many more difficulties during the assembly and a higher number of reads per base is needed to generate a consensus due to the poor sequence quality if compared to the Sanger reads. Nevertheless many of this problems are overcome by the increased coverage itself, and whole genome shot gun sequencing of bacteria using any of the second generation sequencing methods has been demonstrated (Tauch et al. 2008; den Bakker et al. 2010; Technical notes by Illumina).

Even though next generation is not limiting in terms of number of reads to increase the theoretical coverage, too many gaps are still found while sequencing eukaryotic genomes and that is because complex genomes contain many repetitive sequences that are very difficult to align and include in the assembly. To help the process of assembly, reads can be obtained with some long range information. Two common methods traditionally used are: 'double-barreled sequencing', where pairs of reads are obtained from both ends of inserts of various sizes, and 'hierarchical sequencing', where the genome is covered by cloned inserts such as BACs, and then reads are obtained separately from each clone. Paired reads can resolve repeats by jumping across them and disambiguating the ordering of flanking unique regions. Whole genome double-barreled shotgun sequencing has been used successfully to assemble several complex genomes. Hierarchical sequencing relies on clustering reads into small local sets that represent the sequence of one clone, where most of the repeats have a unique copy and therefore assembly is straightforward. In most ap-

plications of hierarchical sequencing a complete physical map of a large set of clones was constructed, covering the genome with redundancy and then, a minimal tiling subset of those clones was selected for full sequencing. All the second generation sequencing platforms have developed protocols for the realization of mate-paired libraries, where the sequenced reads are localized at both ends of genomic fragments of various size. This important procedure, widely used for the study of structural rearrangements in resequenced genomes, is also of fundamental importance to overcome part of the problems encountered during the assembly of eukaryotic genomes.

In this thesis we describe the sequencing strategy and assembly methodology designed for *de novo* sequencing of a small eukaryotic genome using short reads. Our protocol completely relies on second generation sequencing technology. A robust assembly was obtained using the sequences generated by a Titanium 454 run. Reads were assembled and used to generate large contigs of nuclear, mitochondrial and plastidial genome. Mate-paired libraries were then sequenced using the SOLiD 3 plus for bridging the contigs and obtaining the scaffolds. All the alignments were performed using the PASS tool realized in our group (Campagna et al. 2009), while scaffolds were generated using the pipeline Divorce, specifically evolved for this project. A gap filling was also performed, using the sequences that did not align on the contigs, in order to improve the assembly. Correctness of the assembly was finally checked by PCR using the Promix design tool (Zara et al.). We also worked to produce a physical map exploiting the high throughput of SOLiD 4 system to realize a variation of the Happy Mapping approach (Dear at al. 1993; Jiang et al. 2009; Vu et al. 2010).

References

Mitra R D, Shendure J, Olejnik J, Edyta Krzymanska O, Church G M (2003) Fluorescent in situ sequencing on polymerase colonies. Anal. Biochem. 320, 55–65.

Chin C S, Sorenson J, Harris J B, Robins W P, Charles R C, Jean-Charles R R, Bullard J, Webster D R, Kasarskis A, Peluso P, Paxinos E E, Yamaichi Y, Calderwood S B, Mekalanos J J, Schadt E E, Waldor M K (2010) The Origin of the Haitian Cholera Outbreak Strain. N Engl J Med. [Epub ahead of print]

Pernthaler, A. et al. (2008) Diverse syntrophic partnerships from deep-sea methane vents revealed by direct cell capture and metagenomics. Proc. Natl. Acad. Sci. USA 105, 7052–7057.

Shendure J, Ji H (2008) Next-generation DNA sequencing Nat Biotechnol. 26(10):1135-45.

Kahvejian A, Quackenbush J, Thompson J F (2008) What would you do if you could sequence everything?Nat Biotechnol. 26(10):1125-33.

Campagna D, Albiero A, Bilardi A, Caniato E, Forcato C, Manavski S, Vitulo N, Valle G (2009) PASS: a program to align short sequences. Bioinformatics. 25(7):967-8.

Zara I, Schiavon R, Valle G Promix project http://promix.cribi.unipd.it/cgi-bin/promix/promix_menu.pl

Dear P H, Cook P R. (1993) Happy mapping: linkage mapping using a physical analogue of meiosis.Nucleic Acids Res. 21(1):13-20.

Jiang Z, Rokhsar D S, Harland R M (2009) Old can be new again: HAPPY whole genome sequencing, mapping and assembly.Int J Biol Sci. 5(4):298-303.

Vu G T, Dear P H, Caligari P D, Wilkinson M J (2010) BAC-HAPPY mapping (BAP mapping): a new and efficient protocol for physical mapping.PLoS One. 5(2)

den Bakker H C, Cummings C A, Ferreira V, Vatta P, Orsi R H, Degoricija L, Barker M, Petrauskene O, Furtado M R and Wiedman M (2010) Comparative genomics of the bacterial genus Listeria: Genome evolution is characterized by limited gene acquisition and limited gene loss *BMC Genomics* [Epub ahead of print]

Tauch A, Schneider J, Szczepanowski R, Tilker A, Viehoever P, Gartemann K H, Arnold W, Blom J, Brinkrolf K, Brune I, Götker S, Weisshaar B, Goesmann A, Dröge M, Pühler A (2008) Ultrafast pyrosequencing of Corynebacterium kroppenstedtii DSM44385 revealed insights into the physiology of a lipophilic corynebacterium that lacks mycolic acids J Biotechnol. 136(1-2):22-30

Illumina Technical Notes: De Novo Assembly Using Illumina Reads http://www.illumina.com/Documents/products/technotes/technote_denovo_assembly_ecoli.pdf

1.2. Genes and genomes: from reading to writing

Metagenomics and metabolic diversity in the marine environments

A new powerful lens for viewing the microbial world is now accessible for the scientific community that has the potential to revolutionize our understanding of the entire living world: the generation of large datasets of genome information directly from complex ecosystems using the high-throughput of next generation sequencing technologies. Throughout immeasurable time, microorganisms evolved and accumulated remarkable physiological and functional heterogeneity, and now constitute the major reserve for genetic diversity on earth. Cultivation-independent assessment of microbial communities and their metagenomes, while changing our understanding of microbial diversity and ecology also offers a major resource for bioprospecting; furthermore, this material is a major asset in the search for new biocatalytics (enzymes) for various industrial processes, including the production of biofuels. Currently, there is a global political drive to promote industrial biotechnology as a central feature of the sustainable economic future of modern industrialized societies. This requires the development of novel enzymes, processes, products and applications. Metagenomics promises to provide new molecules with diverse functions.

The potential of marine microorganisms for the discovery of new enzymes and new metabolic strategies has led to numerous new discoveries in recent years. These discoveries have been made from both sequencing of microorganisms directly harvested from different marine environment and from functional approach metagenomic libraries.

Metabolic diversity unravelled by sequencing approaches

Photobiology of marine picoplankton

Sequencing of genome fragments from bacterioplankton revealed a new class of genes of the rhodopsin family (named proteorhodopsin) that had never before been observed in bacteria. When the bacterial proteorhodopsin was expressed in *Escherichia coli*, it functioned as a light-driven proton pump (Béjà et al. 2000). The genomic survey of uncultivated marine bacteria led directly to the discovery of a new type of light-driven energy generation in oceanic bacteria. Later studies showed that optimized spectral 'tuning' of bacterial rhodopsins matches depth-specific light availability (Béjà et al. 2001; Venter at al. 2004). The ability to genetically and functionally manipulate natural and engineered proteorhodopsin variants has important potentials for use in biotechnological applications (Kelemen at al. 2003).

Other modes of bacterial phototrophy have also been shown to be common in ocean surface waters. These represent alternative strategies of light utilization compared with that of well-known oxygenic photoautotrophs. Abundant bacteriochlorophyll-containing aerobic, anoxygenic phototrophic (AAnP) bacteria were first identified in seawater through bio-optical and biophysical measurements. Following these reports, several different types of AAnP were found in DNA samples from Monterey Bay as 'photosynthetic superopers', which encode the photosynthetic reaction centre, and carotenoid and bacteriochlorophyll biosynthetic genes (Béjà et al. 2002). In total, the combined biophysical, genomic and culture-based studies have indicated that bacteriochlorophyll-containing AAnP bacteria are broadly distributed throughout the world's oceans, both taxonomically and geo-

graphically. Together, these genome-enabled analyses of proteorhodopsin and AAnP marine bacteria are changing our views about the nature and prevalence of light-utilization strategies in ocean surface waters.

Methane oxidizing pathways

In deep-sea marine sediments near continental margins, large quantities of methane are stored in reservoirs of solid gas hydrates that congeal in low-temperature, high-pressure environments. The fact that little of this methane escapes into the overlying water column has been a geochemical mystery for some time. Recently, it has been discovered that sediment-dwelling deep-sea microorganisms can consume this methane anaerobically and couple this methane oxidation to sulphate reduction. Although no anaerobic methane-oxidizing microorganisms have been cultivated so far, rRNA surveys (Hinrichs et al. 1999) have revealed the presence of methanogen-related archaea that are responsible for this process. Consortia of these archaeal methanotrophs, along with sulphate-reducing bacteria, can oxidize methane to CO_2 with concomitant sulphate reduction in these deep-sea, methane-rich habitats. Genomic DNA from methanotroph-enriched fractions of the environmental samples was sequenced and analysed with a specific focus on the putative methane metabolism. An entire path of genes homologues of known genes in the methanogenesis pathway mapped specifically to the environmental sequences, supporting a previously hypothesized mechanisms of archaeal methane oxidation. Genes encoding proteins involved in all but one of the seven steps of the methanogenic pathway were found, only one central gene in the methane-metabolizing pathway (mer, methylenetetrahydromethanopterin reductse), which has been found in all other methanogens examined to date, seemed to be missing. The mer gene encodes an enzyme that catalyses a key reductive step in the methanogenic pathway. Alteration of this step could regulate the directional flux of carbon for the environmental archaeal methanotrophs in the oxidative direction. Additionally, the absence of this key gene in the deep sea group might indicate that archaeal methanotrophs have lost the ability to generate methane and are now committed to a methane-consuming lifestyle.

New and essential metabolic strategies

The different environmental challenges within the marine environment have led to different microbes adopting different survival and growth strategies. For example, the photosynthetic pelagic bacterium *Prochlorococcus marinus* MED4 has adopted a minimalist approach, with conditions within its environmental niche varying little, the bacterium has reduced its genome size to 1.66 Mbp to gain a competitive advantage (Dufresne et al. 2005); while *Pelagibacter ubique* from the SAR11 clade of marine alphaproteobacteria has no transposons, extrachromosomal elements, pseudogenes or introns and constitutes the smallest genome known for a free living microorganism, encoding the smallest number of predicted open reading frames (Giovannoni et al. 2005).

Adaptation to the challenges of the deep sea, where nutrient sources can be unpredictable has led to greater adaptability in both signalling pathways involved in the detection of nutrients and greater metabolic capabilities in utilizing them. Likewise, microorganisms from the extremes within the marine environment have adapted to the extremes of tem-

perature and pressure, resulting in novel biochemistry which is able to operate at high or low temperatures and pressures, perhaps uniquely suited for many industrial processes.

Life in the deep-sea can also occur entirely independently of sunlight driven photosynthesis. At hydrothermal vents, for example, chemolithoautotrophic bacteria are the primary producers for these unusual ecosystems, with archaea isolated from these environments able to grow at 121°C. Photosynthetic bacteria entirely dependent on the geothermal radiation from these vents have also been isolated (Beatty et al. 2005). Cold seeps, in which hydrocarbon rich fluids seep from the ocean floor, are also hotspots of diversity, in which the primary producers are chemosynthetic microorganisms (Jorgensen et al. 2007). These discoveries at the extremes of ocean life have led to a significant extension of the known biosphere. The ability of microbial life to thrive throughout the entire marine environment, from deep ocean vents to surface sea ice, is entirely due to the diverse and adaptable biochemistry encoded within their genetic resources.

Functional metagenomics: screening nature for useful functions

For functional metagenomics an environmental sample is collected and then total community DNA is extracted. The isolated DNA is used to generate a metagenomic library that is used to transform a suitable host strain, usually *Escherichia coli*, and individual clones can then be screened for the presence of enzymatic or other bioactivities encoded by the environmental DNA fragment. When coupled with a robust and high-throughput screen, this method is an extremely effective way of isolating novel enzymes from otherwise inaccessible microbes. Examples of enzymes isolated from marine sources using a functional metagenomics approach include esterases, lipases, chitinases, amylases and amidases. The functional based approach has been successfully applied to terrestrial environments, especially soil, with the discovery of new genes for antibiotics, antibiotic resistance and industrial enzymes, while the full potential of marine functional metagenomics has yet to be exploited. Only about 0.0001–0.1% of the bacteria found in water environments are believed to be culturable, making a culture independent approach to harvest the metabolic potential of these organisms very attractive. In addition, data from the Global Ocean Sampling (GOS) expedition indicates that despite current large-scale sequencing efforts the rate of discovery of new protein families from the marine environment is linear, implying that marine microorganisms will continue to be a source of novel enzymes in the foreseeable future (Yooseph et al. 2007). As many of these gene products are entirely novel, their activity cannot be inferred from comparison to known protein databases, thus a functional metagenomics approach has the ability to identify novel genes on the basis of phenotypes which lend themselves to high throughput screens. Despite the undoubted promise of functional metagenomics for the discovery of new enzymes, this approach is limited at the moment by the ability of metagenomic clones to produce active enzymes. Many functional metagenomic approaches rely on the use of *E. coli* as a host for the expression of metagenome encoded proteins. While quite a large number of genes derived from *Enterobacteriaceae* will be readily expressed in the most common *E. coli* host, many genes from more distantly related organisms may not be expressed due to the promoter regions of these genes not being recognized by the *E. coli* transcriptional machinery or be expressed at low levels due to differences in codon usage, for example. Even where transcription and translation of foreign genes results in efficient protein ex-

pression, additional problems can arise when proteins need to be post-translationally modified or exported for activity. For these reasons, the availability of suitable heterologous expression hosts remains a barrier to extracting the maximum information from functional metagenomic analyses. On the other hand, due to advanced screening methodologies and the use of robotic instrumentation, it is now possible to screen large clone libraries for functional activities in a high throughput fashion, in relatively short timescales (Kennedy et al. 2008).

The quest for biofuels fuels genome sequencing

In addition to the gain of information produced by metagenomic approaches, the capability for fast and cheap sequencing of genomes and transcriptomes of biotechnologically relevant microorganisms has encouraged the production of more and more useful data.

Cultivable microorganisms are known that produce a variety of potential energy sources including hydrogen, methane, butanol, lipids and even electric current. By carefully manipulating the availability of nutrients and other environmental conditions, ways can be developed to capture and store the useful byproducts of these microbial cultures. Nevertheless much more needs to be known about how these microorganisms function before we can control and channel the energy sources they produce. Genomics offers an important means of achieving this knowledge and should be considered a strategic approach to addressing the challenge of producing renewable energy.

A list of microbial genome projects completed during 2008 that were encouraged by interest for their biotechnological applications has recently been published (Galperin 2008) and underlies how the pressure for biofuel production is driving basic research toward the investigation of poorly studied microorganisms that presents some useful characteristics. The list includes: marine member of *Bacteroidetes*, that present a combination of heterotrophic metabolism with light energy capture by proteorhodopsin; a number of genomes of green sulfur bacteria, anoxygenic phototrophs that live in strictly anaerobic sulfide-rich environments and are a potential source of biomass for biofuels; the soft-rot ascomycete fungus *Trichoderma reesei*, that is widely used in biotechnology as a producer of various cellulases and hemicellulases and has attracted renewed interest owing to its potential use in the conversion of lignocelluloses to biofuel; two thermophilic chemolitoautotrophs, isolated from hot springs at Yellowstone National Park at 60–75°C and capable of growing in microaerophilic conditions by using reduced sulphur compounds and/or hydrogen as electron acceptors and CO_2 as the source of carbon.

The interest in our project is focused on the extraordinary capacity of *N.gaditana* to grow efficiently in a broad range of light intensities and accumulate lipid droplets in response to environmental signals. To understand the metabolic characteristics of this microalga and to elucidate how the carbon flux through the different possible pathways is controlled we obtained the whole genome and the transcriptome sequences.

From genes discovery to artificial life

During a very inspired talk given in 2005 at TEDglobal, Craig Venter, that was just taking a brake from his round the world metagenomic expedition, made a point about exploiting the knowledge accumulated by sequencing and annotating genomes for making artificial life. The incredible discoveries accumulated during the first metagenomic surveys, when

new protein families and new functions were identified, encouraged researchers to look for even more diversity as an inspiration for the most different biotechnological applications, ranging from biocatalysis, to hydrogen or biofuel production or environmental remediation. As so many different genes were found to enriching the tree of life, an important issue become actual again: is there a smaller set of genes that might sustain a simple form of life? And which are those genes? In order to find an answer to this question and be able to design synthetic microbial organisms able to fulfil the tasks of society, the Craig Venter Institute started what was an adventure that culminated with the realization of the first bacterial cell controlled completely by a 'synthetic genome'. Where the 'synthetic genome' was large fragments of DNA assembled in yeast. Even though assembled in a biological system, still the genes included in the assembled chromosome were artificially designed and were able to sustain a simple form of life. This finding, probably still far from providing a definitive answer to this fundamental question, was nevertheless very encouraging for the production of synthetic life and led Craig Venter to envision a future "Combinatorial Genomics" based on these new synthesis capabilities, the vast gene array repertoires and the homologous recombination. According to Venter using a powerful screening system as with all biology, selection can be achieved and any kind of desired function can be realized. The first tasks? Trying to modify photosynthesis to produce hydrogen directly from sunlight using an oxygen-insensitive hydrogenase; combining cellulases, to break down complex sugars into simple sugars and do ethanol fermentation in the same cell; and more recently engineering photosynthetic bacteria to expel lipid chains and harvest biodiesel from the medium. Unfortunately the task was not quite easy since metabolism seems to be more complicated than expected and not yet completely understood but research in this moment has got powerful instruments to explore many different paths!

References

DeLong E F (2005) Microbial community genomics in the ocean *Nature Reviews Microbiology* 3, 459-469.

Béjà O *et al.* (2000) Bacterial rhodopsin: evidence for a new type of phototrophy in the sea. *Science* 289, 1902–1906.

Béjà O, Spudich E N, Spudich J L, Leclerc M, DeLong E F (2001) Proteorhodopsin phototrophy in the ocean. *Nature* 411, 786–789.

Venter J C *et al.* (2004) Environmental genome shotgun sequencing of the Sargasso Sea. *Science* 304, 66–74.

Kelemen B R, Du M, Jensen R B (2003) Proteorhodopsin in living colour: diversity of spectral properties within living bacterial cells. *Biochim. Biophys. Acta* 1618, 25–32.

Béjà, O. *et al.* (2002) Unsuspected diversity among marine aerobic anoxygenic phototrophs. *Nature* 415, 630–633.

Hinrichs K U, Hayes J M, Sylva S P, Brewer P G & DeLong E F (1999) Methane-consuming archaebacteria in marine sediments. *Nature* 398, 802–805.

Costa V, Angelini C, De Feis I, Ciccodicola A. (2010) Uncovering the complexity of transcriptomes with RNA-Seq J Biomed Biotechnol. 2010:853916. Epub 2010 J.

Venter J C (2005) "DNA and the sea" Pubblc lecture at TEDglobal, Edimburg.

Venter J C (2008) "On the verge of creating synthetic life" Pubblc lecture at TED2008, Monterey, California.

Galperin M Y (2008)The quest for biofuels fuels genome sequencing Environmental Microbiology Volume 10, Issue 10, pages 2471–2475.

Dufresne A, Garczarek L, Partensky F (2005) Accelerated evolution associated with genome reduction in a free-living prokaryote. *Genome Biol. 6*, R14.

Giovannoni S J, Tripp H J, Givan S, Podar M, Vergin K L, Baptista D, Bibbs L, Eads J, Richardson T H, Noordewier M, Rappe M S, Short J M, Carrington J C, Mathur E J (2005) Genome streamlining in a cosmopolitan oceanic bacterium. *Science 309*, 1242–1245.

Beatty J T, Overmann J, Lince M T, Manske A K, Lang, A S, Blankenship R E, Van Dover C L, Martinson T A, Plumley F G (2005) An obligately photosynthetic bacterial anaerobe from a deep-sea hydrothermal vent. *Proc. Natl. Acad. Sci. USA 102*, 9306–9310.*Mar. Drugd.*

Jorgensen B B, Boetius A (2007) Feast and famine--microbial life in the deep-sea bed. *Nat. Rev. Microbiol.* 5, 770–781.

Yooseph S, Sutton G, Rusch D B, Halpern A L, Williamson S J, Remington K, Eisen J A, Heidelberg K B, Manning G, Li W, Jaroszewski L, Cieplak P, Miller C S, Li H, Mashiyama S T, Joachimiak M P, van Belle C, Chandonia J M, Soergel D A, Zhai Y, Natarajan K, Lee S, Raphael B J, Bafna V, Friedman R, Brenner S E, Godzik A, Eisenberg D, Dixon J E, Taylor S S, Strausberg R L, Frazier M, Venter J C (2007) The Sorcerer II Global Ocean Sampling expedition: expanding the universe of protein families. *PLoS Biol.* 5, e16.

Kennedy J, Marchesi J R, Dobson A D (2008) Marine metagenomics: strategies for the discovery of novel enzymes with biotechnological applications from marine environments. *Microb. Cell Fact. 7*, 27.

Kennedy J, Flemer B, Jackson S A, Lejon D P, Morrissey J P, O'Gara F, Dobson H D (2010) Marine Metagenomics: New Tools for the Study and Exploitation of Marine Microbial Metabolism *Mar. Drugs 8*, 608-628.

Pernthaler, A. et al. (2008) Diverse syntrophic partnerships from deep-sea methane vents revealed by direct cell capture and metagenomics. Proc. Natl. Acad. Sci. USA 105, 7052–7057.

1.3. Quest for high efficiency solar energy converters

About 2,500,000 exajoules (EJ) of solar energy reach the Earth's surface every year (Ritchie 2010). Since world energy consumption was estimated as less than 500 EJ per year (values for 2009), two hours of sun light would be more than enough to satisfy one year of human energy demand. This represents an enormous and easily available resource that humans indeed have tried to harness since ancient times, using a range of ever evolving technologies. Nevertheless the achieved capacity for solar energy conversion is still very poor and represents a limit for the direct use of solar radiation to power human activities. Solar radiation, along with secondary solar powered resources such as wind and wave power, hydroelectricity and biomass, account for most of the available renewable energy on Earth, while all kind of fossil fuels can be considered 'concentrated solar energy', since they were produced by photosynthetic conversion of solar energy through a very long period and stored in the form of oil, gas or coal by natural events. Therefore all the energy that we use, both for sustaining ourselves and for powering our activities, comes actually from the sun. The very crucial point though, is that the vast majority of the energy that we use was converted to a usable form by natural photosynthesis and not by human technology. Although photosynthesis is an amazing machinery perfected by natural selection over many centuries for solar energy conversion, it did not evolve to be efficient and its final yield is only a few percent (3%, Grobbellar 2009). If we want to directly exploit the potential of the sun to meet the actual energy demand, we need to be able to design a conversion system whose efficiency is far higher than that of photosynthesis, or to be able to manipulate the metabolism of photosynthetic organisms in order to yield the maximum possible amount of useful molecules out of their culturing. Active solar technologies, designed to convert, store and distribute solar energy, had their early development in the 1860s, driven by the industrial demand and the expectation that coal would soon become scarce. However development of solar technologies stagnated in the early 20th century in the face of the increasing availability, economy, and utility of coal and petroleum. Since then, interest for solar technologies has encountered up and down periods always following the trends of fossil fuels in the market. Around 1990s, global warming concern has moved a great part of the public opinion, official Governments and the lot of the scientific community toward a renewed interest for solar technologies. The solar excitement powered by global warming has led to an incredible increase in scientific and technological knowledge in the past few years, has involved an important and growing fraction of the economy and has heavily involved, for the first time, biotechnology in the quest for advantageous energy converters.

The global warming issue

Changes in climate might have significant implications for present lives, for future generations and for ecosystems on which humanity depends. Consequently, climate change has been and continues to be the subject of intensive scientific research and public debate. According to the preeminent scientific committees, there is strong evidence that the warming of the Earth over the last half-century has been caused largely by human activity, such as the burning of fossil fuels and changes in land use, including agriculture and deforestation. Science of course has never had to do with agreement but rather with repro-

ducible experiments, nevertheless the study of global climate change has still lot to deal with data interpretation and correlations that require specific technical expertise to be interpreted. The prediction of future temperature and other aspects of climate change, especially at the regional scale, are still subject to great uncertainty. This short introduction to the problem attempt to summarize the current scientific evidence on climate change and its drivers. This paragraph attempt to include in the discussion well established statements together with assumptions on which there is wide consensus but continuing debate, and those where there remains substantial uncertainty. The impacts of climate change, as distinct from the causes, are not considered here. This report draws mainly upon recent evidences and builds on the "Fourth Assessment Report of Working Group I of the Intergovernmental Panel on Climate Change (IPCC)", published in 2007, and in "Advancing the Science of Climate Change" issued by the National Research Council in 2010, but it takes also into great account the main scientific critics moved to the documents by a number of scientist working on climate research.

Climate and climate change

The Sun is the primary source of energy for the Earth's climate. Satellite observations show that about 30% of the Sun's energy that reaches the Earth is reflected back to space by clouds, gases and small particles in the atmosphere, and by the Earth's surface. The remainder, about 240 Watts per square metre, when averaged over the planet, is absorbed by the atmosphere and the surface. To balance the absorption of 240 Wm^{-2} from the Sun, the Earth's surface and atmosphere must emit the same amount of energy into space; they do so as infrared radiation. Some of this energy is actually trapped in the atmosphere and reflected back. The Earth's surface is thus kept warmer than it otherwise would be because, in addition to the energy it receives from the Sun, it also receives infrared energy emitted by the atmosphere. The warming that results from this infrared energy is known as the greenhouse effect. Measurements from the surface, research aircraft and satellites, together with laboratory observations and calculations, show that, in addition to clouds, the two gases making the largest contribution to the greenhouse effect are water vapour followed by carbon dioxide (CO_2). There are smaller contributions from many other gases including ozone, methane, nitrous oxide and human-made gases such as CFCs (chlorofuorocarbons). According to the current understanding, we can think of climate changes on a global scale, as driven by processes that either modify the amount of energy absorbed from the Sun, or the amount of infrared energy emitted to space. Climate change can therefore be initiated by changes in the energy received from the Sun, changes in the amounts or characteristics of greenhouse gases, particles and clouds, or changes in the reflectivity of the Earth's surface. The imbalance between the absorbed and emitted radiation that results from these changes is generally referred as "climate forcing". In principle, changes in climate on a wide range of timescales can also arise from variations within the climate system due to, for example, interactions between the oceans and the atmosphere. Such internal variability can occur because the climate is an example of a chaotic system: one that can exhibit complex unpredictable internal variations even in the absence of the climate forcings. There is very strong evidence to indicate that climate change has occurred on a wide range of different timescales from decades to many millions of years; human activity is just a relatively recent addition to the list of potential

causes of climate change. The shifts between glacial and interglacial periods over the past few million years for example are thought to have been a response to changes in the characteristics of the Earth's orbit around the Sun. While these led to only small changes in the total energy received from the Sun, they led to significant changes in its geographical and seasonal distribution. The large changes in climate, in moving in and out of glacial periods, provide evidence of the sensitivity of climate to changes in the Earth's energy balance. Once a climate forcing mechanism has initiated a climate response, this climate change can lead to further changes: for example, in response to a warming, the amount of water vapour is expected to increase, the extent of snow and ice is expected to decrease, and the amount and properties of clouds could also change. Such changes can further modify the amount of energy absorbed from the Sun, or the amount of energy emitted by the Earth and its atmosphere, and lead to either a reduction or amplification of climate change. The nature of the climate system is determined by interactions between the moving atmosphere and oceans, the land surface, the living world and the frozen world. The rate at which heat is moved from the surface to the ocean depths is an important factor in determining the speed at which climate can change in response to climate forcing. Since variations in climate can result from both climate forcing and internal climate variability, the detection of forced climate change in observations is not always straightforward. Furthermore, the detection of climate change in observations, beyond the expected internal climate variability, doesn't necessary implicate the attribution of that change as a real climate anomaly. Current understanding of the physics (and increasingly the chemistry and biology) of the climate system is represented in a mathematical form in climate models, which are used to simulate past climate and provide projections of possible future climate change. Climate models are also used to provide quantitative estimates to assist the attribution of observed climate change to a particular cause or causes. In the arbitrary and scarce predictivity of these models until now lays the great part of the critics moved to the popular theory of "human caused global warming".

Evidences for the actual temperature to be assigned as an event of anomalous climate change

Measurements of surface temperature around the world are available starting from 1850. Analyses of these data, in a number of institutes, try to take into account changing distributions of measurements, changing observation techniques, and changing surroundings of observing stations (e.g. some stations become more urban with time, which can make measurements from them less representative of wider areas). The results of this analysis show that averaged over the globe, the surface has warmed by about 0.8°C (with an uncertainty of about ±0.2°C) since 1850 (IPCC fourth report, 2007). This warming has not been gradual, but has been largely concentrated in two periods, from around 1910 to around 1940 and from around 1975 to around 2000. The warming periods are found in three independent temperature records over land, over sea and in ocean surface water. Even within these warming periods there has been considerable year-to-year variability. The warming has also not been geographically uniform. It is argued by a number of experts (Carter, 2007) that the historical temperature record reconstructed and reported in Figure 1.1 is not valuable since it is affected by a number of systematic uncertainties and therefore

changes of less than 1°C per century may not exceed the true error bars of the true average temperature estimates.

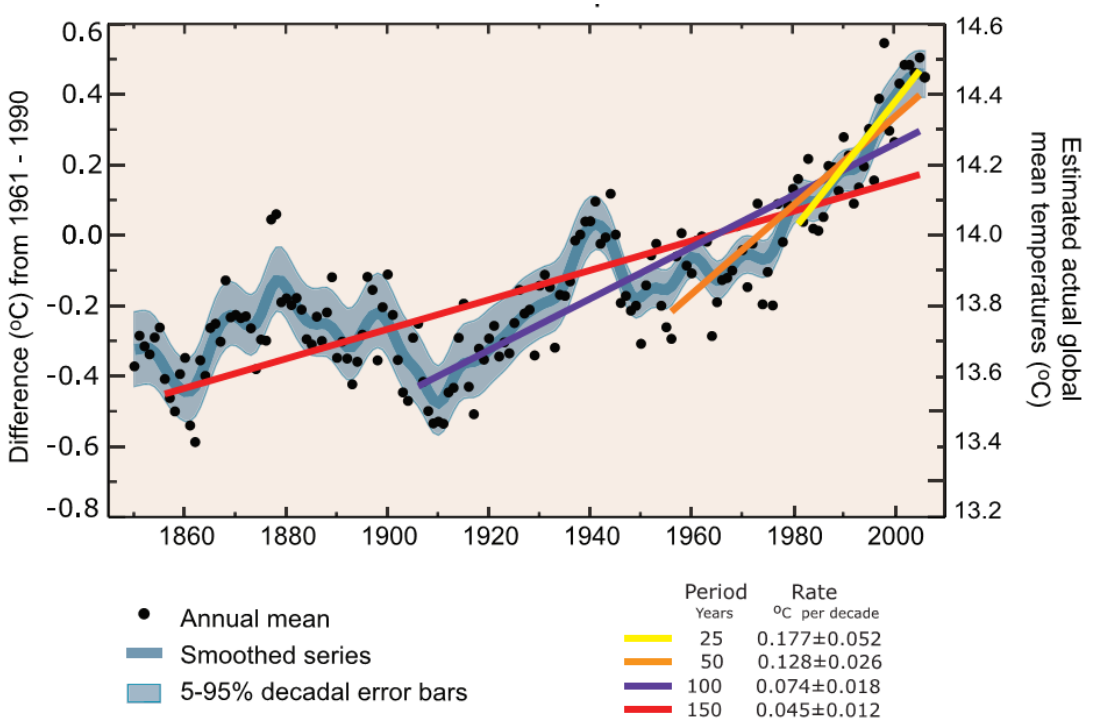


Figure 1.1 Annual global mean observed temperatures (black dots) along with simple fits to the data. The left hand axis shows anomalies relative to the 1961 to 1990 average and the right hand axis shows the estimated actual temperature (°C). Linear trend fits to the last 25 (yellow), 50 (orange), 100 (purple) and 150 years (red) are shown, and correspond to 1981 to 2005, 1956 to 2005, 1906 to 2005, and 1856 to 2005, respectively. Note that for shorter recent periods, the slope is greater, indicating accelerated warming. The blue curve is a smoothed depiction to capture the decadal variations. To give an idea of whether the fluctuations are meaningful, decadal 5% to 95% (light grey) error ranges about that line are given (accordingly, annual values do exceed those limits). From the IPCC fourth report of working group I (2007), chapter 3.

Other aspects of the climate were parallel observed in the past decades that were said to provide much evidence of climate change consistent with the surface temperature changes. This includes increases in the average temperature of both the upper 700m of the ocean and the troposphere, widespread (though not universal) decreases in the length of mountain glaciers and increases in average sea level. There has been an overall decline in the area covered by sea-ice floating on the Arctic Ocean over the past 30 years while there has been a small increase in the area covered by sea-ice around Antarctica.

In any case, as local temperatures are generally a poor guide to global conditions, so are observed variations in global temperature over a period of just a few years, and could be a misleading guide to underlying longer-term trends in global temperature, since evidenced were collected for substantial temperature variations all over the Earth's history. Comparison of figure 1.1, 1.2 and 1.3 gives a good example of this consideration.

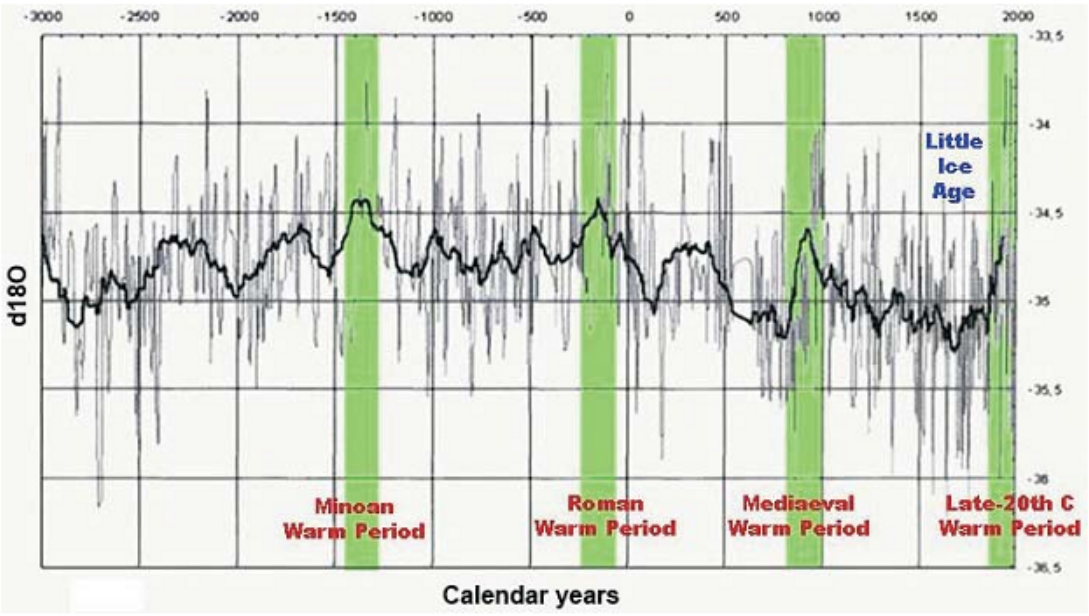


Figure 1.2 Oxygen isotope time series of the 5000 years before 2000. From Carter R M 2007.

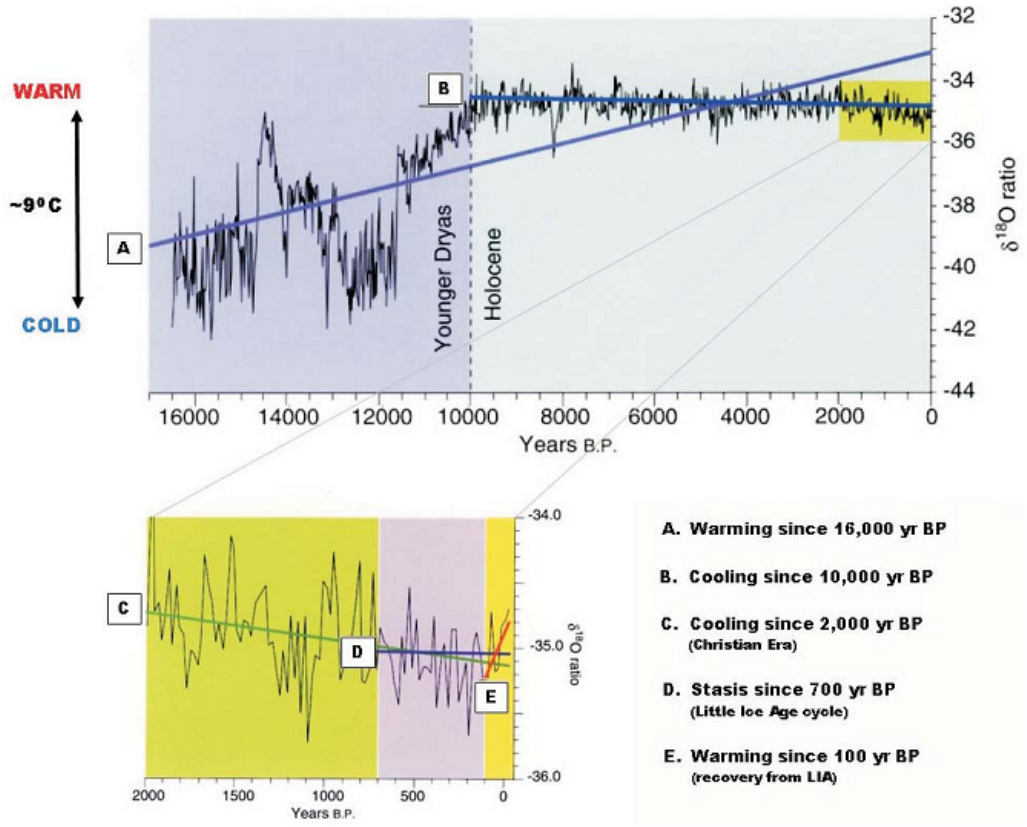


Figure 1.3 Climatic cycling over the last 16000 years indicated by averaged 20-year oxygen isotope ratios from the GISP2 Greenland ice core. Trend lines A-E all extended up to the end of the 20th century, fitted through the data for the last 16000, 10000, 2000, 700 and 100 years respectively. From Carter R M 2007.

As it is shown in figure 1.3 different time intervals can lead to different line trend fits. The same consideration is true for longer time-scale (Figure 1.4) and we need therefore to be very careful prior to attribute a certain trend to an actual climate anomaly. For this porpoise, palaeoclimatic records in chronological order over interannual to millennial time scales have been used in order to examine how the climate system varies and changes over the different time scales and to understand the contributions that lower-frequency

patterns of climate change might make in influencing higher-frequency patterns of variability and change.

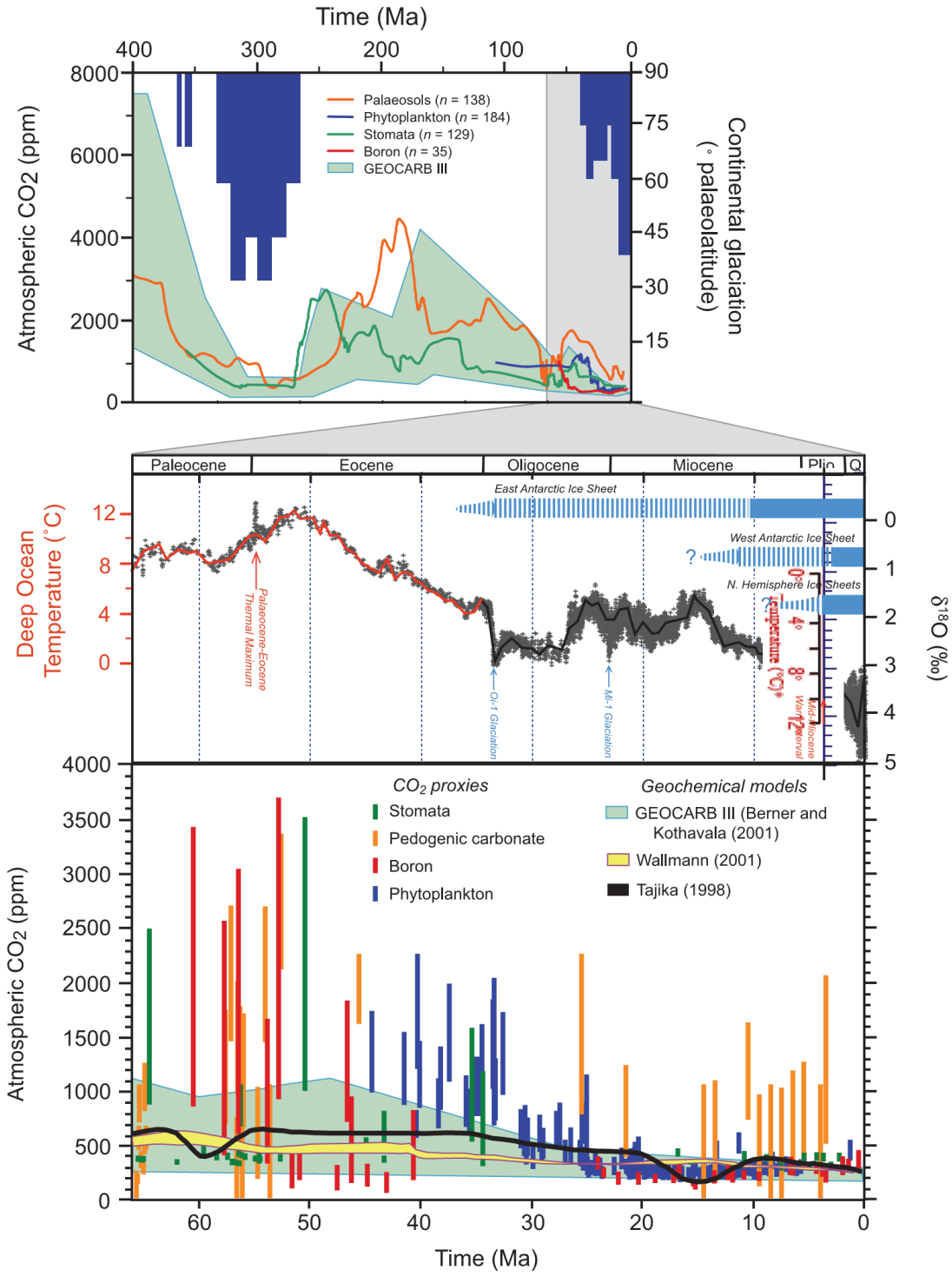


Figure 1.4 (Top) Atmospheric CO_2 and continental glaciation 400 Ma to present. Vertical blue bars mark the timing and palaeolatitudinal extent of ice sheets. CO2 records of the four major proxies are plotted together with data obtained using the geochemical carbon cycle model GEOCARB III. (Middle) Temperatures of the deep ocean as obtained from a global compilation of deep-sea benthic foraminifera 18O isotope records. (Bottom) Detailed record of CO2 for the last 65 Myr from the different proxy methods and from three geochemical carbon cycle models. From the IPCC fourth report of working group I (2007), chapter 6.

As shown in figure 1.4 data obtained using multiple geochemical and biological methods are usually integrated to obtain more reliable plots of temperature and atmospheric composition over the considered time scale. Pre-Quaternary climates prior to 2.6 Ma were

mostly warmer than today and associated with higher CO_2 levels. Looking back in time beyond the reach of ice cores, that is, prior to about 1 Ma, data on greenhouse gas concentrations in the atmosphere become much more uncertain. Of special interest is the reconstructions of the warm climates over the past 65 Million. Four primary proxies were used to estimate pre-Quaternary CO_2 levels. While there is a wide range of reconstructed CO_2 values, magnitudes are generally higher than the interglacial, pre-industrial values seen in ice core data. Changes in CO_2 on these long time scales are thought to be driven by changes in tectonic processes (e.g., volcanic activity source and silicate weathering drawdown). The actual era seems to be one of the relatively cool periods since complex life has been on Earth. If the relationship between CO_2 concentration in the atmosphere and the global temperature is confirmed to follow the popular model of human caused global warming, the actual increases in CO_2 emission could drive a parallel increase of temperature up to the ancient values. It is interesting to note that the warm Paleocene period coincides with the explosion of life immediately following the mass extinction event at the end of the Cretaceous while during the following and still warm Eocene the emerging of the mammals is registered. A proper comparison is inappropriate in this case since global biology and geography were increasingly different further back in time. The sole consideration that we are allowed to make is that such a dramatic change in global temperature will produce for sure deep changes in actual scenarios, nevertheless there are no reason to state that these changes will arm human survival or the development of biodiversity. The relationship between CO_2 and temperature can be traced further back in time as indicated in Figure x (top Panel), which shows that the warmth of the Mesozoic Era (230–65 Ma) was likely associated with high levels of CO_2 and that the major glaciations around 300 Ma likely coincided with low CO_2 concentrations relative to surrounding periods. These considerations led us of course to a fundamental point: what should we infer about the relationship between CO_2 concentration in the atmosphere and global temperature?

Evidences for accumulation of CO_2 due to human activity to be the cause of global temperature increase

Global-average CO_2 concentrations have been observed to increase from levels of around 280 parts per million (ppm) in the mid-19th century to around 388 ppm by the end of 2009. CO_2 concentrations can be measured in "ancient air" trapped in bubbles in ice, deep below the surface in Antarctica and Greenland; these show that present-day concentrations are higher than any that have been observed in the past 800,000 years, when CO_2 varied between about 180 and 300 ppm. Various lines of evidence point strongly to human activity being the main reason for the recent increase, mainly due to the burning of fossil fuels (coal, oil, gas) with smaller contributions from land-use changes and cement manufacture. The evidence includes the consistency between calculations of the emitted CO_2 and that expected to have accumulated in the atmosphere, the analysis of the proportions of different CO_2 isotopes, and the amount of oxygen in the air. These observations show that about half of the CO_2 emitted by human activity since the industrial revolution has remained in the atmosphere. The remainder has been taken up by the oceans, soils and plants although the exact amount going to each of these individually is less well known. Concentrations of many other greenhouse gases have increased. The concentration of me-

thane has more than doubled in the past 150 years; this recent and rapid increase is unprecedented in the 800,000 year record and evidence strongly suggests that it arises mainly as a result of human activity. The plot of CO_2 concentration over time compared with that of temperature anomalies shows an interesting correlation. In this case again we can look up at the measures in different time scales. If we look at the time interval between the first industrial revolution and the present, where the increase in CO_2 concentration is largely attributed to human activity, we can clearly spot an interval between 1940 and 1975 were a relevant increase in CO_2 emission is not accompanied by a parallel rise in temperatures. This has been one of the leading arguments against the largely accepted theory of man made global warming. The data included in the graph are also presented in the Fourth Assessment Report by the IPCC, where it is concluded that "human activities have also caused increased concentrations of fine reflective particles, or 'aerosols', in the atmosphere, particularly during the 1950s and 1960s". As a result "during the 1950s and 1960s, average global temperatures levelled off, as increased on aerosol from fossil fuels and other sources cooled the planet", and the "rapid warming observed since the 1970s has occurred in a period when the increase in greenhouse gases has dominated all over other factors". The IPCC Fourth Assessment Report also pointed out that global "aerosol forcing appear to have decreased after 1980".

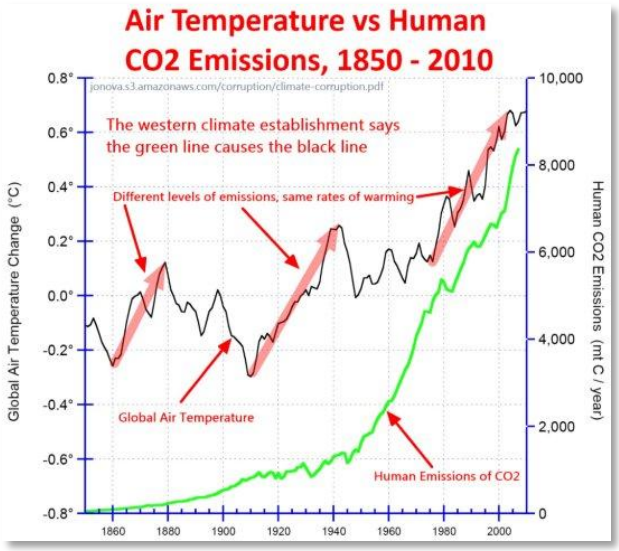


Figure 1.5 The estimate curve of CO_2 concentration in the atmosphere (green line) is plotted without error bars on top of the graph relative to temperature anomalies in the past one and a half century. From Evans 2010.

Palaeoclimatic records document a sequence of glacial-interglacial cycles covering the last 740.000 years in ice cores. The last 430.000 years, which are the best documented, are characterised by 100.000 years glacial-interglacial cycles. A minor proportion (20% on average) of each glacial-interglacial cycle was spent in the warm interglacial mode, which normally lasted for 10.000 to 30.000 years (Figure 1.6). The Holocene, the latest of these interglacials, extends to the present. The ice core record indicates that greenhouse gases co-varied with Antarctic temperature over glacial-interglacial cycles, suggesting a close link between natural atmospheric greenhouse gas variations and temperature. A simple co-variation of two variables does not mean that one is causing the modification of the other one, nevertheless this suggest a possible correlation that is worth to be tested. High-

resolution ice core records of temperature proxies and CO_2 during deglaciation indicates that Antarctic temperature starts to rise several hundred years before CO_2 (IPCC Fourth Assessment Report, 2007). Nevertheless Caillon and coauthors highlighted on a paper, published on Science on March 2003, a lag of 800 years between the temperature rise and the atmospheric CO_2 increase in the period around 240.000 years ago. This argument has been widely used to argue that CO_2 increase might cause temperature variation. The paper concludes that fluctuations in the Earth's orbit initiated the increase in surface temperatures in Antarctica and was followed by a gradual warming of the oceans, which caused substantial release of CO_2. The paper also indicates that the carbon dioxide released added to the atmosphere and contributed to the deglaciation of the northern Hemisphere. Evidences from ice cores might thus indicate an active role for CO_2 in the climate system: changes in CO_2 can lead to climate change and climate change can also alter the concentrations of CO_2, since the amount of carbon held in oceans, soils and plants depends on temperature and other conditions.

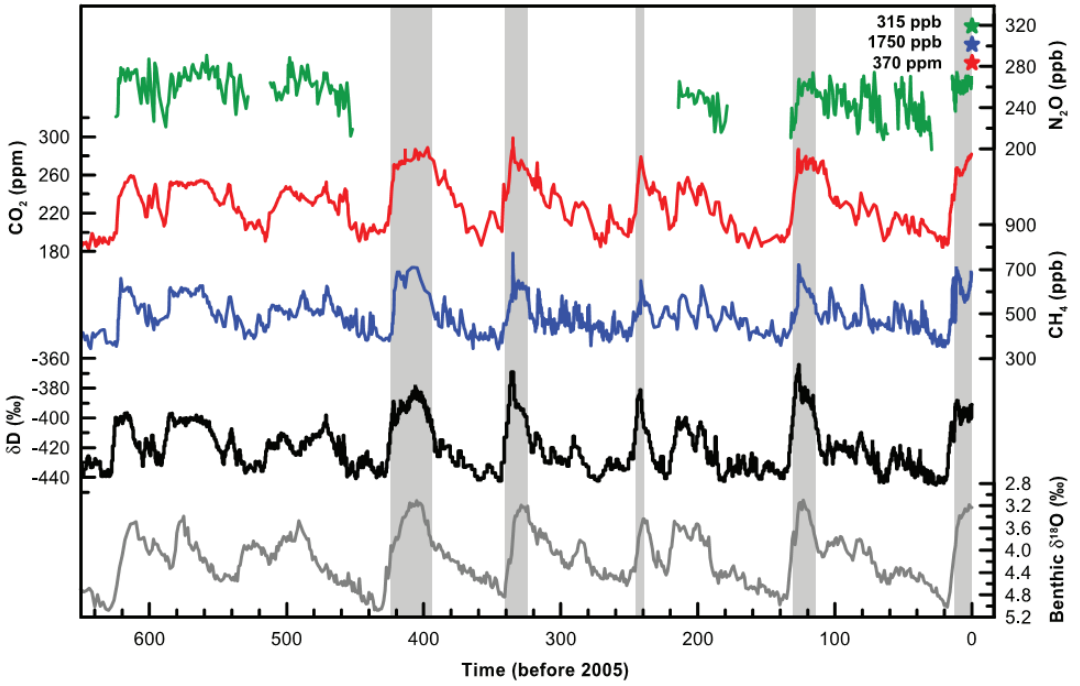


Figure 1.6 Plot of variations of deuterium (δD; black), a proxy for local temperature, and the atmospheric concentrations of the greenhouse gases CO_2 (red), CH_4 (blue), and nitrous oxide (N_2O; green) derived from air trapped within ice cores from Antarctica and from recent atmospheric measurement. From the IPCC fourth report of working group I (2007), chapter 6.

It is my opinion that the all of this indications might provide a number of strong evidences for the construction of a coherent model of the actual observed climate, but do not provide an incontrovertible prove of human emission of CO_2 being the cause of a 'Global Warming'. As Nathan Lewis used to conclude in the presentation become very popular around 2007, "being scientist we have only one way to answer this question, take the chance to make a big experiment: drop CO_2 emission, and see what happens!".

Biomimetic and biotechnological approaches to solar energy conversion

Artificial photosynthesis

Direct conversion of sunlight energy to electricity can be achieved through the use of solar cells. Most popular and widely used solar cells are those based on semiconducting silicon. Solar to electrical conversion efficiency, as high as 24% can be obtained, but the requirement of pure single crystalline form of silicon renders them very expensive. Thin film solar cells made of less expensive materials also allow light energy conversion up to around 15–18% efficiency. Yet another form of photovoltaic solar cell based on dye sensitization is emerging as a even less expensive and still efficient alternative. The actual devises can generates electric power from light without suffering any permanent chemical transformation and yield a conversion efficiency over 11.5%.

Light powered splitting of water to its constituent elements as molecular gases is another important mean for transformation of solar energy into useful products and is exactly what photosynthesis does. Total decomposition of water to H_2 and O_2 both in natural and in artificial systems can be considered into two parts: first, a photochemical or electrochemical component where the required oxidizing or reducing equivalents are generated and a second stage where suitable redox catalysts assist formation of the molecular gases. Most of the efforts till date still revolve around this second component, identifying suitable redox catalysts. Electrolysis of water, for example could be best achieved using a platinum electrode as cathode and a metal oxide such as ruthenium oxide as the anode, but the system is incredibly expensive. Many procedures have been recently evolved based on ruthenium with variable yields and costs (Kalyanasundaram and Graetzel 2010). Nocera and coworkers recently demonstrated (Kanan and Nocera 2008) the functioning of an efficient water-oxidation catalyst formed in situ providing a feasible alternative. Electrolysis of water using an indium tin oxide electrode was examined in aqueous solution in the presence of cobalt and potassium phosphate. Upon applying a voltage to the electrode, cobalt, potassium, and phosphate accumulated on the electrode, forming the catalyst. The catalyst oxidizes water to form oxygen gas and free hydrogen ions.

In addition to photoinduced H_2 production from water, there has been sustained interest to find viable means of reducing CO_2 gas to other C-1 products such as alcohols and aldehydes. Interest for the conversion of carbon dioxide into usable hydrocarbon fuels has actually moved the industrial interest for over a century, and indeed is due to Sabatier the discovery of the reaction that catalyses the reduction of CO_2 to methane in the presence of nickel at high temperature and high pressure. Efforts were made in this recent years to improve this process and use solar radiation as source of energy for the reduction. Meyer and co-workers have summarized (Sutin et al. 1997) the results of their studies of CO_2 reduction using four classes of transition-metal catalysts: metal tetraazamacrocyclic compounds; supramolecular complexes; metalloporphyrins and related metallomacrocycles; $Re(CO)_3(bpy)X$-based compounds where bpy = 2,2o - bipyridine. Carbon monoxide and formate were the primary CO_2 reduction products.

The goal of all the studies attempting artificial photosynthesis is to find alternate means of producing electric power and high-energy fuels from water and abundant and free solar energy. The described strategies are by no mean a reproduction of every single aspect of natural photosynthesis but are rather aimed to mimic the most important features of the

process. None of these systems has achieved to date a realistic result for large scale applications, nevertheless the scientific knowledge in these past few years, since extensive research in this field has been carried on, has made incredible exciting progresses. The all of these studies provided us with a deeper understanding of the most fundamental features that characterize the natural solar energy conversion and promise very encouraging progresses for future applications.

Biofuels

Biofuels include liquid fuels and various biogases, which are in someway derived from cultivable photosynthetic organism and represent a relevant renewable resource since most of the transportation relies on liquid fuels. Vehicles indeed usually require high energy density, as occurs in liquids and solids, and are powered by efficient internal combustion engines that require clean burning fuels to work properly. The fuels that are easiest to burn cleanly are typically liquids and gases. Thus liquids (and gases that can be stored in liquid form) meet the requirements of being both portable and clean burning. Also, liquids and gases can be pumped, which means handling is easily mechanized, and thus less laborious. First generation biofuels are made from sugars, starch and vegetable oils derived from crops or woods. Bioethanol is the most widely produced biofuel and is obtained by the action of microorganisms and enzymes through the fermentation of sugars, starches, or, with more difficulties, cellulose. Ethanol can be used in petrol engines as a replacement for gasoline or it can be mixed with gasoline to any percentage. Green diesel, is a form of diesel fuel which is derived from renewable feedstock rather than the fossil feedstock used for common diesel, and is obtained by traditional fractional distillation of oils. Chemically different is the process trough which biodiesel is produced, from vegetable oils, using transesterification. Recent research gave a relevant contribution to the improvement of biofuel industry by screening and testing non food crops for oil extraction and by implementing cellulose fermentation. Cellulose and lignocellulose indeed are very abundant and cannot be fermented using the traditional procedures unless they are broke down to glucose molecules. In many laboratories various experimental procedures are being developed to release fermentable sugars for ethanol production. Still prizes of renewable fuels are not yet competitive with those of petrol oil and the energy balance for biofuel production leaves some uncertainties. The majority of the recent scientific publications state that the future will be cultivation of genetically modified microalgae for biofuel production.

Engineering of microalgae

Substantial interest is currently devoted by research in utilizing eukaryotic microalgae for the renewable production of several bioenergy carriers, including starches for alcohols, lipids for diesel fuel surrogates, and H_2 for fuel cells. Relative to terrestrial plants, microalgae are more efficient at converting sunlight into chemical energy, and require a smaller footprint and less water for cultivation. Many species of algae thrive in salt water, are productive for the all year and in diverse conditions, and do not accumulate recalcitrant lignocellulosic biomass. While nutrient manipulation approaches have been used to date to enhance the net production of the cultures, research now aims to boost the productivity by developing technologies and model systems for metabolic engineering. Genetic ma-

nipulation techniques have been developed for some species, and are increasingly being applied to optimize biofuel production in several algal systems.

The integration of metabolic pathways is coordinated through complex mechanisms that distribute photosynthetic output into synthesis of proteins, nucleic acids, carbohydrates, lipids, and H_2. A comprehensive understanding of the biosynthesis and degradation of precursors, intermediates, and metabolic end products, and the identification of the regulatory networks that control metabolic flux is central for the design of engineering strategies for optimizing biofuel production. Several recent studies have used 'omics'-based strategies to begin unravelling the regulation and integration of these networks (Miller et al. 2010; Nguyen et al. 2008). The insights gained from these studies and the discovery of novel proteins that are potentially better suited for bioenergy applications are providing promising targets for genetic manipulation to enhance the accumulation of bioenergy carriers. Unfortunately not many genomes of microalgae have been sequenced to date and are available for studying, therefore the majority of the works on proteomics, gene expression and mutagenesis are carried out using the model system *Chlamidomonas reinhardti*. Sequencing of other species of microalgae that have interesting characteristics with respect to energy reach biomolecules accumulation or metabolic regulation, might provide important insights into the key enzymes responsible for the regulation of carbon energy flux and might prove important for the design of optimized strains for large scale bioenergy production.

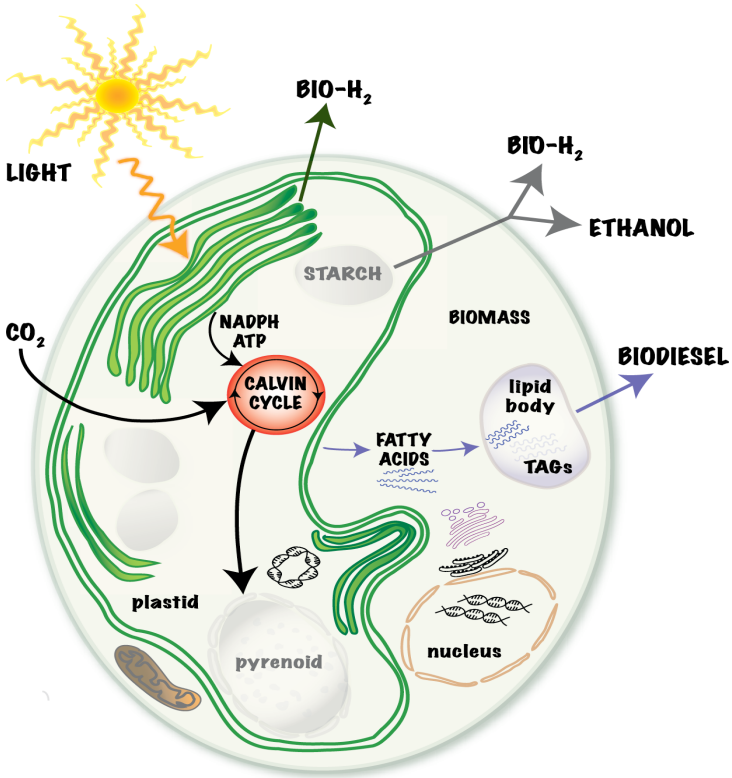


Figure 1.7 Synthetic Draw of the main metabolic pathways leading to fuel production form microalgae. From Berr at al. 2009.

As illustrated in figure 1.7 unicellular algae are capable of synthesizing a range of biofuels. Lipids and carbohydrates represent the main energy storage molecules. A broad understanding of primary metabolism is necessary to manipulate electron flux toward these products or H_2 for bioenergy applications. Complicating these efforts are the distinct met-

abolic processes that occur within algal organelles and the numerous enzyme isoforms present in these cells. These contribute to the complexity of algal metabolism and increase the need to better characterize the systems on a biochemical level.

Improving fuel yield by manipulating photosynthesis

Photosynthetic efficiency is directly related to the size of light-harvesting antennae complexes (LHC). Antennae system controls the amount of energy that reaches the reaction centres in response to energy input variation by channellingß or dissipating the absorbed energy. This system evolved to guarantee a high flexibility in a rapidly changing environment, it was not optimized for high density growth in photobioreactors where it is indeed responsible for much of the energy dispersion. Reduction of the antennae size might enhance the overall photosynthetic efficiency of a culture in a production plant. Random insertion libraries were used to identify mutants with a reduced antenna size in *C.reinhardtii* (Tetali et al 2007).

Improving lipid biosynthesis

Much of our current knowledge on fatty acid biosynthetic enzymes is inferred from genome databases and relatively few studies on algal lipids biosynthesis have been published (Riekhof and Benning 2009). Fatty acid synthesis occurs in the plastid of plants before translocation to the cytoplasm for further assembly into diacylglycerols and triacylglycerols. Many enzymes involved in lipid's biosynthesis are encoded by single genes and are thought to also be targeted to the mitochondria where fatty acids precursors are required to produce essential cofactors for mitochondrial enzyme activity. Under nitrogen deplete conditions, some green algae accumulate high levels of triacylglicerols, and phosphorus and sulphur deprivation induce the conversion of membrane phospholipids to neutral lipids. The regulatory mechanisms in these systems are poorly understood. We can imagine a number of strategies to push lipid's biosynthesis toward more useful yields: we could mutagenize or overexpress the enzymes responsible for the committed step in lipid's biosynthesis; availability of precursor molecules could be some how enhanced; and finally downregulation of catabolic or competing pathways might determine an increased accumulation of fatty acids. It is also determining for industrial scale plants to constitutively enhance lipid's biosynthesis to avoid culturing in nutrient deprived conditions. In order to face this task, the molecular switches responsible for sensing nutrient deprivation and activate the stress response must be identified and mutagenized. In addition, lipids isolated from microalgae are variable and frequently composed of triacyilglicerols and polyunsaturated fatty acids that are prone to undesirable oxidation reactions affecting downstream biofuel applications. Thus, in order to improve lipids processing and downstream usage, saturation profiles could be altered through the introduction or regulation of desaturases and fatty acids chain length can be optimized using thioesterases (Radackovitz et al. 2010). The complexity of lipid metabolism in algae is illustrated by recent large-scale mutant screening in a *C. reinhardtii* insertional library, which identified 80 mutants with altered FAS activity (Beer at al. 2009).

Enhancing H_2 evolution

The metabolic flexibility of some photosynthetic microalgae enables them to survive periods of anaerobiosis in the light by developing a particular photofermentative metabolism.

The latter entails compounds of the photosynthetic electron transfer chain and an oxygen-sensitive hydrogenase in order to reoxidize reducing equivalents and to generate ATP for maintaining basal metabolic function. This pathway results in the photoevolution of hydrogen gas by the algae. A decade ago, Melis and coworkers managed to reproduce such a condition in a laboratory context by depletion of sulphur in the algal culture media, making the photoevolution by the algae sustainable for several days (Melis et al. 2000). This observation boosted research in algal H_2 evolution. A feature, which due to its transient nature was long time considered as a curiosity of algal photosynthesis, suddenly became a phenomenon with biotechnological potential. Although the Melis procedure has not been developed into a biotechnological process of renewable H_2 generation so far, it has been a useful tool for studying microalgal metabolic and photosynthetic flexibility and testing future H_2 production procedures. Most of the critical steps and limitations of H_2 production by sulphur deprivation have been studied in the model organism *Chlamydomonas reinhardtii*, by introducing various changes in culture conditions and making use of mutants issued from different screens or by reverse genomic approaches. The enzyme responsible for hydrogen evolution is HydA hydrogenase, that requires anaerobiosis and reducing equivalents (either light or starch) to work. Sulphur deprivation produces anaerobiosis by impairing photosynthesis and thus ceasing energy conversion. Optimization of H_2 photoproduction will require identification of O_2-tolerant hydrogenases able to work while photosynthesis is still productive. One approach to address this problem is gene shuffling, which has been used to generate a diverse recombinant hydrogenase library to screen for enhanced O_2 tolerance or stability (Nagy et al. 2007). A more recent strategy has been the search of natural diversity for a more suitable HydA isoform (Boyd et al. 2009). DNA extracted from microbial mats that inhabit saline environments and that are exposed to supersaturating concentrations of O_2 during peak photosynthesis, contained a diversity of deduced HydA amino acid sequences, resulting in a near doubling of the known diversity of this protein encoding gene. Further ecological screening for new H2 producing algal species will be likely carried on in the near future in different environments using molecular approaches combined with hydrogen production tests.

References

Ritchie R J (2010) Modelling photosynthetically active radiation and maximum potential gross photosynthesis. Photosynthetica, in press.

Grobbelaar J U (2009) Upper limits of photosynthetic productivity and problems of scaling. J Appl Phycol 21:519-522.

Miller R, Wu G, Deshpande R R, Vieler A, Gärtner K, Li X, Moellering E R, Zäuner S, Cornish AJ, Liu B, Bullard B, Sears BB, Kuo M H, Hegg E L, Shachar-Hill Y, Shiu S H, Benning C (2010) Changes in transcript abundance in Chlamydomonas reinhardtii following nitrogen deprivation predict diversion of metabolism Plant Physiol. 154(4):1737-52. Epub 2010 Oct 8.

Nguyen A V, Thomas-Hall S R, Malnoë A, Timmins M, Mussgnug J H, Rupprecht J, Kruse O, Hankamer B, Schenk P M. (2008) Transcriptome for photobiological hydrogen production induced by sulfur deprivation in the green alga Chlamydomonas reinhardtii. Eukaryot Cell. 7(11): 1965-79. Epub 2008 Aug 15.

Riekhof W R, Benning C (2009) Glycerolipid biosynthesis. In The Chlamydomonas Sourcebook Organellar and Metabolic Processes, edn 2. Edited by Stern DB. In The Chlamydomonas Sourcebook, Vol. 2. Edited by Harris EE. The Chlamydomonas Sourcebook. Vol. 2Academic Press; 41-68.

Radakovits R, Jinkerson R E, Darzins A, Posewitz M C (2010) Genetic engineering of algae for enhanced biofuel production. Eukaryot Cell. 9(4): 486-501. Epub 2010 Feb 5. Review.

Radakovits R, Eduafo P M, Posewitz M C (2011) Genetic engineering of fatty acid chain length in Phaeodactylum tricornutum. Metab Eng. 13(1): 89-95. Epub 2010 Oct 27.

Nagy L E, Meuser J E, Plummer S, Seibert M, Ghirardi M L, King P W, Ahmann D, Posewitz M C (2009) Application of gene-shuffling for the rapid generation of novel [FeFe]-hydrogenase libraries. Biotechnol Lett 2007, 29:421-430.

Boyd E S, Spear J R, Peters J W: [FeFe]-hydrogenase genetic diversity provides insight into molecular adaptation in a saline microbial mat community. Appl Environ Microbiol 75(13): 4620-3.

Kalyanasundaram K and Graetzel M (2010) Artificial photosynthesis: biomimetic approaches to solar energy conversion and storage Current Opinion in Biotechnology 21: 298–310.

Kanan M W and Nocera D G (2008) In Situ Formation of an Oxygen-Evolving Catalyst in Neutral Water Containing Phosphate and Co2+ Science 321: 1072-1075.

Sutin N, Creutz C, Fujita E (1997) Comments Inorg Chem, 19: 67-92. A comprehensive look at various electrochemical and photochemical approaches to reduction of CO_2 to useful fuels.

Durkin M (2007) “The great global warming swindle” BBC documentary

Gore A and Guggenheim D (2006) “An inconvenient truth” Paramount Classics

Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change (2007) Cambridge University Press

The Royal Society (2005) “A guide to facts and fictions about climate change” The Royal Society Press

Ward B (2007) “Five major misrepresentation of the scientific evidence in the DVD version of ‘The great global warming swindle’ ”

National Academy of sciences (2010) “Informing an effective response to climate change” The National Academy Press

Carter R (2007) “The mite of dangerous human caused climate change”

Carter R (2010) “Climate: the counter-consensus” Indipendent Minds

Carter R (2007) “Climate change - is CO_2 the cause?” Australian public lecture

Oreskes N (2007) “The American denial of global warming” public lecture in Perspectives in ocean science

Lewis N (2007) “Powering the Planet” public lecture at “Solar Energy and Artificial Photosynthesis” hosted by Royal Society, London

Moore T (2007) “Global Warming and Artificial Photosynthesis” public lecture at “Solar Energy and Artificial Photosynthesis” hosted by Royal Society, London

Royal Society (2010) “Climate change: a summary of the science” Royal Society Press

1.4. Project outline

Nannochloropsis is a genus comprised of very small (less than 5µm) coccoid unicells and is known primarily from the marine environment although many fresh water species have been identified (Fawley and Fawley 2007). The majority of the species that have been described cannot be discriminated by either light or electron microscopy, and were assigned primarily by DNA sequencing of ribosomal or RBCL genes. The interphase cells contain a single yellow-green parietal plastid surrounded by four membranes. It is widely accepted that such plastids originate from secondary symbiogenesis (Cavalier-Smith 2003).

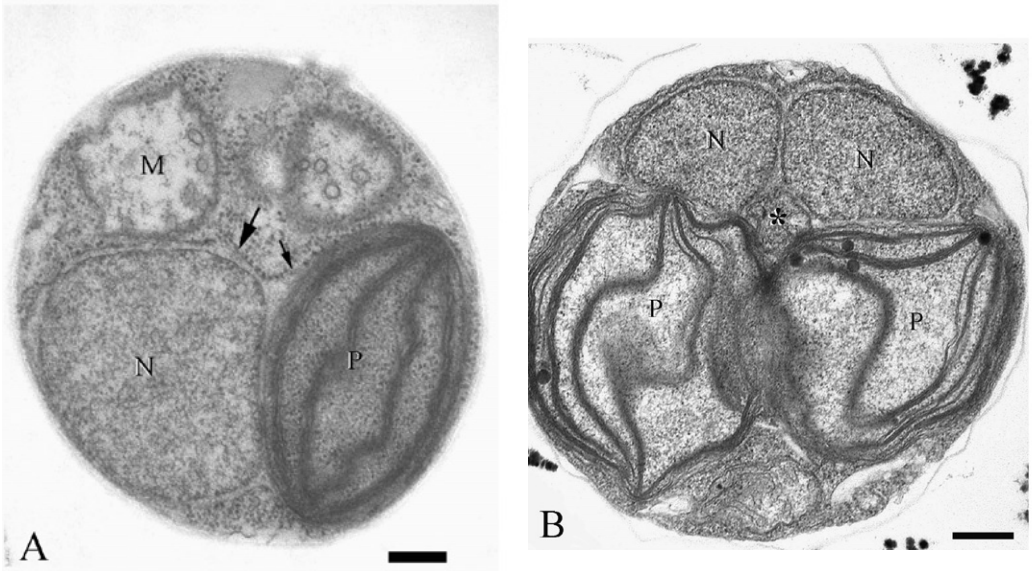


Figure 1.8 Ultrastructure of *Nannochloropsis oculata* from Murakami et al. 2009. A. An interphase cell. A continuous of membranes (indicated by the arrows) forms a sac enclosing the nucleus (N) and the single plastid (P). M: mitochondria. B. A telophase cell. Condensed chromatin areas appear within the duplicated nuclei. Two daughter plastids visible. Bars represent 200 nm.

Some isolates of this genus are important food organisms for aquaculture because they reproduce very rapidly and possess fatty acids that are not found in other types of phytoplankton (Apt and Behrens 1999). These species were also suggested by several research groups as a promising source of oil for biofuel production (Boussiba et al. 1987; Hodgson et al. 1997; Rodolfi et al. 2009) due to high content of saturated and monounsaturated fatty acids accumulated within triacylglycerols under stressful conditions.

It is a widespread notion in the literature that many species of microalgae are able to accumulate relevant amount of lipids in response to nitrogen starvation. Amongst the marine algae that seem to yield the highest amounts of lipids per culture volume in nitrogen deprivation are different species of Nannochloropsis (Rodolfi et al. 2009). Boussiba and coworkers were able to obtain already in 1987 cultures of Nannochloropsis salina able to yield lipids accumulation up to 70% of their overall dry biomass in laboratory conditions (Boussiba et al 1987). Cultures were reported to have a different lipid content in the different stages of growth and the maximum accumulation was reached in the stationary phase. Moreover a continuous increase in the lipid content was observed during the all stationary phase. Interpretation of the results obtained under outdoor conditions was a bit more complicated since nitrogen induced lipid accumulation

resulted different in the different seasons. Although temperature was proved to considerably affect the lipid content, other environmental factors could not be excluded. As reported in the paper, other groups indeed observed variable results and further studies suggested that the reported increase in lipid content caused by N starvation could have been affected by a combination of several factors, such as light intensity, pH, age of the culture and CO_2 concentration, which may have not been always under perfect control. The process of lipid biosynthesis and accumulation seems to be controlled by various factors in a complicated way that is not yet completely understood. Despite the publication of several works in the past decade reporting measures of cellular lipid content in various species of *Nannochloropsis* grown under different laboratory and outdoor conditions, we did not found in the literature any further experiment aimed to unravel the complicated controls responsible for this phenomenon at the molecular level. While many different protocols were tested for the manipulation of the culturing conditions in order to increase the net amount of lipids produced per culture volume, nothing was done in *Nannochloropsis* to identify the enzymes and the molecular switches involved in lipid biosynthesis in order design informed metabolic engineering of either *Nannochloropsis* or other model organisms of algal growth. In recent years commercial interest for a number of compounds, mainly lipids, that can be extracted from microalgae has led to examination of a growing number of, often unusual, algae. A major trend in these researches has been the identification of the various proteins involved in the production of very long-chain polyunsaturated fatty acids such as arachidonic, eicosapentaenoic and docosahexaenoic acids (Guschina and Harwood 2006). Many works also report information about the environmental factors, such as light, temperature or minerals that led to the accumulation of those products. Still none of those studies describes the genetic characteristic of the species able to accumulate the major amount of lipids and to suggest which are the enzymes involved in the translation of the environmental stresses into lipid synthesis.

Being *Nannochloropsis* an organism of special interest for the study of the metabolic processes that led to the extraordinary lipid accumulation observed in microalgae we decided to apply second generation sequencing technologies in order to obtain the whole genome sequence of the species *Nannochloropsis gaditana*. The genomic sequence and the gene content of *Nannochloropsis* are fundamental information for comparative genomics in order to characterize the metabolic profile of the organism and identifying the special features responsible for its adaptation.

As already mentioned *Nannochloropsis,* as the other Eustigmatophytes, is thought to have originated after two consecutive endosymbiotic events, therefore, first the genome of a photosynthetic bacteria had to reach an equilibrium with the nucleus of the host strain through a process of gene transfer and then, afterwards, both the nuclear and the plastidial genome and to come to an equilibrium with the new host. *Nannochloropsis* would be, to our knowledge, the first secondary endosymbiont that lakes the nucleomorph, to be sequenced to date. Defining the genetic content of both the two genomes of *Nannochloropsis* will help elucidate the loss of genetic information in the genome of the endosymbiont following the secondary endosymbiotic event and the exchange of genetic information among the genomes.

To bring even further our analysis of this organism and characterize its gene content, we decided to exploit the potential of the SOLiD system for producing profiles of the gene expression in different conditions. Due to the enormous amount of short reads that can be obtained in every single run it is possible to literally count the number of occurrence that align to a certain region of the genome, obtaining a precise annotation of the gene structure and a quantification of the expression at the level of single exon and with a high dynamic range. We therefore decided to grow *Nannochloropsis* in nutrient sufficient and nitrogen starvation media and to prepare the samples for SOLiD sequencing using two parallel approaches in order to recover the maximum possible information: we used oligodT coated beads for capturing the polyAmRNA codified by the nucleus; and we also attempted a protocol for ribosomal RNA subtraction in order to sequence all the expressed genes codified in the nucleus as well as in the chloroplast, allowing at the same moment the recovery of non coding RNAs. Developing an understanding of coordinate expression of genes encoded on the nuclear and plastidial genomes will increase our understanding of the roles of the various compartments in the cellular processes, the communications between the different genetic compartments of the cell, and the ways in which proteins and metabolites are exchanged among these compartments. This information is of special interest in our case since the initial steps of fatty acids biosynthesis carried out in the chloroplast are followed by translocation of the precursors in the *citosol* for further assembly. In addition to careful annotation of the intron exon profile and transcription start of the expressed genes in the different samples, it is also very important to identify the enzymes that are differentially expressed in conditions leading to lipids accumulation or not. To this aim we tried to reproduce the growth and lipid curves of *Nannochloropsis gaditana* in nutrient starvation and nutrient sufficient medium and we extracted the RNA for sequencing. Expression profiles were compared in order to obtain a complete picture of the differences between the cultures lacking nitrogen and able to accumulate lipids and the cultures in optimal growth conditions at the same stage. This information will help to hypothesis the list of the genes involved in signal transduction and metabolic pathways regulation and will open the way to metabolic engineering of *Nannochloropsis* and other interesting microalgae for enhancing lipid production.

In order to provide the scientific community with the necessary back ground information on *Nannochloropsis* genome to enable further studies at the molecular level, we plan to publish on the web a user friendly interface to browse the genome, the annotated genes and the data obtained in the different experiments of RNA sequencing.

References

Cavalier-Smith T (2003) Genomic reduction and evolution of novel genetic membranes and protein-targeting machinery in eukaryote—eukaryote chimaeras (meta-algae). Philos Trans R Soc Lond B 358: 109—134

Murakami R and Hashimoto H (2009) Unusual Nuclear Division in *Nannochloropsis oculata* (Eustigmatophyceae, Heterokonta) which May Ensure Faithful Transmission of Secondary Plastids. Protist, Vol. 160, 41—49

Fawley K P and Fawley M V (2007) Observations on the Diversity and Ecology of Freshwater *Nannochloropsis* (Eustigmatophyceae), with Descriptions of New Taxa Protist, Vol. 158, 325—336

Apt K E, Behrens P W (1999) Commercial developments in microalgal biotechnology. J Phycol 35: 215—226

Boussiba S, Vonshak A, Cohen Z, Avissar Y, Richmond A (1987) Lipid and biomass production by the halotolerant microalga Nannochloropsis salina. Biomass 12:37–47

Hodgson P, Henderson R, Sargent J, Leftley J (1991) Patterns of variation in the lipid class and fatty acid composition of Nannochloropsis oculata (Eustigmatophyceae) during batch culture. J Appl Phycol 3:169–181

Rodolfi L, Zittelli G C, Bassi N, Padovani G, Biondi N, Bonini G, Tredici M R (2009) Microalgae for oil: strain selection, induction of lipid synthesis and outdoor mass cultivation in a low-cost photobioreactor. Biotechnol Bioeng 102:100–112

Guschina I A and Harwood J L (2006) Lipids and lipid metabolism in eukaryotic algae Progress in Lipid Research 45 160–186

2. Materials and methods

All the experimental procedures used for the production of data about *Nannochloropsis* are reported in great detail in this chapter, in order to be reproduced in any laboratory. A few of the described protocols do not refer to any of the experiments reported in the results. In those cases the results produced were of no relevance for the discussion of the results, nevertheless the procedures are reported in the list of methods since they could be useful for future applications. In the majority of the cases indeed the reported procedures were obtained after careful testing and optimization of the experimental conditions specifically for *Nannochloropsis*.

2.1. Microalgae strains and Propagation

General culturing conditions

Microalgal strains and growth conditions

Microalgal species routinely used in this work are listed in Table 2.1. Strains reported were selected after screening of a number of samples obtained from different culture collections. Samples were examined for their capability for fast growth in the laboratory and for the presence of contamination. Species were also selected for which literature already existed concerning the levels of lipid biosynthesis.

Strain	Reference/Supplier
Nannochloropsis salina	WoRMS (2009). *Nannochloropsis salina* D.J. Hibberd, 1981 /obtained from SAG culture collection, strain number 40.85
Nannochloropsis gaditana	WoRMS (2009). *Nannochloropsis gaditana* Lubián, 1982 /obtained from Cristian Gomis (Biofuel Systems, S.L.) after 3years propagation from the original strain obtained from the Oceaographic Centre of Mazarrón (Murcia) and isolated in San Fernando de Cádiz lagoon
Nannochloropsis oculata	*Nannochloropsis oculata* (Droop) Hibberd (Botanical Journal of the Linnean Society, 82: 93-119, 1981 / obtained from CSIRO collection, strain number CS-189
Nannochloropsis limnetica	obtained from SAG culture collection, strain number 18.99

Table 2.1 Nannochloropsis species and strain references.

Propagation of algal cultures

Microalgae were cultured in f/2 medium (Sea Salts 32g/l (Sigma Aldrich, Italy); 1X Guillard's (F/2) marine water enrichment (Sigma Aldrich, Italy)) with the addition of 17mM Sodium Nitrate. Were indicated antibiotics were added to the culture to prevent contamination (see Table 2.2 for details). In order to keep pH stable during growth 40mM Tris pH=8 was also added to the medium (Rocha et al., 2003). Medium was sterile filtered using disposable pressure filter units 0.2 µm pore size (Sartorius, Italy). Culture were grown at room temperature in gentle agitation (~ 50rpm in orbital shaker) under continuous light irradiance ~100µE. Propagation in agar plates was obtained using the same medium composition with the addition of 8g/L plant agar (Duchefa Biochemichemie BV, Micropoli, Italy). A sea salt solution 2 times concentrated containing already the buffer and adjusted to pH 8.0 was prepared in 500ml mQ water and sterilized by filtering. Solution was subsequently mixed with autocleaved plant agar 8 grams in 500ml. This strategy yielded a clear solution. It was important to set light irradiance between 50µE and 20µE for optimal growth.

Antibiotics	Optimal concentration	Combinations	Supplier
Kanamycin	50 µg/ml	1	Sigma Aldrich, Italy
Streptomycin	100 µg/ml	1	Sigma Aldrich, Italy
Chloramphenicol	50 µg/ml	1	Sigma Aldrich, Italy
Erithromycin	100 µg/ml		Sigma Aldrich, Italy
Zeocin ™	50 µg/ml		Duchefa Biochemichemie BV, Micropoli, Italy
Gentamicin	50 µg/ml		Sigma Aldrich, Italy
Hygromycin	30 µg/ml		Sigma Aldrich, Italy
Paromomycin	100 µg/ml	1	Sigma Aldrich, Italy

Table 2.2 Antibiotics used for propagation either in liquid medium or in plates. Antibiotics were used singularly or in combinations where indicated. Concentration reported were the optimal ones, obtained as Minimum Inhibitory Concentration (*MIC*) for the contaminants grown in LB medium and maximum non toxic amount for algal cultures in f/2.

Obtaining of axenic cultures of Nannochloropsis *gaditana*

Fresh microalgae in active growth were streaked in agar f/2 to obtain single separate colonies. A number of isolated colonies was peaked from the plate and diluted in 100µl liquid medium to yield single starting cultures. Medium contained 100µg/ml streptomycin, 50µg/ml kanamycin and 50µg/ml chloramphenicol. Once the liquid cultures had reached a light green colour they were progressively diluted to 10ml in fresh f/2, always in the presence of antibiotics. 10µl from each of the cultures were plated in: f/2 agar medium; f/2 agar plus single antibiotics; standard LB; LB plus single antibiotics. This procedure allowed to check for the presence of contaminants and to verify if those contaminants were sensitive to the antibiotics tested. Cultures were selected that gave rise to homogeneous green colonies in f/2 and yielded completely clean LB plates after 3-4days. Starting liquid culture were propagated in liquid f/2 in the presence of antibiotics, occasionally streaking a sample in both LB and agar f/2 to check for contamination. Cultures were also repeatedly observed by optical and fluorescence microscopy to confirm for the presence of a unique or a prevalent cell type. The most resistant contaminants were grown in LB medium, total nucleic acid were extracted and DNA coding for ribosomal RNA was amplified by PCR using universal primers. Universal primers are suitable for amplification of rDNA from any source and indeed were designed for experiments of metagenomics. These primers were a kind gift of Dr Ivano Zara and Dr Riccardo Rosselli. Sequencing of 16S and 18S DNA allowed identification of the contaminant microrganisms and the design of a strategy to remove them from the cultures. Sensitivity to specific antibiotics was tested on plates both in the optimal medium for the contaminant and in the medium routinely used for propagation of the algae. Different antibiotic concentrations were tested in order to establish the minimum inhibitory concentration (MIC) when the information was not available. Smallscale liquid cultures of microalagae were also assessed for sensitivity to the antibiotics at different possible concentrations prior to set up the cultures for propagation and biomass production.

Primer name	Sequence
G18s1 for	CCTGCCAGTAGTCATACGCT
G18s1 rev	TTGGATGTGGTAGCCGTCTC
G18s2 for	GATTCCGGAGAGGGAGCCTG
G18s2 rev	TGCTTTCGCAGTAGTTCGTC
G18s3 for	CAGAGGTGAAATTCTTGGAT
G18s3 rev	CACCCATAGAATCAAGAAAG
G16s1 for	AATACCGCGTGGGGGATGAAGA
G16s1 rev	GCACCACCTGTAGAAGCGGAATAA

Table 2.3 List of primers used for amplification of the conserved regions of the genes coding for ribosomal RNA. Primers labelled with a G were specifically designed for *Nannochloropsis gaditana* based on the rDNA sequences deposited at the NCBI and the work of Andreoli et al. 1999.

Growth curves

Growth curves were set up in culturing flasks in a controlled room where temperature, humidity and agitation were fixed constants while light intensity and medium composition could vary. Aeration was guaranteed through a cotton lid.

A hemocytometer was used every day for determination of the number of cells per unit volume. Measures were made systematically once a day all along the growth period. When it proved necessary cell suspension was diluted before counting and final concentration was obtained multiplying by the dilution factor (i.e. cells number/ml = X_m x 10^4 x df).

Optical density at 750nm was also measured daily using a spectrophotometer (see dedicated section for details). A graphic was obtained for correlation between optical densities and cell counts for each of the examined species in the different intervals of values.

Relative content of chlorophyll *a* was estimated through the growth curve. Microalgae where collected allowing the culture through a disposable filtering unit (regenerated cellulose membrane, 0.2 μm pore size, Sartorius, Italy) and pigments were extracted soaking the filter in formaldehyde for at least 48 hours. Chlorophyll was quantified by spectrophotometry measurement as described in section xy. Chlorophyll content was normalized to OD_{750}.

Lipid content was also measured by checking NileRed fluorescent staining at different stages and photosynthetic activity was monitored at the different conditions by pulsed amplitude modulated chlorophyll *a* fluorescence (PAM) measurements. Growth curves were usually carried on for around 20 days, time needed to reach the steady phase.

References

Andreoli C, Bresciani E, Moro I, Scarabel L, La Rocca N, Dalla Valle L, Ghion F (1999) "A Survey on a Persistent Greenish Bloom in the Comacchio Lagoons (Ferrara, Italy)" Botanica Marina 42 (5)

2.2. Cell breakage for preparation of purified fractions

Protoplast generation

Protoplast were generated as previously described by Higashiyama and Yamada (1991) with minor variations. Cells were grown to late logarithmic phase, collected by centrifugation at 1000g for 5 minutes and resuspended in protoplast forming medium (0.1 M Sodium Citrate, 0.7 M Manitol, 60mM EDTA, 4% cellulase Onozuka R-10, 2% mecerozyme R-10, adjusted to pH 6.0 using HCl). Each 100ml of initial culture volume was roughly resuspended in 1ml of protoplast forming solution. Mixture was incubated at 37°C for 3 hours. After treatment with enzymes, protoplasts were collected by centrifugation for 5 min at 800g. Cells were then washed twice in 0.1 M Sodium Citrate, 0.7 M Manitol, 60mM EDTA, pH 6.0 to completely remove the digestion enzymes. Protoplast formation was then checked by fluorescent microscopy. Cell wall indeed could be selectively stained using a fluorescent brightener (FB28 from Sigma) and visualized using the "DAPI filter" of a fluorescent microscope. Protoplast therefore emitted only in the red channel, due to chlorophyll autofluorescence, while whole cell gave both the red chlorophyll signal and the blue signal of the stained cell wall. Samples were prepared for observation as described in the dedicated section.

Cell breakage by sonication

1g of cell pellet was resuspended in 10 ml of lysis buffer (20 mM Tris base, 1 mM EDTA and 5 mM Dithiothreitol, pH 8.0). Cell suspensions in 50 ml glass beakers were subjected to sonication. Thermal effects were minimised by placing the sonication sample in an ice bath. Ultrasound was applied using a probe-type sonicator (Cole-Parmer Ultrasonic Processor) with an operating frequency of 20 kHz. The 4 mm diameter probe was immersed about 5mm below the surface of the sample to be sonicated and subjected to 5 s on/ 5 s off pulses for 4 min at 20% amplitude for two times consecutively, chilling the sample in ice in between.

Cell breakage using Covaris technology

The Covaris acoustic transducer operates at 500 kHz with a wavelength of ~1 mm, unlike conventional sonicators which have a wavelength of ~100 mm. The Covaris device was set up in a cold bath at 4°C and de-gassed for 30 min before acoustic treatment was applied. 1g of *Nannochloropsis* cell pellet was resuspended in 10 ml of lysis buffer (20 mM Tris base, 1 mM EDTA and 5 mM Dithiothreitol, pH 8.0). The Covaris settings for lysis of a 3 ml cell suspension were *20% Duty Cycle*, *500 Cycles/ Burst* and *Power Tracking* mode. These settings were applied to 3 ml aliquots in glass vials and acoustic treatment was carried out over a time course of 1, 2, 5, 10, 15 and 20 min. 200 µL samples were transferred into eppendorfs at the different acoustic times. Three distinct experiments were set up at different Intensities : 10, 6, 2.

Mechanical breakage at low temperature

Cells were grown to late logarithmic phase, collected by centrifuging at 2325g for 8 minutes at 4°C and flash frozen in liquid nitrogen. Pellets were then transferred using a spatula into a prechilled mortar. Cell wall was broken mechanically by pestling the cell

suspension together with quartz powder (Sigma) in the mortar in the presence of liquid nitrogen. The frozen powder was then transferred to a sterile falkon type tube and extracted in the desired buffer. After extraction unbroken cells and the inert quartz powder were removed by centrifuging for 10 minutes at 9000xg.

References

Higashiyama T, Yamada T (1991) "Electrophoretic karyotyping and chromosomal gene mapping of Chlorella" Nucleic Acids Research, Vol. 19, No. 22, 6191-6195

Honjoh K, Suga K, Shinoliara F, Maruyama I, Mlyamoto T, Ilatano S, IIo M (2003) "Preparation of Protoplasts from Chlorella vulgaris K-73122 and Cell Wall Regeneration of Protoplasts from C. vulgaris K-73122 and C-27" J. Fac. Agr., Kyushu Univ., 47 (2), 257-266

Hiu J, Cheung Y (2005-2006) Evaluation of the Utility of Adaptive Focused Acoustics Within a Discovery Research Department.GSK Industrial Placement Report.

2.3. Pigments and lipids analysis

Nile Red assay for estimation of average lipid content per cell

Nile red has a peak excitation at 485nm. When excited at this or greater wavelengths fluorescence emission can be registered in the interval between 530 and 590nm. In particular around 560, peak increases sensibly when nile red is found in a lipophilic environment, and it is therefore commonly used to stain lipid droplets, also inside leaving cells. Density of the cultures was assessed either counting the cells by hemocytometer or measuring the absorbance OD_{750nm}. Cultures were diluted to 10^6cells/ml in mQ water. 10µl of nile red from a 0.5mg/ml stock in 100% acetone were added to 2ml of diluted cell suspension. Samples were incubated for 10 minutes at 37°C in the dark and then measured using a fluorometer (see dedicated section for major details). This system was robust for comparison of lipid content in different samples of the same algal species while it could not be considered a method for absolute measuring of lipid content.

Determination of chlorophyll *a* concentration

Chlorophyll *a* concentration in liquid cultures of *Nannochloropsis* was routinely estimated using the method of Lichtenthaler and Welburn (1983). Liquid cultures were centrifuged at 36.000xg for 5 minutes at room temperature, pellet was respuspended in a minimum volume and sonicated (see dedicated section for details). After sonication, suspension was centrifuged, and supernatant, containing the cellular soluble fraction, was decanted. Cell extracts were diluted 100 times in 100% methanol. Extraction was carried out for 5 minutes and samples were centrifuged for 2 minutes at maximum speed in a microcentrifuge. Supernatant was transferred to a disposable cuvette and measured with a Lambda Bio 40 UV/Visible spectrometer (Perkin Elmer) calibrated with 100% methanol between the wavelengths of 600 and 750 nm. The chlorophyll a content was then estimated using the formula: $(A_{666}-A_{750}) \times 12.61 =$ [chlorophyll *a*] in g ml^{-1}.

Alternatively, for estimation of chlorophyll *a* content in liquid cultures, cells were collected by centrifugation (microfuge, 36.000xg for 10 minutes) and pellet was resuspended in 100% dimethylformammide according to Moran and Porath (1980). Chlorophyll *a* was extracted for at least 48 hours at 2°C in the dark and samples were subsequently centrifuged (microfuge, 36.000xg for 5 minutes). The spectrophotomer was calibrated with 100% DMF. Chlorophyll a concentration was obtained according to the following (Porra *et al.*, 1989):

[chlorophyll *a*] in µg/ml = $12 \times A_{664} - 3{,}11 \times A_{647}$.

Determination of carotenoid concentration

Aliquots of the cultures were spun down and the pellet obtained was freeze-dried and resuspended in 2.5 ml of 80% methanol to extract pigments. Chlorophyll and total carotenoids concentrations in the supernatant were determined spectrophotometrically using the equations proposed by Wellburn:

Chlorophyll a (µg/ml) = 12.21 (A663) - 2.81 (A646)

Chlorophyll b (µg/ml) = 20.13 (A646) - 5.03 (A663)

Carotenoids (µg/ml) = (1000A470 - 3.27[chl a] - 104[chl b])/227

Spectroscopic techniques

Pulsed amplitude modulated chlorophyll *a* fluorescence measurement (PAM)

DUAL PAM 100(Waltz) was used to obtain information about the functionality of PSII in different moments of the culture growth and in different growth conditions. We focused on the photosynthetic parameters F_o and F_v/F_m. Samples were dark adapted for 30 minutes before measuring. Two protocols were routinely applied: Slow Kinetics or SP Kinetics to calculate F_o and F_v/F_m and Light Curve to get the ETR (electron transport rate) and the NPQ (non photochemical quenching). All the measures were performed using a Measuring Light of 42 µE, an Actinic Light of 660 µE and a Fluorescence Saturating Pulse of 6000 µE. The Far Red Light was always switched on to avoid state transitions.

Absorbance spectroscopy

All the absorption measurements were recorded at room temperature using a double beam spectrophotometer (Lambda Bio 40 UV/Visible spectrometer, Perkin Elmer). Quartz cuvettes were routinely used for all porpoises.

Fluorescence spectroscopy

Fluorometric measurements were carried out using a DM 45 Spectrofluorimeter (Olis) at room temperature. Plastic cuvettes were routinely used.

References

Moran R, Porath D (1980) Chlorophyll Determination in Intact Tissues Using N,N-Dimethylformamide. Plant Physiol. 65(3):478-479.

Lichtenthaler H K, Welburn A R (1983) Determination of total carotenoids and chlorophyll a and b of leaf extracts in different solvents. Biochem. Soc. Trans. 603, 591-2.

Porra R J, Thompson W A, Kriedmann P E (1989). Determination of accurate extinction coefficients and simultaneous equations for assaying chlorophylls a and b extracted with four different solvents: verification of the concentration of chlorophyll standards by atomic absorption spectroscopy. Biochem Biophys Acta 975: 384-394

2.4. Microscopy analyses

Fluorescent and confocal microscopy

Samples preparation

Cell wall staining

Samples were incubated for staining in 0.05% FB28 in phosphate buffer saline (PBS) for 10min in the dark and then washed 3 to 4 times in the same buffer prior to proceed with observation (HONJOH et al. 2003).

Lipid bodies staining

Cells were harvested by centrifuging at 1000xg for 5-10 minutes and resuspended in phosphate buffered saline (PBS). Cells were mixed with Nile red at a final concentration of 2.5 g/ml (from a stock of 50g/ml in methanol), followed by a 10-min incubation in the dark. Stained cells were then mixed 1:1 with 2% low-temperature melting point agarose kept at 38°C and applied to the microscope slide.

Image capturing

Images were captured using a confocal system TCS SP5 on a DMI6000 microscope (Leica microsystems). For combined chlorophyll autofluorescence and Nile red fluorescence of polar lipids, the 488nm argon laser was used in combination with a 560 to 585nm filter for neutral lipid specific detection of Nile red fluorescence, while for chlorophyll autofluorescence a 647 to 753nm filter was used. Samples were viewed with a 63 oil immersion lens objective. Digital zoom was applied to magnify the imagines.

For fluorescence microscopy analysis, a Leica DMR DFC 480 microscope system was used. Nile red fluorescence in neutral lipids was detected using a UV light source with a 500±10 nm/542±10 nm excitation/ emission filter set. For imaging of the stained cell wall, signal of the fluorescent brightener FB28 was detected using a DAPI filter set with an excitation wavelength of 390±10nm and emission wavelength of 460±10nm.

Optical microscopy

Microalgae were observed using an optical microscope Olimpus BX40 using the 40X and 60X immersion objectives. Cultures were diluted if needed and directly applied to the microscope slides.

2.5. Molecular biology techniques

Standard buffers and solutions

Standard buffers and solutions were prepared, unless otherwise stated, according to Sambrook and Russel (2001). mQ filtered water nuclease free (Sigma) was routinely used to prepare all buffers and solutions. Chemicals, organic solvents and enzymes were analytical grade reagents and purchased from Sigma Aldrich Company, New England Biolabs, Promega Corporation and Invitrogen. Where necessary buffers, solutions, media and other materials were sterilised by autoclaving for at least 40 min at 121 °C (130 kPa), or in case of thermo labile reagents by filtration through 0.2-m syringe tip or bottle top filters (Nalgene). Antibiotics and IPTG were prepared using mQ water and kept as frozen stocks and stored at -20 °C until required.

Estimation of DNA concentration and quality

The concentration and quality of nucleic acid preparations were determined with a Nanodrop instrument (Nanodrop1000, Thermo Scientific) at a wavelength of 260 nm (A_{260}). An A_{260} of 1.0 is equivalent to a concentration of approximately 50 μg ml^{-1} of double-stranded DNA, 33 μg ml^{-1} of single-stranded DNA or 40 μg ml^{-1} RNA (Sambrook et al., 1989). The degree of contamination in the preparations could be estimated by measuring the A_{260}/A_{280} ratio and A_{260}/A_{230} ratio. Values above 1.95 for the measured A_{260}/A_{280} and A_{260}/A_{230} suggested a clean sample, whereas lower values indicated the presence of contaminants.

The concentration and quality of DNA preparations were also visually estimated after agarose gel electrophoresis in the presence of Ethidium Bromide under UV illumination. The signal for the DNA with the unknown concentration was compared to the intensity of a marker DNA with a known DNA concentration.

Moreover, in order to quantify the RNA contamination in DNA samples and parallel the DNA contamination in RNA samples, solutions were also examined using Qbit fluorometric quantitation kits (Qbit 1.0 fluorometer, Invitrogen), which allow the registering of different signals from the two nucleic acids using specific fluorescent probes. Samples were prepared for dsDNA broad range assay and RNA assay following the manufacturer instruction. Fluorometric assay yielded a quantification of each of the nucleic acids in the samples and could be compared with the data obtained using the spectrophotometer.

Phenol chloroform extraction routinely performed

Nulclease-Free Water was added to the samples where needed to reach a minimum volume of 100μl. Extraction was carried out by adding an equal volume of buffer saturated phenol and inverting the tube to mix the two phases or vortexing where allowed. Acid phenol was used when RNA was purified (phenol solution saturated with 0.1 M citrate buffer, pH 4.3 for molecular biology, Sigma) while basic phenol (phenol solution equilibrated with 10 mM Tris HCl, pH 8.0, 1 mM EDTA, for molecular biology, Sigma) was routinely used for DNA preparations. Phases were separated in phase lock assemblies (PRIME) when possible. Aqueous phase was collected and transferred to a new tube. A second extraction was performed using chloroform:isamyl alcohol 24:1. Phases were sepa-

rated again in phase lock assemblies when possible, according to the manufacturer indications. Aqueous supernatant obtained after this second extraction was routinely precipitated using ethanol and salt at -20°C. Where indicated in the methods the first extraction was performed using half of the initial volume of phenol and half of chloroform:isoamyl alcohol 24:1. This type of extraction was again followed by a second chloroform:isoamyl alcohol extraction.

Ethanol salt precipitation

RNA was routinely precipitated by adding 0.1 volume of 3 M sodium acetate and 2.5 volumes of ethanol. DNA was precipitated by adding 0.1 volume of 3 M sodium acetate and 2 volumes of ethanol. Samples were incubated at –20°C from 30 minutes to overnight and pelleted by centrifugation in a microcentrifuge for 30 minutes at 12.000xg at 4°C. Supernatant was pipetted off and pellet was washed by adding 70% ethanol and flicking the microcentrifuge tube to move the pellet. After a 15-30 minutes centrifugation at room temperature at 24°C, pellet was dried under the hood and resuspended in mQ nuclease free water. When detergent was present in the initial solution 0.2M (final concentration) sodium chloride was used instead of the sodium acetate to avoid detergent precipitation and trapping of the nucleic acid in the precipitate.

Agilent bioanalyzer

Readings of RNA were done using Agilent bioanalyzer 2100 by the Microcribi service using either nano or pico chips according to the concentration of the samples submitted (see Table 2.4 for details about concentration range). For DNA analysis, and in some cases also RNA analysis, Agilent readings were performed by BMR genomics using the following chips: DNA 7500, DNAHS, RNA6000pico (samples' concentration ranges are indicated again inTable 2.4). Samples were quantified using Nanodrop and diluted in mQ nuclease free water for submission.

Chip format	Total RNA	mRNA	DNA
NANO	Range 50-500 ng/µl	Range 25-250 ng/µl	
PICO and 6000pico	Range 200-5000 pg/µl	Range 500-5000 pg/µl	
DNA 7500			2-50 ng/µl
DNAHS			10-500 pg/µl

Table 2.4 Concentration range for agilent bioanalyzer chips

DNA and RNA purification

Co-isolation of high-quality DNA and RNA from *Nannochloropsis*

Nucleic acids were purified from *N.gaditana* as previously described by La Claire and Harrin (1997) with minor variations. 1,5 liters of microalgae in their late exponential growth phase were harvested by centrifuging, to obtain a pellet of 4 grams roughly. Algal cells were ground to a fine powder pestling together with quartz powder as already described. Powder was transferred to a 50ml falkon tube and resuspended in 20ml extraction buffer (0.1M Tris-HCl pH 8.5, 0.1 M EDTA pH 8.0, 0.2 M NaCl, 2.5%(w/v) SDS, 1mg/ml ProteinaseK (added right before use)). Mixture was then incubated on a rocking platform for 15

minutes at room temperature. Sarkosyl from a 30% stock was added to a final concentration of 2% and suspension was mixed for at least 60 minutes. Mixture was centrifuged at 9,500xg at room temperature for 10 min and supernatant was transferred to a new tube. 1/2 volume of phenol was added to supernatant, vortexed and mixed for 15 minutes. 1/2 volume of chloroform:isoamyl alcohol (24:1) was added only at this point, vortexed and mixed for an additional 20 minutes. Upper aqueous phase was recovered after centrifuging at 500xg for 5-15 min at room temperature and re-extracted with an equal volume of chloroform:isoamyl alcohol. 0.6 volumes of isopropanol were added to the recovered upper phase, mixed thoroughly and stored at –20°C for 90 minutes to precipitate nucleic acids. DNA and RNA were recovered by centrifuging at 12,000xg for 30 minutes at 4°C, supernatant was discarded and pellet rinsed in 5ml of 70% ethanol. After centrifuging pellet was drained from residual ethanol under the hood and resuspended in 200µl of cold mQ water. Centrifuge tube wall were rinsed with another 200µl of mQ water and combined with the resuspended pellet. Isolated nucleic acids were extracted again using 1/2 volume of cold phenol vortexing for 1 minute and 1/2 volume of cold chloroform:isoamyl alcohol, vortexing again.

Mixture was then centrifuged at 12,000xg for 5 min (4°C) in Phase Lock Assemblies (PRIME). Upper phase was decanted and extracted with an equal volume of cold chloroform:isoamyl alcohol, gently mixed and re-centrifuged in a Phase lock Assembly. Upper phase was transferred to a new tube and 1/10 of the volume of 3M sodium acetate was added together with 2 volumes of ethanol, mixed well and store at –20°C for 60 minutes. After centrifuging at 12,000xg for 30minutes at 4°C, pellet was overlayed with 200µl of 70% ethanol and centrifuged again for 15 minutes. Pellet was air dried in sterile transfer hood for 10 min and resuspended in 400µl of cold mQ water. To isolate high molecular weight RNA, 200 µl of 8M LiCl were added, mixed thoroughly and solution placed at 4°C overnight. RNA was recovered by centrifugation at 12,000xg for 30 minutes at 4°C. RNA containing pellet was resuspended in 100µl of cold mQ water. Supernatant, which contained DNA and tRNA, was also collected. 1/10 of the volume of 3M sodium acetate and 2 volumes of ethanol were added to both DNA and RNA solutions for precipitation. Samples were kept at –20°C overnight, centrifuged at 16,000xg for 30 minutes at 4°C and washed with 100µl of 70% ethanol. Obtained pellets were air dried in sterile transfer hood for 15 min and resuspended in 50µl of cold mQ water. Both the preparations were routinely assessed for quality and concentration using respectively Nanodrop and Qbit and stored either at 4°C or -20°C.

When purification was carried out to obtain RNA for library construction and sequencing, acidic phenol was used (phenol solution saturated with 0.1 M citrate buffer, pH 4.3 for molecular biology, Sigma) while in all the other cases standard basic phenol was used (phenol solution equilibrated with 10 mM Tris HCl, pH 8.0, 1 mM EDTA, for molecular biology, Sigma).

Cesium Cloride centrifugation for nuclear and plastidial genome isolation and further DNA purification

Volume of the DNA sample was adjusted to 4.62 ml with TE buffer containing 1 g/ml CsCl. 80 µl of ethidium bromide from a 10mg/ml stock solution was also added and solution mixed well. Solution was transferred to a 30 ml corex tube and spun at 15,000 x g for 15

minutes at 20°C to remove unsoluble particles, while DNA was still in solution. CsCl supernatant from above was tranferred to the gradient tubes using a sterile siringe, tube was sealed following the manufacturer instructions. Centrifugation was carried for 17 hours at 47,000 rpm in a beckman VTi 65.2 rotor at 20°C. Bands were visualized with long wave UV light and removed using an 18 gauge needle attached to a 1 ml syringe. Collected DNA was extracted in an equal volume of butanol (saturated with TE) repeatedly until all of the dye was removed. Aqueous phase containing DNA was then diluted in three volumes of sterile mQ water and mix thoroughly, overlayed with 8.5 volumes of cold 100% ethanol and stored overnight at 4°C. Samples were centrifuged for 20 min at 20.000xg in corex tubes, pellet was washed twice first in 100% ethanol followed by 70% ethanol wash and air dried prior to resunspend in 50-100 µl of sterile mQ water.

Messenger RNA enrichment and ribosomal RNA subtraction

PolyA+ mRNA recovery using Dynabeads

100 µg of total RNA were resunspended in lysis/binding buffer (mM Tris-HCl, pH 7.5; 500 mM LiCl; 10 mM EDTA, pH 8; 1% LiDS; 5 mM (DTT)) to a final volume of 400µl. 200µl of Dynabeads Oligo (dT)25 (Invitrogen) were transferred from the stock tube to a RNase-free 1.5 ml microcentrifuge tube. Tube was placed on a magnet for 30 seconds and the clear supernatant was removed. Beads were then washed by resuspending in 400µl of fresh lysis/binding buffer. Lysis/binding buffer was removed from the pre-washed Dynabeads by placing on the magnet until the suspension was clear and sample in lysis/binding buffer was added. After pipetting to resuspend the beads completely, the sample was incubated in continuous agitation for 10 minutes at room temperature to allow the polyA tail of the mRNA to hybridize to the oligo(dT)25 on the beads. Eppendorf tube was then placed on the magnet for 2 minutes. Supernatant was collected and precipitated overnight by adding 2,5 volumes of ethanol and 1/20 volume of 3M sodium acetate. Beads/mRNA complex was washed two times with 800µl of Washing Buffer A at room temperature, using the magnet to separate the beads from the solution between each washing step. Beads/mRNA complex was then washed twice in 800 µl of Washing Buffer B at room temperature again using the magnet to separate the beads from the solution. mRNA was then eluted by adding 25µl of fresh elution buffer and placing at 80°C for 10minutes in the thermocycler. Supernatant was recovered and placed to a new tube. Elution was then repeated on the beads following the same procedure. The two supernatant were pooled, placed again on the thermocycler for a couple of minutes and magnet was applied to remove eventual residual beads from the mRNA solution. In some cases the eluted mRNA was diluted again in lysis/binding buffer and the all procedure was repeated identical in order to improve the preparation and remove all the ribosomal RNA present in the sample. Obtained mRNA was quantified using Nanodrop and Qbit and was also analyzed on electrophoresis using Agilent bioanalyzer.

mRNA-ONLY™ Prokaryotic mRNA Isolation Kit: Terminator Exonuclease

The following reaction component were combined in a sterile (RNase-free) 0.2 ml tube:

Component	Amount (µl)
RNAse free Water	Up to 20µl
mRNA-ONLY Prokaryotic 10X Reaction Buffer	2,0
Ribo Guard RNAse inhibitor	0,5
Total RNA Sample (1-2,5 µg)	x
Terminator Exonuclease (1 unit/ µl)	1,0
Total reaction volume	20,0

Reaction was incubated at 42°C for 30 minutes in a thermocycler (with heated lid at 52°C) and terminated by phenol extraction (phenol was used equilibrated at pH 4.5) and ethanol precipitation. RNase-Free Water was added to the reaction to a total volume of 200 µl. Extraction was carried out in an equal volume of buffer saturated phenol and phase were separated in a phase lock system. Aqueous phase was collected and transferred to a new RNase-free tube. After a second extraction using chloroform the mRNA was precipitated using ethanol and salt and finally resuspended in 5µl RNase-Free Water.

mRNA-ONLY™ Prokaryotic mRNA Isolation Kit: Poly(A) Polymerase

The following protocol was designed to produce a poly(A)-tail length of ~150-200 b on the entire reaction product of a standard mRNA-ONLY reaction. In a sterile (RNase-free) 0.2ml tube, the following reaction components were combined:

Component	Amount (µl)
RNAse-free Water	Up to 20µl
Poly(A) Polymerase 10X reaction buffer	2,0
10mM ATP	2,0
Ribo Guard RNAse inhibitor (optional)	0,5
RNA substrate	x
Poly(A) Polymerase (4unit/ µl)	1,0
Total reaction volume	20,0

Reaction was incubate at 37°C for 20minutes in a thermocycler (with heated lid at 47°C) and terminate by phenol extraction and ethanol precipitation as described previously for the terminator exonuclease reaction. RNA was resuspended in 5µl of RNase-Free Water.

Ribosomal RNA subtraction using biotinylated probes

PCR amplification of rRNA genes

This step creates sample-specific amplicon pools that are used as template for *in vitro* transcription using T7 RNA polymerase, to produce anti-sense rRNA probes complimentary to rRNA in the total RNA extract.

Primers list:

Primer name	Sequence
16S for	AGAGTTTGATCCTGGCTCAG
16S rev	**GCCAG**TGAATTGTAATACGACTCACTATA**GG**TACGGCTACCTTGTTACGACTT
18S for	ATCTGGTGATTCTGCCAG
18S rev	AATTATAATACGACTCACTATACCTTCCGCAGGTTCACCTAC
23S for	GAACTGAAACATCTTAGTA
23S rev	**GCCAG**TGAATTGTAATACGACTCACTATAAGCCGACATCGAGGTGCCAAAC
28S for	ACCCGCTGGATTTAAGCATA
28S rev	AATTATAATACGACTCACTATAGATTCTGACTTAGAGGCGTTCAG
18S rev2	**GCCAG**TGAATTGTAATACGACTCACTATA**GG**CCTTCCGCAGGTTCACCTAC
23S rev2	**GCCAG**TGAATTGTAATACGACTCACTATA**GG**CCGACATCGAGGTGCCAAAC
28S rev2	**GCCAG**TGAATTGTAATACGACTCACTATA**GG**ATTCTGACTTAGAGGCGTTCAG
5.8S for	TATGGATCAAGGAAGTAGTC
5.8S rev	**GCCAG**TGAATTGTAATACGACTCACTATA**GG**ATCGTAAATCGATCATAACA
5S for	CCACCTGATCCCATTCCGAA
5S rev	**GCCAG**TGAATTGTAATACGACTCACTATA**GG**CGATGACCTACTCTCACATG

Table 2.5 Primers list. T7 promoter is underlined in the above table. The 5' bases upstream of the T7 promoter (bold characters) facilitate RNA polymerase binding. Transcription efficiency increases if the first bases downstream of the promoter (bold characters) are GG in the transcribed sequences (CC in the template strand). See Delong et al. (1999) and Stewart et al. (2010) for details on primer design.

Ribosomal RNA from *N.gaditana* were amplified using the forward-revers primer couples listed in Table 2.5. Phusion Taq polymerase was used following the protocol described in the dedicated section. Since high yields were necessary for the following step (*in vitro* transcription) four-five 100µl reactions were pooled for each of the ribosomal amplicon. PCR products had to be purified from salts and residual nucleotides prior to proceed with the next step. It was also important to concentrate the sample to guarantee a good yield of the *in vitro* transcription reaction. PCR products were either purified using a commercial kit based on spin column chromatography (pure link PCR purification kit, Invitrogen) following the manufacturer instruction or via phenol/chloroform extraction followed by ethanol and salt precipitation. After resuspension in RNAse free mQ water, DNA was quantified using Nanodrop and brought to a final concentration of 300-500ng/µl.

In vitro transcription of biotinylated probes

In vitro transcription of biotin-labelled anti-sense RNA probes was carried out using the MEGAscriptTM High Yield Transcription kit (Ambion) following the manufacture instruction with minor variations (that were introduced according to Stewart *et al*. 2010). Separate reactions were prepared for all the probes.

Reagent	Amount (µl)
PCR Amplicons (300-500ng/µl)	1.00
ATP	2.00
GTP	2.00
CTP	1.50
UTP	1.50
Biotin-11-CTP (10 mM, Roche)	3.75
Biotin-16-UTP (10 mM, Roche)	3.75
10X transcription buffer	2.00
SUPERase RNase inhibitor (Ambion)	0,50
T7 RNA Polymerase	2.00
Total volume	20.00

Reagents were mixed in the listed order at room temperature (not on ice, as spermidine in the reaction buffer can cause DNA precipitation). Reaction was incubated at 37°C gently mixing (300rpm) on the thermomixer for 6 hours. After incubation, 1 µl DNase I (included in the MEGAscript kit) was added to remove the DNA template and incubated at 37°C for additional 30 minutes. Synthesized RNA was purified using the MEGAclearTM kit and eluted in 50µl elution solution. RNA concentration was then quantified using Nanodrop and obtained probes were stored at -80°C.

rRNA subtraction with biotinylated arRNA

This step was used to bind the biotinylated antisense ribosomal RNA (arRNA) obtained to the rRNA in the total RNA sample. The labelled ds-rRNAs were then removed via hybridization to streptavidin-coated magnetic beads (Dynabeads Myone StreptavidinC1, Invitrogen), followed by magnetic separation. The probe to RNA ratio was kept at 1:1 in 1X sodium chloride-citrate (SSC) and 20% formamide. Formamide was important to denature the RNA avoiding the use of high temperatures. Concentration >20% formamide were found to inhibit the probe-bead binding, and < 20% to allow non-specific binding (non-target RNA to beads). The procedure obviously involved a significant reduction in RNA concentration therefore as much as 20 µg total RNA was used for each reaction. Prior to proceed with hybridization streptavidin coated beads were prepared. 1.5ml of beads were used for each reaction (20 µg total RNA). Beads were bound to the magnet for 2 minutes, supernatant was pipetted off and discarded while beads were resuspended in 1.5 ml of 0.1 N NaOH to deactivate RNAses eventually associated to the beads suspension. NaOH was removed and beads were washed twice in an equal volume of 1X SSC buffer, always removing the liquid after magnetization. On the third wash beads-buffer suspension was immersed on ice until hybridization. Hybridization reactions were sat up as follows:

Component	Amount (µl)
Template Total RNA Sample (20 µg)	X
Probe 16 S (20 µg)	20
Probe 18 S (20 µg)	20
Probe 23 S (20 µg)	20
Probe 28 S (20 µg)	20
Probe 5 S (3.5 µg)	30
Probe 5.8 S (4 µg)	30
Superase RNAse inhibitor (Invitrogen)	4
20X SSC buffer	10
100% Formamide	40
mQ RNAse free Water	Up to 200µl

Reactions were incubated in the thermocycler under the following conditions: 5 minutes at 70°C rampingdown to 25°C using a decrement of 2.7°C each minute. After hybridization, reactions were equilibrated at room temperature for 2-5 minutes and then immediately bound to the beads. SSC buffer was removed from the beads suspension after magnetization, reactions were brought to 1.5 ml using a solution of 1X SSC buffer 20% formamide at room temperature and mixed with the dried beads. Beads and hybridized RNA were incubated for 10 minutes at room temperature occasionally flicking to mix. After a quick spin, beads were captured using the magnet and non rRNA containing supernatant was transferred to a new tube. An equal volume of 1X SSC buffer was added and magnetization was repeated on the solution to remove eventual residual beads. In order to remove the formamide that could interfere with downstream utilization of the samples, obtained RNA were repeatedly extracted using acid phenol (Sigma) and then precipitated by ethanol and salt at low temperature. Samples were resuspended in 20µl of mQ RNAse free water and quantified using Nanodrop. Samples were also run on Agilent bioanalyzer to check the electrophoretic profile.

RNA manipulations

mRNA deCAPping

Decapping was performed using a Tobacco Acid Pyrophosphatase (TAP) from Ambion, supplied between the components of the kit 'RLM-RACE'. Reaction was assembled as follows:

Component	Volume (µl)
mQ RNAse free water	up to 20µl
10X reaction buffer	2.0
mRNA (700ng)	x
TAP	2.0
Final volume	20.0

Reaction was carried out for one hour on the thermomixer at 37°C gently mixing (300rpm). mRNA was then purified by phenol chloroform extraction followed by salt ethanol precipitation. Pellet was resuspended in 11µl. 1µl was diluted 1:2 and used for Nanodrop quantification. DecCAPped mRNA was checked for integrity by running the diluted sample in the Agilent Bioanalyzer chip.

Preparation of CAP-ligated sample

mRNA was bound to Dynabeads Oligo (dT)25 (Invitrogen) using its polyA tail. All the reactions were carried out on the sample conjugated to the magnetic beads. 250µl of beads were prepared as already described in the section 'PolyA+ mRNA recovery using Dynabeads', according to the manufacturer indications. Sample (around 1.5µg) was diluted 50 times in lysis/binding buffer to reach a final volume of 500µl and incubated with the dry beads for 5 minutes in continuous agitation. After a complete series of washes (2 washes using a 500µl of wash buffer A and 2 washes using 500µl of wash buffer B) the mRNA-beads conjugate was washed in 250µl of 1X Antarctic Phosphatase Buffer (NEB). Dry beads were then resuspended in the following solution:

Component	Volume (µl)
mQ water	24.0
10X Antarctic Phosphatase Buffer (NEB)	3.0
Antarctic Phosphatase (NEB)	3.0
Total volume	30.0

Reaction was incubated in gentle agitation (300rpm) for 30 minutes at 37°C. Enzyme activity was then heat inactivated for 5minutes at 65°C and beads washed in 500µl of wash buffer B to further denaturate the enzyme. After magnetization and removal of the wash buffer B, one more wash was performed using 250µl of 1X TAP buffer (Ambion). Dry beads were resuspended in:

Component	Volume (µl)
mQ water	20.0
10X TAP Buffer (Ambion)	2.7
TAP enzyme (from RACE kit, Ambion).	4.0
Total volume	27.0

Reaction was incubated for 60 minutes at 37°C in continuous mixing (300rpm). A wash was performed using washing buffer B as described before, and beads were resuspended in 1X ligase buffer. After magnetization and buffer removal beads were resuspended in:

Component	Volume (µl)
mQ water	17.0
DMSO	2.0
CAGE adaptor	1.0
10X T4 RNA ligase buffer (NEB)	2.5
T4 RNA ligase (NEB)	2.5
Total volume	27.0

Ligation was carried out at 16°C for all the night long in the orbital shaker to avoid beads precipitation. CAGE adaptor (Table 2.6) was an RNA oligo used for 5' tagging.

Adaptor name	Sequence
CAGE Adaptor	ATGGACCAG

Table 2.6 CAGE adaptor oligoribonucleotide

Reaction was stopped by adding 700µl of lysis/binding buffer followed by a complete series of washes. Dry beads were then resuspended in 30µl of elution buffer (cold) and incubated at 65°C for 5 minutes. Tube was immediately placed on the magnet to recover the supernatant containing the eluted sample. Elution was repeated a second time using 10µl of elution buffer following the same procedure of the first elution. Sample was quantified and assessed for quality by Nanodrop measurement and run on Agilent bioanalyzer chip to check the profile.

DNA manipulations

Amplification of DNA fragments by polymerase chain reaction (PCR)

DNA polymerase enzymes

The thermo stable DNA polymerases used in this study were: GOTAQ (Promega) and PHUSION (New England Biolabs). GOTAQ DNA polymerase was used for routine screening. PHUSION DNA polymerase was used to amplify DNA fragments for high fidelity cloning and sequencing and produced blunt-ended PCR products.

Conditions for standard PCR

Either the Mastercycler gradient (Eppendorf) or the X T gradient (Biometra) PCR machine was used to amplify a desired DNA fragment using different DNA templates and the primers listed in tables specific for each experiment. A typical 25 µl reaction mixture, in which 0,2µl of GOTAQ DNA polymerase were used, contained: 5 µl of 5x reaction buffer supplied by the manufacturer (0.5 M KCl, 0.1 M Tris/HCl pH 8.3, 7.5 mM $MgCl_2$), 0,5 µl of 10 mM dNTP mixture (Invitrogen; final [0.2 mM] for each nucleotide: dATP, dCTP, dGTP and dTTP), 1 µl of 5 µM forward primer (Invitrogen; final [0,1 µM]), 1 µl of 5 µM reverse primer (Invitrogen; final [0,1 µM]) and 1-5 µl template DNA. This reaction mixture was made up to 25 µl with sterile mQ water, mixed and briefly centrifuged. When possible Green Buffer, containing already the loading dyes (Xylene cianolo e tartrazina) for the

subsequent electrophoresis, was used. The lid of the PCR machine was heated during the program to prevent sample evaporation and condensation in the lid of the tube. A standard PCR program consisted of an initial denaturation step at 94 °C for 2 min and 35 subsequent cycles of 94 °C for 30 sec (denaturation), from 46 to 60 °C for 30 sec (primer annealing) and 72 °C for 1 to 6 min (primer extension; 1 min per 1 kb). The final extension step was performed at 72 °C for 10 min. The reaction mixture and the PCR program were varied when the standard procedure did not yield an optimum amplification.

PCR conditions for the PHUSION DNA polymerase

In order to quickly amplify high fidelity, blunt-ended DNA fragments, PHUSION DNA polymerase was used in PCRs, that varied from the standard PCRs in the following parameters. The typical reaction mixture had a volume of 20 µl and contained 0.2 units of PHUSION DNA polymerase, 4 µl of 5x PHUSION HF buffer (as supplied by New England Biolabs; no information about its composition available), 0.4 µl of 10 mM dNTP mixture (Invitrogen; giving a final concentration of 0,2mM for each nucleotide: dATP, dCTP, dGTP and dTTP), 0.4 µl of 5 µM forward primer (Invitrogen; final [0,1 µM]), 0.4 µl of 5 µM reverse primer (Invitrogen; final [0,1 µM]) and ~ 100 pg template DNA. This reaction mixture was made up to 20 µl with sterile mQ water, mixed and briefly centrifuged. The PCR program consisted of an initial denaturation step at 98 °C for 30 s and 35 subsequent cycles of 98 °C for 5-10 s (denaturation), from 50 to 70 °C for 10 sec (primer annealing) and 72 °C for 15 s to 1.5 min (primer extension; 15-30 s per 1 kb). The final extension step was performed at 72 °C for 5 min. The reaction mixture and the PCR program were varied when the standard procedure did not yield an optimum amplification.

Agarose gel electrophoresis

Agarose gel electrophoresis allows the separation of DNA fragments according to their sizes. Gels were prepared with molecular grade agarose (Sigma; final [0.5-1.5 % (w/v)]) dissolved in Tris-acetate-EDTA buffer (TAE; 40 mM Tris-acetate, 1 mM EDTA pH = 8.0) and Ethidium bromide DNA stain (final [1 µg ml^{-1}]). Samples were loaded in the wells after the addition of 6x loading buffer (30 % (w/v) phicoll, 0.25 % (w/v) orange, 0.25 % (w/v) xylene cyanol; final [1x]). The gels were run in TAE-buffer at 80 to 100 V in a horizontal gel apparatus. DNA could be visualised by using a UV transilluminator and photographs were taken. To estimate the size of unknown DNA fragments a DNA marker was loaded in one lane of the gel. We routinely used the Generuler series marker (Fermentas) or occasionally other ladders either from New England Biolabs or Promega. Specific indication about the ladder used will be always indicated in the gel pictures.

DNA Sequencing

Sanger sequencing

DNA sequencing of PCR amplificates was performed using the Sanger method by BMR genomics. Linear DNA necessary for sequencing reaction was usually 20 ng/Kbase. The reaction mixture that was send off usually contained also 3,2 pmol of a primer. The all volume was heat dried at 65°C.

454

Roche GS FLX 454 pyrosequencing runs were performed by BMR genomics using the Titanium series machine. Samples were prepared with the purity and concentration requested and libraries were directly prepared by the service.

SOLiD

SOLiD runs were performed in our group. Libraries were prepared with the precious help of the Lab Personnel dedicated to the instrument.

Libraries preparation for DNA and RNA sequencing

Full length cDNA library preparation and sequencing

In order to obtain the full length cDNA for 454 sequencing we decided to use the SMARTer technology for double strand cDNA (ds-cDNA) synthesis of the full length transcripts. ds-cDNA were produced using the In-Fusion SMARTer cDNA Library Construction Kit from Clontech according to the manufacturer instruction. Library production was repeated using three alternative primers for first strand synthesis starting from the polyA tail listed in Table 2.7. This new primer set was designed in order to avoid long stretches of homopolymers, that are not resolved by pyrosequencing and led to loss of sequences. In order to eliminate at least part of the polyA/polyT sequence at the end of each full length ds-cDNA, a BpmI recognition site was inserted in the primers immediately following the polyT in direction of the 5'. BpmI is able to cut 16 bases downstream the recognition site. In this way the best part of the polyT stretch can be removed. In order to remove all the sequences where the enzymatic cut was not successful as well as the short oligodT obtained after the cut, the primer 'DP' was functionalized with a biotin at the 5' extremity. Bind of the biotin to streptavidin-coated beads allows to purify the sample from both the short oligodT and the full length ds-cDNA still harbouring an intact polyA/polyT tail. Digestion of the library obtained using the primer 'DP' was performed using BpmI enzyme from NEB according to the manufacturer instructions. Enzyme was heat inactivated and mixture was incubated with streptavidin-coated magnetic beads (Dynabeads Myone StreptavidinC1, Invitrogen). Hybridization to the beads was carried out as indicated by the manufacturer. Supernatant containing the purified library was recovered and precipitated using ethanol and salt. Purified library was quantified using Nanodrop and assessed for length distribution by Agilent bioanalyzer.

Alternative to BpmI digestion, polyT stretch could be interrupted by inserting a G or a C in between the consecutive T of the primer (in bold in Table 2.7). In this way a library was directly produced lacking the polyA/polyT stretch. Libraries produced were purified by size exclusion chromatography using the CHROMASPIN DEPC-1000 columns supplied in the kit and then quantified using Nanodrop and run on both gel electrophoresis and Agilent Bioanalyzer chip.

Primer name	Sequence
DP	Biotin-spacer-**CTGGAG**TTTTTTTTTTTTTTTTTTTTVN
T7-BpmI-brok-dT-failsafe	GGCCAGTGAATTGTAATACGACTCACTATAGGG**CTGGAG**TTTT-GTTTTTTTTTCTTTTTTTTTTVN
T7-BpmI-brok-dT-power	GGCCAGTGAATTGTAATACGACTCACTATAGGG**CTGGAG**-TTTTTTTTTCTTTTTTTGTTTTTTTVN

Table 2.7 List of primer reverse used for cDNA synthesis. In order to avoid loss of information during 454 sequencing due to homopolymers, the polyA tail could be either interrupted or removed. Primers were designed harbouring a BpmI recognition site (in bold in the table) or interrupted polyT series (brok-dT primers). In two cases the full sequence necessary to promote transcription using T7 RNA polymerase (underlined in the table) was also added to the polyT for further uses of the synthesized cDNA.

Libraries preparation for SOLiD sequencing

Libraries were prepared following very carefully the instruction supplied by Applied Biosystems for each of the SOLiD System version in use at the moment of library preparation.

Two mate-pairs DNA libraries were prepared for sequencing using SOLiD 3 plus. Starting DNA was purified as described in the dedicated section and sheared using Hydroshear (Digilab) to reach a smear size centred on 3Kb. DNA preparation from *Nannochloropsis gaditana* was thawed from -20°C to 4-8°C in the fridge, freshly diluted to ~27ng/µl in mQ water Sigma and kept in the fridge until the moment of use. A first aliquot of 8µg DNA in 300µl was sheared applying the following setting: shearing code 9, 20 cycles. Obtained smear size was checked by gel electrophoresis in 0.7% agarose. Ethidium bromide was added only after run to stain the gel, for a correct estimation of the DNA size. If band resulted quite compact and was centred between the desired marker bands, more aliquots were sheared immediately after, using the same set up, and were checked again by gel electrophoresis. Sheared DNA was then pooled and precipitated overnight using ethanol and salt. In some cases, obtained samples were further purified by silica gel chromatography using the specific kit supplied by Applied Biosystems. DNA was finally checked for quality and concentration using Nanodrop and library preparation was carried out following the SOLiD protocol without modifications. According to the procedure, after CAP adaptors ligation, library had to be purified from gel in order to remove the unbound adaptors and eventually, if desired, to further selected the size. An agarose gel 0.8% in TAE was prepared using LMP Agarose (Invitrogen) and run for 3 hours at 4°C. In this case again ethidium bromide was added only after run to stain the DNA and was not included in the gel. Two bands were excised corresponding to the reference ladder bands at 1.5-3Kb and 3-5Kb. Obtained DNA was purified from agarose by column chromatography using the kit supplied by Applied Biosystems, quantified using Nanodrop and visualized by running a small aliquot on electrophoresis. Following this step two parallel library preparations were carried on again following carefully the manufacturer protocol. Results were checked after each step by spectrophotometric and fluorometric analysis and by electrophoresis on agilent chips. Enriched beads were finally obtained for sequencing run. Prior to load the run, small-scale runs in which the sole first ligation is carried out, were set up in order to check for quality and quantification of the enriched beads.

8 mate pair libraries were prepared for sequencing of expressed genes with the SOLiD 4 using the 'SOLiD Total RNA Seq Kit'. Libraries were prepared starting from different input RNA prepared as described in the specific sections following the procedure recommended

by the kit supplier. In a number of cases obtained libraries had to be further purified prior to proceed with emulsion PCR since unspecific PCR amplificates were produced during the amplification step. Libraries were purified by electrophoresis run on polyacrylamide gel as suggested in the kit protocol. cDNA in gel separation and excision were performed as described in the following paragraph, while purification of DNA from the gel following this step was carried out according to the kit instructions. Results were checked after each step by spectrophotometric and fluorometric analysis and by electrophoresis on agilent chips. As for the DNA libraries, enriched beads were produced for sequencing of the 8 libraries. In this case again small-scale control runs were sat up in order to check the beads before loading the full run.

Size selection of the amplified cDNA via Polyacrylamide gel electrophoresis

Stock solutions

10X TBE BUFFER pH 8.3 108 g Trizma base 55 g Borica Acid 9.3 g Na2EDTA dH2O to 1 L	Polyacrylamide 40 % Polyacrylamide solution from Sigma acrylamide/bisacrylamide=37.5/1	H_2O mQ H_2O DNAse-RNAse grade from Sigma
APS 10% (w/v) ammonium persulphate solution	TEMED N,N,N′,N′-Tetramethylethylenediamine for electrophoresis, ~99% (Sigma)	Hi-Density TBE Sample Buffer 18 mM Tris base 18 mM Boric acid 0.4 mM EDTA (free acid) 3% Ficoll Type 400 Bromophenol Blue Orange G

Gel recipe

Reagent	Volume
H_2O	11.1 ml
Polyacrylamide solution	2.4 ml
10X TBE buffer	1.5 ml
10% APS	150 µl
TEMED	15 µl

Gel caster and glassware were cleaned using ethanol and rinsed in abundant mQ water prior to use. Gel solution was prepared in a 50ml falkon tube, immediately poured on the gel caster and let polymerize for at least 30 minutes. All the volumes were handled using sterile pipettes or filter tips. Once the gel was polymerized (the few ml left over on the falkon tube were used as a control) comb was removed, electrophoresis unit was assem-

bled using a thick glass on the second gel slab and both the upper and the lower chamber were filled with 1X TBE buffer. Gel was then ready for loading.

Gel setup

Gel was run on a Hoefer SE 250 mini-vertical gel electrophoresis unit. Plates used were 10cm high per 8cm length, giving a gel size of roughly 8cmX7cm final. Gel sandwich was assembled using 1,5mm thick spacers and 12 well comb.

Running parameters

Voltage	Current	Expected time length	End of the run
100V constant	Start: 12-15 mA End: 6-15 mA	120 minutes	gel was run until the bromophenol blue tracking dye reached the bottom of the gel.

Loading

Sample buffer was added to each of the samples prior to loading using the stock Hi-Density TBE Sample Buffer as a 5X. Samples were loaded on the gel leaving one empty well between different samples to avoid contamination. As a reference molecular weight markers were also loaded after addition of the same 5X Hi-Density TBE Sample Buffer to the unstained ladder solution.

Gel staining

Gel was stained using SYBR GOLD according to the manufacturer instruction.

Band excision

After staining gel was leaned on the plastic film and moved to the UV-transilluminator plate. Bands were cut between 200 and 300 bp reference markers.

Chromosomes separation by Pulsed Field Gel Electrophoresis

Plugs preparation:

Method 1

Protoplasts were generated as previously described but the enzymatic treatment was carried out in agarose plugs. After resuspension in the protoplast forming medium 1ml of low melting point agarose solution (10mM Tris-HCl pH 7.5, 0.125M EDTA, 1% Low Melting Point Agarose) kept at 50°C was added. Mixture was immediately poured into plug molds (disposable plug mold, Biorad) and left solidifying. Plugs were kept at 37°C for 3 hours and then immersed in solubilization solution (0.1M Tris-HCl pH 8.5, 0.1 M EDTA pH 8.0, 0.2 M NaCl, 2.5%(w/v) SDS, and 1mg/ml Proteinase K) and left overnight in the thermomixer at 37°C and 400rpm. Plugs were finally rinsed several times in 1ml 0.5M NaEDTA pH8.0 for all the day long and stored indefinitely at 4°C in 0.5M NaEDTA pH 8.0.

Method 2

Cells were grown to late logarithmic phase harvested by centrifuging and broken mechanically by pestling in a mortar as described previously but without the addition of quartz powder. Since the usual amount of starting material used for this experiment was relatively poor, mashed cells were thawed in the mortar and then collected by pipetting. Initial amount of cells was variable in the various experiments, nevertheless in all the cases 500µl of low melting point agarose solution (10mM Tris-HCl pH 7.5, 0.125M EDTA, 1% Low Melting Point Agarose) at 50°C was added. Cell suspension was immediately poured into plug molds (disposable plug mold, Biorad). After solidification plugs were immersed in solubilization solution (0.1M Tris-HCl pH 8.5, 0.1 M EDTA pH 8.0, 0.2 M NaCl, 2.5%(w/v) SDS) and left overnight in the thermomixer at 37°C and 300rpm. Solubilization buffer was then removed and plugs were treated with proteinase K for 3 hours (0.1 M NaCl, 10 mM Tris pH 8.0, 1 mM EDTA, 0.5% SDS, and 1mg/ml Proteinase K). Plugs were finally rinsed several times in 1ml 0.5M NaEDTA pH8.0 and then stored at 4°C in 0.5M NaEDTA pH 8.0.

Electrophoretic runs:

Low molecular weight chromosomes (from 0.1 to 1 Mb)

1.2% agarose for pulsed field in 0.5X TBE buffer – 120ml gel

Running buffer was 0.5X TBE 12°C constant.

Running program:

5.1 V/cm voltage gradient

34 h run time

60s initial switch

120s final switch

High molecular weight chromosomes (from 1 to 5 Mb)

0.8% agarose for pulsed field in 1X TAE buffer – 120ml gel

Running buffer was 1X TAE 14°C constant.

Running program:

5.1 V/cm voltage gradient

48h run time

500s switch

Reference ladders:

DNA size standards routinely used for pulsed field runs were commercial chromosomal prep-arations from S.cerevisiae (range from 2000Kb to 200Kb), H.wingei (range from 3000Kb to 1000Kb), S.pombe (range from 6000Kb to 3000Kb). All the ladders were purchased from Biorad.

References

La Claire J W II, Harrin D L (1997) Co-isolation of high-quality DNA and RNA from coenocytic green algae. Plant Molecular Biology Reporter 15:263-272.

Fain S R, Druehl L D, Baillie D L (1988). Repeat and single copy sequences are differentially conserved in the evolution of kelp chloroplast DNA. J. Phycol. 24(3):292-302.

Herrin D, Worley T (1990) A rapid procedure for the isolation of chloroplast DNA from Chlamydomonas using the TL-100 ultracentrifuge. Plant Mol. Rep. 8(4):292-296.

Sambrook, Russell (2001) *Molecular Cloning: A Laboratory Manual* (3rd ed.). Cold Spring Harbor Laboratory Press

2.6. Protein biochemistry techniques

All protein biochemistry techniques were performed at room temperature or, where necessary, on ice or at 4° C and under dim light conditions, in order to minimise protein damage.

Determination of protein concentration

The protein concentration of a sample was determined with the BCA protein assay (Pierce) according to the manufacturer's instructions. Some of the buffer used are not compatible with this protein concentration determination method. In many cases protein concentration was indirectly estimated by measuring chlorophyll concentration in the solubilised sample prior to load on SDS-PAGE.

Acetone protein precipitation

In order to precipitate proteins from a sample, four times the sample volume of ice-cold, 100 % acetone were added, the mixture vortexed and incubated for at least 60 min at -20 °C. In many cases precipitation was allowed forall the night long. Precipitated proteins were pelleted in a microfuge (maximum speed, 10 min, RT) and resuspended in an appropriate volume of a desired buffer.

Polyacrylamide gel electrophoresis (PAGE)

Polyacrylamide gel electrophoresis (PAGE) was performed to separate complex protein mixtures according to their size differences into distinctive bands in a polyacrylamide gel matrix (Laemmli, 1970).

One-dimensional sodium dodecyl sulphate gel electrophoresis (1-D SDS-PAGE)

Samples for 1-D SDS-PAGE analyses were prepared by adding 3x SDS sample buffer (3x concentrate: 187 mM Tris/HCL pH = 6.8, 6 % (w/v) SDS, 30 % (v/v) glycerol, 0.1 % (w/v) bromophenol blue and 20 % (v/v) mercaptoethanol or 300mM DTT added freshly before each use; final [1x]) and incubation of the mixture for 45 min at room temperature under dim light. Unsolubilised material was pelleted in a microfuge (maximum speed, 5 min, RT) and samples were either flash frozen in liquid nitrogen for storage at –20 °C or loaded into the stacking gel wells. The gels were self cast, 1.5 mm thick and routinely run in the Hoefer Mighty Small II vertical gel system (Hoefer, Italy). The separation gel was typically a 10 or 12.5 % (w/v)continuous polyacrylamide (PAA; from a 40 % acrylamide/bisacrylamide = 37.5/1 stock solution), 6 M urea, 0.375 M Tris/HCl pH = 8.3, 0.01 % (v/v) N,N,N',N' tetramethylenediamine (TEMED); 0.1 % (w/v) ammonium persulphate (APS) SDS-PAGE gel. A 5 % (w/v) PAA (from a 40 % acrylamide/bisacrylamide = 37.5/1 stock solution) stacking gel containing 0.125 M Tris/HCl pH = 6.8, 0.01 % (v/v) TEMED and 0.1 % (w/v) APS was poured on top of the separation gel. The gels were run at RT with a constant current of 30 mA for a period of 3 h using a EV 265 CONSORT POWER SUPPLY (Hoefer, Italy) in a variant of the Laemmli running buffer (25mM Tris, 190 mM glycine, 0.1 % (w/v) SDS, pH = 8.3) (Laemmli, 1970). The obtained gels were either stained with Coomassie or silver or used for immunoblotting analyses.

Coomassie-brilliant-blue-R-250-staining of polyacrylamide gels

Gels were incubated for at least two hours, but preferably overnight, under gentle shaking in staining solution (40 % (v/v) ethanol, 10 % (v/v) acetic acid and 0.2 % (w/v) Coomassie-brilliant-blue-R-250). Background as well as stained proteins could be destained by incubation in the first destaining solution (40 % (v/v) ethanol and 10 % (v/v) acetic acid). Incubating the gel in the second destaining solution (10 % (v/v) acetic acid) would not destain proteins, but would eventually completely remove any background staining.

Silver staining of polyacrylamide gels

Polyacrylamide gels were silver-stained when the sensitivity of the Coomassie staining method described in the previous section proved to be insufficient. The method devised by Blum et al. (1987) is summarised in Table 2.8.

Step	Solution composition	Amount per 200ml solution	Incubation time
FIX	40 % (v/v) methanol 10 % (v/v) acetic acid	80ml 20 ml	90 min
WASH	30 % (v/v) ethanol	60 ml	3 x 20 min
PRETREAT	0.02 % (w/v) sodium thiosulphate ($Na_2S_2O_3$ 5x H_2O)	0,04 g	1 min
WASH	mQ water	n.a.	3X20 sec
STAIN	0.2 % (w/v) silvernitrate ($AgNO_3$) 0.02 % (v/v) formaldehyde (from 37 % (w/v) stock solution)	0,4g 40µl	20 min
WASH	mQ water	n.a.	3x20 sec
DEVELOP	3 % (v/v) sodium carbonate (Na_2CO_3) 0.05 % (w/v) formaldehyde (from 37 % (w/v) stock solution 2 % (v/v) pretreat stock solution	6g 10µl 4ml	bye eye
STOP	40% methanol x% acetic acid	80ml x ml	at least 20 min
WASH	40% methanol	80ml	indefinitely

Table 2.8 Silver staining protocol for polyacrylamide gels.

Protein molecular weight markers for PAGE

The protein molecular weight markers that were used in this work are listed in the following Table 2.9.

Low Molecular Weight Markers	
protein	Size (KDa)
Phosphorylase B	97.0
Albumin	66.0
Ovoalbumin	45.0
Carbonic anhydrase	30.0
Trypsin inhibitor	20.1
Alpha-Lactalbumin	14.4

Table 2.9 Protein molecular weight markers. Low molecular weight calibration kit for SDS-PAGE (LMW; Amersham Biosciences, Italy).

Western-blotting and immunodecoration

SDS-PAGE gels were used for immunoblotting analyses using a tank blot method (Mighty Small Transfer Tank; Oefer, Italy) (Burnette,1981; Towbin et al., 1979). Protein transfer was performed for at least 90 min at a constant voltage of 50V in transfer buffer (3 mM Na_2CO_3 , 10 mM $NaHCO_3$ and 20 % (v/v) methanol) onto either nitrocellulose membrane (Hybond-C Extra, Transfer medium nitrocellulose membrane; GE Healthcare, Italy) or PVDF membrane (Amersham Hybond™-P, hydrophobic polyvinylidene difluoride (PVDF); GE Healthcare, Italy).

After the proteins were blotted, the membrane was blocked for 1 h with 1x TBS (50 mM Tris-HCl pH 7.4, 150 mM NaCl) supplemented with 10 % (w/v) milk powder. The membrane was then washed a few times with 1x TBS and incubated for 1 hour and a half at RT on a rocking shaker with the primary antibody of choice (listed in Table 2.10). Primary antibody was diluted as indicated in table x in TBS-T (standard TBS supplemented with tween 20 to a final concentration of 0,05%). Membrane was then washed three times for 5 min with 1x TBS-T and subsequently incubated for 1 h at RT with the appropriate secondary antibody (anti-rabbit IgG, horse radish peroxidase conjugate or anti-mouse IgG, horse radish peroxidase conjugate, 1:10000 dilution in 1x TBS-T for revelation using the chemioluminescence method; anti-rabbit IgG, alcaline phosphatase conjugate, 1:10000 dilution in 1x TBS-T for revelation using the chromagen reagents; GE Healthcare, Italy). Unbound secondary antibody was removed by washing the membrane three times for 5 min with 1x TBS-T. When secondary antibody was detected using the enhanced chemiluminescence procedure (ECL; Durrant, 1990; Schneppenheim et al., 1991), membrane was further washed in TBS twice and then incubated for 5 minutes in a 1:1 mixture of ECL reagents [A (100 mM Tris/HCl pH = 8.3, 0.4 mM p-coumaric acid (90 mM stock solution in DMSO), 2.5 mM luminol (250 mM stock solution in DMSO)) and B (100 mM Tris/HCl pH = 8.3, 100 mM H_2O_2); ECL Western Blotting Substrate, Pierce, Italy]. To obtain the optimum chemiluminescent signal, it is important to omit Tween 20 from the last two washing steps, as the detergent seems to inhibit the peroxidase activity. Membrane was then put into an A4 reinforced pocket and exposed to a X-Ray film (Kodak® BioMax™ light film, 24x12 cm; SIGMA Aldrich, Italy) from 1 sec to 10 minutes. The film was developed according to the manufacturer's instructions. Dry membranes were stored between Whatman paper at RT and could be probed again with another antibody if necessary.
For revelation using the chromagen reaction, when secondary antibodies conjugated to

alchaline phosphatase were used, membranes were incubated in a freshly prepared mixture of nitroblue tetrazolium (NBT, 0.35 mg/ml) and 5-bromo-4-chloro-3-indolyl phosphate (BCIP, 0.18 mg/ml) in 0.1 M Tris buffer, pH 9.5, containing 0.1 M NaCl and 5mM $MgCl_2$. BCIP stock solution was prepared by dissolving 0.5 grams of the disodium salt in 10 ml H_2O mQ, while NBT was prepared in glass vials by dissolving 0.5 grams in 10 ml 70% dimethylformamide. Stock solutions were stored separately at 4°C in dark containers and mixed at the moment of use. Reaction was carried on checking bye eye until the desired colouration was reached and stopping in a slightly acid solution.

Antibody name	Working dilution	Secondary antibody
Anti Li818	1:2000	anti-rabbit IgG
D1 ciano	1:1000	anti-rabbit IgG
D2 mix	1:1000	anti-rabbit IgG
LSU	1:2500	anti-rabbit IgG
Cyt F	1:500, 5% milk	anti-rabbit IgG
alpha-ATPase	1:500, 5% milk	anti-rabbit IgG

Table 2.10 Primary antibodies used in this work.

References

Laemmli U K (1970) Cleavage of structural proteins during the assembly of the head of bacteriophage T4. Nature 227, 680-5.

Blum H, Beier H, Gross, H J (1987) Improved silver-staining of plant proteins, RNA and DNA in polyacrylamide gels. Electrophoresis 8, 93-9

Burnette W N (1981) "Western blotting": Electrophoretic transfer of proteins from sodium dodecyl sulfate-polyacrylamide gels to unmodified nitrocellulose and radiographic detection with antibody and radioiodinated Protein A. Anal. Biochem. 112, 195-203.

Towbin H, Staehelin T, Gordon J (1979) Electrophoretic transfer of proteins from polyacrylamide gels to nitrocellulose sheets: Procedure and some applications. Proc. Natl. Acad. Sci. USA 76, 4350-4.
Durrant I, Benge L C, Sturrock C, Devenish A T, Howe R, Roe S, Moore M, Scozzafava G, Proudfoot L M, Richardson T C, et al. (1990). The application of enhanced chemiluminescence to membrane-based nucleic acid detection. Biotechniques 8, 564-70.

Schneppenheim R, Budde U, Dahlmann N, Rautenberg P (1991) Luminography - a new, highly sensitive visualization method for electrophoresis. Electrophoresis 12, 367-72.

3. Results

Despite the interest for the industrial and biotechnological applications, *Nannochloropsis* is a completely new organism for molecular biology. In this report we present a characterization of the organism through confocal and fluorescent microscopy. All the experimental procedures necessary to the study of *N.gaditana* were tested and optimized. Some of the interesting observations collected during the experimental procedures' set up are reported in this chapter. The sequencing strategy designed for obtaining the whole nuclear genome is described and the results obtained are reported together with the preliminary data of genome assembly. Finally, the experiments performed for characterization of the transcriptome and identification of the genes involved in lipids metabolism are explained and the results obtained are reported and commented.

3.1. Culturing of microalgae

Obtaining of axenic *Nannochloropsis*

The difficulty often encountered in the isolation of axenic cultures of marine microorganisms is well reflected by the great variety of purification methods described in the literature and by the difficulty in finding samples from the culture collections completely free of contaminants. We considered the purity of the strains a prerequisite for the studies on nutrition and growth in response to variable conditions and we were also aware of the importance of avoiding DNA contaminations in the samples that were to undergo ultra deep sequencing. The sensitivity of the method indeed, is such that small amounts of contaminant DNA could be possibly sequenced. Due to the absence of a reference genome for the assembly of the reads of *Nannochloropsis*, filtering of the contaminant sequences would represent a difficult task. In order to be able to produce reliable and reproducible growth curves of *Nannochloropsis*, where al the registered variations could be attributed to the species of interest and the different parameters could be kept under complete control, we tested all the strains of interest available from the culture collections for purity e we used dilution and antibiotics (following the procedure described in materials and methods) to obtain pure cultures. A number of species of *Nannochloropsis* are reported in the literature to yield similar growth curves and similar lipid production profiles. In particular *N.salina*, *N.gaditana* and *N.oculata* are three very common halotolerant species while *N.limnetica* lives in fresh water. We were able to obtain pure cultures form all the four strains and, after reproducing the growth profiles found in the literature, we decide to focus our attention on *N.gaditana*.

The strain of *N.gaditana* available in the lab, was the sole obtained from a company, the Biofuel Systems S.L., and was already in use in a pilot plant for the production of biofuel in Alicante. The researchers in Alicante selected the strain after a systematic testing of all the cultures available at the Oceaographic Centre of Mazarrón and found that it was the most suitable in their hands for growth and lipids accumulation in outdoor conditions (personal communication of dr. Cristian Gomis). In order to produce data that could be useful for biotechnological applications, all the work that will be described herein was carried out on the strain of *N.gaditana* (further details on the strain can be found in material and methods).

As mentioned in the dedicated section in material and methods, some of the contaminants most frequently found in the control plates, were isolated and studied in further detail in order to design a specific strategy for their elimination. Identification of the strains was carried out by extraction of the DNA directly form colonies grown on agar plates and amplification of the ribosomal genes using universal primers. Primers used were designed on the conserved regions of the ribosomal operon and were also suitable for metagenomic analysis. Amplified genes were sequenced using the Sanger method and blasted against the nucleotide collection of the NCBI database. One of the most interesting findings of this experiment was the identification of a species of pigmented *Flavobacterium* of the phylum *Bacteroidetes* that was found in the cultures on *N.gaditana*. A picture of the typical colonies obtained while propagating the contaminant in LB agar plates is shown in Figure 3.1. Heterotrophic communities, where heterotrophic bacteria are found in associa-

tion to cyanobacteria and microalgae, have been extensively described in the literature (Hube et al. 2009, Grossart et al. 2005, Justi et al. 2005). Phytoplankton-bacteria interactions described, range from symbiosis to pure parasitism and in many cases it is suggested that this association has the potential to dramatically influence growth and blooming dynamics. The bacterial species most widely found in these marine communities are *Flavobacteria*. As a conclusion we cannot exclude an effect of the presence of this contaminant on the yield of *N.gaditana* in the outdoor systems. Nevertheless, in order to obtain data on the pure microalgae we decided to eliminate the contaminant prior to carry on our experiments.

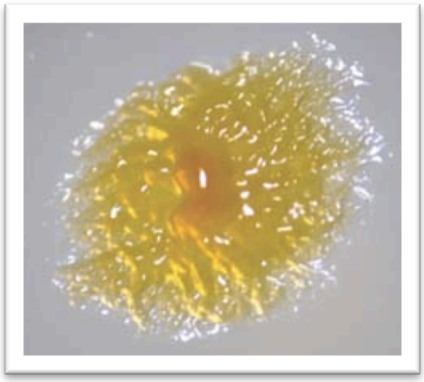


Figure 3.1 Flavobacterium. Picture of one of the colonies obtained by culturing one of the contaminants of *N.gaditana* in LB agar without antibiotics.

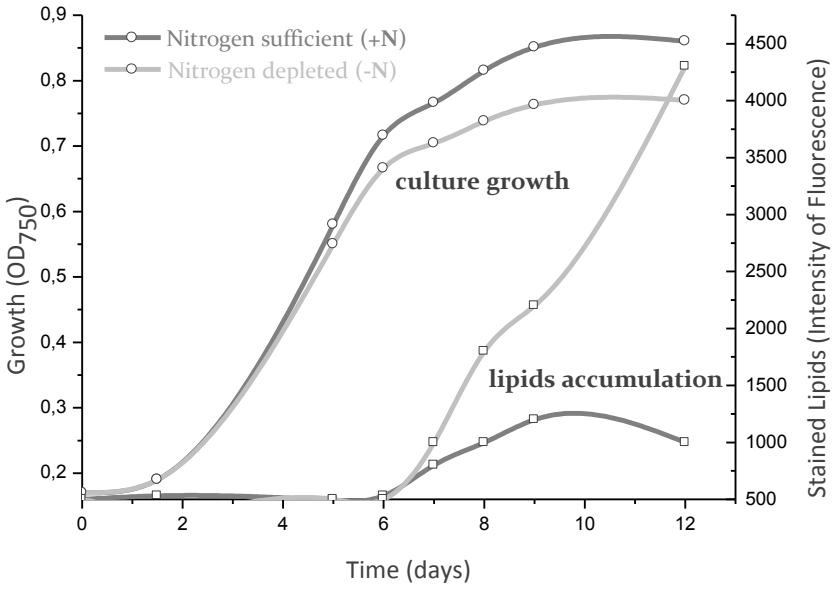
According to the information available about the antibiotic sensitivity of the studied members of this *Genus*, an antibiotic mix was produced and tested for toxicity in microalgal cultures. Moreover, since a number of *Flavobacteriaceae* are known to digest agar and use it as a fixed carbon source plant agar was used for the propagation of the microalgae in plates. The strategy revealed successful and *Flavobacterium* was no longer found on the control plates.

Growth and lipid accumulation

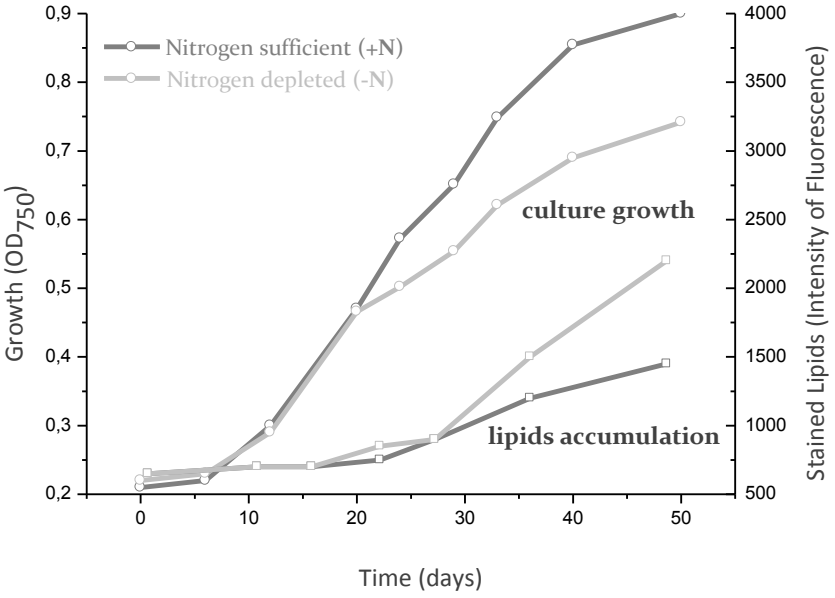
In Nature, the microalgae growth does not wonder about kinetics, being the growth rate just the enough one for species survival. Like any other species, the multiplication rate is highly dependent on environmental conditions, which are not constant in time being dependent on several factors. On the other hand, in artificial microalgae cultivation the goal is to favour the increasing of growth rate as much as possible or to push the metabolic route to follow one direction, if a particular metabolite has to be obtained, like in our case lipids. The number of variables that affect microalgae growth rate is huge, with direct, indirect and cross-effect influences. The few experiments reported in this dissertation were done for checking the effect of nitrogen and carbon limitation on growth and lipids accumulation. These two variables indeed are of particular interest for our porpoises since availability of nitrogen was shown to have the largest effect observed to date on controlling lipids accumulation in microalgal cultures, and carbon availability, on the other hand, has of course a fundamental effect on growth rates. The absolute values achieved for cell densities were not too high since the main interest was not to optimize, but just to set and

to keep stable the other variables also affecting the growth rate, in order to study, at the molecular level, the effects of the two considered variables. As an important variable that affects *Nannochloropsis* growth rate is the pH value and its fluctuation, Tris buffer was always added to the cultures to keep pH constant during the observations. Artificial lamps were used to supply continuous illumination to the cultures and light irradiance was kept as stable as possible during growth to avoid stress signals that might affect lipid metabolism. It must be said that the thermostatic cabinet where algae were grown, wasn't completely isolated form the external light. The commonly used f/2 culture medium for *Nannochloropsis* species was shown to be limiting for both carbon and nitrogen sources, reflecting the actual situation commonly found in the oceanic ecosystems.

The only nitrogen source present in f/2 medium is KNO_3 with a low concentration (0.75 mM). The effect of nitrogen addition ($NaNO_3$ 8.8 mM), keeping constant the supply of atmospheric CO_2 through the cotton plug, was studied. Precultures were sat up in standard f/2 liquid medium, starting from colonies grown on agar plates. Precultures were brought to an OD750 of approximately 0.2 and split into two parallel cultures. One of the cultures was supplemented with nitrogen (nitrogen sufficient culture) while the other was kept in f/2 medium without any further addition (nitrogen depleted culture). Experiments were sat up in order to check if the nitrogen concentration of the precultures had an effect on the following growth curves. It was clearly proved that two days culturing in the new medium were sufficient for complete elimination of the effects of the previous culturing conditions on the observed parameters. The time needed was reduced to one day if the culture diluted in the new medium was in active growth. Figure 3.2, Panel A and Panel B, shows the favourable effect of nitrate addition on cell growth. While growth rates during the lag and early logarithmic phases are comparable between nitrogen sufficient and nitrogen depleted samples, nutrients became limiting earlier in the nitrogen depleted cultures (grown in standard f/2 medium) and biomass accumulation during the stationary phase results lowered. As it must be noted in the plots, during the late logarithmic phase lipids start to accumulate in all the cultures, nevertheless average per cell lipid content is enhanced in the nitrogen deficient samples. Growth curves shown in the two Panels of Figure 3.2 have a similar pattern but the time scale is completely different. While cultures grown during the cold period reached the stationary phase in 10 to 12 days, cultures sat up during in the period between may and august took about 50 days to reach the same amount of biomass. The interval of time necessary to reach the stationary phase in winter was shortened when CO_2 was supplied to the cultures as shown in Figure 3.3.



Panel A Growth curve registered in winter.



Panel B Growth curve registered in summer.

Figure 3.2 Growth curves of *N.gaditana*. Plots were obtained as an average of 3 independent experiments. Growth curves of nitrogen sufficient and nitrogen depleted algae, compared in the two graph, were obtained by splitting a preculture, that had reached an OD750 of approximately 0.2, into two parallel cultures, one of which was supplemented of $NaNO_3$ from a 100X stock solution (nitrogen sufficient sample). Growth was plotted as a function of turbidity, measured as optical density at 750nm. Lipid content was also measured as fluorescent intensity of the nile red conjugated lipids per million of cells. Plot shown in Panel A was obtained in wintertime, while growth curve in Panel B was registered during summer.

Despite the use of the same growth medium and the tight control of temperature and light inside the thermostatic cabinet, for two consecutive years we registered the same reproducible result in cultures growth. During summer growth rate is conspicuously decreased and lipids accumulation does not reach its maximum values in nitrogen starva-

tion. In some cases, during summer, a stress response was evident in the cultures that led to a decrease in chlorophyll content and to lipid accumulation in both nitrogen sufficient and nitrogen depleted samples during the early logarithmic phase. Growth rates were also decreased even further. This behaviour was registered occasionally and was not reproducible. These observations revel the presents of parameters relevant for growth that were not under complete control in or experiments. Similar results were registered before in cultures under out door conditions (Boussiba et al. 1987).

The medium does not have any fixed carbon source because autotrophic organisms are able to use the atmospheric CO_2 and fix it through photosynthesis. Nevertheless, one of the interesting features of microalgae for biomass production is their ability to sustain high concentrations of CO_2 in the medium (if compared to the atmospheric CO_2 partial pressure) and to fix it producing an increased biomass accumulation. This feature is of special importance since CO_2 normally emitted to the atmosphere by producing plants could be channelled into biofuel plants and used to boost the productivity. In the experiment reported in Figure 3.3 we followed the same procedure described for the previous growth curves shown in Figure 3.2 but, in this case, experiments were conducted in glass bubbled tubes, agitated by air flowing containing 5% CO_2. Stationary phase in these cultures was reached in 6 days during winter and the overall biomass yield per volume, in the same conditions of the previous curves, was 1.5 times higher. Here again we observed a decreased growth rate in the late logarithmic phase in the samples in nitrogen starvation comparing to the nitrogen sufficient cultures and a lower biomass yield in the stressed samples.

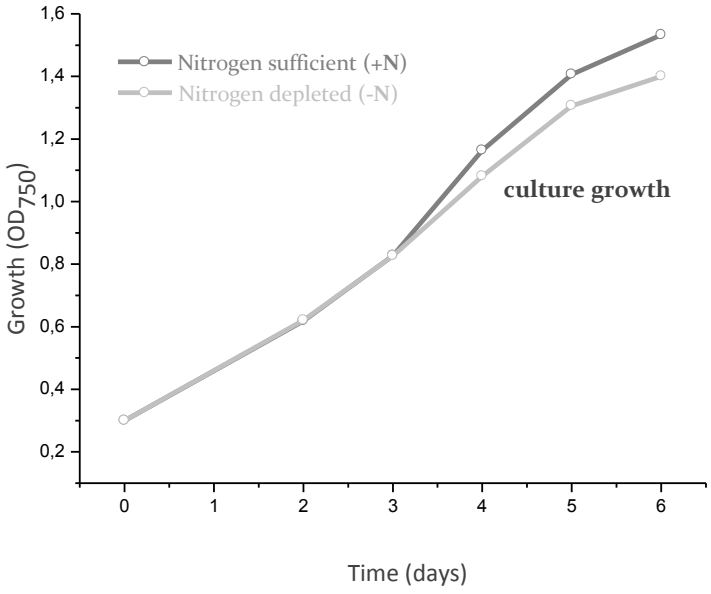


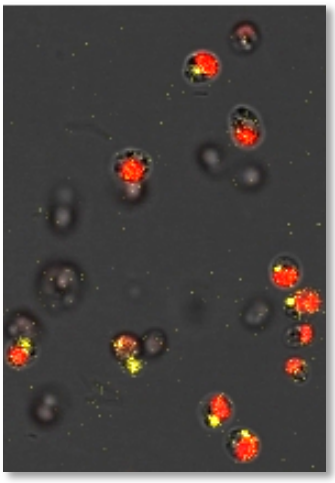
Figure 3.3 Growth curve of *N.gaditana* supplemented with 5% CO2.

Picking the eyes out of *Nannochloropsis*

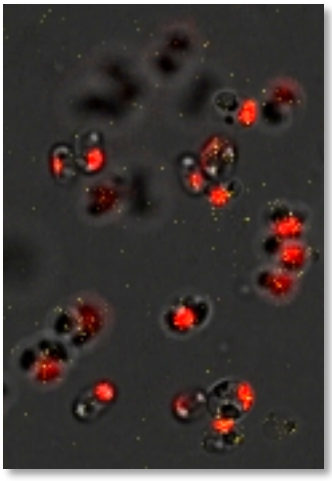
Being naturally coloured microorganisms, microalgae are very suitable for *in vivo* observation using different microscopes. In particular, since they contain both chlorophyll and carotenoids, that have a conspicuous autofluorescence, the main features of their photo-

synthetic apparatus can be easily visualized using confocal and fluorescent microscopes. When we first looked at *N.gaditana* through a confocal microscope, without the application of any staining, two main features become suddenly evident: the presence of a prominent chloroplast, accounting for the largest part of the coccoid cell volume and the appearance of a small bright orange-red spot localized immediately outside the chloroplast. While exiting trough an argon laser at 488nm, chlorophyll fluorescent emission was mainly detected in the red region of the spectrum between 650nm and 700nm, while the signal of the bright spot had its maximum specific emission around 580nm and dropped to zero above 650nm. The combination of the two signals yielded the clearly resolved imagines shown in Figure 3.4. The emission profile of the orange-red spot is consistent with the one registered from carotenoids, allowing us to hypothesize that it can be mainly constituted of agglomerates of these pigments. *Eustigmatophytes* indeed, and *N.gaditana* was assigned to this class, are characterized by the presence in the zoospores of an eye-spot (called '*eustigma*') harbouring carotenoid containing globules. The orange-red eye-spot was clearly visible in the majority of the cells in the nitrogen sufficient cultures, as shown in Figure 3.4, Panels A and C, while its presence was not registered in the nitrogen depleted culture as it can be noticed in Panels B and D of Figure 3.4. Presence of an eye-spot was assigned to date, in the *Eustigmatophytes* class, to motile asexual spores only, while it was not seen in vegetative cells before. In our observations, presence of a motile flagellum in the cells containing the cytoplasmic eye-spot was not detected and moreover eye-spot was visible in the very large majority of the cells imagined in the nitrogen sufficient samples, suggesting that, if the eye-spot is a specific feature of the zoospores, all the cells in the culture are actually zoospores. Given the poor resolution that our system allows for these cells, whose average diameter is below the 2µm, we cannot exclude any of the hypotheses. Moreover we were not able to notice in our observations any morphological feature that allowed us to distinguish between spores and vegetative cells in the observed samples.

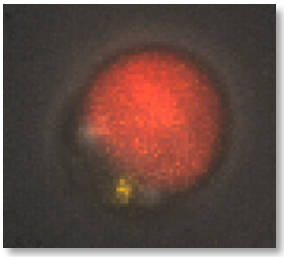
Cells imagined in the nitrogen depleted samples presented, on average, an evident reduction of the chloroplast size as it can be evaluated from the comparison of imagines in Panels B and D of Figure 3.4 with the corresponding Panels A and C. This result is consistent with the dramatic decrease in chlorophyll content per cell, observed, by spectrophotometric measurement, in the sample in nitrogen limitation and illustrated in Figure 3.5. Nitrogen depletion during batch culture of microalgae was reported to decrease the intracellular chlorophyll content also in early works (Eppley and Renger 1974), and the decrease in chlorophyll content was connected to a diminishing of the chloroplast size before (Vesk and Jeffrey 1977). As already mentioned the orange-red body was not visible in the cells grown in nitrogen limitation. Moreover we measured at the same time the average carotenoid content per cell of the two samples and the result, shown in Figure 3.5., highlights a decrease in carotenoid content in the nitrogen depleted cells down to 50% of the total amount present in the nitrogen sufficient cultures. This result is consistent with the data reported by Forján and coworkers (2007) for *N.gaditana*. The result supports the hypothetical attribution of the orange-red eyespot as a conglomerate of carotenoids. At the moment we cannot exclude that such a decrease is due to the absence of zoospores in the nitrogen depleted samples.



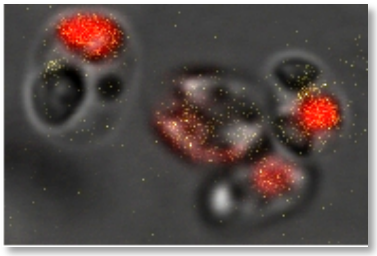
Panel A Nitrogen sufficient culture low digital zoom



Panel B Nitrogen depleted culture low digital zoom



Panel C Nitrogen sufficient culture high digital zoom



Panel D Nitrogen depleted culture high digital zoom

Figure 3.4 Confocal microscopy images of *N.gaditana* in nitrogen sufficient and nitrogen depleted conditions. Combined orange and red autofluorescent signals were captured. No stain was applied to the samples. The figure is a comparison of the two main features observed in *N.gaditana* looking through the confocal microscope, in the two conditions under study: the conspicuous chloroplast size, which accounts for the greatest part of the cellular volume in nitrogen sufficient samples, shrinks in the nitrogen depleted cultures; an orange-red eye-spot is clearly visible in the majority of the imagined cells of the nitrogen sufficient sample which is not visible in the parallel stressed cells.

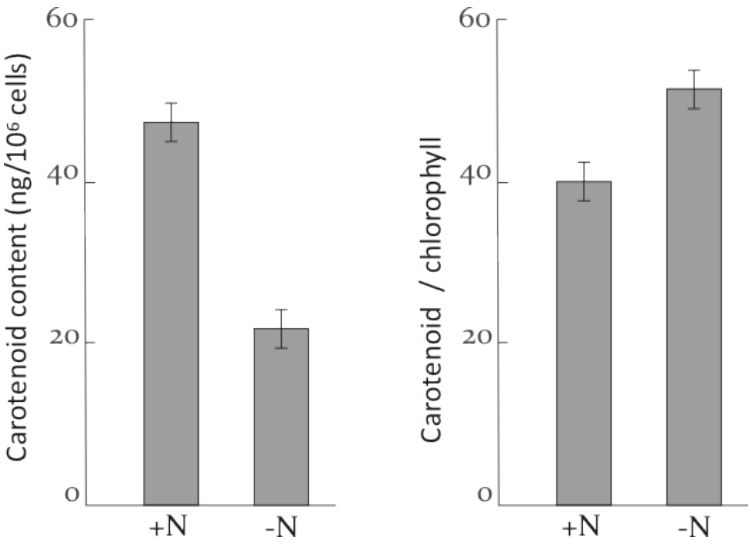
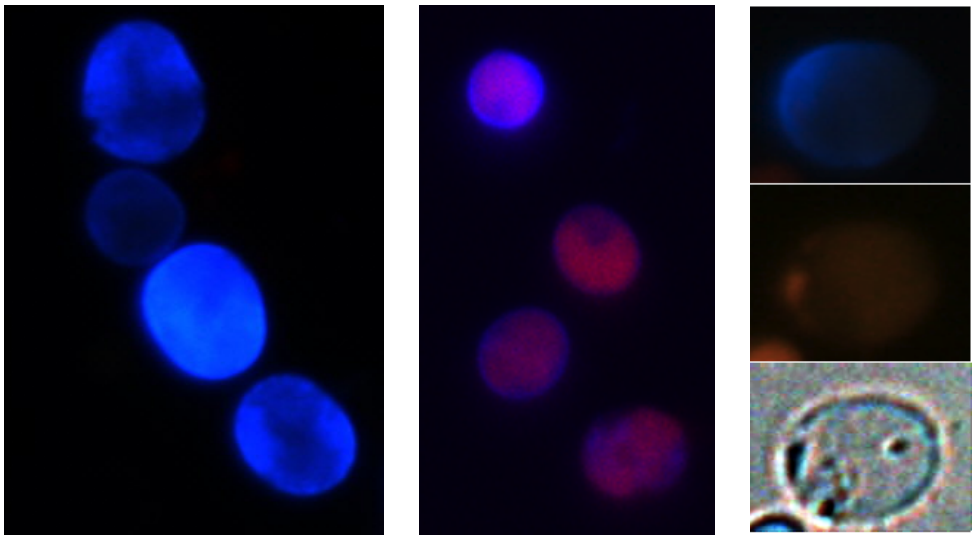


Figure 3.5 Measure of the average per cell content of carotenoids and chlorophyll in two parallel nitrogen sufficient and nitrogen depleted cultures observed under the confocal microscopy.

N.gaditana cell wall was also studied in order infer information about the chemical composition and attempt a cell wall digestion for protoplast production. Cell wall was successfully stained using fluorescent brightener 28, a water soluble dye that exhibits selective binding to the cell wall of fungi, algae and higher plants, due to its high affinity to cellulose fibbers and chitin and the reported binding to acidic polysaccharides, characteristic of the cell wall of many algal species. Cells were imagined using a fluorescent microscope equipped with a DAPI filter set for detection of the fluorescent staining signal, result is shown in Figure 3.6 Panel A. The coccoid cellular shape is evident in the homogeneous stained cells. A parallel detection of chlorophyll autofluorescent signal was performed on the same samples and colour combined imagines were realized (Figure 3.6, Panel B). In this case again, the presence of a prominent chloroplast accounting for the largest part of the cell volume, can be appreciated. Panel C of Figure 3.6 shows a succession of imagines of the same cell shouted using the DAPI filter (top), red filter (middle) and visible light bright field set up (bottom). As it can be noticed chlorophyll fluorescence results diffused and chloroplast shape is not so well resolved as it is in the confocal microscope imagines. Nevertheless the presence of regions with higher pigment density is clearly in evidence in both the red florescent signal and the imaging using visible light.



Panel A Cell wall stain.

Panel B Cell wall and chlorophyll, colour combine.

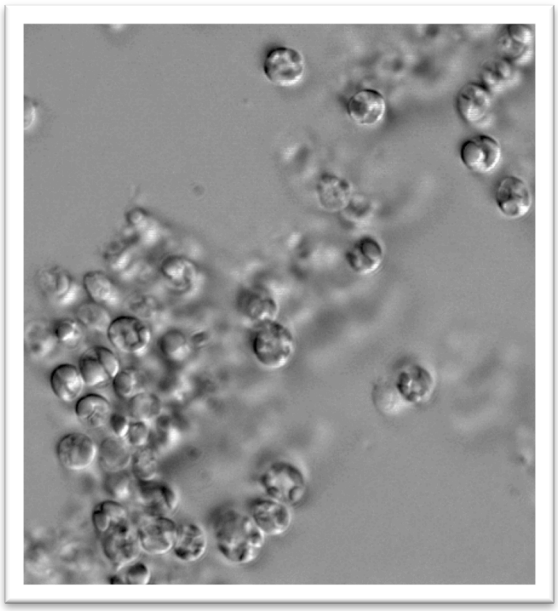
Panel C Succession of imagines of the same cell. Cell wall, chlorophyll, whole cell.

Figure 3.6 Fluorescent microscopy imaging of *N.gaditana* cells grown in nitrogen sufficient medium and stained using FB28. Cell wall was successfully stained using a fluorescent brightener able to bind cellulose, chitin and acidic polysaccharides. Coccoid cell shape is evident in the stained cells in Panel A. A merge of the chlorophyll signal and FB28 signal was realized in Panel B, where dimension of the bean shaped chloroplast is outstanding, filling the majority of the cell volume. In Panel C a succession of imagines of the same cell obtained registering the blue (top) and red (middle) fluorescent signals and the bright field imagine (bottom) shows the presence of pigment dense regions inside the cell.

Two different illumination set ups were used for visible light imaging of *N.gaditana* cells (Figure 3.7). The bright field illumination, as already pointed out commenting Figure 3.6 Panel C, shows the presence of regions with different absorbance of light in the sample, that we attributed as differences in pigment density distribution. Picture shown in Figure 3.7 Panel B, was shouted using a phase contrast illumination and provides a course grain

description of the main morphological features of the microalga. The volumes occupied by the bean shaped chloroplast and the nucleus can be gathered from the imagine.

Panel A Bright field illumination imaging of *N.gaditana*, high digital magnification. The white bar on the bottom right measures 1µm.



Panel B Phase contrast illumination imaging of *N.gaditana.*

Figure 3.7 Bright field and phase contrast illumination imagines of *N.gaditana*

Lipid bodies

When stressed by nutrient limiting conditions, algae synthesize large amounts of neutral lipids and accumulate them mostly in discrete cytosolic droplets referred to as lipid bodies. While some information is available about oil bodies in higher plants, particularly in seeds, knowledge of structure and function of algal lipids is in general very poor. The oil bodies of seeds of higher plants serve as a source of energy during germination whereas lipid bodies in algae are divided, not consumed during regreening after nitrogen limitations (Shifrin and Chisholm 1981). *N.gaditana*, as shown in the previous paragraphs, responds to nitrogen limitation decreasing the amount of pigments and accumulating conspicuous amounts of lipids (Figure 3.2). We followed the variations in cellular lipid content measuring the variations in fluorescence emission of nile red stained cells in parallel cultures in nitrogen depleted and nitrogen sufficient conditions and moreover we also observed the stained samples through a confocal microscope in order to localize the accumulated lipids inside the cells. Figure 3.8 shows the fluorescence profile of the observed samples. Cultures were diluted to 1 million cells per millilitre, stained using nile red and fluorescence was measured using the fluorometer. Nitrogen depleted samples (indicated in figure with the symbol –N) produced a peaked curve centred on 565nm, characteristic of nile red bound to neutral lipids. The parallel nitrogen sufficient sample (+N) presents a lower value of fluorescence in the same wavelength interval and the peaked curve, if present, is below the noise signal. This measure gives us an indication about the relative lipid concentrations of the two samples. The measure was taken in the two samples during the stationary phase, when lipids had already started to accumulate.

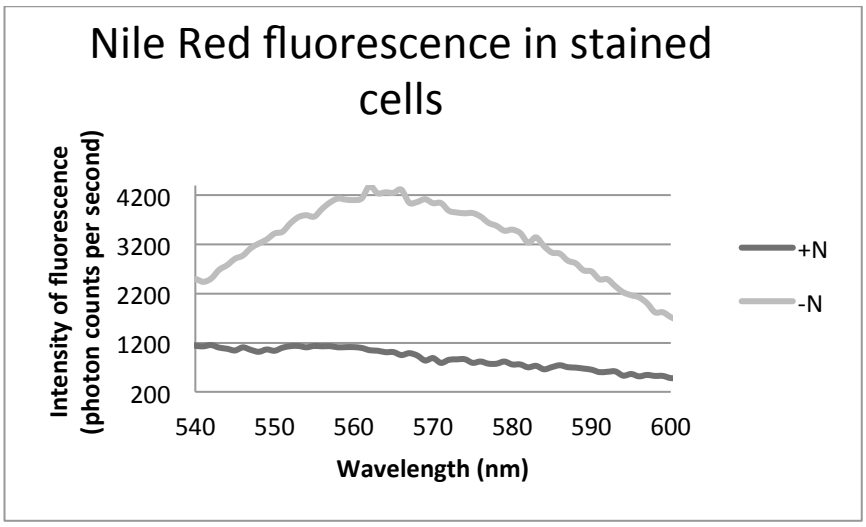
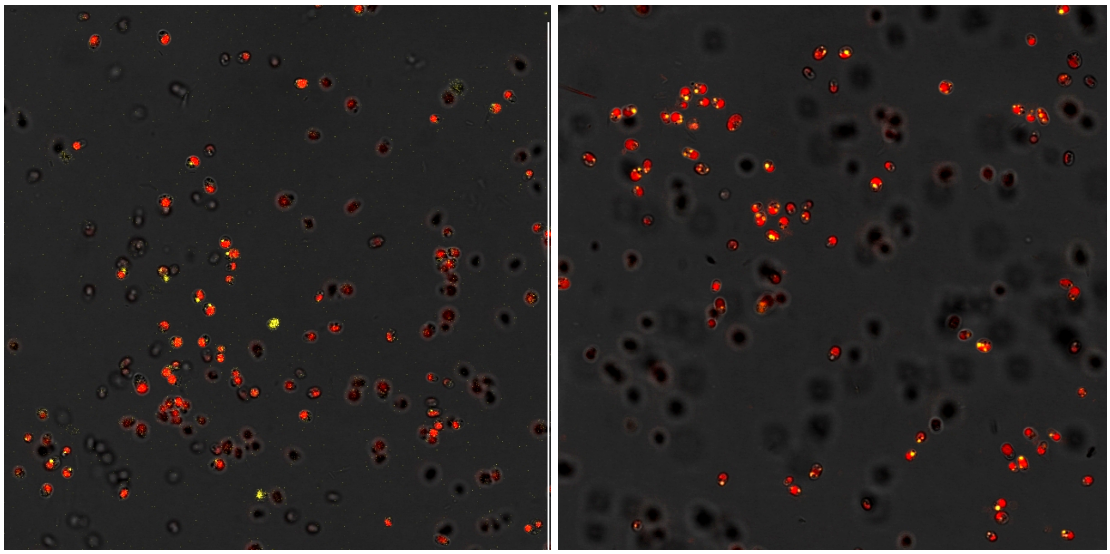


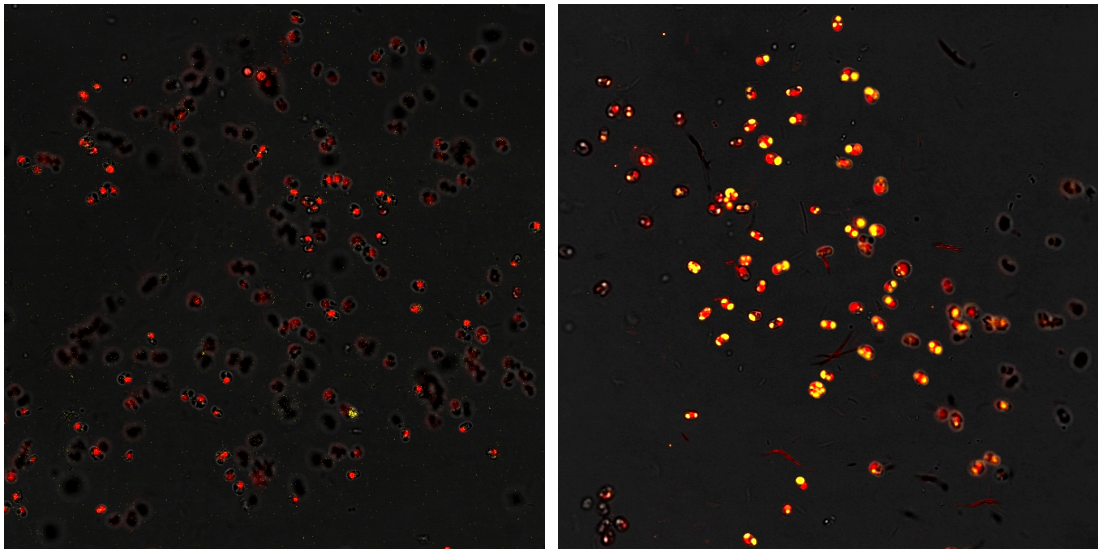
Figure 3.8 Measure of lipid content in cells grown in nitrogen sufficient (+N) and nitrogen depleted (-N) medium. Cultures were diluted to 1 million cells per ml and stained using nile red as described in materials and methods. Dye emission increases linearly with the number of bound lipids. Graph shows the plot registered by the fluorometer in the region around 560nm where the dye increases its fluorescent emission after binding to neutral lipids. The –N trace shows a clear signal of binding of the fluorescent stain to lipids, represented by the peak at 565nm. The same amount of nile red in the +N samples yields a lower fluorescent signal and peak is not clearly distinguishable from the background noise. The difference is due to the availability of lipids in the samples and provides a comparative measure of differences in lipids accumulation in two parallel samples.

The two samples were then observed under the confocal microscope registering the combined signals emitted in the yellow and red portion of the spectrum. Resulting imagines are presented in Figure 3.9. Unstained cells were used as a control. Unstained nitrogen sufficient samples (Panel A) showed a main signal in the red channel and yellow spots are also visible, as already noticed in Figure 3.4, accounting for the cytosolic carotenoid body, probably a photoreceptor. When the parallel stained sample (Panel B) is observed, the overall signal in the yellow channel slightly increases and the number of yellow spots increases to two or three per cell in a few cases. The production of lipids is indeed registered even in the nitrogen sufficient samples (as evident in Figure 3.2) measured as an increased value of fluorescence at 565nm, even though a proper peak is not visible. Nevertheless lipid production is at lest 4 times lower than in the nitrogen depleted cultures. Nitrogen depleted cells are shown in Panels C and D. As expected we did not register any signal from the yellow channel in the unstained sample (Panel C), while the parallel stained cells (Panel D) yielded a bright yellow signal due to multiple prominent lipid bodies accumulated inside the majority of the cells. Imagines shown in Figure 3.9 were taken using the 63X oil immersion objective and thus no further magnification could be applied. Nevertheless we decide to apply the digital zoom to the high resolution data transmitted from the microscope and realize magnified imagines to allow a better description of the visualized lipid bodies. Results of the analysis are shown in Figure 3.10.



Panel A Nitrogen sufficient culture, low digital zoom, no stain

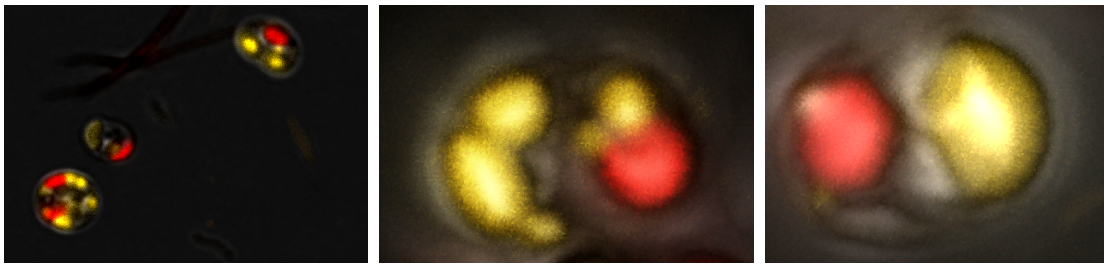
Panel B Nitrogen sufficient culture, low digital zoom, nr stain



Panel C Nitrogen depleted culture, low digital zoom, no stain

Panel D Nitrogen depleted culture, low digital zoom, nr stain

Figure 3.9 Imaging of combined yellow and red fluorescent signals in nitrogen sufficient and nitrogen depleted cultures. Samples in Panels B and D were stained using Nile Red (nr stain), while A and C were the correspondent unstained controls. Yellow signal is registered in A, where is attributed to eyespot autofluorescence, as well as in B and D, where it is mainly due to lipid bodies staining.



Panel A Nitrogen depleted culture, high digital zoom, nr stain

Panel B Nitrogen depleted culture, high digital zoom, nr stain

Panel C Nitrogen depleted culture, high digital zoom, nr stain

Figure 3.10 Lipid bodies imaging. Combined yellow and red fluorescent signals in nitrogen depleted cultures. Magnified details. The red signal corresponds to the chloroplast, while the yellow to the lipid bodies.

Lipid bodies, in yellow, are cytoplasmic, most being closely appressed to the reduced size chloroplast (red), they vary in size and abundance in the various cells observed. Each lipid body maintains an integral spherical shape, suggesting that each is surrounded by a membrane and/or a coating material. To our knowledge *Nannochloropsis* lipid bodies were not characterized in the literature before.

References

Boussiba S, Vonshak A, Cohen Z, Avissar Y, Richmond A (1987) Lipid and biomass production by the halotolerant microalga Nannochloropsis salina. Biomass 12:37–47

Hube A E, Heyduck-so B and Fischer U (2009) Phylogenetic classification of heterotrophic bacteria associated with filamentous marine cyanobacteria in culture. Systematic and Applied Microbiology 32 : 256–265

Grossart H P, Levold F, Allgaier M, Simon M and Brinkhoff T (2005) Marine diatom species harbour distinct bacterial communities. Environmental Microbiology 7(6): 860–873

Jasti S, Sieracki M E, Poulton N J, Giewat M W and Rooney-Varga J N (2005) Phylogenetic Diversity and Specificity of Bacteria Closely Associated with Alexandrium spp. and Other Phytoplankton. Appl Environ Microbiol. 71(7): 3483–3494.

Eppley R W and Renger H E (1974) Nitrogen assimilation of an oceanic diatom in N-limited countinuous culture. Journal of Phycology 10: 15-23

Vesk.M. and Jeffrey.S.W. (1977) Effect of blue-green light on photosynthetic pigments and chloroplast structure in unicellular marine algae from six classes. Journal of Phycology 13, 280-288.

Forján E, Garbayo I, Casal C and Vílchez C (2007) Enhancement of carotenoid production in Nannochloropsis by phosphate and sulphur limitation. Communicating Current Research and Educational Topics and Trends in Applied Microbiology A. Méndez-Vilas (Ed.)

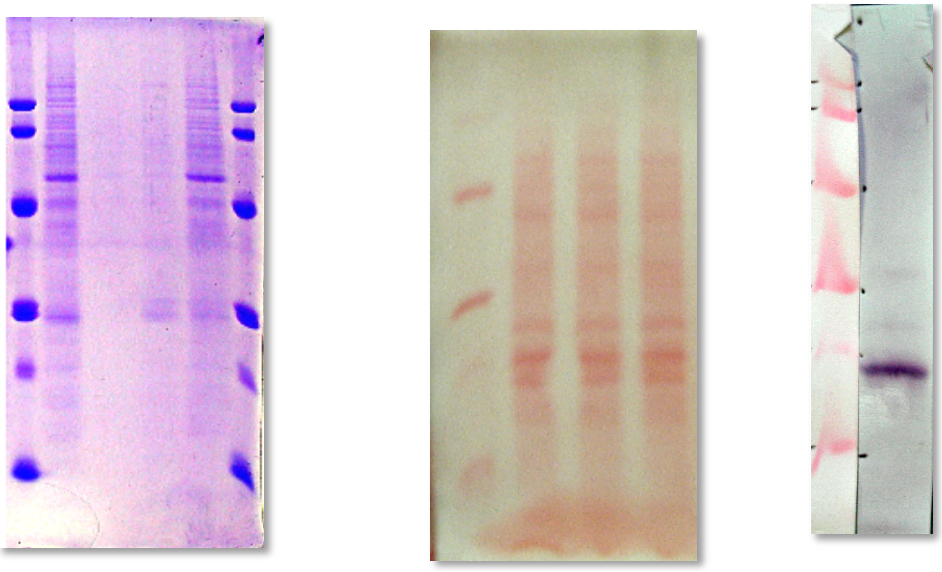
Shifrin N S and Chisholm S W (1981)Phytoplancton lipids: interspecific differences and effects of nitrate, silicate and light-dark cicles. Journal of Phycology 17: 374-384.

3.2. Cell rupture for preparation of organelles, proteins and DNA

As already shown in Figure 3.6 *Nannochloropsis* is embedded in a thick cell wall and moreover a system of four membranes engulfs chloroplast and nucleus in a continuum of membranes connected to the endoplasmic reticulum. This complex membrane system is a trace of the two endosymbiotic processes that generated the *Eustigmatophytes*. In order to purify nucleic acids and proteins as well as organelles it is necessary to brake or remove these envelopes without degrading the macromolecules of interest. A number of protocols were test to establish a procedure for DNA and RNA purification that will be described in the next paragraphs.

Sonication

Cultures were concentrated by centrifuging and cell pellet was resuspended in buffer. Half of the sample was centrifuged again and two separated fractions were obtained: pellet and supernatant. The remaining half was sonicated as described in 'materials and methods' and, after sonication, sample was centrifuged and two fractions were again separated: pellet and supernatant. The four obtained fractions were solubilized in SB buffer and loaded on SDS-PAGE in order to quantitate the amount of soluble proteins extracted by sonicating the cell suspension. As shown in Figure 3.11, system was effective in extracting proteins and we did not found any evidence of degradation in the gel. Procedure was therefore repeated and total extracts of *N.gaditana* were separated by SDS-PAGE, blotted and tested for interaction with different antibodies specific for the photosynthetic apparatus. The immunodecoration analysis was carried out in order to collect preliminary information about the characteristics of the photosynthetic apparatus of this microalgae, which is distantly related to the model organism *C.reinhardtii*. The most interesting finding was the detection of a specific interaction with the antibody raised against the LI818 (LhcSR) protein of *C.reinhardtii*. This protein is a member of the light-harvesting complex superfamily involved in photoprotection and was recently found also in the diatom *T.pseudonana* (Zhu and Green, 2010) which is more closely related to *N.gaditana* than the green algae *C.reinhardtii*. LI818 seems to be responsible for activation of the response to changes in illumination, which is mainly mediated by PsbS in higher plants. The immunoblot is shown in Figure 3.11 in Panel C.



Panel A SDS-PAGE, Coomassie stain, loading: LMW markers; cellular pellet before sonication; supernatant before sonication; pellet after sonication; supernatant after sonication; LMW markers

Panel B Western blot on nitrocellulose membrane, Ponceau Red stain, loading: LMW markers; supernatant after sonication

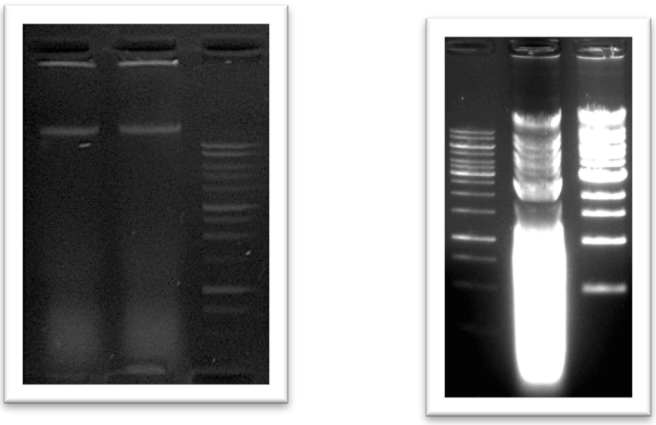
Panel C Immunodecoration using antiLi818 antibody, revelation by alkaline phosphatase, loading: LMW markers; supernatant after sonication

Figure 3.11 Cell rupture by sonication and total cellular extract analysis by SDS-PAGE. Fractions obtained after sonication were solubilized in SB buffer and loaded in SDS-PAGE. In Panel A the effective yield of cell breakage was checked loading pellet and supernatant before and after sonication. The majority of the proteins are found in the pellet before the sonication and in the soluble supernatant after the treatment. Panel B is a stained western blot membrane where total cellular extracts were loaded after sonication, for further immunodecoration analysis. Nitrocellulose membrane was stained using Ponceau red. Panel C shows the signal obtained from alkaline phosphatase revelation of the interaction between *N.gaditana* proteins and the anti Li818 antibody.

DNA was also purified from the cellular extracts obtained after sonication as described in 'materials and methods'. Obtained DNA was further treated using RNAse, in order to remove residual RNA eventually present in the preparation, and loaded on agarose gel electrophoresis for checking the quality. Reaction was carried out following the manufacturer recommendations. A parallel control experiment was carried out where a mixture of DNA ladder and RNA was incubated with the RNAse solution in the same conditions applied to the DNA purified from *N.gaditana*, in order to have a positive control on the RNAse activity and exclude the presence of contaminant DNAse activities. The results are shown in Figure 3.12 where in Panel A the DNA preparation from *N.gaditana* is visible and in Panel B the control is reported. In both the lanes (pre and post treatment) of the gel showed in Panel A, a smear of low molecular weight DNA is visible, probably due to mechanical fragmentation or degradation during the purification phase. The presence of contaminant RNA can be indeed excluded since the RNAse was active in the tested conditions and there wasn't any evidence of unspecific DNAse activity in the control experiment.

As a conclusion sonication was effective in breaking the cells and proteins were not disturbed, as expected, by the treatment with ultrasounds. DNA extracted from these samples was anyway fragmented to a certain extent into low molecular weight pieces and was not considered suitable for library preparation and sequencing. Moreover the obtained DNA was a mixture of nuclear, chloroplastic and mitochondrial genomes and the frag-

mented molecules found in solution could not be suitable for further purification of the three genomes based either on size or on density.



Panel A loading: DNA purified from N.gaditana; DNA purified from N.gaditana and treated with RNAse A; 1Kb DNA ladder PROMEGA

Panel B loading: 1Kb DNA ladder PROMEGA; 1Kb DNA ladder plus RNA; 1Kb DNA ladder plus RNA treated with RNAse A

Figure 3.12 Agarose gel electrophoresis of DNA extracted after sonication. DNA was extracted from cellular lysates after sonication and loaded on agarose gel electrophoresis prior and after RNAse treatment (Panel A). In Panel B a control was loaded, where a parallel RNAse treatment was carried out on a solution of DNA and RNA to check for effective activity of the enzyme and the absence of contaminant DNAses activities in the solution. A smear of degraded or fragmented DNA is visible in the extraction from *N.gaditana* that could not be attributed to unspecific DNAse activity in the solution or partial removal of contaminant RNA.

Covaris

The Covaris acoustic transducer operates at 500 kHz with a wavelength of ~1 mm, unlike conventional sonics which have a wavelength of ~100 mm. The low frequency energy of the sonicator is not focused due to its longer wavelength, resulting in scattering and uncontrollable energy transfer that produces local increases in temperature and macromolecules degradation. In contrast, AFA wavelengths are much shorter and scaled to the process, allowing energy to be focused into a localised area of a sample and avoiding secondary drawbacks. Covaris was therefore applied to the cell pellets in order to break the cell wall and release the cellular content avoiding DNA fragmentation. Different protocols were applied for cell breakage and results were tested again on SDS-PAGE electrophoresis after solubilisation of the obtained fractions in SB buffer (Figure 3.13). As evidenced in the picture, the amount of released material was too poor, even after the application of the strongest protocol permitted.

Panel A Intensity of treatment= 2

Panel B Intensity of treatment= 6

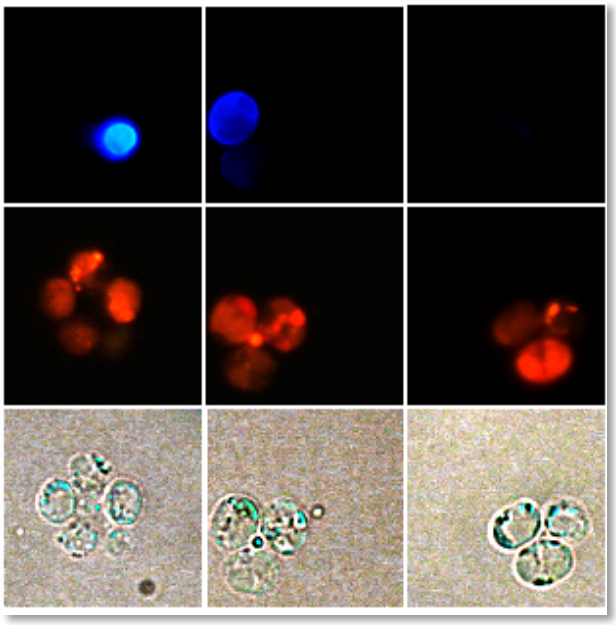
Panel C Intensity of treatment= 10

Figure 3.13 SDS-PAGE analysis of total cellular extracts after cell rupture using Covaris. Treatment intensity was 2, 6 and 10 in Panel A, B and C respectively. The loaded lanes from left to right in all the gels were respectively pellet and supernatant after the treatment carried out for 2, 5, 10, 15 and 20 minutes. Proteins were visualized by silver staining he gel. The amount of proteins extracted and found in the soluble fraction grew progressively from intensity 2 to 10, but even in the strongest set up, the amount of material extracted accounted just for a small fraction.

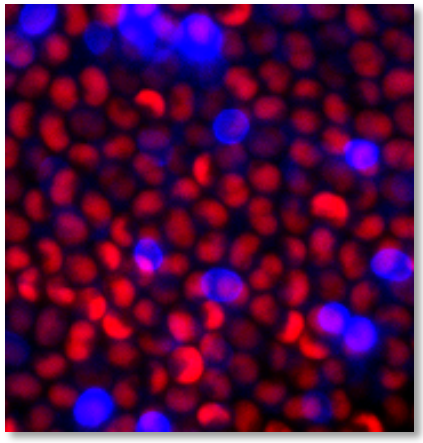
Further nucleic acid purification was not even attempted due to the scarce yield of cell breakage while applying the Covaris technique on *N.gaditana*.

Protoplast generation

Cell wall digestion followed by mild membrane solubilisation was tried in order to purify the organelles and extract intact genomic DNA from each of the separated organelles. Cellulase and macerozyme were applied to cell wall degradation. Choice of the enzymes was done in agreement with the expected composition of algal cell wall and according to the results obtained by cell wall staining using FB28. Different digestion protocols were tested and we checked the results by imaging the treated samples using a fluorescent microscope after staining with FB28. Results are shown in Figure 3.14. Digestion was never complete but cell wall was no longer visible in the majority of the cells. Cells were harvested by centrifuging after treatment and the pellet was subjected to staining and observation. The presence of green material in the supernatant suggested a partial degradation of the protoplast in the solution, in spite of the presence of manitol to keep the osmolarity. Parallel control experiments were always carried out staining and checking a sample of cells before the enzymatic treatment in order to have a reference of the staining efficiency. The number of protoplasts and whole cells was counted in control and treated samples, in order to have an approximate estimation of the protocol efficiency. As it can be noticed in Figure 3.14 the assignment of a cell as protoplast or whole cell implied a certain degree of arbitrariness. Moreover unstained whole cells could also represent a portion of the population of cells assigned as protoplasts.



Panel A magnified details of three groups of cells after enzymatic treatment for cell wall removal. Top: stained cell wall (emission in the blue channel); middle: chlorophyll autofluorescence; and bottom: bright field imagine.



Panel B colour combine of blue and red emission. Cells giving the sole red signals are protoplasts, while cell yielding both a blue and a red emission are whole cells with partially digested or integral cell wall.

Figure 3.14 Fluorescent microscope imaging of *N.gaditana* after enzymatic treatment for cell wall digestion. Cells were centrifuged after treatment to remove enzymes and free organelles, stained using FB28 and imagined through a fluorescent microscope registering the blue and red emission and taking a bright field image as a reference. A colour combine of a dense sample is shown in Panel B while in Panel A three groups of cells are shown in major detail. Digestion was effective although not complete. As evident from Panel B a certain arbitrary is necessary for counting the protoplast visible in the sample.

Sample	Percentage of protoplasts
Control	35 ± 20
Enzyme treated	70 ± 15

Table 3.1 Protoplast counts in control and treated samples. Given numbers are an average of three independent experiments performed following the procedure that yielded the most promising results. As evidenced by the great uncertainty of the value, yield was very variable in the different experiments.

Variability of the results could be probably attributed to the variable conditions of the input cultures, as suggested by the counting in Table 3.1, where the range of values obtained in the control experiments was also very variable. Cells were harvested every time during the logarithmic phase, but we did not perform any further control on the cultures. Nevertheless the procedure proved effective in removing the cell wall for prompting further purification of the desired cellular fractions. Enzymes used for cell wall digestion were anyway supplied lyophilized from a partially purified preparation and thus they could not represented a guaranty for the purification of *N.gaditana* DNA free of contaminants for the ultra deep sequencing. We had therefore some concern about the application of the described procedure to nucleic acid extraction. We extracted anyway the genomic DNA and we checked the preparation for quality and concentration using Nanodrop, Qbit and agarose gel electrophoresis. Moreover we tried to quantify the relative amounts of nuclear and plastidial DNA by PCR amplification of the ribosomal genes localized exclusively in each of the two genomes. Portions of the genes coding for 18S and 16S ribosomal RNA

were amplified using primers specifically designed on *N.gaditana* available sequences. Results are shown in Figure 3.15, where the PCR products obtained after 15, 20, 25, 30 and 35 cycles were loaded in order to obtain an approximate estimation of the amount of template present in the preparation. Both the two genomes were present, apparently in equimolar ratio.

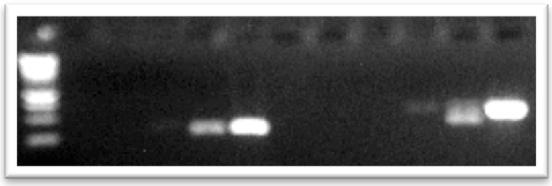


Figure 3.15 PCR amplification of 18S and 16S ribosomal genes in DNA extraction from *N.gaditana*. Agarose gel electrophoresis of PCR amplificates. Loading: 1Kb DNA ladder NEB, 18S amplificates after 15, 20, 25, 30 and 35 cycles; empty well; 16S amplificates after 15, 20, 25, 30 and 35 cycles. The two templates seem to be present in the preparation in equimolar ratio.

We tried a further organelles separation, after solubilisation of the obtained protoplasts, in order to try a separate purification of the plastidial and nuclear genomes. Results are described in the next paragraph.

Finally cell wall digestion following the procedure described in this paragraph resulted extremely useful for chromosomal DNA preparation in agar plugs for analysis via pulsed field gel electrophoresis.

Chloroplast purification

Chloroplast purification was performed according to the protocol published by Mason and coworkers in 2006 for the preparation of intact chloroplasts from *C.reinhardtii*. Protoplasts were checked by fluorescent microscopy after enzymatic treatment and we proceeded with breaking the cell-wall-deficient cells by passage through a narrow syringe needle and we purified the chloroplasts by differential centrifugation followed by discontinuous Percoll gradient centrifugation. The separation that we obtained is shown in Figure 3.16, two discrete green bands were found where isolated chloroplast and still intact protoplast were expected to settle. We harvested the two fractions and checked them by fluorescent microscopy again after staining with FB28. As shown in Figure 3.16 both the two fractions contained intact protoplasts and whole cells, while purified chloroplast were not visible. On the other hand, it was not possible to control the preparation through specific reaction with marker antibodies since antibodies raised against nuclear or cytoplasmic proteins of *Nannochloropsis* are not available. Forced passage through a narrow needle was clearly not sufficient for breaking the envelope of membranes that engulfs both nucleus and chloroplast. Separation of the protoplast population into two discrete fractions, with different densities, was probably due to the diverse amount of lipids accumulated in the various cells that produces an effect on the cell density. We did not further stain the two fractions to check for the presence of lipid bodies for testing our hypothesis. Nevertheless it would be surely interesting to perform this experiment. We hypothesis indeed that one of the functions of the lipid bodies, accumulated in microalgae in response to nutrient deprivation, is the regulation of cell density in order to induce a movement of the cells inside the column of water in order to reach more nutrient rich regions.

We tried then a solubilisation of the membranes prior to load on the percoll gradients. Different detergents were tested and a gradient of detergent concentrations was tried. The result was either a completely inefficient removal of the membrane system or a successful solubilisation of all the membranes including the ones surrounding the organelles and therefore not suitable for organelles purification. We therefore decide to proceed with total nucleic acid extractions and to try a separation of the genomic DNA of nucleus and chloroplast by centrifugation through caesium chloride gradients.

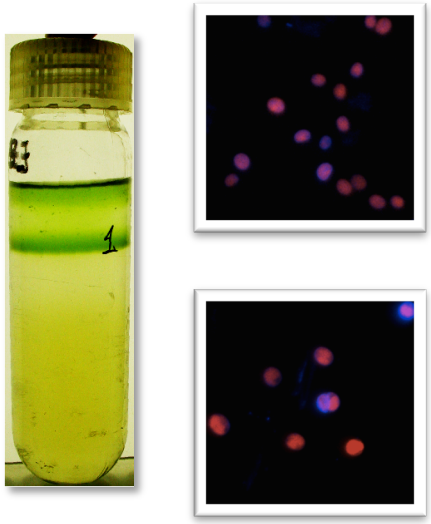


Figure 3.16 Percoll gradient separation of the fractions obtained after cell wall digestion. Samples obtained after cell wall digestion wee loaded on a percoll gradient and the two fractions obtained were checked by fluorescent microscopy after staining using FB28 (colour combine shown on the right).

Grinding in liquid nitrogen

Whole cells were harvested by centrifuging, liquid supernatant was carefully removed and dry pellet was flash frozen in liquid nitrogen. The frozen samples were grinded to a fine powder in a mortar always keeping the temperature low by adding liquid nitrogen. Quartz powder was added to help mechanical fragmentation. This system is very effective in fracturing cell wall and breaking to a certain extent the envelope of membranes, while keeping undisturbed nucleic acids and proteins. It is anyway necessary to perform a proper solubilisation if purification of macromolecules has to be performed. In order to purify nucleic acids, high concentration of detergents was used in order to completely solubilize the membranes, preserving at the same time the integrity of high molecular weight DNA fragments in the dense solution. Strategy proved successful for the purification of both DNA and RNA and further details about the preparations will be given in the next paragraphs. Obtained DNA included nuclear, plastidial and mitochondrial genomes as shown in Figure 3.17 and Figure 3.18, we therefore tried to separate the genomes by caesium chloride gradient centrifugation.

Caesium chloride gradients

Genomic DNA preparations, obtained after grinding the whole cells in liquid nitrogen and solubilisation in high concentration of detergents, were subjected to caesium chloride gradient centrifugation as described in 'material and methods'. We could not relay on any information about the DNA sequence while planning our gradients, we could not know, for example, weather the chloroplast genome was rich in GC and therefore denser, as it

was found in other organism. We therefore decided to apply a procedure based on the BAC purification protocol, where BACs are assumed to be circular and intact, while genome to be fragmented to a certain extent. Being circular the BACs will result denser than the genomic DNA after interaction with ethidium bromide. BAC size is usually around 120Kb. We expected our chloroplast genome to be approximately 100-150Kb and to be intact in our purification. Gradients were run and two discrete bands were visible that settled where genomic DNA and chloroplast were expected according to the applied protocol. Bands were harvested and purified from caesium and ethidium bromide. Purified bands were used as a template for PCR amplification of the ribosomal genes 18S and 16S. The eukaryotic and prokaryotic genes for ribosomal RNA were used as markers for the nuclear and plastidial genomes respectively. The genome content of the two purified bands was compared to that of the initial genomic DNA preparation. Results are shown in Figure 3.17 and Figure 3.18. Both the 18S and 16S genes were present in all the three examined preparations while the amplificates were proved to be specific PCR products due to the absence of detectable amplificates in the negative controls. Semiquantitative PCRs were performed in order to provide a relative quantification of the templates in the three samples. PCRs were sat up in 100µl and aliquots were sampled after 10, 15, 20 and 25 cycles to be loaded on the gel. Appearance of a detectable band after a different number of amplification cycles provides a relative estimation of the specific template concentration in the examined sample. We expected to detect enrichment in 18S and therefore nuclear genome in one of the two bands obtained after caesium chloride gradient and 16S (chloroplast genome) in the other, in comparison to the initial purification. The result is clearly visible in Figure 3.18: the three preparations harbour the same relative amounts of nuclear and plastidial genomic copies, with the two purified bands being more diluted than the initial preparation. Such a result might be addressed to the mechanical fracture of the circular plastidial genome and the settling of the different linear fragments of both nuclear and plastidial DNA according to their base composition. DNA preparation protocol could not be further modified. The need for strong treatments for cell wall breaking and membranes solubilisation had to be coupled to the preservation of intact macromolecules. We therefore decided to carry out the libraries preparation for genomes sequencing, using the purified DNA harbouring nuclear plastidial and mitochondrial DNA all together and to work on the data elaboration in order to recover all the useful information and to produce a draft assembly of all the three genomes.

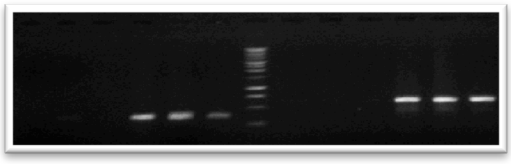


Figure 3.17 Agarose gel electrophoresis of PCR amplificates of 18S and 16S genes, using total DNA extractions and caesium chloride purified fractions as templates. In the gel 6 PCR amplificates of 18S gene were loaded followed by the 1Kb DNA ladder and 6 amplificates of the 16S gene. The 6 PCR amplificates of each gene were obtained using the following templates: mQ water, *Synechocystis* genomic DNA, *E.coli* genomic DNA, *N.gaditana* genomic DNA extraction, light band separated by caesium gradient centrifugation and heavy band separated by caesium gradient centrifugation. All the amplifications were carried out for 35 cycles. PCR amplificates were specific and all the three preparations from *N.gaditana* contained both nuclear and plastidial DNA.

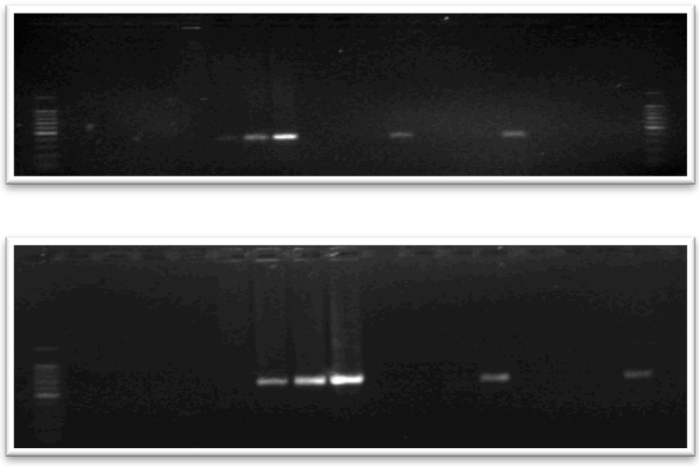


Figure 3.18 Agarose gel electrophoresis of semiquantitative PCR amplificates of 18S and 16S genes, using total DNA extractions and caesium chloride purified fractions as templates. In the gel 4 PCR amplifications of the 18S gene were loaded on the top part of the gel between two lanes of 1Kb DNA ladder. The bottom part shows the 4 amplifications of the 16S gene loaded in between the 1Kb markers. The 4 PCR amplificates of each gene were obtained using the following templates: mQ water, *N.gaditana* genomic DNA extraction, light band separated by caesium gradient centrifugation and heavy band separated by caesium gradient centrifugation. All the amplifications were loaded after 10, 15, 20 and 25 cycles. PCR amplificates were specific and all the three preparations from *N.gaditana* seem to contain the same amount of nuclear and plastidial DNA.

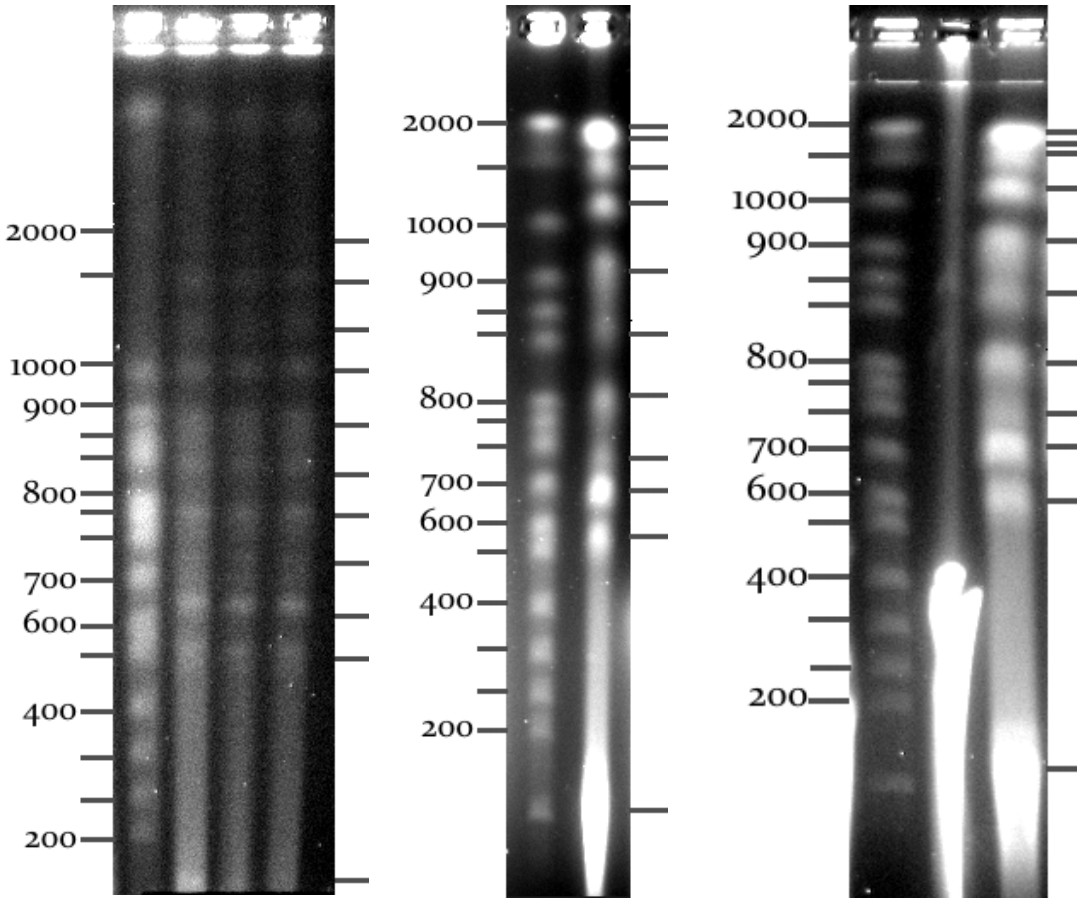
References

Mason C B, Bricker T M and Moroney J V (2006) A rapid method for chloroplast isolation from the green alga Chlamydomonas reinhardtii. Nature Protocols 1, 2227 – 2230

Zhua S H and Green B R (2010) Photoprotection in the diatom Thalassiosira pseudonana: Role of LI818-like proteins in response to high light stress. Biochimica et Biophysica Acta-Bioenergetics 1797(8): 1449-1457

3.3. Electrophoretic karyotyping

N.gaditana has been used for quite a long time in industrial applications and has been intensely studied in the recent period for cultivation in photobioreactors. Nevertheless it is a completely new organism for molecular biology and biochemistry, as it can be gathered from the previous paragraphs. Despite being a microorganism of certain interest, due to its position in the phylogenetic tree, it has not been extensively studied so far and we lack the more basic information as well as the experimental procedures. When we decided to sequence the genome of *N.gaditana* the size of its genetic complement was also unknown. The sole information that we found in the literature was a measure due to Veldhuis and coworkes (Veldhuis et al 1997) that, according to the values of fluorescence, detected after staining the nucleus with fluorescent dyes, produced an estimate of the DNA content of the examined cells. They indicated an estimated genome size between 30 and 40Mb for the three species of *Nannochloropsis* examined. We therefore decided to purify the intact chromosomes of *N.gaditana* and to produce an electrophoretic karyotyping in order to obtain a more reliable estimate of the genome size and to assign a number of chromosomes to our organism of interest. Since the number and the size of the chromosomes was completely unknown we started to explore the regions that could be covered with accuracy using the commercial chromosomes preparations. Chromosomal preparations from *S.cerevisiae* (range from 2000Kb to 200Kb), *H.wingei* (range from 3000Kb to 1000Kb) and *S.pombe* (range from 6000Kb to 3000Kb) were used. Electrophoretic run parameters were adjusted to the set up suggested by the supplier. While we could not detect the presence of chromosomes at molecular weights higher than 3000Kb, 10 bands were clearly visible in the region comprised between 200Kb and 2000Kb, as shown in Figure 3.19. Slightly different running parameters were applied in order to better focus the visualized bands, as shown in the three Panels of Figure 3.19. The overall molecular weight of the bands detected accounts for approximately 12Mb, a genome size similar to that of *S.cerevisiae*, which is also a unicellular eukaryotic microorganism. Nevertheless we cannot exclude the presence of more than one chromosome per visible band that could justify a genome size up to 3 times bigger. Despite our efforts for better resolving each of the bands to obtain a more precise estimate of number and size of the single chromosomes, we did not succeed in improving the resolution of our PFGE. This was most probably due to a scarce quality of the chromosomes preparations, that we could not further improve, in spite of the application of a number of protocols found in the literature and adjusted for our sample. On the other hand the natural chromosomes size could be actually very similar and the visualized bands represent the best resolution allowed by the applied experimental procedure.



Panel A Loading: *S.cerevisiae* chromosomes preparation; *N.gaditana* chromosomes preparation progressively diluted in the three lanes.

Panel B Loading: *S.cerevisiae* chromosomes preparation; *N.gaditana* chromosomes preparation.

Panel C Loading: *S.cerevisiae* chromosomes preparation; genomic DNA preparation and chromosomes preparation from *N.gaditana*.

Figure 3.19 PFGE of N.gaditana chromosomes. Three pulsed field gel electrophoresis are presented in this picture run with slightly different parameters but always focused in the region comprised between 200 and 2000 Kb. *S.cerevisiae* chromosome preparation was used as a molecular weight marker. In all the three analysis 10 chromosomes were identified accounting for approximately 12Mb and the chloroplast genome was visible at the bottom of the gel. Nuclear genome size was estimated as approximately 30Mb. The region at higher molecular weights was also explored but we could not find any further chromosome. We therefore deduce that a number of bands visible in the PFGEs presented contains more than one chromosome and was not properly resolved.

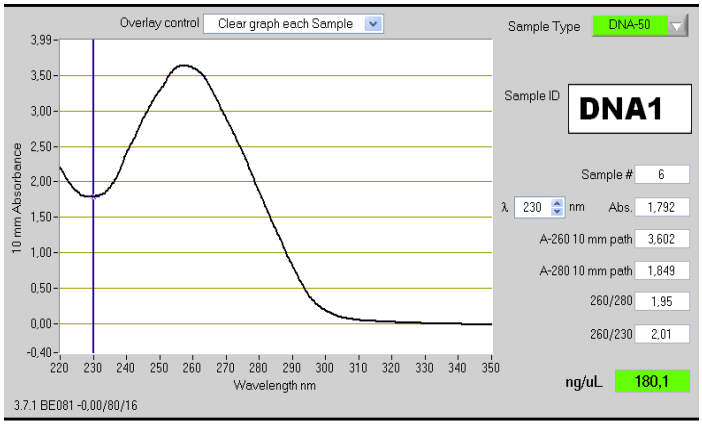
References

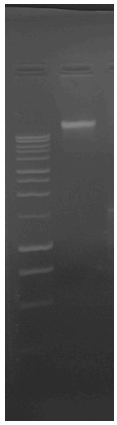
Veldhuis M W, Cucci T L and Sieracki M E (1997) cellular DNA content of marine phytoplankton using two new fluorochromes: taxonomic and ecological implications. J. Phyml. 33: 527-541

3.4. Nucleic Acid Extraction and sequencing

Genomic DNA purification

We purified the genomic DNA for sequencing applications following the procedure described in 'materials and methods'. Cells were broken by grinding in liquid nitrogen and nucleic acids were extracted after solubilisation in high detergents concentration. This protocol typically yielded between 100µg and 150µg of pure DNA per litre of culture in exponential phase. The introduction of a proteinase K treatment before phenolic extraction resulted critical for the yield of the preparation, while the sole phenol extractions were already sufficient for an efficient deproteinization. High molecular weight RNA was removed from the preparation by lithium chloride precipitation, while the small RNA and tRNA, that are not efficiently precipitated by lithium salts, were digested using RNAse A when necessary. DNA was assessed for quality and concentration by spectrophotometric measurement. On average, our preparation yielded 260/280 ratios around 1.9, meaning a protein contamination smaller than 30%. 260/230 ratio was used as a secondary measure of DNA purity, since EDTA, carbohydrates and organic compounds all absorb at 230nm. Values above 2 were routinely registered. A second measure of the concentration was taken using the Qbit fluorometer, that allows removing the unspecific signals due to free nucleotides and RNA that contribute to the 260nm peak registered by the spectrophotometer. The average of four Qbit measures of the sample shown in Figure 3.20 was 175ng/µl, indicating the absence of degradation and RNA contamination. The result was confirmed by gel electrophoretic analysis of the samples, where a unique high molecular weight non resolved band was always visible.





Panel A Concentration and quality parameters obtained by Nanodrop measurement.

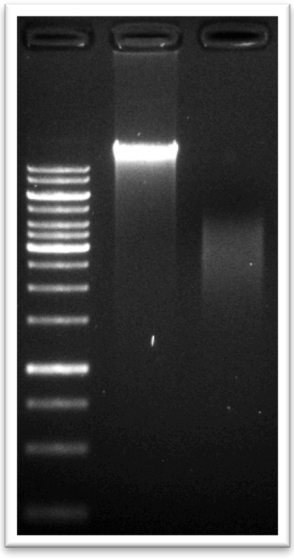
Panel B loading: 1Kb ladder; genomic DNA from *N.gaditana.*

Figure 3.20 Spectrophotometric and electrophoretic profiles of a genomic DNA purification from *N.gaditana*. Preparations routinely obtained had a DNA to protein ratio (measured as 260/280) above 1.9 and a 260/230 around 2. Measured concentration was usually consistent with that obtained from Qbit if RNA was completely removed during preparation. Absence of detectable amounts of RNA was confirmed by gel electrophoresis (showed in Panel B). The purified genomic DNA yielded a high molecular weight non-resolved band above the 10Kb DNA marker band.

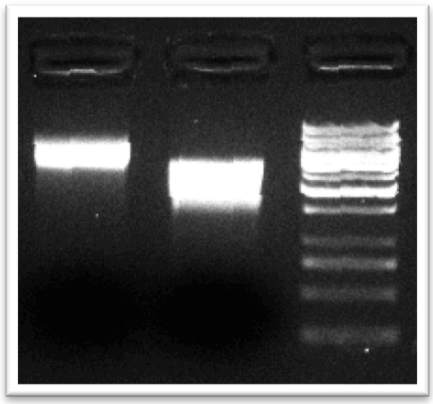
These genomic DNA preparations were used for library construction and sequencing using the Roche 454 and the SOLiD. High purity was not strictly required but strongly recommended for the success of the library preparation procedure.

SOLiD mate-pairs library preparation

Two mate-pairs libraries were realized for sequencing using the SOLiD 3plus system. The two libraries were constructed to have an average distance size between the two mate-pairs of 4 and 2 respectively. Genomic DNA was fragmented using the Hydroshear and we obtained a smear comprised between the 5Kb and 2Kb shown in Figure 3.21, in Panel A. Fragmented DNA was precipitated to concentrate and then further purified by column chromatography, since some impurities could be detected by spectrophotometric measurement after fragmentation. DNA was end repaired in order to obtain blunt ends and purified by column chromatography prior to proceed with adaptors ligation. End repairing procedure was repeated twice in order to improve the yield of the step, which proved critical in our experience. SOLiD adaptors were ligated at the two extremities of the DNA fragments and we used column chromatography purification to remove buffer and enzyme from the solution. We then run a preparative gel electrophoresis to separate the DNA fragments present in the solution according to their size. Two bands were cut from 5Kb to 3Kb and from 3Kb to 1.5Kb respectively. This step allowed a further size selection of the fragments in order to obtain a sharper size distribution of the distances between the two mate-pairs. DNA extracted from the gel was assessed by gel electrophoresis and result is shown in Figure 3.21 in Panel B.



Panel A Gel electrophoresis of genomic DNA before and after shearing. Loading: 1Kb DNA ladder Generuler; genomic DNA preparation form *N.gaditana*; sheared DNA.



Panel B Bands obtained after by size selection and purification. Loading: 5-3Kb band; 3-1.5Kb band; 1Kb DNA ladder Promega.

Figure 3.21 Gel electrophoresis of sheared and further size selected DNA fragments, after binding of the adaptors. Genomic DNA was sheared and purified. Results were checked by gel electrophoresis (Panel A). The smear of fragmented genomic DNA was centred on the 3Kb marker band and was detectable between the 5Kb and the 2Kb reference bands. Obtained smear was ligated to the SOLiD adaptors. DNA fragments obtained after adaptor ligation were purified from gel electrophoresis in order to remove the residual unligated adaptors. Band excised was further cut in two bands and purified. Result is shown in Panel B. The procedure yielded two sharper smears and thus a narrower interval of distances between the mate-pairs of the library.

The amount of DNA recovered after each of the steps is reported in Table 3.2. It was of the maximum importance to keep under control the quantities through the procedure, in order to reach the circularization step with a sufficient amount of DNA. Circularization is the most important and most delicate step of the library construction, allowing to bring together the two extremities of the DNA fragments, distant in our case from 5Kb to 1.5 Kb, and to ligate them to a special adaptor harbouring the sequencing primers. The step has a poor yield and the amount of circularized DNA obtained is crucial to carry on the library construction and to obtain a library with low redundancy values.

Step during the preparation	Total amount of DNA (µg)	Recovered DNA
Purified genomic DNA	80	100 %
DNA purified after shearing	60	75 %
DNA purified after two rounds of end repairing	41	51 %
DNA purified after adaptor ligation	35	44 %
DNA purified from gel (sum of the two libraries)	19.5	24 %
Library 5-3 Kb	7.5	9 %
Library 3-1.5 Kb	12	15 %

Table 3.2 Amount of DNA recovered after each of the initial steps prior to proceed with circularization.

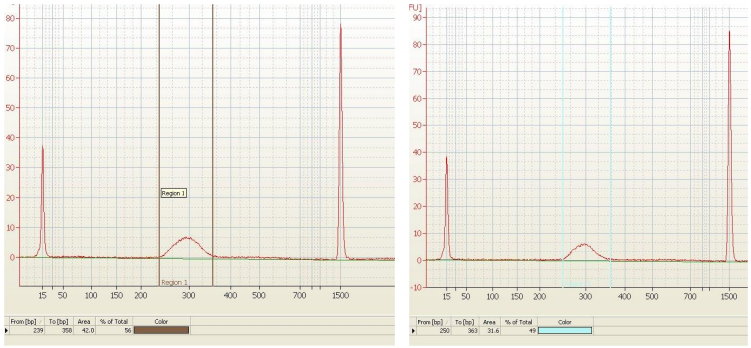
Library 5-3Kb

In the mate pair procedure the extremities of the DNA fragments that successfully ligated a sticky-end SOLiD adaptor are ligated at the two opposite extremities of a small adaptor during circularization. Ligation, catalysed by a common T4 ligase, leaves a nick in one of the two strands after ligation. The nick can be "translated" a number of bases downstream using a reaction under kinetic control and this same nick, after translation, will be the starting point for directional DNA degradation. The procedure was carried out in order to obtain approximately 50 bases of each of the extremities bound to the central adaptors. The amount of DNA obtained after degradation of the exceeding part and purification was 1µg for this library. The extremities of the obtained DNA constructs were repaired to obtain blunt ends and two adaptors were ligated at the two extremities, necessary for library amplification, emulsion PCR and selection of the correct construct for sequencing reaction on the SOLiD slide. 10 amplification cycles were necessary for library amplification prior to proceed with the emulsion PCR step. The amplified library was run on preparative gel electrophoresis in order to purify from the residual primers. 120ng of purified library were obtained and a small aliquot was run on Agilent chip in order to confirm the correct size of the obtained constructs (Figure 3.22, Panel A).

Library 3-1.5Kb

For the small library we obtained 3.2µg of DNA constructs after circularization, nick translation and directional degradation of the DNA. 8 amplification cycles were necessary in this case for library amplification prior to proceed with the emulsion PCR step. After

primers removal via in gel size selection and purification, approximately 100ng of library were obtained. Agilent analysis confirmed the correct size distribution of the obtained library as shown in Figure 3.22, Panel B.



Panel A Library 5-3Kb **Panel B** Library 3-1.5Kb

Figure 3.22 Agilent analysis of the amplified libraries prior to proceed with emulsion PCR. The two vertical lines, in different colours in the two panels, indicate the interval of sizes where the amplified libraries should lye. The obtained distributions were correct in both cases.

SOLiD sequencing

Emulsion PCR was performed using the two prepared libraries as templates in order to produce DNA coated beads that could be deposited on the SOLiD slide for sequencing. The enriched beads were quantified and checked by WFA, which is a small scale run performed prior to load the proper sequencing run, in order to confirm the quality and the quantification of the beads. We performed the proper sequencing on a SOLiD slide that was divided in four quadrants using the apposite mask. The two libraries were loaded each in two quadrants referred to as A and B. The output result of the sequencing is illustrated in Table 3.3, where the number of reads per quadrant is reported together with the quality parameters of the obtained sequences. A total of 14.5 gigabases was sequenced, realizing a theoretical coverage of 500X, assuming the genome of *N.gaditana* as big as 30Mb.

Sequence set	Number of reads	Minimum	Maximum	Average	Median
3-1.5Kb A	74.749.807	3.9	28.6	15.3	15.3
3-1.5Kb B	69.418.621	4.9	29.9	17.7	18.2
5-3Kb A	68.334.726	4.9	30.3	18.1	18.9
5-3Kb B	78.164.673	4.8	29.9	17.4	18.0

Table 3.3 Number of reads per quadrant and quality parameters. Total amount of reads obtained per quadrant is reported in the second column. Values of quality were extrapolated for a subset of 90.000 reads and some of the parameters are reported in table: minimum registered quality, maximum quality obtained, average quality and median of the quality distribution.

The quality of the bases was also plotted as a function of the position of each of the sequenced base on the read. Plot is shown in Figure 3.23. As expected the quality decreases going from the first to the 50th sequenced base. The quality decrement was measured as approximately 35% from the first to the last position, and it is on line whit the expectations on the SOLiD system performances declared by the manufacturer. The loss of in-

formation at the end of the reads could be partially restored by the application of a specific informatics tool evolved by Life Science for the recovery of the missing colours from SOLiD reads. When a read colour is not assigned, a dot is found in the read that can be rebuilt in a number of cases from the whole dataset. The process is called spectral correction and the programme used was 'saec_solid_mp'. As expected the plot of the differences between the initial reads and the spectral corrected ones, as a function of the position indicated a broader application of the correction to the bases found in the terminal positions. Nevertheless the application of the spectral correction did not produce a dramatic change in the data that were then used for assembly.

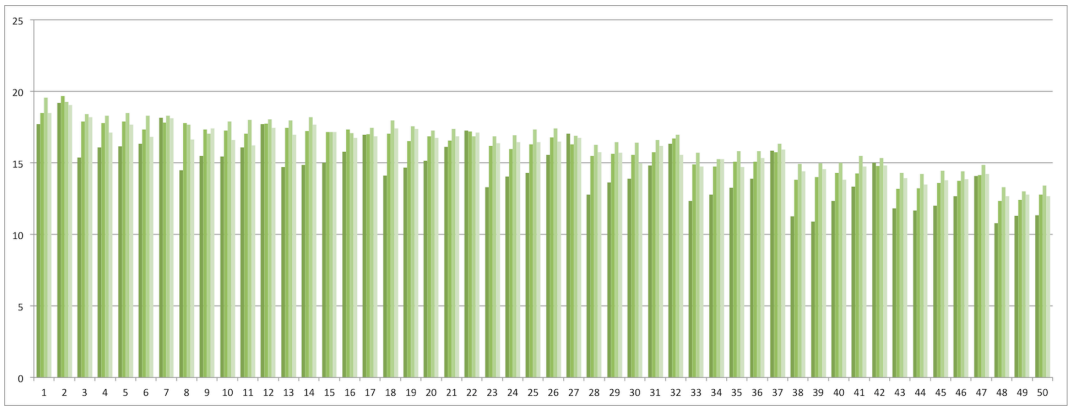


Figure 3.23 Average quality at each position of the sequence. The four bars at each position represent each of the four libraries. The quality decreases as expected going from the first to the last sequenced base. The quality decrement is approximately of 35%.

454 sequencing of genomic DNA fragments libraries

Being *N.gaditana* a completely new organism for genomic research, and due to the absence of a sequenced genome that could be used as a reference for the assembly of the short reads, we decided to realize a 454 run. The number of bases sequenced per run by the 454 was indeed much smaller, but the average size of the reads, 400bp, allowed a much easier assembly of the first genome draft. A fragments library was realized by BMR Genomics, starting from 25µg of purified DNA. The library was sequenced using a full slide. Approximately 250 megabases were sequenced, realizing a theoretical coverage of 8X. The length distribution of the obtained reads was plotted and it is shown in Figure 3.28. The distribution curve was centred on 400bp although imbalanced towards smaller reads sizes.

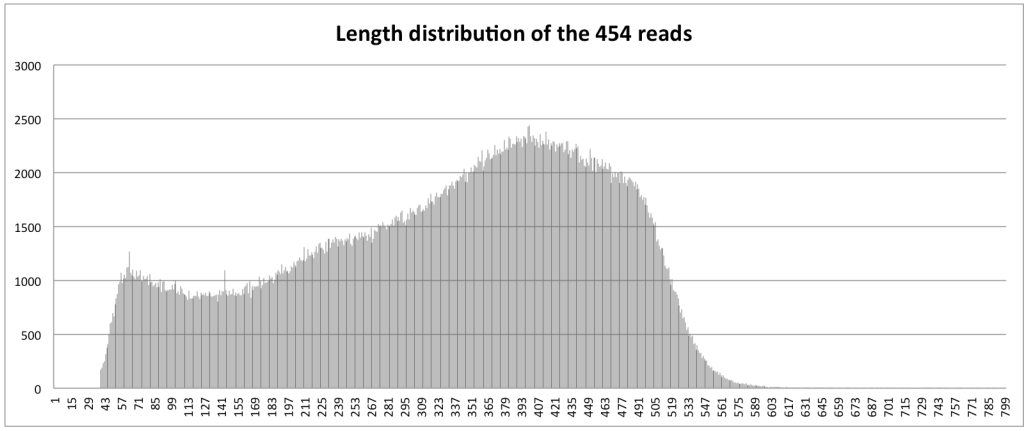


Figure 3.24 Length distribution of the reads obtained by 454 sequencing. On the x axis read length in base pairs (bp) while on the y axis number of reads. Distribution curve is centred on 400bp and is imbalanced towards the smaller reads size.

Assembly

As already mentioned, we decided to apply a hierarchical approach to the assembly of the genome of *N.gaditana*, the general scheme is summarised in Figure 3.22. The relatively long reads produced by the 454 sequencing could be first assembled into a robust population of contigs. The high coverage of short mate-pairs reads could be then aligned to the contigs and the aligned instances could be classified into groups. The most interesting groups for assembly are the unique pairs in, the unique pairs out and the unique single pairs. The definition 'unique pairs in' designs the population of pairs that align only in one position on the contigs and where both the mate-pairs are aligned to the same contig. While these sequences are of no use for the bridging of contigs together into scaffolds, they are of fundamental importance for the experimental measure of the average distance between the two mate-pairs of the sequenced libraries. The obtained distance distributions are important for the estimation of the gap size between two bridged contigs. The 'unique pairs out' indeed are all those instances that align in a unique position on the contig population and have the two mate-pairs aligned on different contigs. The alignment of a sufficient number of mate-pairs on two distinct contigs constitute and evidence of the vicinity of those two contigs in the genome and allows the assembly of the two contigs into a scaffold. A gap in the sequencing is present between the two consecutive contigs of the scaffold. The gap size can be estimated according to the distance distribution produced by the alignment of the 'unique pairs in'. Moreover a third population was mentioned, the 'unique single pairs' that includes the subset of instances where one of the mate-pairs aligns on a unique position in the contigs population while the other does not align at all. If a sufficient number of 'unique single pairs' aligns on one of the extremes of a contig, all the respective not aligned mate-pairs can be used for a *de novo* assembly. The new small contigs assembled using the short reads help, in many cases, filling the gap between two contiguous contigs in the scaffold. If the number of obtained scaffolds is sufficiently small they represent already a genome draft. In order to improve the draft or to reach the goal of a finished genome, scaffolds can be further mapped into chromosomes using the physical map as a reference. We decided to produce a physical map of the genome of *N.gaditana* using a modified version of the 'Happy Mapping' approach (Dear at al. 1993; Jiang et al. 2009; Vu et al. 2010). A BAC library was produced using the purified DNA from *N.gaditana*. BACs were purified and pooled in subpopulations containing less than 50% of the genome. Pools were then digested using a restriction enzyme with a frequency of cut of around 300bp and digestion products were selected by specific ligation to a biotinylated adaptor for recovery. After fragmentation each subpopulation was tagged and sequenced using the SOLiD v4 system. The work is still in progress, nevertheless this approach will produce a number of tag sequences ordered on the chromosomes and will therefore provide a reference physical map for the assembly of our scaffolds into chromosomes. On the other hand, the set up of a protocol for pulsed field gel electrophoresis separation of the chromosomes, allows also the assignment of the scaffolds and the mapped scaffolds to the singles chromosomes, which are at least 15.

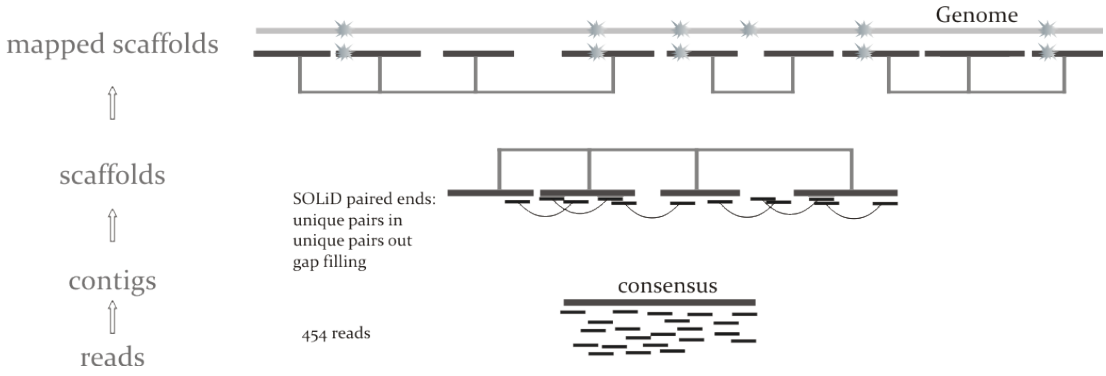


Figure 3.25 General scheme of the hierarchical approach for assembly using next generation sequencing. 454 reads are first assembled into contigs. Mate-pairs sequences are then used to put together the contigs and fill the gaps between the ordered contigs. Obtained scaffold can be further assembled into chromosomes with the help of a physical map.

454 sequences were assembled using the 'Newbler' assembler provided by Roche. Parameters were adjusted to work with low coverage sequencing. The different Newbler versions, released during the last year, were every time tested on our sequences, in order to check if the starting assembly could be improved. The application of the different versions did not produce substantial differences. About 93% of the produced sequences were included in the assembly, indicating a scarce frequency of low complexity or repeated regions. About 150Kbases were assembled in 6 contigs with coverage higher than 150X: 4 large contigs (> 30Kb) and two smaller contigs (1Kb and 2.5Kb respectively). The high coverage contigs were blasted against the NCBI database and were found to contain the genes coding for the prokaryotic ribosomal operon, the rubisco large subunit (RBCL), the psaA subunit of Photosystem II and the cytochrome oxidase subunit I (COX I). Moreover the single contigs showed extended similarity to either mitochondrial or plastidial genomes of sequenced organism such as the *Heterokontophyta H.akashiwo* and the diatom *T.pseudonana*. The single contigs could be then assigned as plastidial or mitochondrial, nevertheless they were not sufficient to assemble the full genome drafts of the two organelles that remain incomplete. At the moment, an analysis was carried out using the data obtained after the alignment of the SOLiD mate-pairs on these high coverage contigs, suggesting that at least one of the large contigs should be broken in more than one point, since high similar sequences incorrectly collapsed in the same contig during the assembly. The breakage of the contig might help completing the assembly of the two organelles' genomes, which have a high enough coverage. The remaining sequences were assembled in 11700 large contigs with an average coverage of 7X, the average contig size was 2500, while 50% of the contigs were bigger than 4500bp.

SOLiD data were first trimmed and filtered and then aligned on the produced contigs, a summery of the results obtained is provided in Table 3.4. After spectral correction, sequences were filtered for quality and trimmed in the same step by "PASS-PAIR" (Campagna et al. 2009). The trimming process consists on a shortening of the reads, after removal of the low quality portion of each read while the rest of the sequence is recovered. This system allows avoiding filtration of sequences with an average low quality when the low quality portion is localized, and the rest of the sequence has a quality above the fixed threshold. Following this strategy the loss of useful information is reduced and at the same time the sole high quality sequences are kept for a robust alignment. In our case, as shown in Table 3.4, approximately 21% of the reads were eliminated. Of the remaining

reads 67% was successfully aligned on the contigs, making about 55% of the initial reads obtained from sequencing.

Sequence set	Filtered	Aligned	% Aligned (filt)	% Aligned (tot)
1.5-3 kb For (A)	13290103 (24%)	33375006 / 54921894	60.77%	49%
1.5-3 kb For (B)	13439413 (24%)	38013535 / 55918324	67.98%	55%
1.5-3 kb Rev (A)	16014482 (31%)	35682949 / 52197515	68.36%	52%
1.5-3 kb Rev (B)	10420041 (18%)	41376522 / 58937696	70.20%	60%
3-5 kb For (A)	10159400 (16%)	41476953 / 64534508	64.27%	56%
3-5 kb For (B)	16538367 (27%)	39692841 / 61522574	64.52%	51%
3-5 kb Rev (A)	10340714 (16%)	46457338 / 64353194	72%	62%
3-5 kb Rev (B)	10728512 (16%)	46194717 / 67332429	68.6%	59%
Average	21%		67%	55%

Table 3.4 Number of filtered reads for each set and percentage of alignment. For each of the sets of sequences the number of filtered reads, the number of aligned reads and the percentage of aligned reads are reported in the table. Percentages are calculated as, either percentage of reads obtained after filtering that was aligned (% aligned filt), or percentage of the total reads obtained after sequencing that was successfully aligned on the contigs (%aligned tot).

Aligned reads were further classified into groups, as reported in Table 3.5, and the 'unique pairs in' were used to estimate the mate-pairs distance distribution. A plot of the size distribution of all the contigs (including the very small ones, that were considered below the threshold) is provided in Figure 3.26 where it is compared to the average distance between the mate-pairs of the two SOLiD libraries.

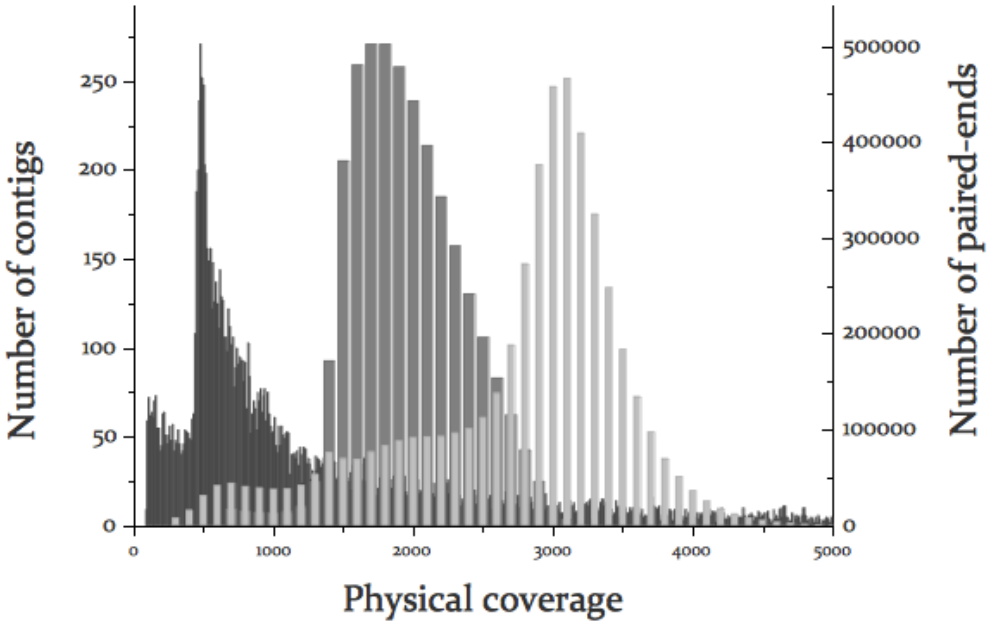


Figure 3.26 Length distribution of the three sequences clusters used for scaffold production. In the x axis length is given in bp while in the y axis the number of instances is reported. In dark grey length distribution of all the contigs produced by the Newbler assembly is plotted. The actual distribution included instances much longer than 5000 that were not included in this plot. In grey, distribution of the distances between the mate-pairs of the SOLID library 3-1.5Kb, while in light grey the same distribution is reported for the 5-3Kb library. Distributions of the mate-pairs distances were plotted for a subset of 5000000 unique pairs in for each of the libraries.

As clearly shown in Figure 3.26 a number of contigs presents a size equal or bigger than the distance between the mate-pairs, thus allowing to calculate and plot the distributions of the distances between the mate-pairs of the SOLiD libraries shown in the graph. The majority of the contigs instead has a size smaller than the distance covered by the mate-pairs, thus making the libraries rather useful for their connection into scaffolds. The use of two short distance mate-pairs libraries makes it possible to obtain a fine resolution ordering of the short contigs, produced by the low coverage 454 sequencing. Moreover, the distance distribution of the two SOLiD libraries is only partially superposed and therefore allows, on one hand, to confirm some of the connections using the two independent datasets, while, on the other hand, to cover a broader distance range avoiding information redundancy. Together with the 'unique pairs in' mentioned in this paragraph, the 'unique pairs out' were also obtained. The number of sequences obtained for each group of mate-pairs is reported in Table 3.5, where the groups interesting for our assembly are evidenced in bold. The 'unique pair out' aligned to every single contig were used by "Divorce" for generating the scaffolds. Divorce is a tool evolved by Andrea Telatin in our group, and not yet published, that tacking as input the list of unique pairs out generated by both the two libraries and aligned at each of the extremes of a contig, expands the connection toward the next contig until it can find a number of 'unique pairs out' above the fixed threshold to support the process. Parameters were adjusted to obtain a reliable and robust result while avoiding to push the scaffolding further at the expenses of accuracy. A subset of the scaffolds was manually assembled in parallel to check the reliability of the program and some of the junctions were amplified by PCR to confirm the ordering of the contigs generated by the program.

Mate-pairs groups	3-1.5Kb	5-3Kb
Unique pairs in	**24 594 353**	**24 362 389**
Not unique pairs in	118 112	61 026
Unique wrong d	319 769	99 258
Unique wrong s	12 949	13 161
Not unique wrong d	4 201	2 913
Not unique wrong s	8 893	7 030
Unique single	**47 812 440**	**55 014 700**
Not unique single	5 748 316	4 078 892
Unique pairs out	**47 242 584**	**66 847 510**
One not unique pairs out	9 255 622	7 499 808
Both not unique pairs out	503 061	430 776
Discarded pairs	297 631	276 325

Table 3.5 Mate-pairs groups. Mate-pairs were grouped according to the alignment of each of the paired sequences. The groups that are important for the assembly are evidenced in bold in the table.

After obtaining a first set of scaffolds, the SOLiD mate-pairs were aligned again, this time on the scaffolds, and a classification of the couples of pairs analogous to the one showed in Table 3.5 was obtained. Some of the sequences that aligned in more than one position

on the contigs, yielded unique alignments on the scaffolds subset, and could be then included in the second run of the procedure for connecting together the scaffolds. 'Divorce' was indeed run a second time. The result obtained is shown in Figure 3.27. 167 large scaffolds were generated including 18.7Mb of the nuclear genome, 50% of the scaffolds were bigger than 50Kb with the biggest one being 1Mb. We defined 'large scaffolds' all the connected sequences of contigs containing more than 10 contigs, irrespectively to their size. As evidenced by the graph shown in Figure 3.27 the first 20 contigs included already 10Mb. The plot reveals the generation of a number of big scaffolds accounting for 1/3 of the hypothesized 30Mb genome, followed by a number of relatively small scaffolds accounting for the remaining 1/3 included in the large scaffolds subset. Approximately 13Mb of nuclear genome were still contained in the remaining small scaffolds and contigs that were not incorporated into the large scaffolds category.

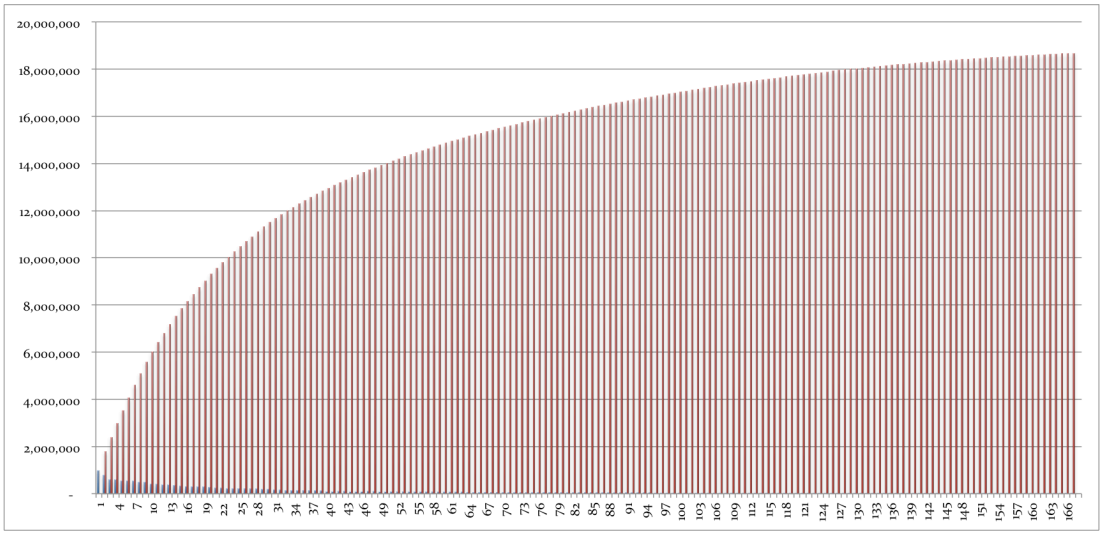
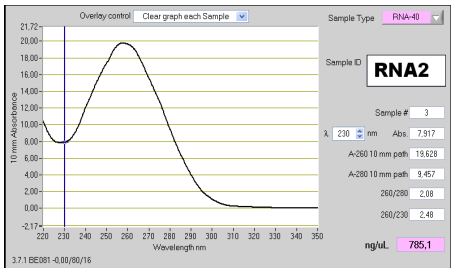


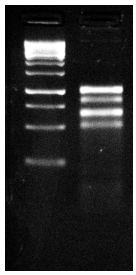
Figure 3.27 Plot of the 167 large scaffolds ordered according to their size. In the x axis the ordering number of each contig, while in the y the number of bp is reported. Each of the scaffolds is represented by a bar in the graph. Scaffolds are ordered from the largest to the smallest as clearly shown by the blue bars. In blue, bars represent the number of bases included in each of the scaffolds, while in red the bases included are progressively summed going from the largest to the smallest scaffold. As it is visible 10Mb were already included in the first 23 scaffolds, while the remaining 8.7Mb were included in approximately 140 smaller scaffolds. In this graph the sole large scaffolds were represented.

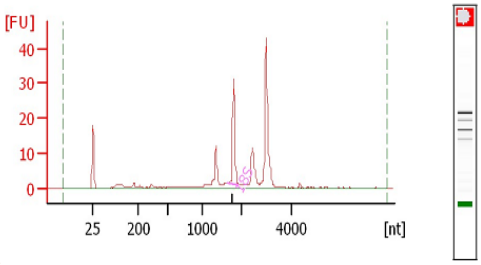
RNA purification

We purified the RNA for sequencing applications following the same procedure reported for genomic DNA preparations and described in 'materials and methods'. When we performed RNA preparations we always used phenol equilibrated in acid buffer for all the steps of the purification. RNA was isolated from the total nucleic acid preparations by precipitation in lithium chloride. The protocol typically yielded between 250µg and 400µg of pure RNA per litre of culture in exponential phase. RNA was assessed for quality and concentration by spectrophotometric measurement. On average, our preparation yielded 260/280 ratios around 2.1, meaning a protein contamination smaller than 30%. 260/230 ratio was used, also in this case, as a secondary measure of RNA purity, and values above 2.3 were routinely registered. A second measure of the concentration was taken using the Qbit fluorometer, that allows removing the unspecific signals due to free nucleotides and contaminant DNA that contribute to the 260nm peak registered by the spectrophotometer. The average of four Qbit measures of the sample shown in Figure 3.28 was 653ng/µl, indicating the absence of degradation and the presence of DNA contamination. The result

was confirmed by gel electrophoretic analysis of the samples, again shown in Figure 3.28, where the 4 bands of ribosomal RNA (rRNA) were clearly resolved, but high molecular weight non resolved bands were also visible in the majority of the cases, accounting for the contaminant genomic DNA. Integrity of the RNA was routinely checked by Agilent bioanalyzer and a typical profile obtained for the total RNA is shown in Figure 3.28, Panel C. Only the rRNA could be detected as expected. As already mentioned the visible bands are four since both the eukaryotic (coded in the nucleus) and the prokaryotic (coded in both chloroplast and mitochondrion) rRNA are expressed. Sharpness of the bands provides an indication of the integrity of all the RNA in the sample. We never spotted evidences of degradation in our preparations.







Panel A Spectrophotometric analysis of purified RNA using Nanodrop.

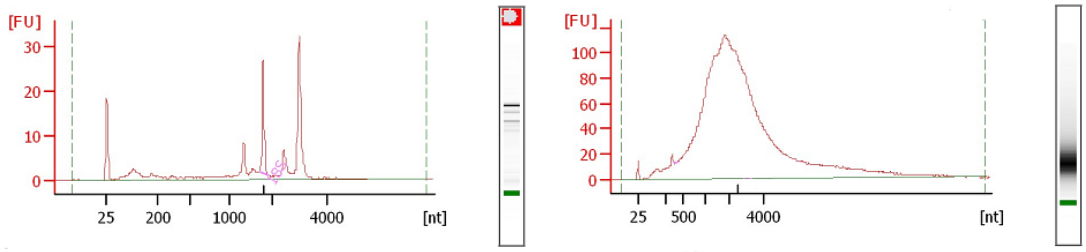
Panel B Loading: 1Kb DNA ladder, total RNA extracted from *N.gaditana*

Panel C Agilent bioanalyzer nano chip. Total RNA from *N.gaditana* was loaded. X axis values are given in nucleotides while in the y axis the intensity of the signal is reported. Eukaryotic and prokaryotic rRNA are visible.

Figure 3.28 Total RNA isolation form *N.gaditana*: spectrophotometric and electrophoretic analysis.

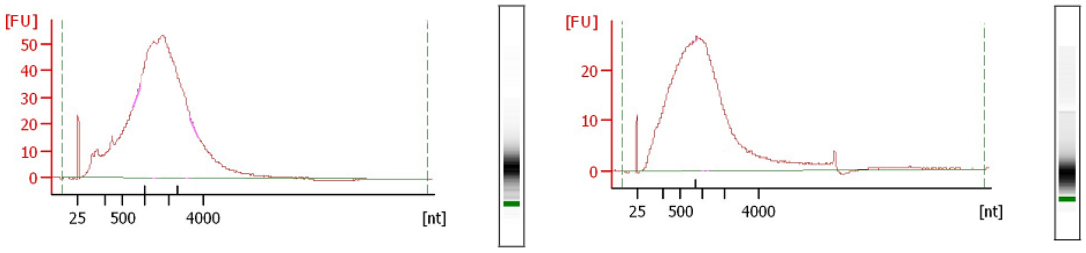
Transcriptome analysis: annotation and 5'capturing

Ultra deep sequencing is a fantastic tool for genome annotation, since the high number of short reads provides both the coverage and the resolution to annotate with a high accuracy the expressed regions of the genome. Exons and introns can be assigned at single base resolution and differentially expressed exons as well as alternative transcription start sites can be spotted in the sample. Moreover the SOLiD platform allows obtaining strain specific information, which is fundamental for an accurate annotation of the genes coded in the two strains and most of all for the correct quantification of the expression of the different exons, when the genes coded in the two opposite directions partially overlap. In order to produce these data and make sense of the sequenced genome, we grew our microalgae in two different mediums, we harvested the cells at different stages of the culture growth and we pooled all the samples before extracting the RNA. This procedure allowed us to increase the number of expressed genes, since the sample was a pool of algae in different conditions. Total RNA was extracted (profile is shown in Figure 3.29 in Panel A) and messenger RNA (mRNA) was purified hybridizing its polyA tail to polyT-coated-beads as described in 'materials and methods'. Profile of the purified mRNA is shown in Figure 3.29, in Panel B. rRNA was no longer visible, distribution was centred, as expected, around 3000bp, and there was no evidence of degradation. An aliquot of the sample was further processed following the SOLiD protocol for library preparation and sequencing and the results obtained will be described in the next paragraph.



Panel A Total RNA from *N.gaditana.*

Panel B PolyA+mRNA from *N.gaditana.*



Panel C PolyA+mRNA from *N.gaditana* after deCAPping.

Panel D PolyA+mRNA from *N.gaditana* after adaptor ligation at the 5' end.

Figure 3.29 Analysis of the samples produced for transcriptome sequencing using the SOLiD4 platform. The same agilent bioanalyzer nano chip was used for running the four samples. Samples in Panel B, C and D are represented using a logarithmic scale for the x axis values. X axis values are given in nucleotides while in the y axis the intensity of the signals is reported.

The remaining aliquot of the purified mRNA was further processed prior to proceed with library preparation. A careful look at the typical graphs obtained after SOLiD sequencing while plotting the number of aligned sequences per nucleotide, reveals a systematic loss of information while approaching the 5' end of a coding region. Correct assignment of the transcription start is the prerequisite for further studying the promotorial regions involved in transcription activation and regulation. Moreover annotation must be precise if switching of the transcription start site is involved in regulating gene expression. In order to avoid the common loss of information and to improve our annotation we decided to set up two strategies for 5' capturing. The first procedure that we applied was based on removal of the protecting CAP, which is found at the 5'end of all eukaryotic transcripts, using a pyrophosphatase. Removal of the CAP should increase the yield of SOLiD adaptors ligation at the 5' end of the transcript, increasing the coverage of the region after sequencing. Sample was prepared starting from the isolated mRNA shown in Panel B of Figure 3.29 and was check for integrity by agilent bioanalyzer. Result is shown in Panel C of the same figure. Since we captured the mRNA using its 3' polyA tail and mechanical fragmentation might occur to the sample during the preparation, loss of coverage at the 5' extreme might be due to the preferential recovery of the 3' fragments that are bound to the magnetic beads, while the corresponding 5' fragments are lost in the supernatant. In order to tag the real 5', independently from their coverage, and promote a correct genome annotation we applied a procedure described in 'materials ad methods' consisting in a removal of the 5' terminal phosphate from all the fragmented mRNA, followed by a removal of the protecting CAP from the intact full length transcripts that leaves a phosphate at the ex-

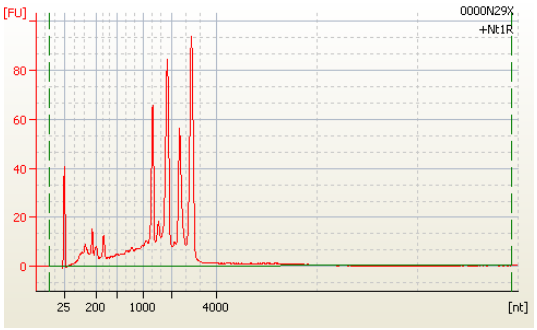
treme, and finally followed by the specific ligation of a RNA adaptor to the 5' ends of the intact mRNA. The adaptor could be used as a tag for the identification of the real 5' ends even though the technique used for mRNA isolation produces a systematic loss of coverage at the 5'. The tagging system does not produce alterations of the single mRNAs abundance in the sample and it is effective even at low efficiency, since few tags are enough to assign the tagged site as a transcription start. The sample obtained by applying the described procedure to the mRNA sample shown in Panel B in Figure 3.29, is shown in Panel D of the same figure. A slight decrease in the average size of the transcripts was registered. This was probably due to the mechanical fragmentation during the increased number of passages necessary for sample production.

Transcriptome analysis: differentially expressed genes involved in lipids accumulation

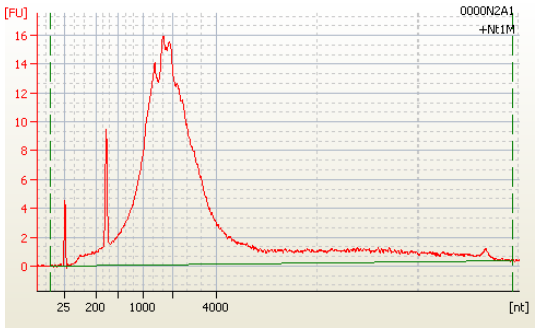
As discussed in the introduction, one of the features that make *N.gaditana* an organism of special interest for both biology and biotechnology, is the capacity for accumulation of a conspicuous amount of lipids in response to a number of different stresses. Nutrient limitations and in particularly nitrogen deprivation are well studied conditions for the analysis of lipids production and we showed in this dissertation that while inducing lipids accumulation the deprivation also led to changes in photosynthetic metabolism and pigments biosynthesis. Control of the lipid metabolism at the molecular level is quite an unexplored field in photosynthetic microorganisms, and we will surely learn a lot from the study of gene expression in the conditions in which metabolism supports the accumulation of lipids. It is noteworthy, anyway, that while studying the response to nitrogen limitation we will likely register the alterations of genes involved in a number of different metabolic processes probably, but not necessary, related to each other. In order to cast a light on the alteration of gene expression while nutrients are limiting and either lipids biosynthesis is promoted or lipids degradation is blocked, we planned and realized two experiments of RNA-Seq that will be described in this paragraph.

The first and relatively simple experiment was the extraction of RNA from parallel cultures growth in the same conditions, but containing different amounts of nitrogen salts in the medium. The growth curves obtained were analogous to the one showed in Figure 3.2, Panel A, and cells were harvested for RNA extraction at the beginning of the stationary phase when the average per cell lipids content was already clearly different between the two samples. Cultures used for this experiment were grown in winter. Total RNA was extracted from both the parallel cultures and results are shown in Figure 3.30 Panels A and C. mRNA was selected as described before using the polyA tail of the eukaryotic transcripts and the agilent run for assessing the messengers preparations is visible in Figure 3.30, Panel B and C. While selecting mRNA using the polyA tail we did not have the guaranty to recover the plastidial transcripts that, according to the literature, are not usually polyadenilated. Nevertheless mitochondrial mRNAs were found to harbour a polyA tail in many organisms and we could not exclude that the messenger RNA from algal chloroplasts were also polyadenilated. This data is of special importance since the lipids metabolism is partially localized in the chloroplast and moreover the abundance of the enzyme that catalyses one of the committee steps of lipids biosynthesis, the Acetyl-CoA carboxylase (ACC), is known to be regulated at the level of the plastidial subunit in plants

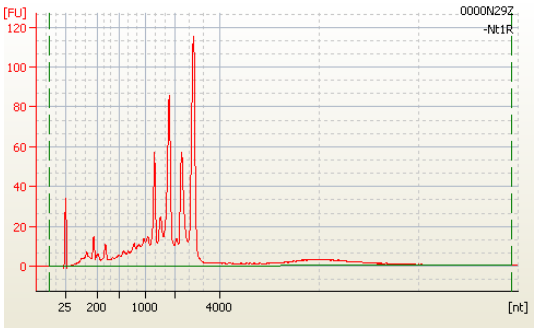
(Guschina and Harvood 2006). The characterization of the changes of gene expression in both nucleus and chloroplast is therefore of particular interest. In a recent publication concerning the study of gene expression in *C.reinhardtii* in response to nitrogen deprivation (Miller at al. 2010), the authors pointed out that they probed only the expression of nuclear genes while the study of the expression levels of genes for organelle-encoded proteins relevant to respiration or photosynthesis, could help elucidating the many interrogatives that remained unresolved concerning lipids metabolism. Furthermore the specific selection of messenger RNA leaves out all the regulatory non coding RNAs that might play an important role in regulation of gene expression. We therefore decided to attempt a second strategy based on ribosomal RNA subtraction rather than RNA selection.



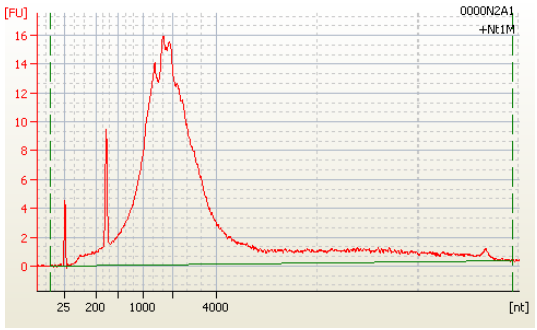
Panel A Total RNA from *N.gaditana* in nitrogen sufficient medium.



Panel B PolyA+mRNA from *N.gaditana* in nitrogen sufficient medium.



Panel C Total RNA from *N.gaditana* in nitrogen depleted medium.

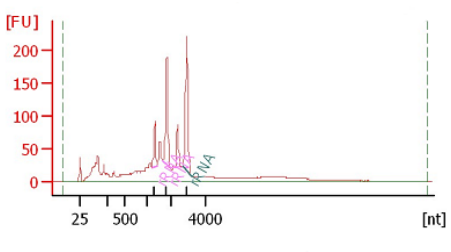


Panel D PolyA+mRNA from *N.gaditana* in nitrogen depleted medium.

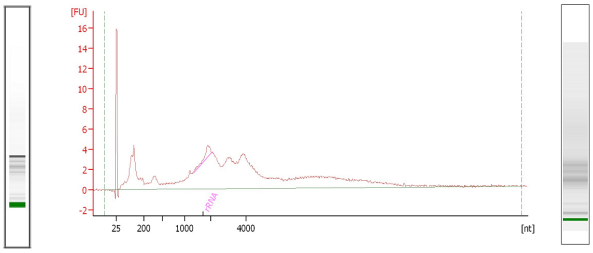
Figure 3.30 Analysis of the samples produced to study the differential gene expression using the SOLiD4 platform. The same agilent bioanalyzer pico chip was used for running the four samples. X axis values are given in nucleotides while in the y axis the intensity of the signals is reported.

Cultures were sat up during summer in nitrogen sufficient and nitrogen depleted medium and yielded growth curves similar to the one shown in Figure 3.2, Panel B. Cells were harvested at the end of the logarithmic phase from both the parallel cultures when there was no evidence of lipids accumulation and a slight decrease in chlorophyll and carotenoid content could be already registered in the nitrogen depleted samples. Samples were pooled and RNA was extracted. After a few days, cultures reached the stationary phase and lipids accumulation was evident in both the parallel cultures. Nevertheless lipid content was enhanced in the nitrogen depleted samples, that showed, in addition, the already described metabolic alterations that occur in nitrogen limitation. Cells were then

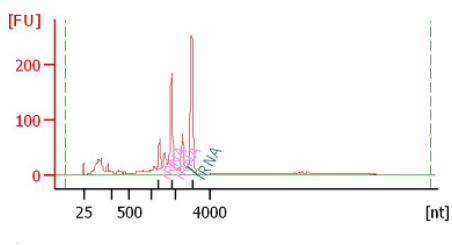
harvested from both the two parallel samples and RNA was extracted this time keeping separate the two cultures. Extracted RNA was analysed by chip electrophoresis and the obtained profiles are shown in Figure 3.31 in Panels A, C and E.



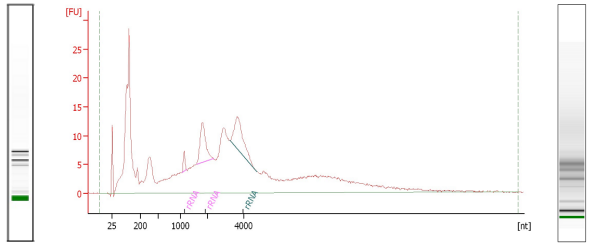
Panel A Total RNA from *N.gaditana* in nitrogen sufficient medium.



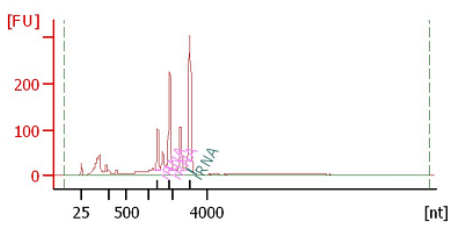
Panel B Ribominus RNA from *N.gaditana* in nitrogen sufficient medium.



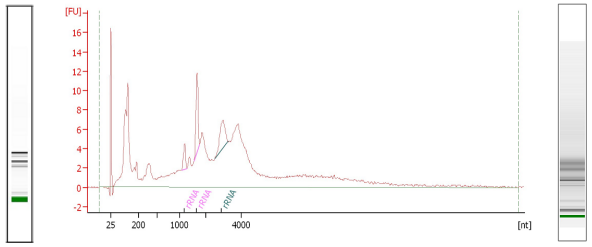
Panel C Total RNA from *N.gaditana* in nitrogen deficient medium.



Panel D Ribominus RNA from *N.gaditana* in nitrogen deficient medium.



Panel E Total RNA from *N.gaditana* pool.



Panel F Ribominus RNA from *N.gaditana* pool

Figure 3.31 Analysis of the ribominus samples produced to study the differential gene expression using the SOLiD4 platform. Agilent bioanalyzer pico chip were used for running the four samples. X axis values are given in nucleotides while in the y axis the intensity of the signals is reported. In all the three samples residual rRNA is visible and a peak at low molecular weight is also registered whose attribution is still doubtful.

In this second experiment rRNAs were subtracted from the total RNA following the procedure described by DeLong and coworkers (Stewart et al. 2010) with some *ad hoc* variations (the overall protocol is described in detail in 'materials and methods'). This procedure was chosen after testing the different strategies for rRNA subtraction or degradation described in detail in 'materials and methods'. The procedure reported here was the one that offered the best result in terms of integrity of the recovered RNA and removal of the rRNA. Yield of the different procedures was also taken into account. All the ribosomal RNAs of *N.gaditana* were amplified using specific primers designed on the sequences that we produced. Reverse primers harboured a sequence for binding of the T7 RNA Polymerase and a transcription enhancing sequence. Antisense ribosomal RNAs

(arRNA) were synthesized *in vitro* using the amplified genes harbouring the T7 recognition site as a template. During synthesis biotinylated ribonucleotides were included in the arRNA. The produced arRNA were hybridized to the rRNA of the total RNA samples at low temperature and in the presence of formamide. After hybridization the double stranded rRNA were removed using streptavidin-coated magnetic beads while the unbound fraction was recovered and used for sequencing. The 'ribominus' RNA was analysed on agilent chip and results are shown in Figure 3.31 in Panels B, D and F. As it can be seen ribosomal RNAs were not completely removed but nevertheless the samples were greatly enriched in non-ribosomal RNAs comparing to the respective initial samples, where the rRNA accounted for over 95% of the overall RNA content. A peak at low molecular weight was visible, that was anyway too sharp to be clearly interpreted as an evidence of degradation. Size distribution of the RNAs obtained was comparable to that observed in Figure 3.29 and Figure 3.30. The SOLiD slide was divided in 8 quadrants using the dedicated mask and all the described samples were directionally sequenced. A summery of the samples sequenced using the SOLiD4 system is provided in Table 3.6 prior to proceed with report of the results obtained with construction of the libraries.

Sample name	Short description
mRNA	pool of algae grown in different conditions; purified using polyA tail
deCAPped mRNA	pool of alga; purified using polyA tail; CAP removed
CAPligated mRNA	pool of alga; purified using polyA tail; adaptor ligated at the 5' of full length mRNAs
polyAmRNA+N	Nitrogen sufficient medium; no lipids; purified using polyA tail
polyAmRNA-N	Nitrogen deficient medium; lipids; purified using polyA tail
ribominus_mRNA+N	+N medium; low lipids; recovered after rRNA subtraction
ribominus_mRNA-N	-N medium; high lipids; recovered after rRNA subtraction
ribominus_mRNApool	-N and +N pool; no lipids; recovered after rRNA subtraction

Table 3.6 RNA Samples described in the current paragraph and sequenced using the 8 quadrants of a solid slide.

Paired-ends libraries preparation

Paired-ends libraries were prepared for sequencing of all the 8 samples of RNA described in the previous paragraphs. The firsts step of the 'whole library transcriptome procedure' is the chemo-enzymatic fragmentation of RNA using the RNAseIII enzyme, provided by the manufacture with a fragmentation buffer that contains salts able to promote chemical breakage of the RNA. The fragmentation system should in theory interrupt the RNA molecule stochastically, while in practise all the fragmentation systems were proven to be biased towards certain patterns. The preference for certain sites during fragmentation is in great part responsible for the irregular profiles of coverage obtained after alignment of the produced reads on the reference genome. Fragmentation protocol using the RNAseIII in its buffer is reproducible but conditions must be optimized for every single sample in order to obtain a correct size distribution of the fragments. Aliquots of the purified samples were then used to set up a number of digestion trials. Each trial was purified and assessed for size distribution using the agilent bioanalyzer. Results are not reported, but the all

procedure was carried on until at least 50ng of fragmented RNA of the correct size were produced. As summery of the amount of fragmented RNA of the correct size produced is reported in Table 3.7.

Sample name	Initial mRNA amount (ng)	Amount of correctly fragmented and purified RNA (ng)	Amplification cycles
mRNA	200	80	18
deCAPped mRNA	370	110	18
CAPligated mRNA	630	120	18
polyAmRNA+N	146	54	18
polyAmRNA-N	65	24	18
ribominus_mRNA+N	2500	50	18
ribominus_mRNA-N	1700	50	18
ribominus_mRNApool	1200	50	18

Table 3.7 Summery of the yields of the various steps during library preparation for RNA-Seq using the SOLiD4

In the paired-ends procedure RNA is fragmented to 150-200bp in order to sequence 50b at one side and 25b at the opposite extreme without overlapping. Fragmented RNA was hybridized to the SOLiD adaptor mix and then ligation was carried out according to the manufacturer instructions, followed by reverse transcription. The all of these steps were performed consecutively in the parallel samples. At the end of the protocol for cDNA synthesis, the mixture was run through a silica gel column for purification and then loaded on a precast preparative gel for size selection. After size selection and purification, the produced cDNA was PCR amplified using the apposite primers provided by the manufacturer to introduce the sequences necessary to initiate the sequencing reaction. Between 15 and 18 PCR cycles were necessary to produce the amount of DNA necessary to continue the procedure, nevertheless we decided to apply 18 amplification cycles to all the samples (as reported in Table 3.7) in order to uniform the redundancy of the libraries and obtain a more reliable data when comparing the level of expression of the single genes in the different samples. After PCR amplification a number of samples analysed by agilent bioanalyzer showed the presence of a low molecular weight band due to primers self amplification, that could interfere with the following sequencing reactions. We therefore decided to further purify the amplified libraries by performing a size selection using preparative polyacrylamide gel electrophoresis, described in detail in 'materials and methods'. Removal of the unspecific DNA was checked by assessing again the purified libraries by agilent. All the obtained libraries were used for emulsion PCR, beads enrichment, control WFA and finally sequencing. The sequencing reaction is carried out twice on the same samples, to obtain the 50b + 25b at the opposite extremes of the DNA fragments bound to the sequencing beads.

SOLiD sequencing of the transcriptome

Sequencing of the paired-ends libraries of the 8 transcriptome samples, using a full SOLiD 4 slide, yielded about 100 million sequenced tags per RNA sample. The number of reads obtained is plotted in Figure 3.32.

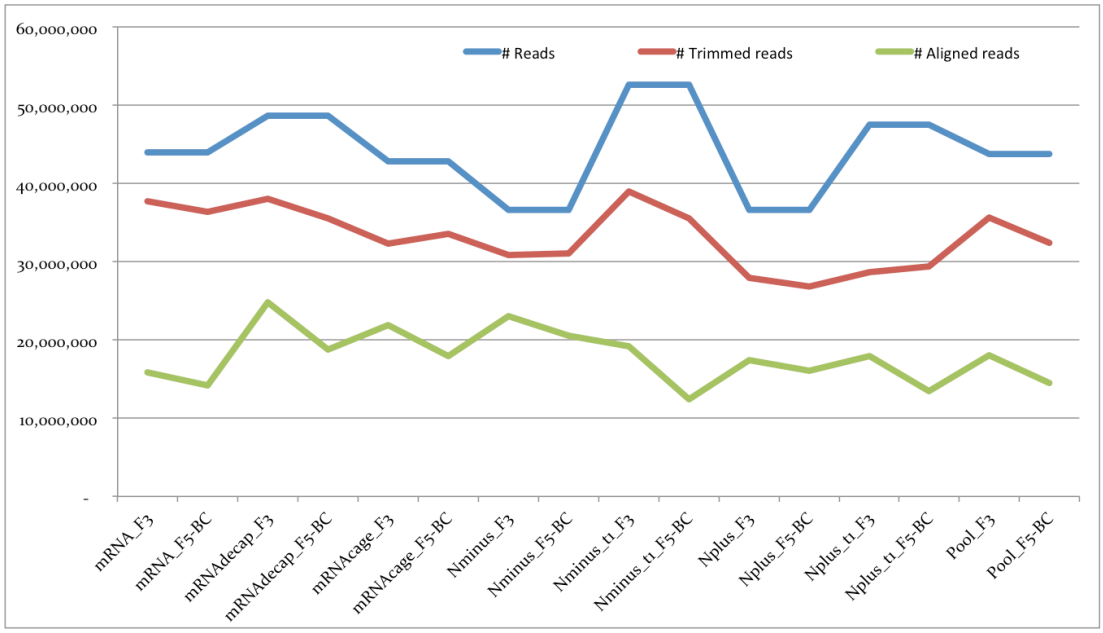


Figure 3.32 Prospect of the alignment of the SOLiD paired-ends reads of the transcriptome. In the x axis the various forward and reverse output sequences from the 8 samples indicated with a different nomenclature, in the y axis is the number of instances.

Moreover, the read pairs denominated F3 and F5BC were obtained by sequencing the two extremes of a same molecule, providing additional physical coverage and further information for genome annotation. The two extremes indeed are localized at a given distance and therefore harbour information about the structure and the structural variations of the transcriptome.

Sequences were trimmed, as already described for the DNA mate-pairs, and aligned to the large contigs. The number of trimmed and aligned reads for each F3 and F5BC sample is reported in Figure 3.32. About 35% of the reads produced were succesfully aligned. A plot of the distance distribution between the corresponding paired-ends of each sample was produced. Analogous results were obtained for all the libraries. To provide an example, the distribution of distances between the paired-ends of the library 'CAPligated mRNA' is reported in Figure 3.33. After alignment of the paired-ends to the contigs, scaffolds were produced where contigs were connected to each other by the paired-ends. The new scaffolds produced provided a further confirm for the already assembled genomic scaffolds and, in some cases, supplied the evidences for bridging the genomic scaffolds together and extending the assembly. Nevertheless it is difficult to translate in numbers the extent of the contribution of the paired-end to the scaffolding since the analysis of the data is just at a preliminary stage.

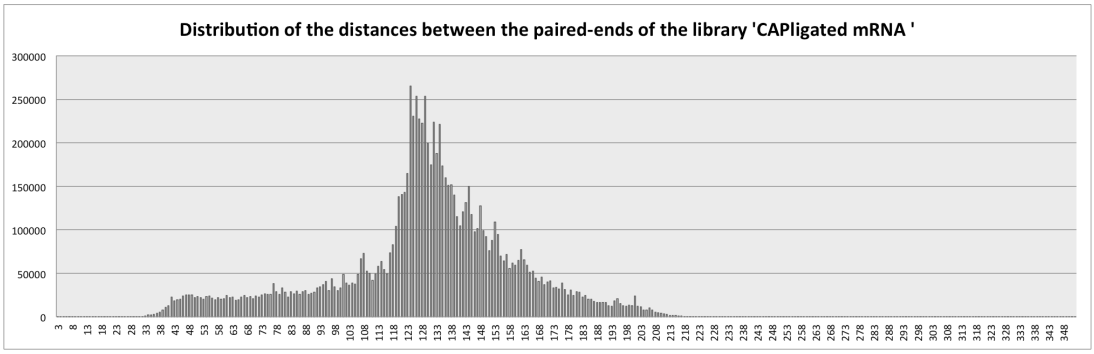
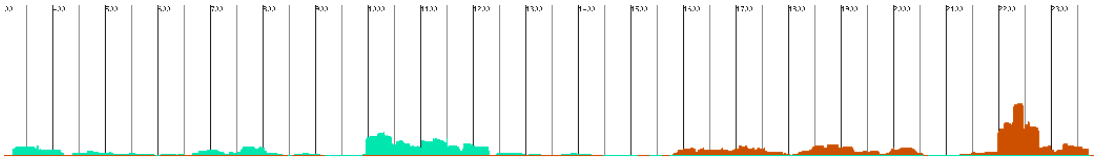
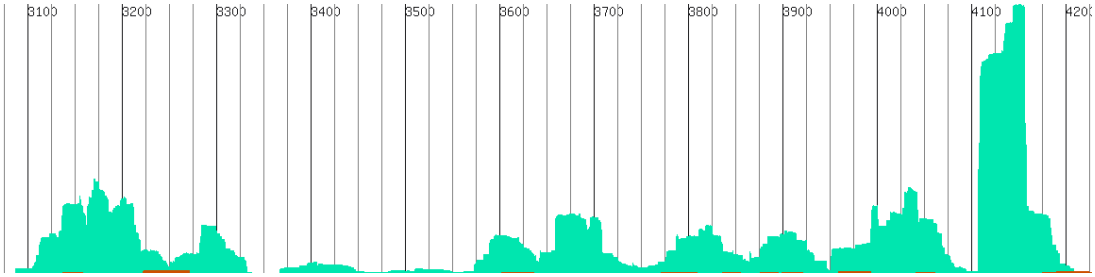


Figure 3.33 Plot of the distance between the paired-ends of a library. Distance is reported in the x axis, while number of instances for each distance unit is scaled to the y axis values.

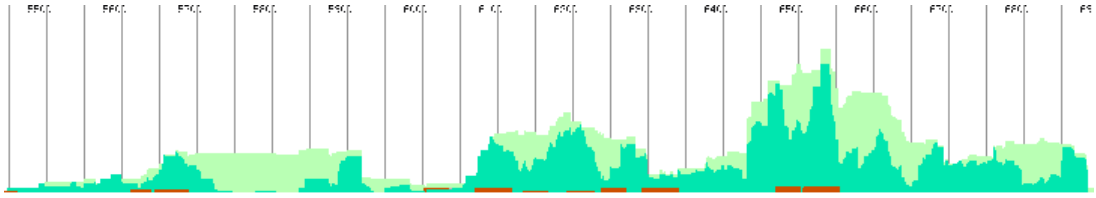
As already mentioned the analysis of the data obtained by transcriptome sequencing are at a very preliminary stadium, sequences were aligned and information concerning the correlation between the paired-ends couples was introduced. Graphs of the alignments were produced for each contig, where the number of aligned sequences for each base of the genome was plotted. The plots reported in Figure 3.34, are details of a summery of the expressed sequences found by each of the three samples sequenced for genome annotation and 5'recovery (mRNA, deCAPped mRNA and CAP ligated mRNA). Bars were scaled to the most abundant transcript found in each contig. Strand specific information allowed to annotate the expressed regions coded by the two opposite strands and to find in some cases partially overlapped genes. As it can be seen in Panel C of Figure 3.34 information concerning the sequence coverage of each base could be used together with the information about the physical coverage per base in order to produce a reliable description of the introns exons pattern, avoiding misattributions due to the loss of coverage in determinate regions. A comparison between the various samples produced for 5' recovery is not yet available and moreover we are still working on the analysis of our samples using an already tested pipeline of programmes for gene prediction and genome annotation. As a consequence a complete list of the genes identified in *N.gaditana* is not yet available for further studies. It is worth mentioning anyway that in our original experiment design we planned to sequence a full length cDNA library using the 454, in order to produce a course grain annotation of the genome and to add the information obtained using the SOLiD short reads for improving the annotation and for studying the differences in gene expression in the various samples. A full length library was produced, as described in the dedicated section in 'materials and methods'. However 454 sequencing of the library was not successful, yielding aborted reads, and was of not use for our porpoises. We worked on the trouble shouting and repeated the library preparation, nevertheless, on the mean time, alignment of the SOLiD sequences showed that the information provided by 3 quadrants of a SOLiD 4 slide loaded with mRNA libraries, was already enough for running our pipeline and obtain the desired information. Profiles analogous to those shown in Figure 3.34 were obtained also for the contigs assigned to the chloroplast and the mitochondrial genomes while aligning the mRNA obtaining by selection of the polyA tail, indicating that probably in *Nannochloropsis* at least part of the genes coded in the organelles harbour a polyA sequence at the 3' of the transcripts.



Panel A Genes coded in the two opposite strands are represented in green and red in the figure.



Panel B The coverage per base of the aligned sequences shows the modulation in expression of the different regions of the genome



Panel C In light green the physical coverage is reported together with coverage per base of the aligned sequences showed in green. The comparison of the two traces helps identifying the introns.

Figure 3.34 Coverage distribution of the aligned transcriptome sequences on the reference genome. Strand specific sequences were aligned on the contigs and coverage per base of the sequenced reads was plotted on the reference DNA. Details of the graph obtained are shown in this figure. Expressed regions coded by the two opposite strands are shown in green and red. Different levels of expression in the different regions can spotted in the graphs. Graph shown in Panel C in particular shows the identification of introns from the comparison between the traces of the physical coverage and those of the sequence coverage thanks to the information available from the paired-ends.

References

Dear P H, Cook P R. (1993) Happy mapping: linkage mapping using a physical analogue of meiosis.Nucleic Acids Res. 21(1):13-20.

Jiang Z, Rokhsar D S, Harland R M (2009) Old can be new again: HAPPY whole genome sequencing, mapping and assembly.Int J Biol Sci. 5(4):298-303.

Vu G T, Dear P H, Caligari P D, Wilkinson M J (2010) BAC-HAPPY mapping (BAP mapping): a new and efficient protocol for physical mapping.PLoS One. 5(2)

Campagna D, Albiero A, Bilardi A, Caniato E, Forcato C, Manavski S, Vitulo N, Valle G (2009) PASS: a program to align short sequences. Bioinformatics. 25(7):967-8.

Miller R, Wu G, Deshpande R R, Vieler A, Gärtner K, Li X, Moellering E R, Zäuner S, Cornish A J, Liu B, Bullard B, Sears B B, Kuo M H, Hegg E L, Shachar-Hill Y, Shiu SH and Benning C. (2010) Changes in transcript abundance in Chlamydomonas reinhardtii following nitrogen deprivation predict diversion of metabolism. Plant Physiol. 154(4):1737-52.

Guschina I A and Harwood J L (2006) Lipids and lipid metabolism in eukaryotic algae Progress in Lipid Research 45 160–186.

Stewart F J, Ottesen E A and DeLong E F (2010) Development and quantitative analyses of a universal rRNA-subtraction protocol for microbial metatranscriptomics. The ISME Journal 4(7):896-907.

4. Conclusions and perspectives

In this report we presented the sequencing of nuclear and organelles' genomes of *N.gaditana* using two single run of second generation sequencing. The nuclear genome was assembled in a good quality genome draft; nevertheless the assembly could be further improved if a proper tool was applied for the managing of all the information harboured by the SOLiD short reads. At the moment, in our group, people are working at the realization of a pipeline of programmes for the assembly of eukaryotic genomes using the complementary information obtained from the use of assembled contigs and mate-pair short reads based on the De Bruijn graph. The application of this suit of programmes will surely improve the assembly without the need for higher sequencing coverage. Moreover the realization of a reference physical map would further improve the assembly and would make it possible to assign each scaffold to a chromosome. As already mentioned in this report we produced a library of BACs, harbouring 120Mb inserts, and accounting for approximately 40X physical coverage of the nuclear genome. BACs were purified, quantified and distributed in 48 BAC-pools, each accommodating approximately 35% of the genome. Pools were checked by PCR amplification of unique regions of the genome to confirm the genome content. In order to further confirm the genome content of each pool and to quantify the amount of eventually present contaminant sequences of the host strain of *E.coli*, we realized a DNA microarray and all the pools are tested at the moment. Protocols were set up for the automation of all the steps of the procedure necessary to produce 48 bar coded SOLiD libraries, each harbouring approximately 1650otags. The statistical association of the single tags between each other produces a map of distances between the tags that represents the physical map of the genome. The programs necessary for analysis of the data were already produced and tested. Sequencing using 3/8 of a SOLiD slide while increase the sequencing coverage of the genome while providing at the same time the information necessary for the production of a reference physical map with an estimated resolution of 2500b. As it can be gathered from all the reported considerations, even though we produced the majority of the data necessary for the sequencing and assembly of the genome, the data managing is still at a preliminary phase and will need more work prior to release the genome for the scientific community. Being assembly and genome annotation still incomplete it was not possible to perform a proper analysis of comparative genomics and to produce a list of classified genes involved in the metabolism of *Nannochloropsis*. The all of these analyses will be carried on in the near future. Nevertheless, it is important to remark that the data produced are of great interest for the study of the molecular processes involved in the evolution of the secondary endosymbiosis as well as for the study of the control of the energy carbon metabolism in simple photosynthetic organisms. This project conjugated the potential of new generation sequencing techniques with the study of two important biological issues such as the evolution of the genomes of nucleus and chloroplast after the endosymbiotic process and the regulation of the basic metabolism of a cell. It is worth reminding that *Nannochloropsis* is an organism of increasing interest for the application of technologies aimed to extract biofuel after mass cultivation of microalgae. While the design of photobioreactors with different characteristics has been successfully realized in the past few years and efficient chemical processes for the

transesterification of triacylglycerols into diesel oil were realized, procedures for production of transgenic microalgae with enhanced lipids content are still lacking. While the genome of the model organism *C.reinhardtii* has been available since 2007 and protocols for efficient transformation were obtained and are widely used, a genome reference for genetic modification of *Nannochloropsis*, which is the organism of election for large scale production of biofuel, was still missing to date. Following the production of the genome sequence two projects will be soon activated in our group for genetic manipulation of *N.gaditana*. A new PhD student already started the design of promoter and reporter gene cloning for testing of the available protocols of transformation on our organism of interest. The set up of a protocol for stable insertion of exogenous DNA inside the genome of *Nannochloropsis* will open the possibility for experiments of targeted mutagenesis. A second project is based on ethyl methane sulfonate-induced mutagenesis to obtain mutant strains with increased lipid content or constitutive lipids accumulation. Single cells could be selected by fluorescence-activated cell sorting after staining using nile red and genome of the interesting mutants obtained could be easily resequenced using a fraction of the SOLiD slide for characterization of the produced changes. Random mutagenesis in *Nannochloropsis* was already reported (Chaturvedi and Fujita, 2006).

Once the assembly and annotation will be concluded a user friendly interface for browsing the available data will be published on the Web, with the possibility for all the scientist working on the molecular biology of *Nannochloropsis*, to had their own data and make them available for the scientific community. To this aim a genome browser will be realized and all the data of gene expression will be published on the browser together with the genome sequence and the list of annotated genes.

The experiments realized for characterization of the transcriptome provided an important contribution for the annotation of the genome and for the selection of candidate genes relevant for the study of lipid metabolism. While biological replicates of each of the samples were pooled for production of paired-end libraries, each of the sequencing experiments was realized only ones. The absence of experimental replicates increases the risk to analyze data affected by a random bias. Sequencing will be therefore soon repeated increasing the number of conditions sampled and realizing 3 independent experimental replicates for each of the sampled conditions.

References

Chaturvedi R and Fujita Y (2006) Isolation of enhanced eicosapentaenoic acid producing mutants of Nannochloropsis oculata ST-6 using ethyl methane sulfonate induced mutagenesis techniques and their characterization at mRNA transcript level. Phycological Research 54(3):208–219.