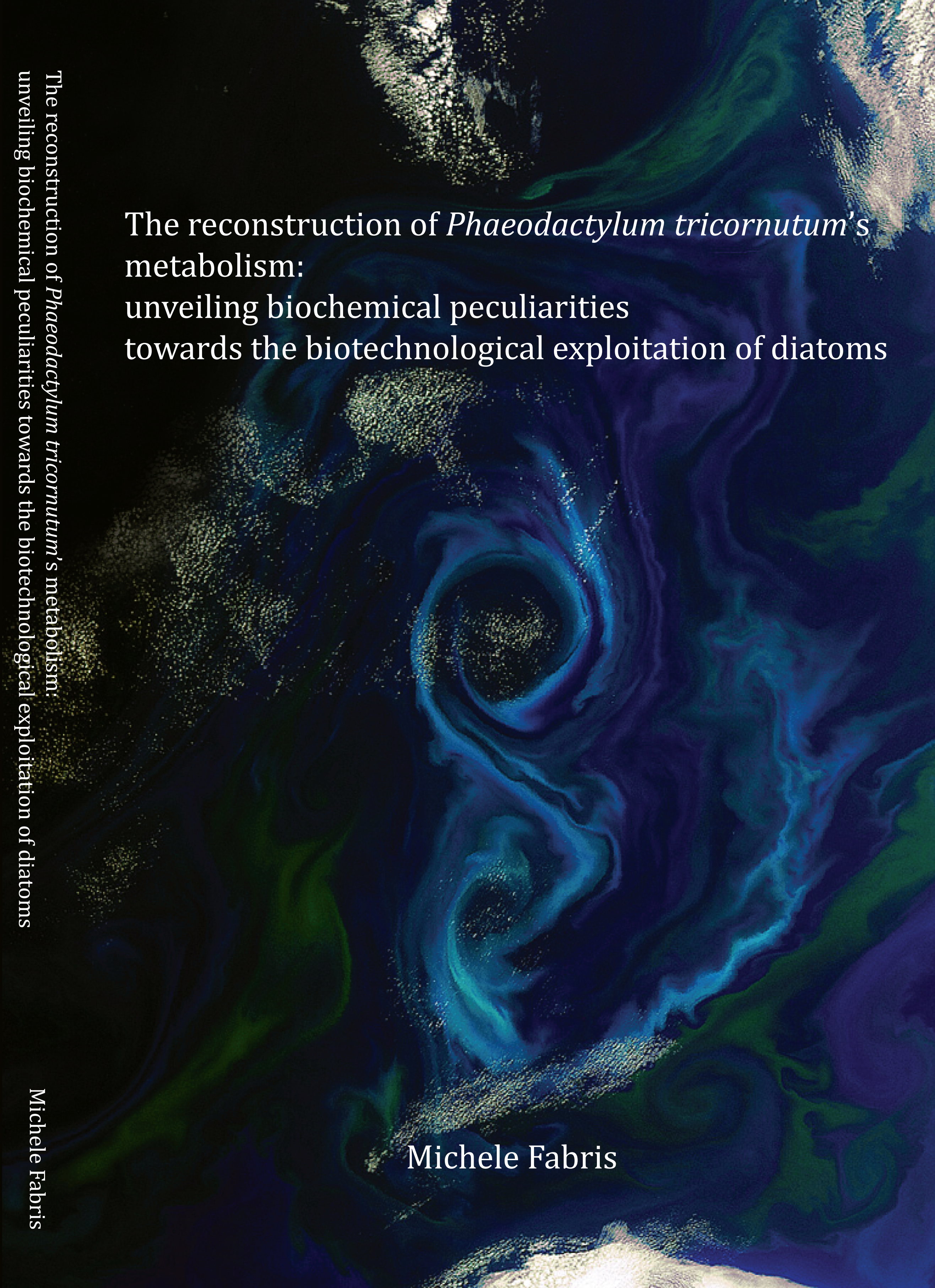
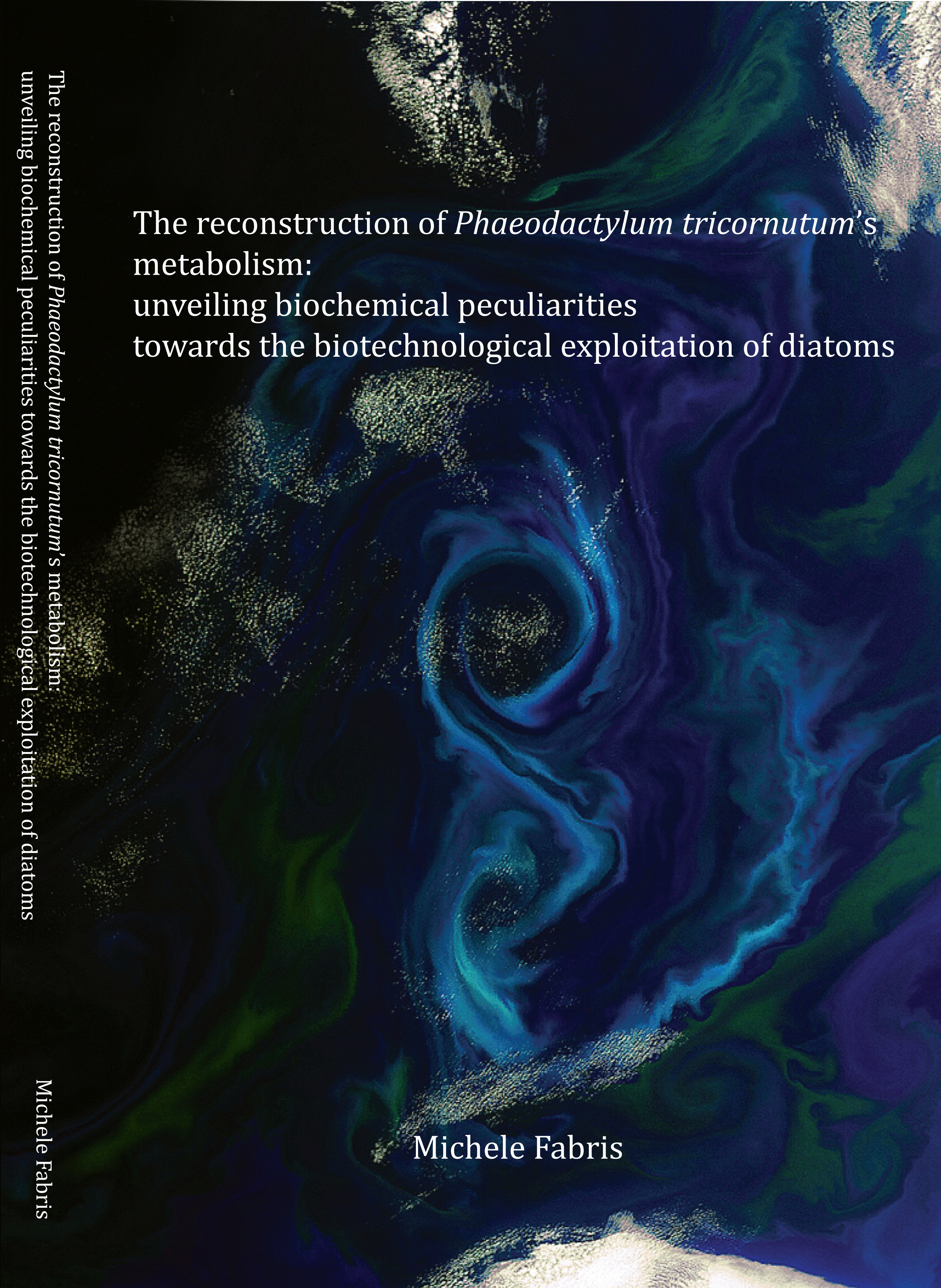


The reconstruction of *Phaeodactylum tricornutum*'s metabolism: unveiling biochemical peculiarities towards the biotechnological exploitation of diatoms

Michele Fabris

The reconstruction of *Phaeodactylum tricornutum*'s metabolism:
unveiling biochemical peculiarities towards the biotechnological exploitation of diatoms

Michele Fabris



The reconstruction of *Phaeodactylum tricornutum*’s metabolism: unveiling biochemical peculiarities towards the biotechnological exploitation of diatoms

Michele Fabris

The reconstruction of *Phaeodactylum tricornutum*’s metabolism:
unveiling biochemical peculiarities towards the biotechnological exploitation of diatoms

Michele Fabris

UNIVERSITEIT GENT

FACULTY OF SCIENCES

Ghent University - Faculty of Sciences
Department of Plant Biotechnology and Bioinformatics
VIB - Department of Plant Systems Biology

The reconstruction of *Phaeodactylum tricornutum*'s metabolism: unveiling biochemical peculiarities towards the biotechnological exploitation of diatoms

Michele Fabris

Thesis submitted in fulfillment of the requirements
for the degree of Doctor (PhD) in Sciences: Biotechnology and Biochemistry

Academic year: 2013-2014
Promoter: Prof. Dr. Alain Goossens
Co-promoter: Prof. Dr. Wim Vyverman

This work was conducted at the VIB Department of Plant Systems Biology, Ghent University and at the Department of Biology – Laboratory of Protistology and Aquatic Ecology, Ghent University.

Members of the Examination Board

Prof. Dr. Wout Boerjan (chairman)
Department of Plant Biotechnology and Genetics, Ghent University

Prof. Dr. Alain Goossens (promotor)
Department of Plant Biotechnology and Genetics, Ghent University

Prof. Dr. Wim Vyverman (co-promotor)
Department of Biology, Ghent University

Dr. ir. Gino Baart
Department of Microbial and Molecular Systems, Centre of Microbial and Plant Genetics, KU Leuven

Prof. Dr. Peter Kroth*
Faculty of Biology, University of Konstanz, Konstanz (Germany)

Prof. Dr. ir. Imogen Foubert*
Subfaculty of Science, KU Leuven, Kortrijk

Prof. Dr. Lieven de Veylder*
Department of Plant Biotechnology and Genetics, Ghent University

Dr. Steven Maere*
Department of Plant Biotechnology and Genetics, Ghent University

*Members of the Reading Committee

Cover image:

A phytoplankton bloom swirls a figure-of-8 in the South Atlantic Ocean about 600 km east of the Falkland Islands. Image acquired on 2 December 2011 by MERIS instrument mounted on Envisat satellite at a resolution of 300 m (European Space Agency)

“*You miss one hundred percent of the shots you don't take*”

Wayne Gretzky

Contents

Chapter 1

Introduction

Author contribution
MF wrote the chapter.

Abstract

Diatoms are one of the dominant populations of phytoplankton and significantly contribute to the global primary productivity and marine geochemical cycles. During their evolution, two distinct endosymbiotic events and numerous gene transfers caused the incorporation of genes of different origin in their genomes. The strong selective pressure exerted by the marine environment carved a peculiar metabolism that embraces biochemical features from all domains of life. Additionally, diatoms have numerous industrial and biotechnological applications linked to their metabolic capabilities, as they are attractive candidates for the sustainable production of bioenergy and high value bio products. However, a broader knowledge on their metabolism is required to better understand diatom's ecology and evolution, as well as to develop strategies for a cost-effective industrial exploitation. Although the sequencing of the genomes of the model species *Thalassiosira pseudonana* and *Phaeodactylum tricornutum* have recently revealed many uncommon metabolic traits, diatom metabolism is still largely uncharacterized. This chapter gives a descriptive overview on diatom's origin, evolution and biotechnological applications; additionally it proposes systems biology as a valid approach to maximize the metabolic information obtainable from diatoms, introducing the scopes and objectives of this dissertation.

Phytoplankton

Photosynthesis sustains aerobic life on Earth and keeps the planet alive. After the origin of life, the rise of unicellular photosynthetic organisms represented one of the most important events occurred in nature. Similarly, to the primitive forms of life, photosynthesis evolved in water, when the first cyanobacteria started converting the abundant CO_2 into organic carbon, chemical energy and oxygen using sunlight as source of energy. Microalgae evolved from it and contributed significantly to the process of oxygenation of the atmosphere, making possible the rise of aerobic life on Earth.

Phytoplankton is a heterogeneous group of photosynthetic microorganisms, both prokaryotic and eukaryotic, mainly composed by cyanobacteria, diatoms, dinoflagellates and coccolithophores. Phytoplankton has an enormous impact on Earth's biosphere. It has been estimated that 1% of Earth's photosynthetic biomass is composed by phytoplankton but it produces up to the 45% of the oxygen available in the atmosphere (Falkowski *et al.*, 2004).

Diatoms

Estimated to account for approximately 200.000 different species, diatoms are the dominant population of eukaryotic phytoplankton (Armbrust, 2009) and one of the most ecologically significant group of organisms on Earth. Diatoms are unicellular eukaryotic microalgae, belonging to the Stramenopiles phylum, which includes, among the others, brown algae and oomycetes.

Characterized by recurrent genetic rearrangements, diatoms possess an extraordinary adaptability, which led them to colonize every aquatic photic zone of Earth. They thrive in oceans, lakes, rivers, ponds and even under a thick layer of polar ice (Charvet *et al.*, 2012; Janech *et al.*, 2006). Diatoms are particularly successful in upwelling coastal marine regions, which are characterized by frequent pulses of nutrient availability (Tozzi *et al.*, 2004). Due to their widespread distribution, diatoms have important ecological effects on global scale. They harbor one of the most efficient photosynthetic machinery in nature (Giordano *et al.*, 2005) that allows to constantly sequester huge amounts of atmospheric CO_2 and generate as much biomass as all the terrestrial rainforests combined (Field *et al.*, 1998). Combining atmospheric carbon dioxide and water with the energy conveyed by sunlight, diatom biomass releases 20-25% of the total amount of oxygen present in the atmosphere (Field *et al.* 1998), corresponding to almost half of the total phytoplankton production. The carbon adsorbed from the atmosphere and fixed by diatoms accounts for the 40% of the total organic carbon present in the oceans (Nelson *et al.*, 1996), significantly sustains the marine food web. Moreover, diatoms are enclosed in elaborated silica shells, which have several biological functions, only partly elucidated (Sims *et al.*, 2006) making diatoms one of the major contributors in the geochemical balance of silica in the marine ecosystem. Their impact in cycling oceanic silica is on a such large scale that it has been estimated that one atom of silicon, before sinking to the seafloor, is incorporated 40 times in diatoms cell (Tréguer *et al.*, 1995). Consequently, the amount of silica utilized and cycled by diatoms is so large that some oceanic locations are characterized by huge deposits of silica reaching the thickness of 1 km , originated by sinking diatoms over millions of years (Sims *et al.*, 2006).

Origin and evolution of diatoms

Diatoms evolved from two different endosymbiotic events, which significantly shaped their genetic setup and separated them from red, green algae and land plants. The first endosymbiosis occurred approximately 1,5 billion years ago, when a cyanobacterium

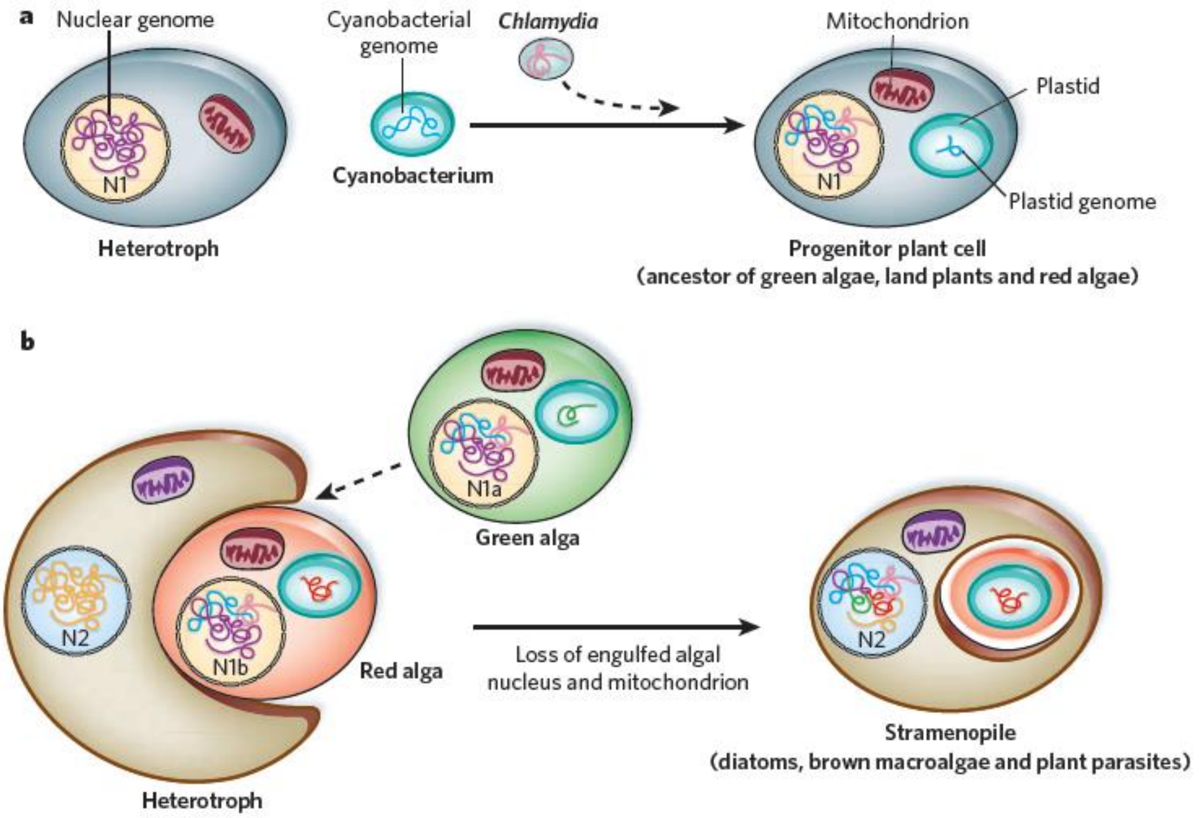


Figure 1.1 Endosymbiotic origin of Stramenopiles. (a) Primary and (b) secondary endosymbiotic event. a) Origin of the ancestor of plant and green/red algae, through the engulfment of a cyanobacterium and possibly a chlamydia. Genes from both invaders were progressively transferred to the host nucleus (N1). The cyanobacterium conserved some genes and became the progenitor of plant and green/red algal chloroplasts. (b) Origin of diatoms and Stramenopiles: the photosynthetic eukaryote, presumably differentiated as a red alga and was engulfed by a eukaryotic heterotroph. The algal nucleus (N1) progressively disappeared, genes were transferred to the host nucleus (N2) and the red algal symbiont was reduced to the present chloroplast of Stramenopiles, characterized by four membranes. (From: Armbrust, 2009)

was included in a marine eukaryotic heterotrophic cell. This caused a progressive gene transfer from the cyanobacterium's genome to the host nucleus, giving origin to the ancestor of red algae, green algae and land plants. It has been suggested that at this stage, a chlamydial endosymbiosis might have occurred, significantly contributing to the gene transfer (Huang and Gogarten, 2007). Approximately 500 millions of years later, a red alga that evolved from such primordial organism, was subsequently incorporated by another eukaryotic cell, causing the consequent gene transfer towards the nucleus and progressive reduction of the primitive alga to the present plastid of Stramenopiles (Prihoda *et al.*, 2012). As indelible sign of the particular series of endosymbiotic events, four membranes surround the plastid of Stramenopiles (Figure 1.1). Estimates collocate the appearance of the first primitive diatoms in the Triassic period, about 250 millions of years ago, according to molecular phylogenetic dating analysis (Sorhannus, 2007), although the most ancient diatom micro fossils are dated back 190 million years ago, in the early Jurassic (Sims *et al.* 2006). Diatoms and other eukaryotic phytoplanktonic groups, such as coccolithophores, started spreading about 205 million years ago (Falkowski *et al.*, 2005). It is only after the mass extinction event of 65 million years ago, which considerably reduced the amount of living species on Earth, that diatoms begun to dominate phytoplankton. Diatoms survived this dramatic event nearly unaffected and expanded their dominance by colonizing new ecological niches due to the reduction of competitors (Armbrust, 2009). The great expansion of diatoms in the subsequent period is believed to have significantly contributed to a drastic decline of atmospheric CO_2, which facilitated the global cooling that characterized the Eocene–Oligocene transition, approximately 33.5 million years ago (Rabosky and Sorhannus, 2009).

A genetic patchwork

Diatoms are phylogenetically divided into four groups: radial centrics, multipolar centrics, araphid pennates and raphid pennates diatoms, reflecting morphological differences but also substantial genetic separation. In the last decade, the number of molecular studies on evolution, biology, physiology and biochemistry of diatoms increased significantly, with the first genome sequencing projects. So far, five different diatom genomes have been sequenced: two centric diatoms belonging to the group of Thalassiosirales, *Thalassiosira pseudonana* (Armbrust *et al.*, 2004) and *Thalassiosira oceanica* (Lommer *et al.*, 2012); three pennate diatoms, *Phaeodactylum tricornutum* (Bowler *et al.*, 2008), *Fragilariopsis cylindrus* (http://genome.jgi-psf.org/Fracy1/Fracy1.home.html) and *Pseudo-nitzschia multiseries* (http://genome.jgi-psf.org/Psemu1/Psemu1.home.html). The analysis of the relatively small genomes of *T. pseudonana* (34.5 B) and *P. tricornutum* (27.4Mb) revealed an extraordinary mosaicism of genes of different origin and a remarkable divergence from other microalgal species. Moreover, the comparison of the two genomes indicated that the separation between pennate and centric diatoms that occurred approximately 90 million years ago (Bowler *et al.,* 2008) is characterized by a remarkable genetic divergence., In fact, *P. tricornutum* only shares 57% of its genes with *T. pseudonana*. The extent of such separation is comparable to the genetic differences developed between the groups of fish and mammals in over 550 million years of divergent evolution (Bowler *et al.*, 2008). This suggests that the frequency of genetic rearrangements and the evolution rate are unusually high in diatoms, mainly due to the high number of transposable elements present in their genome (Bowler *et al.,* 2008). In *P. tricornutum*, this is underscored by the striking presence of genes deriving from other organisms, mainly bacteria (approximately 500), presumably due to recurrent Horizontal Gene Transfers (HGT), which in diatoms take place with extraordinary frequency (Bowler *et al.*, 2008). Differently, the presence in *T. pseudonana* of 108 proteins of clear red algal origin and, surprisingly, of other 806 showing similarities to animal genes but not to plant or algal genes, is ascribable to the genetic re-organization occurred as consequence of endosymbiotic events (Endosymbiotic Gene Transfers, EGTs) (Armbrust *et al.*, 2004). During diatom evolution, HGT events were enhanced by close interactions with other marine organisms, principally bacteria, to which diatom physiology and biology is tightly bound (Bruckner *et al.*, 2011; Amin *et al.*, 2012). The presence of active copia-like retro-transposons activated in particular growth conditions suggests that the genome of diatoms is constantly re-shaped by the environment (Oliver *et al.*, 2010) . Over millions of years, similar phenomenona contributed to the evolution of a significant number of diatom-specific genes, which encode proteins that do not correspond to any known reference and, in *P. tricornutum,* represent the 44% of the total number (Maheswari *et al.*, 2010). One of the main consequences of such uncommon evolution path and high genetic dynamism is that diatoms developed a unique metabolism, greatly optimized to face drastically changing conditions, by exploiting metabolic traits from all domains of life.

An unprecedented metabolism

In the oceans, conditions change very rapidly and a flexible metabolism is a key factor for the ecological success of diatoms in such variable environment (Allen et al. 2011, Kamp et al. 2011, Fabris et al. 2012). The ability of non-motile photosynthetic algae to thrive in harsh conditions and to outcompete other species is tightly related to intrinsic features of the central metabolism, which include strategies of uptake of nutrients, generation of energy, synthesis of precursor and cell division. The central metabolism of diatoms is characterized by a set of features that has no equals. For example, diatom photosynthesis is distinguished by an exceptional efficiency in converting solar energy into biomass (Wagner *et al.*, 2006; Giordano *et al.*, 2005). This is made possible by the combined presence of the isoform 1D of Rubisco, which has the highest level of affinity for carbon dioxide (Giordano *et al.*, 2005) and the presence of different carbon concentrating mechanisms (CCMs) that involve HCO_3^- transporters (Nakajima *et al.*, 2013), carbonic anhydrases (Haimovich-Dayan *et al.*, 2013; Kroth *et al.*, 2008; Armbrust *et al.*, 2004) and systems that possibly involve the silica frustule as buffering agent (Kröger and Poulsen, 2008). Despite the presence of genes putatively involved in the plant-like C4 metabolism (Haimovich-Dayan *et al.*, 2013; Kroth *et al.*, 2008; Bowler *et al.*, 2008; Armbrust *et al.*, 2004; Fabris *et al.*, 2012), Rubisco's efficiency does not seem to be related to them and it has been proposed that they may be involved in energy dissipating futile cycles (Haimovich-Dayan *et al.*, 2013). For the same purpose, diatoms use a very resourceful mechanism to disperse excessive photosynthetic energy, based on particular non-photochemical quenching and a cyclic de-epoxidation of xanthophyll's (Lepetit *et al.*, 2012). At the same time, this allows a very efficient use of electrons with minimal losses and an optimal protection for the photosystems, which entails a considerable synthesis-cost if damaged (Hildebrand *et al.*, 2013). In addition, a peculiar light-harvesting complex stress-related (LHCSR) protein has been recently identified and characterized as a pivotal factor for the rapid adaptation to sudden changes of light regimes (Bailleul *et al.*, 2010). The uncommon metabolic organization of diatoms is also reflected by the unusual subcellular localization of metabolic pathways such as the oxidative pentose phosphate (OPP), which in diatoms is situated in the cytoplasm. As such, it is physically separated from the plastidic Calvin-Benson cycle that, additionally, is regulated differently than in other photosynthetic organisms (Kroth *et al.*, 2008; Michels *et al.*, 2005; Gruber *et al.*, 2009). Similarly, the synthesis of nucleotides, which in plants usually is localized in the plastids, in diatoms occurs in the cytoplasm and specific uncommon transporters are used for the import of the synthetized molecules to the plastid (Ast *et al.*, 2009). Even conserved biochemical mechanisms, such as the glycolytic process, show peculiarities in diatoms. From the analysis of *P. tricornutum* genome, emerged a certain redundancy of glycolytic enzymes

(Kroth et *al.*, 2008). Precisely, glycolytic isozymes had been found to be putatively localized in the cytoplasm, in the chloroplast, possibly enabling the generation of ATP in the dark (Ast et *al.,* 2009), and in the mitochondrion (Kroth *et al.*, 2008). As described in Chapter 3 of this thesis, the lower half of the Embden-Meyerhof-Parnas glycolysis identified in the mitochondrion connects to a functional eukaryotic Entner-Doudoroff pathway, so far identified only in prokaryotes. Moreover, *P. tricornutum* also harbors a phosphoketolase pathway, which represents a catabolic by-pass of the pentose phosphate pathway (Fabris *et al.*, 2012). This pathway possibly confers to the cell the ability to form additional ATP by the breakdown of C5 sugars and its conservation is generally restricted to some bacterial group and some fungal species (Fabris *et al.* 2012). This redundancy of glycolytic pathways presumably allows cells to faster adapt to sudden changes of nutrient availability (Fabris *et al.*, 2012) or to generate important cofactors such as NADPH (Chavarría *et al.*, 2012). Diatoms use chrysolaminaran, a β(1,3) β(1,6) polysaccharide, as short-term storage of carbon, whereas lipids are used for long-term storage (Armbrust *et al.*, 2004). The breakdown of odd-numbered chains of fatty acids (FAs) occurs through a methylmalonyl–coenzyme A pathway, typical for metazoans but not used by plants and green algae (Schönknecht *et al.*, 2013). Doubtless, one of the most striking metabolic features of diatoms is represented by the presence of the whole set of genes necessary to enable a complete urea cycle (Armbrust *et al.,* 2004; Bowler *et al.,* 2008), absent in other eukaryotic photoautotrophs. Its characterization allowed to elucidate its role in the efficient recycling of intracellular carbon and nitrogen and showed that this metabolism has a central role on growth performance of diatoms (Allen *et al.*, 2011). Transported by marine currents and welling events, diatoms can sink and remain for long periods in dark and anoxic zones in the deep ocean. To survive prolonged dark periods, they evolved the ability to respire nitrate performing the dissimilatory nitrate reduction to ammonium (DNRA) (Kamp *et al.*, 2011). This metabolic solution is present only in a few prokaryotes and fungal species, but uncommon in other photosynthetic organisms. The biosynthesis of sterols and their precursors, essential components of the cellular membrane, is another example of unusual biochemical organization. As described in Chapters 5, the metabolism of sterols is derived by a combination of the pathway used by plants with that of fungi, producing a hybrid biochemical route that involves novel enzymes. Finally, the presence of a silica frustule is believed to be one of the factors that contributed to diatom's evolutionary success, as it has been proposed that its synthesis might require less cellular resources compared to an organic cell wall (Raven, 1983), providing mechanical protection, as well as conferring improved light harvesting capacity and acting as auxiliary component of carbon concentrating mechanism (Kröger and Poulsen, 2008). For the biosynthesis of such cellular structures, diatoms have evolved very efficient silica uptake systems, metabolism and transport (Armbrust *et al.,* 2004), as well as sophisticated biomineralization and deposition mechanisms, (Poulsen and Kröger 2008).

Research on diatom metabolism is currently unveiling unique intriguing features, which directly reflect the peculiar evolutionary history of diatoms. To date, available knowledge allowed only fragmentary reconstructions of the metabolic capabilities and organization of diatoms. A better, comprehensive understanding of diatom biochemistry will lead to the discovery of novel biochemical mechanisms, metabolic pathways and enzymes. This will help to better define the evolution path of diatoms, which still presents some unclear steps (Lohr *et al.*, 2012) and to further unveil key factors of their ecological success. Moreover, detailed mapping the cellular metabolism of diatoms will serve as basis to explore unknown aspects of it, such as cell-cell communications and signaling (Adolph *et al.*, 2004), symbiosis and formation of biofilm

(Bruckner *et al.*, 2011) and secondary metabolism, which has relevant practical implications such as the production of toxic compounds (Vanelslander *et al.*, 2012; Schnetzer *et al.*, 2007).

Biotechnological and industrial applications of diatoms and microalgae

Aside from playing a central role in the global ecosystem, diatoms and phytoplankton influence human society and its daily activities based on fossil oil. Burial of dead phytoplanktonic biomass over millions of years is at the origin of the formation of present petroleum reserves, to which diatoms contributed significantly (Falkowski *et al.*, 2004). Additionally, the constant sequestration of atmospheric CO_2 operated by diatoms and other phytoplankton considerably mitigates the effects of anthropogenic atmospheric pollution and deriving greenhouse gases. It has been estimated that more than 30% of the CO_2 released by human activities in the past century, has been adsorbed by marine phytoplankton (Sabine *et al.*, 2004).

Diatoms, like other microalgae produce a large amounts of saturated fatty acids and monoenoic fatty acids, which are attractive precursors for the production of biodiesel (Bozarth *et al.*, 2009). Although bioenergy processes from diatoms and other oil producing microalgae are not yet economically competitive, their theoretical energy potential (9–154 kW ha $^{-1}$ d^{-1}, depending on oil content) is significantly higher than that of *Saccharum sp.* (sugar cane) and *Elaeis guineensis* (palm) (Chisti, 2007; Chisti, 2008; Demirbas, 2009), making diatoms a promising renewable and carbon-neutral alternative to petroleum fuels for the future. As benefit, they do not require fertile land; therefore they do not compete with crops designated for food. For their cultivation, sea- or freshwater and sunlight are sufficient and, despite being aquatic organisms, microalgae require less water than land crops (Rodolfi *et al.*, 2009). Some diatom and microalgae species are able to grow and thrive even in sewage and industrial wastewater, rich in phosphates, nitrates, sulphates, metals and CO_2, resulting promising bioremediation agents (Hildebrand *et al.*, 2012). After the extraction of oils and other compounds, residual biomass can be used as animal feed, fertilizer and as substrate for bacterial anaerobic digestion by bacteria for the production of biogas (Chisti, 2007; Zamalloa *et al.*, 2012).

Currently, a vast number of microalgal species are object of extensive research designed to identify suitable strains for biofuel and bioenergy production. A recent review compared diatoms productivity performances to those of the most common green algal strain used in the biofuel research field (Hildebrand *et al.*, 2012). The overall higher growth rates, lipid yields, the better tolerance to harsh environments and rapidly changing conditions, conferred by their unique metabolism and subcellular organization, collocate diatoms among the best candidates for such applications (Hildebrand *et al.*, 2012).

Diatoms and micro algae have potential applications in several biotechnological fields other than bioenergy production. Commercial interests are bound to the metabolic capabilities of specific microalgal strains, such as the natural or engineered synthesis of valuable compounds (Fernie *et al.*, 2012; Goulitquer *et al.*, 2012). For example, diatoms are efficient producers of omega-3 polyunsaturated fatty acids such as eicosapentaenoate (EPA) and dodecahexanoate (DHA) (Uttaro, 2006; Fernández *et al.*, 2003; Tonon *et al.*, 2002; Truksa *et al.*, 2008; Napier, 2002), which have recognized positive effects on brain, vision and cardiac function (EFSA, 2011). Carotenoids are relevant and valuable antioxidants. Particularly, fucoxanthin has raised interest for its pharmaceutical potential, linked to promising preliminary results on its bioactivities as antioxidant, anti-inflammatory, anticancer, anti-obese, antidiabetic, anti-angiogenic, and antimalarial (Peng *et al.*, 2011; T.,-W., Chung *et al.*, 2013). Additionally, sterols deriving from algae and plants, are often biologically active compounds and important molecules with demonstrated positive effects on human health as cholesterol lowering agents and for the maintenance of normal prostate size and function (EFSA, 2010).

Recently, microalgae entered the synthetic biology era, which until a few years ago was restricted to some model species of bacteria and yeasts (Wang *et al.*, 2012). Engineered microalgae have shown their potential as solar-powered cell-sized-factories for the production of therapeutics and recombinant enzymes antibodies (Tran *et al.*, 2013; Tran *et al.*, 2009; Hempel, Lau, *et al.*, 2011), theoretically lowering the current high production costs (Gimpel *et al.*, 2013). Recently, researchers have succeeded in producing bulk chemicals such as poly-3-hydroxybutyrate (PHB), precursor of bioplastic, designing diatoms as possible clean feedstock alternative to petroleum (Hempel *et al.*, 2011). The nanotechnological sector actively seeks for important breakthroughs on silica metabolism and deposition mechanism of diatoms, which recently produced promising results (Poulsen *et al.*, 2013) opening towards ambitious possible applications (Hildebrand, 2003; Kröger and Poulsen, 2008), among which drug delivery systems (Zhang *et al.*, 2013; Gordon *et al.*, 2009), engineered biosensors (Choi et al., 2011) and enzyme immobilizers (Poulsen *et al.*, 2007).

Despite remarkable theoretical potential and promising lab-scale results, the scale up of microalgal cultivations and process is still a bottleneck for many industrial applications. For example, recent life cycle assessments (LCA) underscored the fact that the production of biofuel from algal biomass is economically unfavorable (Passell *et al.*, 2013). The production of biofuels from algal biomass is more expensive than the current production cost of fossil-oil derived fuels. Currently, the cost for producing algal biofuels is estimated to range from $1.68/l to $75/l and will need to be lowered to less than $1/l to compete with fossil oil-derived fuels production costs (Chisti, 2013). On this regard, much effort is currently being put in optimizing several aspects of algal biotechnology, concerning all the facets of the industrial exploitation of microalgae. Culture conditions, biomass harvesting, oil extraction, refinery and processing methods are being investigated in many labs in order to address relevant issues to meet the productive requirements necessary to make the sustainable production of algal biofuel economically feasible.

Important insights on diatom metabolic capabilities and regulation mechanisms will result by extending the available knowledge on their metabolism, providing hints to improve growth conditions and processing methods and, ultimately, deliver strategies for the generation of metabolic engineered microalgal strains, specifically optimized for improved performances, in particular in oil productivity, photosynthetic efficiency, carbon and other nutrients uptake and auto-flocculation.

Metabolic systems biology of microalgae

Metabolism is an interconnected and tightly coordinated network of enzymes, molecules and reactions and an integrative systems-biology approach is the ideal way to study it. A genome-scale metabolic network is defined as the set of reactions and biochemical conversions that occur in an organism. The reconstruction of genome-scale metabolic networks is a multi-step procedure, which can be achieved by using different approaches and protocols (Thiele and Palsson, 2010). In the process of reconstruction, the information relative to metabolism and biochemical capabilities are inferred from the genetic sequence of a given organism. In the genome, the presence of genes encoding specific enzymes reflects the ability of the organism of performing one or more corresponding biochemical reactions. Once organized, refined and properly connected, this large group of gene-enzyme-reaction associations represents the *in silico* projection of the metabolic architecture of the organism and provides a wealth of extensive information (Figure 1.2). Genome-scale metabolic networks are extensively used for a number of different purposes, such as metabolic pathway analysis and characterization, comparative and evolutionary studies, metabolic engineering and modeling. The reconstruction of metabolic networks has become an established practice during the last years, supported by the growing number of available sequenced genomes. Many lab- or commercially relevant organisms have been object of genome-scale metabolic network reconstructions in the last decade. Best examples are to be found in the prokaryotic lineage, where the lower level of cellular, metabolic and genomic complexity, allows more accurate reconstructions (McCloskey *et al.*, 2013). Requiring a different effort, many genome-scale metabolic networks have been developed for eukaryotic model organisms, such as fungi (Förster *et al.*, 2003; Chung *et al.*, 2013; Vongsangnak *et al.*, 2013), plants (Zhang *et al.*, 2005; Poolman *et al.*, 2009; Radrich *et al.*, 2010) and animals (Romero *et al.*, 2005). However, the reconstruction of genome-scale metabolic networks and models of marine photosynthetic microorganism, is only a recent achievement. Several metabolic networks have been developed for the model green alga *Chlamydomonas reinhardtii* (Boyle and Morgan, 2009; May *et al.*, 2009; Dal'Molin *et al.*, 2011; Chang *et al.*, 2011), for two species of *Ostreococcus*, namely *O. tauri* and *O. lucimarinus* (Krumholz *et al.*, 2012) and, in the present work, for the diatom *P. tricornutum* (Fabris *et al.* 2012, Chapter 3). One metabolic network of *C. reinhardtii* (May *et al.* 2009) and that of *P. tricornutum* (Fabris *et al.* 2012) have been converted into Pathway/Genome Databases (PGDBs), an interactive format based on the MetaCyc family of metabolic databases (Caspi *et al.*, 2010). Such databases are accessible online through a web-interface and allow comparative studies, the download of selected sets of genes, reactions and pathways and the upload and visualization of high throughput experiments (Karp *et al.*, 2002).

One of the limits of the reconstruction of genome-scale metabolic models of microalgae is represented by the lower amount of information available, compared to other "land"

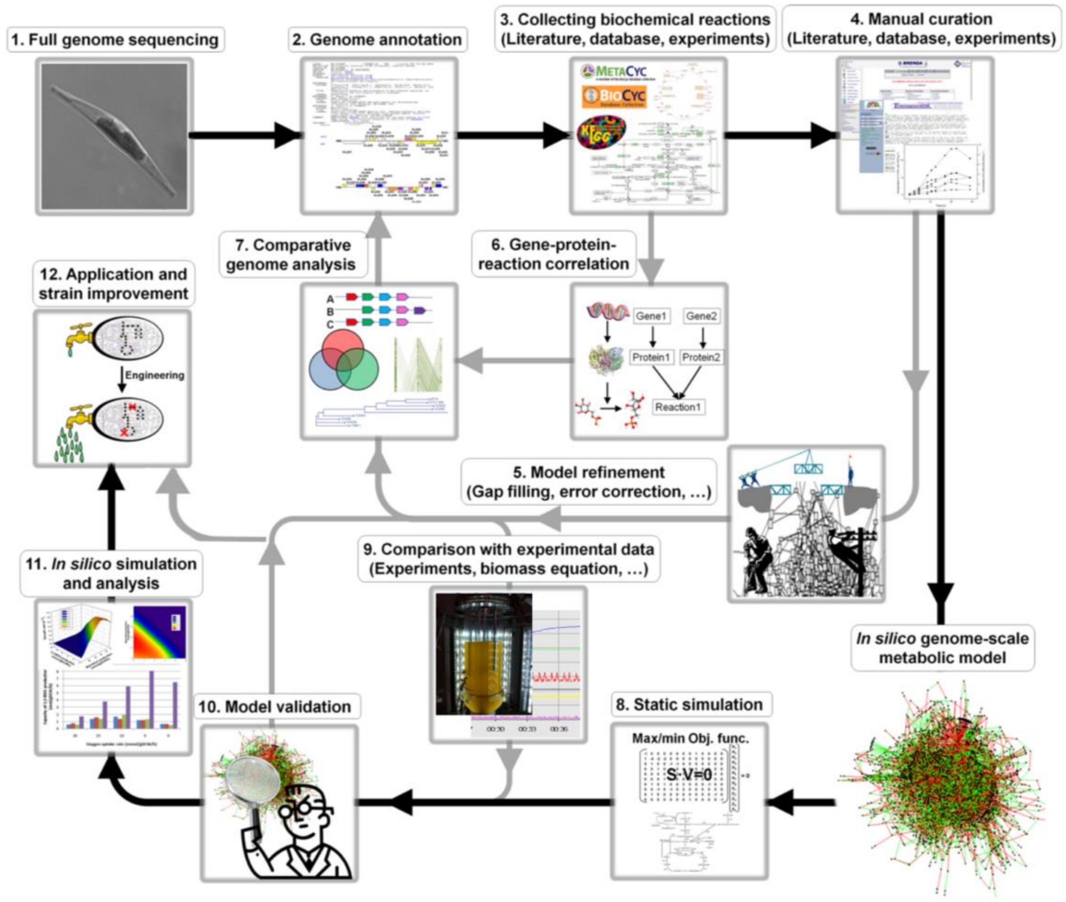

1. Full genome sequencing
2. Genome annotation
3. Collecting biochemical reactions (Literature, database, experiments)
4. Manual curation (Literature, database, experiments)
METACYC
BIOCYC
KEGG
BRENDA
7. Comparative genome analysis
6. Gene-protein-reaction correlation
A
B
C
Gene1
Gene2
Protein1
Protein2
Reaction1
12. Application and strain improvement
Engineering
5. Model refinement (Gap filling, error correction, ...)
9. Comparison with experimental data (Experiments, biomass equation, ...)
11. *In silico* simulation and analysis
In silico genome-scale metabolic model
10. Model validation
8. Static simulation
Max/min Obj. func.
S·V=0

Figure 1.2. Reconstruction of genome-scale metabolic networks and possible uses. (1) The genomic sequence is annotated (2) and mined for genes encoding metabolic enzymes. (3/6) Reactions and pathways are connected and organized with the help of literature and databases. (4/5) The network is refined manually: unneeded or missing reactions/pathways/enzymes are removed or added, respectively. The network can be used for comparative studies (7) or to design metabolic engineering strategies (12). The metabolic network can be translated in a stoichiometric matrix and used for metabolic modeling (8/10/11), validated by experimental data (9) and used to predict metabolic behaviors (11) or metabolic engineering targets (12). (Adapted from Park *et al.*, 2013) organisms. This is primarily due to the limited number of sequenced genomes, insufficient gene functional annotation, hence lack of reliable reference organisms. Poor functional genome annotation negatively affects the output of the metabolic network, resulting in missing reactions and even entire pathways. This was recently highlighted by the coupled reconstruction and comparison of the metabolic networks of *Ostreococcus tauri* and *Ostreococcus lucimarinus,* which were functionally annotated through different methods (Krumholz *et al.*, 2012). Consequently, the more conservative annotation of the *O. tauri* genome resulted in the prediction of fewer reactions (Krumholz *et al.,* 2012). To address similar issues, metabolic network reconstructions often involve an additional computational step, through which the functional annotation of the genome is further improved and refined. This can be fully automated (Moriya *et al.*, 2007; Moriya *et al.*, 2010; Lopez *et al.*, 2011) or semi-automated (Horan *et al.*, 2008; Reed *et al.*, 2006; May *et al.*, 2009; Fabris *et al.*, 2012). The annotation process is often based on computational comparison between the genome of the organism of interest with others, generally well characterized. The choice of the best reference organism(s) is crucial for the quality and completeness of the final output (the metabolic network). In the case of diatoms or other microalgae, this step presents some complications. Although the use of phylogenetically closely related organisms is generally recommended, genomes of close relatives of diatoms and other microalgae are poorly annotated. Therefore, the use of distantly related model organisms becomes necessary. However, the inclusion of both phylogenetically closely and distantly related organisms in set of references to be used for the comparative functional annotation analysis has advantages, as it may allow the identification of unexpected metabolic pathways (Fabris *et al.* 2012). Once generated, the draft reconstruction needs to be carefully reviewed and adjusted. On this regard, manual curation is a fundamental step and determines the final accuracy of the network.

The type and the extent of the required manual curation depend on the intended use of the metabolic network. For example, for modeling purposes, many details have to be adjusted manually, such as reaction stoichiometry, charge of compounds, reaction direction, reversibility and redundancy. Differently, these elements are less important for more descriptive uses of metabolic networks, which might instead require manual effort in determining specific variation of metabolic pathways, redundancy of isozymes and filling pathway holes. The determination and analysis of orthology and inparalogy relationships provides a significant help in determining the occurrence of reactions and false-positive predictions, as well as in identifying possible candidates for reaction that are not associated to any enzyme. Bioinformatics approaches have been developed to assist this crucial phase of network reconstruction, allowing to find candidate genes for specific gaps in the pathways (Karp *et al.*, 2002) or to identify mis-annotations in the reconstructed network (Liberal and Pinney, 2013). To further improve resolution and accuracy, several of the developed genome-scale metabolic networks of microalgae have been reconstructed in combination with proteomics, metabolomics (May *et al.*, 2008) and experimental transcript verification (Manichaikul *et al.*, 2009; Chang *et al.*, 2011).

The limited knowledge on the subcellular localization of pathways, enzymes and metabolites, represents a hurdle in metabolic networks reconstruction. Localization information can be inferred by comparative analysis using other organisms as reference, such as plants, for which the localization of enzymes and reactions is known (Dal'Molin *et al.*, 2011). However, this approach can easily lead to incorrect prediction, due to the phylogenetic distance between organisms. Web-based algorithms offer accurate prediction of the localization of proteins in land species (Emanuelsson *et al.*, 2007), but targeting sequences are significantly different in microalgae, especially in heterokonts, given the peculiar structure of chloroplasts (Gruber *et al.*, 2007). Much work is being done to elucidate metabolic pathways localization and enzyme targeting sequences and mechanisms in diatoms (Gruber *et al.*, 2009; Weber *et al.*, 2009) and significant progresses has been done with the development of specific algorithms for the *in silico* prediction subcellular localization of proteins in heterokonts (Gschloessl *et al.*, 2008) and in green algae (Tardif *et al.*, 2012).

The reliability and complexity of a reconstructed genome-scale metabolic network increases when more information on specific aspects of the metabolism of the organism of interest becomes available. Therefore, the curation of a genome-scale metabolic network is an ongoing task that requires continuous updates and refinements, often resulting in a labor-intensive challenge. This goal is achievable by joining efforts with extensive collaborations, as in the case of the research consortia that constantly curate databases of well-characterized metabolic networks (www.ecocyc.org, www.plantcyc.org, www.humancyc.org) .

A model species, *Phaeodactylum tricornutum*

The raphid pennate diatom *Phaeodactylum tricornutum* represents the only species of the family Phaeodactylaceae and it belongs to the order Naviculales. This diatom is routinely used as lab-strain because it possesses a number of natural features and scientific resources that made of this organims an ideal model for the study of diatoms. Consequently, a large amount of knowledge had been generated on this species over the years.

As a unique feature among diatoms, the cell wall of *P. tricornutum* is poorly silicified and the requirement of silica is facultative. In laboratory conditions, this species is routinely cultured in absence of silica. *P. tricornutum* is a pleiomophic species and its cellular morphology changes in response to environmental stimuli (De Martino *et al.*, 2011). Precisely, *P. tricornutum* alternates fusiform, triradiate or oval morphotypes. While the first two confer a planktonic lifestyle, the latter is associated to a benthic behavior and to the formation of biofilms (Figure 1.3) (De Martino *et al.*, 2011).

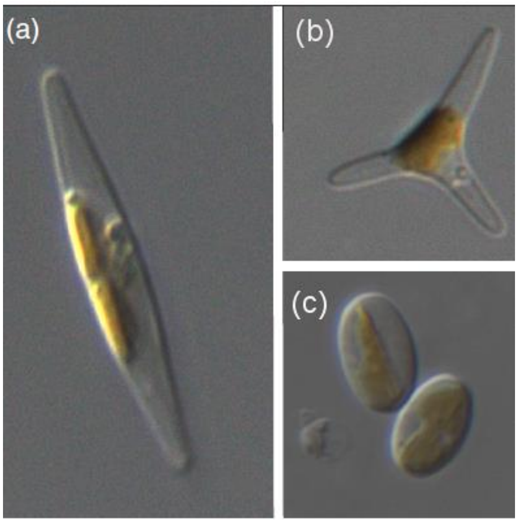


Figure 1.3 Light micrographs of the three morphotypes of *P. tricornutum*. (a) Fusiform, (b) triradiate and (c) oval. (From Vardi *et al.*, 2008)

Unlike other species, in which the vegetative division is alternated with the formation of meiotic auxospores, formed once the asexual divisions reduced the silica frustule exceeding a certain size threshold, *P. tricornutum* only reproduces vegetatively by mitosis. The growth rate of *P. tricornutum* is variable and depends on culture conditions. However, in average it is elevated, as it reaches one cellular division a day. As many other diatoms, *P. tricornutum* is a good oil producer (Harwood and Guschina, 2009; Domergue *et al.*, 2002; Meyer *et al.*, 2004) and represents an optimal model organisms for metabolic studies on lipid accumulation. In particular, *P. tricornutum* produces considerable amounts polyunsaturated fatty acids (PUFAs), mainly EPA, which reach the 5% of the total cellular dry weight (Bozarth *et al.*, 2009), making *P. tricornutum* an suitable strain for the industrial production of PUFAs. *P. tricornutum* has been recognized as bioremediation agent (Torres *et al.*, 1998) and it is currently used by industrial companies as fish-feed and for cosmetic applications (Astaxa, Germany,; Truth in Aging, USA).

In the recent years, many research resources for the molecular study of *P. tricornutum* became available, such as the genomic sequence (Bowler *et al.*, 2008), an Expressed Sequence Tags (ESTs) database relative to 16 different growth conditions (Maheswari *et al.*, 2010), established molecular tools (Siaut *et al.*, 2007) that include genetic transformation protocols through DNA bombardment (Falciatore *et al.*, 1999) and electroporation (Niu *et al.*, 2012) which allow heterologous gene expression, gene overexpression and gene silencing (De Riso *et al.*, 2009) and even genome engineering through TALEN™ nucleases. The latter made possible the development of optimized strains for biofuel applications (Cellectis Press Release) and also as possible resource for recombinant protein production (Hempel *et al.*, 2011). In conclusion, *P. tricornutum* holds a wealth features and biotechnological resources that make it an ideal model organism for the reconstruction and the study of diatom metabolism. Therefore, *P. tricornutum* strain CCAP 1055/1 was used as object of this work.

References

Adolph, S., Bach, S., Blondel, M., Cueff, A., Moreau, M., Pohnert, G., Poulet, S.A., Wichard, T. and Zuccaro, A. (2004) Cytotoxicity of diatom-derived oxylipins in organisms belonging to different phyla. *J. Exp. Biol.*, **207**, 2935–46.

Allen, A.E., Dupont, C.L., Oborník, M., *et al.* (2011) Evolution and metabolic significance of the urea cycle in photosynthetic diatoms. *Nature*, **473**, 203–7.

Amin, S. a, Parker, M.S. and Armbrust, E.V. (2012) Interactions between diatoms and bacteria. *Microbiol. Mol. Biol. Rev.*, **76**, 667–84.

Armbrust, E.V. (2009) The life of diatoms in the world's oceans. *Nature*, **459**, 185–92.

Armbrust, E.V., Berges, J. a, Bowler, C., *et al.* (2004) The genome of the diatom Thalassiosira pseudonana: ecology, evolution, and metabolism. *Science*, **306**, 79–86.

Ast, M., Gruber, A., Schmitz-Esser, S., Neuhaus, H.E., Kroth, P.G., Horn, M. and Haferkamp, I. (2009) Diatom plastids depend on nucleotide import from the cytosol. *Proc. Natl. Acad. Sci. U. S. A.*, **106**, 3621–6.

Bailleul, B., Rogato, A., Martino, A. De, Coesel, S., Cardol, P. and Bowler, C. (2010) An atypical member of the light-harvesting complex stress-related protein family modulates diatom responses to light.

Bowler, C., Allen, A.E., Badger, J.H., *et al.* (2008) The Phaeodactylum genome reveals the evolutionary history of diatom genomes. *Nature*, **456**, 239–44.

Boyle, N.R. and Morgan, J. a (2009) Flux balance analysis of primary metabolism in Chlamydomonas reinhardtii. *BMC Syst. Biol.*, **3**, 4.

Bozarth, A., Maier, U.-G. and Zauner, S. (2009) Diatoms in biotechnology: modern tools and applications. *Appl. Microbiol. Biotechnol.*, **82**, 195–201.

Bruckner, C.G., Rehm, C., Grossart, H.-P. and Kroth, P.G. (2011) Growth and release of extracellular organic compounds by benthic diatoms depend on interactions with bacteria. *Environ. Microbiol.*, **13**, 1052–63.

Caspi, R., Altman, T., Dale, J.M., *et al.* (2010) The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res.*, **38**, D473–9.

Cellectis Press Release Cellectis has successfully engineered the genome of photosynthetic algae with a view to biofuel production. Available at: http://www.cellectis.com/media/press-release/2013/cellectis-has-successfully-engineered-genome-photosynthetic-algae-view-biof.

Chang, R.L., Ghamsari, L., Manichaikul, A., *et al.* (2011) Metabolic network reconstruction of Chlamydomonas offers insight into light-driven algal metabolism. *Mol. Syst. Biol.*, **7**, 518.

Charvet, S., Vincent, W.F., Comeau, A. and Lovejoy, C. (2012) Pyrosequencing analysis of the protist communities in a High Arctic meromictic lake: DNA preservation and change. *Front. Microbiol.*, **3**, 422.

Chavarría, M., Nikel, P.I., Pérez-Pantoja, D. and Lorenzo, V. de (2012) The Entner-Doudoroff pathway empowers Pseudomonas putida KT2440 with a high tolerance to oxidative stress. *Environ. Microbiol.*

Chisti, Y. (2008) Biodiesel from microalgae beats bioethanol. *Trends Biotechnol.*, **26**, 126–31.

Chisti, Y. (2007) Biodiesel from microalgae. *Biotechnol. Adv.*, **25**, 294–306.

Chisti, Y. (2013) Constraints to commercialization of algal fuels. *J. Biotechnol.*, **167**, 201–214.

Chung, B.K.-S., Lakshmanan, M., Klement, M., Ching, C.B. and Lee, D.-Y. (2013) Metabolic reconstruction and flux analysis of industrial Pichia yeasts. *Appl. Microbiol. Biotechnol.*, **97**, 1865–73.

Chung, T.-W., Choi, H.-J., Lee, J.-Y., ***et al.*** (2013) Marine algal fucoxanthin inhibits the metastatic potential of cancer cells. *Biochem. Biophys. Res. Commun.*

Dal'Molin, C.G.D.O., Quek, L.-E., Palfreyman, R.W. and Nielsen, L.K. (2011) AlgaGEM--a genome-scale metabolic reconstruction of algae based on the Chlamydomonas reinhardtii genome. *BMC Genomics*, **12 Suppl 4**, S5.

Demirbas, A. (2009) Progress and recent trends in biodiesel fuels. *Energy Convers. Manag.*, **50**, 14–34.

Domergue, F., Lerchl, J., Zähringer, U. and Heinz, E. (2002) Cloning and functional characterization of Phaeodactylum tricornutum front-end desaturases involved in eicosapentaenoic acid biosynthesis. *Eur. J. Biochem.*, **269**, 4105–4113.

EFSA (2011) Scientific Opinion on the substantiation of health claims related to docosahexaenoic acid (DHA), eicosapentaenoic acid (EPA) and brain , eye and nerve development (ID 501 , 513 , 540), maintenance of normal brain 4294), maintenance of normal cardia. *EFSA J.*, **9**, 1–30.

EFSA (2010) Scientific Opinion on the substantiation of health claims related to plant sterols and plant stanols and maintenance of normal blood cholesterol 3140), and maintenance of normal prostate size and normal urination (ID 714 , 1467 , 1635) pursuant to Arti. *EFSA J.*, **8**, 1–22.

Emanuelsson, O., Brunak, S., Heijne, G. von and Nielsen, H. (2007) Locating proteins in the cell using TargetP, SignalP and related tools. *Nat. Protoc.*, **2**, 953–71.

Fabris, M., Matthijs, M., Rombauts, S., Vyverman, W., Goossens, A. and Baart, G.J.E. (2012) The metabolic blueprint of Phaeodactylum tricornutum reveals a eukaryotic Entner-Doudoroff glycolytic pathway. *Plant J.*, **70**, 1004–14.

Falciatore, a, Casotti, R., Leblanc, C., Abrescia, C. and Bowler, C. (1999) Transformation of Nonselectable Reporter Genes in Marine Diatoms. *Mar. Biotechnol. (NY).*, **1**, 239–251.

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

Falkowski, P.G., Katz, M.E., Milligan, A.J., Fennel, K., Cramer, B.S., Aubry, M.P., Berner, R. a, Novacek, M.J. and Zapol, W.M. (2005) The rise of oxygen over the past 205 million years and the evolution of large placental mammals. *Science*, **309**, 2202–4.

Fernández, F.G.A., Alias, C.B., Pérez, J.A.S., José, M., Sevilla, F., González, M.J.I. and Grima, E.M. (2003) Production of 13 C polyunsaturated fatty acids from the microalga Phaeodactylum tricornutum. , 229–237.

Fernie, A.R., Obata, T., Allen, A.E., Araújo, W.L. and Bowler, C. (2012) Leveraging metabolomics for functional investigations in sequenced marine diatoms. *Trends Plant Sci.*, **17**, 395–403.

Field, C.B., Behrenfeld, M.J., Randerson, J.T. and Falkowski, P. (1998) Primary Production of the Biosphere: Integrating Terrestrial and Oceanic Components. *Science (80-.).*, **281**, 237–240.

Flamholz, a., Noor, E., Bar-Even, a., Liebermeister, W. and Milo, R. (2013) Glycolytic strategy as a tradeoff between energy yield and protein cost. *Proc. Natl. Acad. Sci.*, 2–7.

Förster, J., Famili, I., Fu, P., Palsson, B.Ø. and Nielsen, J. (2003) Genome-scale reconstruction of the Saccharomyces cerevisiae metabolic network. *Genome Res.*, **13**, 244–53.

Gimpel, J. a, Specht, E. a, Georgianna, D.R. and Mayfield, S.P. (2013) Advances in microalgae engineering and synthetic biology applications for biofuel production. *Curr. Opin. Chem. Biol.*, **17**, 489–95.

Giordano, M., Beardall, J. and Raven, J. a (2005) CO2 concentrating mechanisms in algae: mechanisms, environmental modulation, and evolution. *Annu. Rev. Plant Biol.*, **56**, 99–131.

Gordon, R., Losic, D., Tiffany, M.A., Nagy, S.S. and Sterrenburg, F. a S. (2009) The Glass Menagerie: diatoms for novel applications in nanotechnology. *Trends Biotechnol.*, **27**, 116–27.

Goulitquer, S., Potin, P. and Tonon, T. (2012) *Mass spectrometry-based metabolomics to elucidate functions in marine organisms and ecosystems.*,.

Gruber, A., Vugrinec, S., Hempel, F., Gould, S.B., Maier, U.-G. and Kroth, P.G. (2007) Protein targeting into complex diatom plastids: functional characterisation of a specific targeting motif. *Plant Mol. Biol.*, **64**, 519–30.

Gruber, A., Weber, T., Bártulos, C.R., Vugrinec, S. and Kroth, P.G. (2009) Intracellular distribution of the reductive and oxidative pentose phosphate pathways in two diatoms. *J. Basic Microbiol.*, **49**, 58–72.

Gschloessl, B., Guermeur, Y. and Cock, J.M. (2008) HECTAR: a method to predict subcellular targeting in heterokonts. *BMC Bioinformatics*, **9**, 393.

Haimovich-Dayan, M., Garfinkel, N., Ewe, D., Marcus, Y., Gruber, A., Wagner, H., Kroth, P.G. and Kaplan, A. (2013) The role of C4 metabolism in the marine diatom Phaeodactylum tricornutum. *New Phytol.*, **197**, 177–85.

Harwood, J.L. and Guschina, I. a (2009) The versatility of algae and their lipid metabolism. *Biochimie*, **91**, 679–84.

Hempel, F., Bozarth, A.S., Lindenkamp, N., Klingl, A., Zauner, S., Linne, U., Steinbüchel, A. and Maier, U.G. (2011) Microalgae as bioreactors for bioplastic production. *Microb. Cell Fact.*, **10**, 81.

Hempel, F., Lau, J., Klingl, A. and Maier, U.G. (2011) Algae as protein factories: expression of a human antibody and the respective antigen in the diatom Phaeodactylum tricornutum. *PLoS One*, **6**, e28424.

Hildebrand, M. (2003) Biological processing of nanostructured silica in diatoms. *Prog. Org. Coatings*, **47**, 256–266.

Hildebrand, M., Abbriano, R.M., Polle, J.E., Traller, J.C., Trentacoste, E.M., Smith, S.R. and Davis, A.K. (2013) Metabolic and cellular organization in evolutionarily diverse microalgae as related to biofuels production. *Curr. Opin. Chem. Biol.*, 1–9.

Hildebrand, M., Davis, A.K., Smith, S.R., Traller, J.C. and Abbriano, R. (2012) The place of diatoms in the biofuels industry. *Biofuels*, **3**, 221–240.

Horan, K., Jang, C., Bailey-Serres, J., *et al.* (2008) Annotating genes of known and unknown function by large-scale coexpression analysis. *Plant Physiol.*, **147**, 41–57.

Huang, J. and Gogarten, J.P. (2007) Did an ancient chlamydial endosymbiosis facilitate the establishment of primary plastids? *Genome Biol.*, **8**, R99.

Janech, M.G., Krell, A., Mock, T., Kang, J.-S. and Raymond, J. a. (2006) Ice-Binding Proteins From Sea Ice Diatoms (Bacillariophyceae)1. *J. Phycol.*, **42**, 410–416.

Kamp, A., Beer, D. de, Nitsch, J.L., Lavik, G. and Stief, P. (2011) Diatoms respire nitrate to survive dark and anoxic conditions. *Proc. Natl. Acad. Sci. U. S. A.*, **108**, 5649–54.

Karp, P.D., Paley, S. and Romero, P. (2002) The Pathway Tools software. *Bioinformatics*, **18 Suppl 1**, S225–32.

Kröger, N. and Poulsen, N. (2008) Diatoms-from cell wall biogenesis to nanotechnology. *Annu. Rev. Genet.*, **42**, 83–107.

Kroth, P.G., Chiovitti, A., Gruber, A., *et al.* (2008) A model for carbohydrate metabolism in the diatom Phaeodactylum tricornutum deduced from comparative whole genome analysis. *PLoS One*, **3**, e1426.

Krumholz, E.W., Yang, H., Weisenhorn, P., Henry, C.S. and Libourel, I.G.L. (2012) Genome-wide metabolic network reconstruction of the picoalga Ostreococcus. *J. Exp. Bot.*, **63**, 2353–62.

Lepetit, B., Goss, R., Jakob, T. and Wilhelm, C. (2012) Molecular dynamics of the diatom thylakoid membrane under different light conditions. *Photosynth. Res.*, **111**, 245–57.

Liberal, R. and Pinney, J.W. (2013) Simple topological properties predict functional misannotations in a metabolic network. *Bioinformatics*, **29**, i154–i161.

Lohr, M., Schwender, J. and Polle, J.E.W. (2012) Isoprenoid biosynthesis in eukaryotic phototrophs: a spotlight on algae. *Plant Sci.*, **185-186**, 9–22.

Lommer, M., Specht, M., Roy, A.-S., *et al.* (2012) Genome and low-iron response of an oceanic diatom adapted to chronic iron limitation. *Genome Biol.*, **13**, R66.

Lopez, D., Casero, D., Cokus, S.J., Merchant, S.S. and Pellegrini, M. (2011) Algal Functional Annotation Tool: a web-based analysis suite to functionally interpret large gene lists using integrated annotation and expression data. *BMC Bioinformatics*, **12**, 282.

Maheswari, U., Jabbari, K., Petit, J.-L., *et al.* (2010) Digital expression profiling of novel diatom transcripts provides insight into their biological functions. *Genome Biol.*, **11**, R85.

Manichaikul, A., Ghamsari, L., Hom, E.F.Y., *et al.* (2009) Metabolic network analysis integrated with transcript verification for sequenced genomes. *Nat. Methods*, **6**, 589–92.

Martino, A. De, Bartual, A., Willis, A., Meichenin, A., Villazán, B., Maheswari, U. and Bowler, C. (2011) Physiological and molecular evidence that environmental changes elicit morphological interconversion in the model diatom Phaeodactylum tricornutum. *Protist*, **162**, 462–81.

May, P., Jochristianmpimp-golmmpgde, J.O.C., Kempa, S., Walthermpimp-golmmpgde, D.W., Christian, J.-O. and Walther, D. (2009) ChlamyCyc: an integrative systems biology database and web-portal for Chlamydomonas reinhardtii. *BMC Genomics*, **10**, 209.

May, P., Wienkoop, S., Kempa, S., *et al.* (2008) Metabolomics- and proteomics-assisted genome annotation and analysis of the draft metabolic network of Chlamydomonas reinhardtii. *Genetics*, **179**, 157–66.

McCloskey, D., Palsson, B.Ø. and Feist, A.M. (2013) Basic and applied uses of genome-scale metabolic network reconstructions of Escherichia coli. *Mol. Syst. Biol.*, **9**, 661.

Meyer, A., Kirsch, H., Domergue, F., *et al.* (2004) Novel fatty acid elongases and their use for the reconstitution of docosahexaenoic acid biosynthesis. *J. Lipid Res.*, **45**, 1899–909.

Michels, A.K., Wedel, N., Kroth, P.G. and Germany, A.K.M. (2005) Diatom Plastids Possess a Phosphoribulokinase with an Altered Regulation and No Oxidative Pentose Phosphate Pathway 1. , **137**, 911–920.

Moriya, Y., Itoh, M., Okuda, S., Yoshizawa, A.C. and Kanehisa, M. (2007) KAAS: an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Res.*, **35**, W182–5.

Moriya, Y., Shigemizu, D., Hattori, M., Tokimatsu, T., Kotera, M., Goto, S. and Kanehisa, M. (2010) PathPred: an enzyme-catalyzed metabolic pathway prediction server. *Nucleic Acids Res.*, **38**, W138–43.

Nakajima, K., Tanaka, A. and Matsuda, Y. (2013) SLC4 family transporters in a marine diatom directly pump bicarbonate from seawater. *Proc. Natl. Acad. Sci. U. S. A.*, **110**, 1767–72.

Napier, J. a (2002) Plumbing the depths of PUFA biosynthesis: a novel polyketide synthase-like pathway from marine organisms. *Trends Plant Sci.*, **7**, 51–4.

Nelson, D.M., DeMaster, D.J., Dunbar, R.B. and Smith, W.O. (1996) Cycling of organic carbon and biogenic silica in the Southern Ocean: Estimates of water-column and sedimentary fluxes on the Ross Sea continental shelf. *J. Geophys. Res. Ocean.*, **101**, 18519–18532.

Niu, Y.-F., Yang, Z.-K., Zhang, M.-H., Zhu, C.-C., Yang, W.-D., Liu, J.-S. and Li, H.-Y. (2012) Transformation of diatom Phaeodactylum tricornutum by electroporation and establishment of inducible selection marker. *Biotechniques*, 1–3.

Oliver, M.J., Schofield, O. and Bidle, K. (2010) Density dependent expression of a diatom retrotransposon. *Mar. Genomics*, **3**, 145–150.

Park, J.M., Song, H., Lee, H.J. and Seung, D. (2013) Genome-scale reconstruction and in silico analysis of Klebsiella oxytoca for 2,3-butanediol production. *Microb. Cell Fact.*, **12**, 20.

Passell, H., Dhaliwal, H., Reno, M., *et al.* (2013) Algae biodiesel life cycle assessment using current commercial data. *J. Environ. Manage.*, **129C**, 103–111.

Peng, J., Yuan, J.-P., Wu, C.-F. and Wang, J. (2011) Fucoxanthin, a marine carotenoid present in brown seaweeds and diatoms: metabolism and bioactivities relevant to human health. *Mar. Drugs*, **9**, 1806–28.

Poolman, M.G., Miguet, L., Sweetlove, L.J. and Fell, D. a (2009) A genome-scale metabolic model of Arabidopsis and some of its properties. *Plant Physiol.*, **151**, 1570–81.

Poulsen, N., Berne, C., Spain, J. and Kröger, N. (2007) Silica immobilization of an enzyme through genetic engineering of the diatom Thalassiosira pseudonana. *Angew. Chem. Int. Ed. Engl.*, **46**, 1843–6.

Poulsen, N., Scheffel, A., Sheppard, V.C., Chesley, P.M. and Kröger, N. (2013) Pentalysine clusters mediate silica targeting of silaffins in Thalassiosira pseudonana. *J. Biol. Chem.*, **288**, 20100–9.

Prihoda, J., Tanaka, A., Paula, W.B.M. de, Allen, J.F., Tirichine, L., Bowler, C. and Paula, W.B.M. De (2012) Chloroplast-mitochondria cross-talk in diatoms. *J. Exp. Bot.*, **63**, 1543–57.

Rabosky, D.L. and Sorhannus, U. (2009) Diversity dynamics of marine planktonic diatoms across the Cenozoic. *Nature*, **457**, 183–6.

Radrich, K., Tsuruoka, Y., Dobson, P., Gevorgyan, A., Swainston, N., Baart, G. and Schwartz, J.-M. (2010) Integration of metabolic databases for the reconstruction of genome-scale metabolic networks. *BMC Syst. Biol.*, **4**, 114.

Raven, J.A. (1983) The transport and function of silicon in plants. *Biol. Rev.*, **58**, 179–207.

Reed, J.L., Patel, T.R., Chen, K.H., *et al.* (2006) Systems approach to refining genome annotation. *Proc. Natl. Acad. Sci. U. S. A.*, **103**, 17480–4.

Riso, V. De, Raniello, R., Maumus, F., Rogato, A., Bowler, C. and Falciatore, A. (2009) Gene silencing in the marine diatom Phaeodactylum tricornutum. *Nucleic Acids Res.*, **37**, e96.

Robles Medina, a, Molina Grima, E., Giménez Giménez, a and Ibañez González, M.J. (1998) Downstream processing of algal polyunsaturated fatty acids. *Biotechnol. Adv.*, **16**, 517–80.

Rodolfi, L., Chini Zittelli, G., Bassi, N., Padovani, G., Biondi, N., Bonini, G. and 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**, 100–12.

Romero, P., Wagg, J., Green, M.L., Kaiser, D., Krummenacker, M. and Karp, P.D. (2005) Computational prediction of human metabolic pathways from the complete human genome. *Genome Biol.*, **6**, R2.

Sabine, C.L., Feely, R. a, Gruber, N., *et al.* (2004) The oceanic sink for anthropogenic CO2. *Science*, **305**, 367–71.

Schnetzer, A., Miller, P.E., Schaffner, R. a., Stauffer, B. a., Jones, B.H., Weisberg, S.B., DiGiacomo, P.M., Berelson, W.M. and Caron, D. a. (2007) Blooms of Pseudo-nitzschia and domoic acid in the San Pedro Channel and Los Angeles harbor areas of the Southern California Bight, 2003–2004. *Harmful Algae*, **6**, 372–387.

Schönknecht, G., Chen, W.-H., Ternes, C.M., *et al.* (2013) Gene transfer from bacteria and archaea facilitated evolution of an extremophilic eukaryote. *Science*, **339**, 1207–10.

Siaut, M., Heijde, M., Mangogna, M., Montsant, A., Coesel, S., Allen, A., Manfredonia, A., Falciatore, A. and Bowler, C. (2007) Molecular toolbox for studying diatom biology in Phaeodactylum tricornutum. *Gene*, **406**, 23–35.

Sims, P. a., Mann, D.G. and Medlin, L.K. (2006) Evolution of the diatoms: insights from fossil, biological and molecular data. *Phycologia*, **45**, 361–402.

Sorhannus, U. (2007) A nuclear-encoded small-subunit ribosomal RNA timescale for diatom evolution. *Mar. Micropaleontol.*, **65**, 1–12.

Tardif, M., Atteia, A., Specht, M., *et al.* (2012) PredAlgo: a new subcellular localization prediction tool dedicated to green algae. *Mol. Biol. Evol.*, **29**, 3625–39.

Thiele, I. and Palsson, B.Ø. (2010) A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nat. Protoc.*, **5**, 93–121.

Tonon, T., Harvey, D., Larson, T.R. and Graham, I. a (2002) Long chain polyunsaturated fatty acid production and partitioning to triacylglycerols in four microalgae. *Phytochemistry*, **61**, 15–24.

Torres, E., Cid, A., Herrero, C. and Abalde, J. (1998) Removal of cadmium ions by the marine diatom Phaeodactylum tricornutum Bohlin, accumulation and long term storage. , **63**, 213–220.

Tozzi, S., Schofield, O. and Falkowski, P. (2004) Historical climate change and ocean turbulence as selective agents for two key phytoplankton functional groups. , **274**, 123–132.

Tran, M., Van, C., Barrera, D.J., Pettersson, P.L., Peinado, C.D., Bui, J. and Mayfield, S.P. (2013) Production of unique immunotoxin cancer therapeutics in algal chloroplasts. *Proc. Natl. Acad. Sci. U. S. A.*, **110**, E15–22.

Tran, M., Zhou, B., Pettersson, P.L., Gonzalez, M.J. and Mayfield, S.P. (2009) Synthesis and assembly of a full-length human monoclonal antibody in algal chloroplasts. *Biotechnol. Bioeng.*, **104**, 663–73.

Tréguer, P., Nelson, D.M., Bennekom, A.J. Van, DeMaster, D.J., Leynaert, A. and Quéguiner, B. (1995) The Silica Balance in the World Ocean: A Reestimate. *Science (80-.).*, **268**, 375–379.

Truksa, M., Vrinten, P. and Qiu, X. (2008) Metabolic engineering of plants for polyunsaturated fatty acid production. *Mol. Breed.*, **23**, 1–11.

Uttaro, A.D. (2006) Critical Review Biosynthesis of Polyunsaturated Fatty Acids in Lower Eukaryotes. , **58**, 563–571.

Vanelslander, B., Paul, C., Grueneberg, J., Prince, E.K., Gillard, J., Sabbe, K., Pohnert, G. and Vyverman, W. (2012) Daily bursts of biogenic cyanogen bromide (BrCN) control biofilm formation around a marine benthic diatom. *Proc. Natl. Acad. Sci. U. S. A.*, **109**, 2412–7.

Vardi, A., Thamatrakoln, K., Bidle, K.D. and Falkowski, P.G. (2008) Diatom genomes come of age. *Genome Biol.*, **9**, 245.

Vongsangnak, W., Ruenwai, R., Tang, X., Hu, X., Zhang, H., Shen, B., Song, Y. and Laoteng, K. (2013) Genome-scale analysis of the metabolic networks of oleaginous Zygomycete fungi. *Gene*, **521**, 180–90.

Wagner, H., Jakob, T. and Wilhelm, C. (2006) Balancing the energy flow from captured light to biomass under fluctuating light conditions. *New Phytol.*, **169**, 95–108.

Wang, B., Wang, J., Zhang, W. and Meldrum, D.R. (2012) Application of synthetic biology in cyanobacteria and algae. *Front. Microbiol.*, **3**, 344.

Weber, T., Gruber, A. and Kroth, P.G. (2009) The presence and localization of thioredoxins in diatoms, unicellular algae of secondary endosymbiotic origin. *Mol. Plant*, **2**, 468–77.

Zamalloa, C., Vrieze, J. De, Boon, N. and Verstraete, W. (2012) Anaerobic digestibility of marine microalgae Phaeodactylum tricornutum in a lab-scale anaerobic membrane bioreactor. *Appl. Microbiol. Biotechnol.*, **93**, 859–69.

Zhang, H., Shahbazi, M.-A., Mäkilä, E.M., Silva, T.H. da, Reis, R.L., Salonen, J.J., Hirvonen, J.T. and Santos, H. a (2013) Diatom silica microparticles for sustained release and permeation enhancement following oral delivery of prednisone and mesalamine. *Biomaterials*, 1–10.

Zhang, P., Foerster, H., Tissier, C.P., Mueller, L., Paley, S., Karp, P.D. and Rhee, S.Y. (2005) MetaCyc and AraCyc . Metabolic Pathway Databases. , **138**, 27–37.

Zhu, S.-H. and Green, B.R. (2010) Photoprotection in the diatom Thalassiosira pseudonana: role of LI818-like proteins in response to high light stress. *Biochim. Biophys. Acta*, **1797**, 1449–57.

Chapter 2

Scopes and objectives

Author contribution
MF wrote the chapter.

For hundreds of millions of years, diatoms have evolved under the constant selective pressure of a highly variable environment. Their evolution is characterized by the contribution of two distinct endosymbiotic events and by the occurrence of numerous genetic re-arrangements. This is reflected by the presence of genes of different origin in their genome. As dominant group of eukaryotic phytoplankton, diatoms have enormous impacts both on marine and terrestrial ecology. Additionally, they have high potential in numerous biotechnological applications that range from sustainable bioenergy production to high valuable molecules synthesis. Both ecological success and biotechnological potential of diatoms largely rely on one of the main consequences of their peculiar evolution: their metabolism.

Significant efforts and exciting breakthroughs characterized the study of diatom metabolism during the last years, highlighting an overall superior efficiency and peculiar organization of many biochemical processes. These findings anticipated that diatom metabolism is uniquely shaped and encloses unprecedented features that will be key in reconstructing specific evolutionary steps and understanding the mechanisms of ecological success. Elucidating details the metabolism of diatoms, will also help in generating strategies to optimize certain commercially interesting metabolic traits and overcome limitations that are currently slowing their biotechnological exploitation. When this project started, knowledge on diatom metabolism was poor and fragmentary. Despite the availability of genomic sequences, a systems biology approach had never been applied for the study of diatoms.

The work described in this dissertation aimed to maximize the metabolic information obtainable from the genome of *Phaeodactylum tricornutum* using a systems biology approach to fully map its metabolism. This had two scopes, both oriented towards the ultimate goal of optimizing the lipid productivity in *P. tricornutum*. Firstly, to convert the reconstructed metabolic network into an online metabolic database (DiatomCyc), providing a useful resource for the study of the metabolism of diatoms; secondly, to set the basis for metabolic modeling studies, currently in progress in our group. Originally, the scope of developing a metabolic model was directed to identifying metabolic engineering targets and optimized growth conditions to improve the lipid productivity of *P. tricornutum*. During the course of the project, however, priority was given to the elucidation of novel metabolic features that emerged from the reconstruction of the central metabolic pathways of the diatom, particularly in the carbohydrate and in the isoprenoid metabolism. Chapter 4 describes the discovery and the functional characterization of the Entner-Doudoroff pathway for the first time in eukaryotes, the prediction of the presence of a phosphoketolase pathway and possible implications that additional glycolytic pathways might have in diatoms. Chapter 5 explores biotechnologically relevant pathways such as the mevalonate and sterol biosynthesis, which have been accurately reconstructed *in silico* and experimentally characterized. Their analysis revealed the existence of a multifunctional enzyme in the mevalonate pathway and a peculiar hybrid sterol biosynthetic pathway, characterized by the absence of the key enzyme squalene epoxidase. Moreover, we indicated the existence of a link between the sterol biosynthesis and the accumulation of lipids, offering important insights for the metabolic engineering of diatoms. The obtained results are further discussed in Chapter 6, along with possible future research development, while a summary is provided in Chapter 7. Chapter 8 describes a proof-of-principle of the approach used for the reconstruction DiatomCyc, through which we created CathaCyc, the metabolic database of the medicinal plant *Catharanthus roseus*.

Chapter 3

The metabolic blueprint of *Phaeodactylum tricornutum*

Adapted from the published article:

Fabris, M.; Matthijs, M.; Rombauts, S.; Vyverman, V.; Goossens, A.; Baart, G.J.E.
The metabolic blueprint of *Phaeodactylum tricornutum* reveals a eukaryotic Entner-Doudoroff glycolytic pathway. *The Plant Journal* (2012), 70 (6):1004-14.

Author contribution

MF performed orthology-based annotation, metabolic network reconstruction and manual curation, developed the database and wrote the chapter.

Abstract

Diatoms are one of the most successful groups of unicellular eukaryotic algae. Successive endosymbiotic events contributed to their flexible metabolism, making them competitive in variable aquatic habitats. Although the recently sequenced genomes of the model diatoms *Phaeodactylum tricornutum* and *Thalassiosira pseudonana* have provided the first insights into their metabolic organization, the current knowledge on diatom biochemistry remains fragmentary. By means of a genome-wide approach, we developed DiatomCyc, a detailed pathway/genome database of *P. tricornutum*. DiatomCyc contains 286 pathways with 1719 metabolic reactions and 1613 assigned enzymes, spanning both the central and parts of the secondary metabolism of *P. tricornutum*. Central metabolic pathways, such as those of carbohydrates, amino acids and fatty acids, were covered and the current model of the carbohydrate metabolism was extended.

DiatomCyc is accessible online (http://www.diatomcyc.org), and offers a range of software tools for the visualization and analysis of metabolic networks and 'omics' data. We anticipate that DiatomCyc will be key to gaining further understanding of diatom metabolism and, ultimately, will feed metabolic engineering strategies for the industrial valorization of diatoms.

Introduction

Genome-wide analysis revealed that a remarkable number of the genes of the model diatom *Phaeodactylum tricornutum* have diverse origins (Tirichine and Bowler, 2011; Whitaker *et al.*, 2009). As a consequence, it possesses metabolic pathways from all domains of life, as illustrated by the recently described metazoan-like ornithine–urea cycle (Allen *et al.*, 2012) that is typically absent in plants and green algae. Although several aspects of diatom metabolism have been explored already, a comprehensive overview is still lacking.

Genome-scale databases of metabolism have been constructed for several organisms, including *Escherichia coli* (Keseler *et al.*, 2011), *Homo sapiens* (Romero *et al.*, 2005), *Arabidopsis thaliana* (Radrich *et al.*, 2010) and *Chlamydomonas reinhardtii* (May *et al.*, 2009) and can be used to better understand cellular metabolism (Baart *et al.*, 2008; Chang *et al.*, 2011) in order to develop metabolic engineering strategies (Fong *et al.*, 2005; Smid *et al.*, 2005; Hua *et al.*, 2006), design culture media and processes (Teusink *et al.*, 2005; Baart et al., 2007) , and even to develop online process control (Provost and Bastin, 2004). Hence, the aim of this study was to develop a pathway/genome database (PGDB) of *P. tricornutum* based on its genome sequence and the biochemical literature. Here, we present the features of PGDB, designated DiatomCyc, and discuss some specific metabolic pathways that have been discovered. Mining the *P. tricornutum* genome information led to the identification of unusual metabolic pathways, such as the typically prokaryotic Entner–Doudoroff pathway and a phosphoketolase pathway that, to date has only been found in a few fungi, which will be discussed in more detail in the next chapter.

Results

Creation of DiatomCyc

The available translated genomic sequence of *P. tricornutum* (Bowler *et al.*, 2008) was taken as the starting point for the reconstruction of the first diatom PGDB: DiatomCyc. To complete and improve the current genome annotation, we applied an orthology-based methodology (Notebaart *et al.*, 2006). Twenty-three genomes of organisms (i.e. reference organisms, see Experimental Procedures) were used for high-resolution predictions of translated gene sequence equivalency (Remm *et al.*, 2001; Brien *et al.*, 2005). Eleven out of the 23 organisms have a published genome-scale metabolic network and curated annotation. Gene functions of the reference organisms were conveyed to the corresponding *P. tricornutum* orthologs, for which we employed the 'PHATRDRAFT' terminology, which refers to the common gene identifiers used in databases such as NCBI and KEGG. Gene-to-function and function-to-reaction associations were transferred semi-automatically with the KEGG (Kanehisa and Goto, 2000) and MetaCyc (Caspi *et al.*, 2010) databases as input. The results were imported into PATHWAY TOOLS 15.0 (Karp *et al.*, 2002) and were subsequently refined. As for the other members of the MetaCyc family (Caspi *et al.*, 2010) , including the highly curated databases of *Saccharomyces cerevisiae*, *E. coli, H. sapiens* and *A. thaliana*, DiatomCyc is accessible online (http://www.diatom cyc.org), and offers a range of software tools for the visualization and querying of metabolic networks and the analysis of 'omics' data. DiatomCyc provides different levels of information, from the cellular metabolic overview to single gene, protein, reaction and metabolite information, and includes literature references. Through the web interface, users can query the PGBD through keywords, the implemented BLAST utility and the genome browser. In addition, comparative analysis between different PGDBs can be carried out and the 'Omics Viewer' utility allows the graphical visualization of transcriptomics, proteomics and metabolomics data. Table 1 shows the types of data and the number of entries for each data type included in DiatomCyc, and compares these with data held in other species-specific MetaCyc-based databases, including EcoCyc (Keseler *et al.*, 2011), YeastCyc (Ball *et al.*, 2001), ChlamyCyc (May *et al.*, 2009), AraCyc (Zhang *et al.*, 2005) and HumanCyc (Romero *et al.*, 2005). Currently, DiatomCyc comprises 286 pathways, 1719 metabolic reactions and 1613 enzymes, which is comparable with other eukaryotic PGDBs, spanning both the central and parts of the secondary metabolism of *P. tricornutum*.

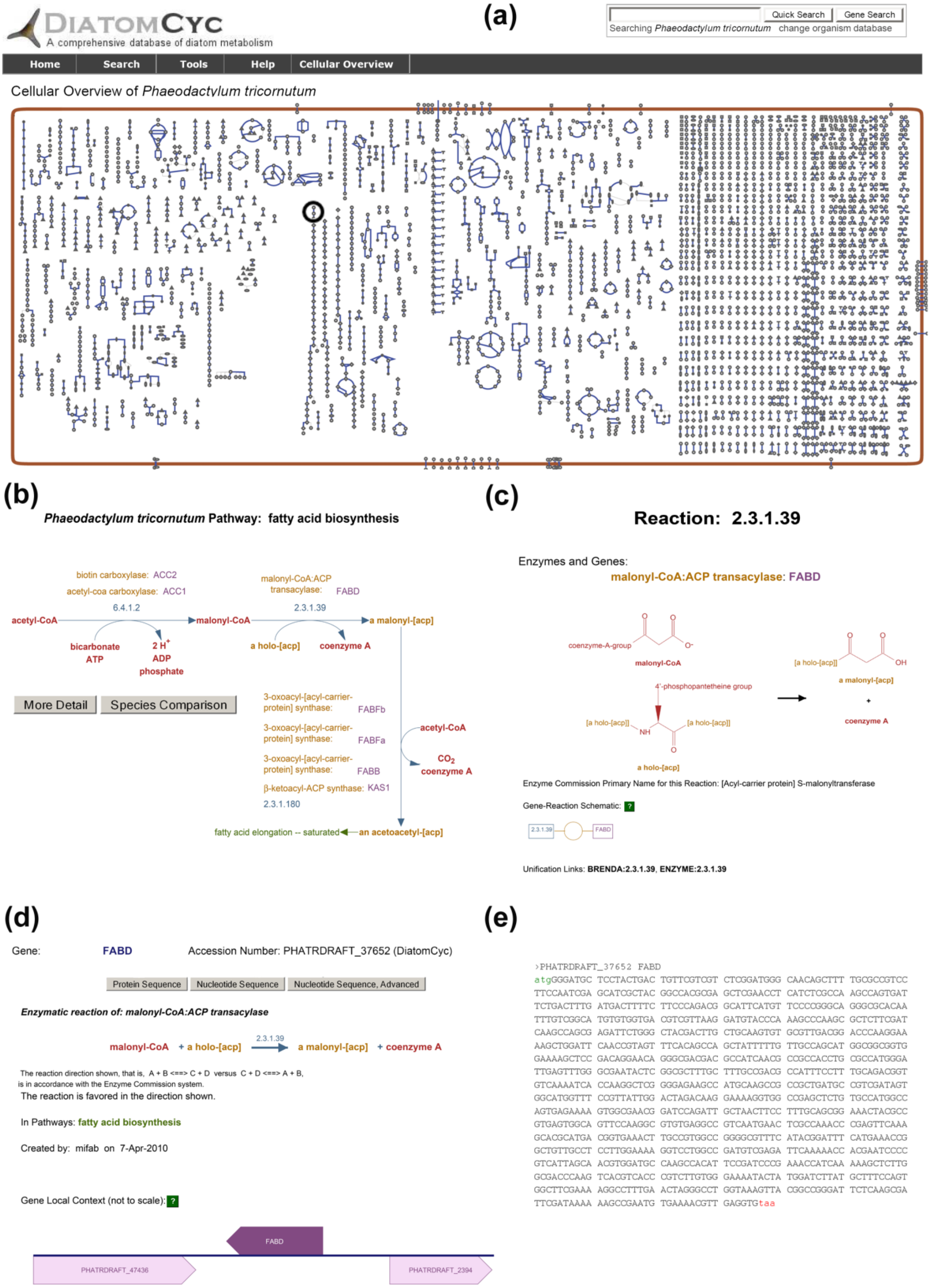
Content	**MetaCyc**	**EcoCyc (15.1)**	**YeastCyc (15.0)**	**ChlamyCyc (2010-03-03)**	**AraCyc (8.0)**	**HumanCyc (15.1)**	**DiatomCyc (1.0)**
Total Genes	-	5305	8069	15025	33602	21086	10647
Protein coding	-	5115	6607	14339	27416	17566	10565
Pathways	1745	281	152	263	393	263	286
Enzymatic reactions	9458	1492	981	1419	2627	1866	1719
Enzymes	7424	1467	708	2851	3203	3623	1613
Compounds	9187	2161	688	1066	2825	1196	1173

Table 3.1 Main contents of DiatomCyc and other PGDBs (version between brackets)

Pathway visualization and manual curation

Given the broad interest in diatoms as potential polyunsaturated fatty acid and biodiesel producers, the metabolic pathways related to lipid biosynthesis (i.e. saturated and polyunsaturated fatty acids, triacylglycerols and phospholipid biosynthesis) are completely charted in DiatomCyc. As described below, the available MetaCyc templates were adjusted or refined manually when required. To illustrate the content of DiatomCyc, the fatty acid biosynthetic pathway was taken as an example. In *P. tricornutum* the most abundant saturated fatty acid is palmitic acid (C16:0) (Siron *et al.*, 1989; Zhukova, 1995; Patil *et al.*, 2006) (Domergue *et al.*, 2002). Unlike mammals that use a single gene product to carry out the entire reaction set necessary to produce fatty acids*, P. tricornutum*, analogously to plants and bacteria, uses discrete proteins encoded by different genes involved in separate steps (http://www. diatomcyc.org/DIATOM/NEW-IMAGE?type=PATHWAY& object=PWY-5156). The pathway can be selected from the cellular overview of DiatomCyc (Figure 3.1a), and can be visualized in detail (Figure 3.1b). All reactions starting from acetyl-CoA to palmitate and stearate are associated with the corresponding enzymes and genes. By clicking on a specific reaction on the pathway (Figure 3.1c), the enzyme(s) and the associated gene(s) are shown, as well as a detailed graphical representation of the reaction. When a gene is selected (Figure 3.1d), different types of information become available. A short description of the gene, its genomic coordinates, its length and the molecular weight of the protein it encodes are visible in the main window. In the genome browser, the gene sequence (Figure 3.1e) and its translation are directly accessible by clicking the corresponding boxes. Furthermore, links to the relevant literature and external resources, such as links from each gene to UniProt, the JGI genome sequence, the expressed sequence tag (EST) database, the KEGG database and NCBI are provided.

The semi-automatic reconstruction of PGDBs is based on known metabolic pathways, often preventing the detection of variations within a given biochemical route. The biosynthesis of eicosapentaenoic acid (EPA) present in DiatomCyc is an example of the


(a)
DIATOMCYC
A comprehensive database of diatom metabolism
Quick Search
Gene Search
Searching *Phaeodactylum tricornutum* change organism database
Home
Search
Tools
Help
Cellular Overview

Cellular Overview of *Phaeodactylum tricornutum*

Phaeodactylum tricornutum **Pathway: fatty acid biosynthesis**

biotin carboxylase: ACC2
acetyl-coa carboxylase: ACC1
6.4.1.2
acetyl-CoA
malonyl-CoA
bicarbonate
ATP
2 H+
ADP
phosphate
malonyl-CoA:ACP transacylase: FABD
2.3.1.39
a holo-[acp]
coenzyme A
a malonyl-[acp]
3-oxoacyl-[acyl-carrier-protein] synthase: FABFb
3-oxoacyl-[acyl-carrier-protein] synthase: FABFa
3-oxoacyl-[acyl-carrier-protein] synthase: FABB
β-ketoacyl-ACP synthase: KAS1
2.3.1.180
acetyl-CoA
CO2
coenzyme A
fatty acid elongation -- saturated
an acetoacetyl-[acp]
More Detail
Species Comparison
(c)

Reaction: 2.3.1.39

Enzymes and Genes:

malonyl-CoA:ACP transacylase: FABD

Enzyme Commission Primary Name for this Reaction: [Acyl-carrier protein] S-malonyltransferase

Gene-Reaction Schematic: ?

Unification Links: **BRENDA:2.3.1.39**, **ENZYME:2.3.1.39**

Gene: **FABD** Accession Number: PHATRDRAFT_37652 (DiatomCyc)

Enzymatic reaction of: malonyl-CoA:ACP transacylase

malonyl-CoA + **a holo-[acp]** $\xrightarrow{2.3.1.39}$ **a malonyl-[acp]** + **coenzyme A**

The reaction direction shown, that is, A + B <==> C + D versus C + D <==> A + B, is in accordance with the Enzyme Commission system.

The reaction is favored in the direction shown.

In Pathways: **fatty acid biosynthesis**

Created by: mifab on 7-Apr-2010

Gene Local Context (not to scale): ?

```
>PHATRDRAFT_37652 FABD
atgGGGATGC TCCTACTGAC TGTTCGTCGT CTCGGATGGG CAACAGCTTT TGCGCCGTCC
TTCCAATCGA GCATCGCTAC GGCCACGCGA GCTCGAACCT CATCTCGCCA AGCCAGTGAT
TCTGACTTTG ATGACTTTTC TTCCCAGACG GCATTCATGT TCCCCGGGCA GGGCGCACAA
TTTGTCGGCA TGTGTGGTGA CGTCGTTAAG GATGTACCCA AAGCCCAAGC GCTCTTCGAT
CAAGCCAGCG AGATTCTGGG CTACGACTTG CTGCAAGTGT GCGTTGACGG ACCCAAGGAA
AAGCTGGATT CAACCGTAGT TTCACAGCCA GCTATTTTTG TTGCCAGCAT GGCGGCGGTG
GAAAAGCTCC GACAGGAACA GGGCGACGAC GCCATCAACG CCGCCACCTG CGCCATGGGA
TTGAGTTTGG GCGAATACTC GGCGCTTTGC TTTGCCGACG CCATTTCCTT TGCAGACGGT
GTCAAAATCA CCAAGGCTCG GGGAGAAGCC ATGCAAGCCG CCGCTGATGC CGTCGATAGT
GGCATGGTTT CCGTTATTGG ACTAGACAAG GAAAAGGTGG CCGAGCTCTG TGCCATGGCC
AGTGAGAAAA GTGGCGAACG GATCCAGATT GCTAACTTCC TTTGCAGCGG AAACTACGCC
GTGAGTGGCA GTTCCAAGGC GTGTGAGGCC GTCAATGAAC TCGCCAAACC CGAGTTCAAA
GCACGCATGA CGGTGAAACT TGCCGTGGCC GGGGCGTTTC ATACGGATTT CATGAAACCG
GCTGTTGCCT CCTTGGAAAA GGTCCTGGCC GATGTCGAGA TTCAAAAACC ACGAATCCCC
GTCATTAGCA ACGTGGATGC CAAGCCACAT TCCGATCCCG AAACCATCAA AAAGCTCTTG
GCGACCCAAG TCACGTCACC CGTCTTGTGG GAAAATACTA TGGATCTTAT GCTTTCCAGT
GGCTTCGAAA AGGCCTTTGA ACTAGGGCCT GGTAAAGTTA CGGCCGGGAT TCTCAAGCGA
TTCGATAAAA AAGCCGAATG TGAAAACGTT GAGGTGtaa
```

Figure 3.1 ***description on the next page***

Figure 3.1 ***previous page.*** Screenshots from DiatomCyc describing the pathway for the biosynthesis of fatty acids in *Phaeodactylum tricornutum*, and illustrating the different levels of information. (a) Cellular overview of the complete metabolism of *P. tricornutum*. Part of the pathway, shown in more detail on the information page (b) is encircled. (b) Details of genes and proteins associated with corresponding reactions. A general description of the pathway and links to the literature are provided (not shown in the figure). Comparative analysis can be carried out by the occurrence of the same pathway in different organisms (Species comparison). (c) Selection and visualization of a single reaction. (d) Gene information page that, in turn, includes links to external databases and literature references, and information relative to gene length and protein size. The genomic localization of the protein is graphically represented and the genomic coordinates are indicated. (e) Genome browser, protein and nucleotide sequences are accessible by clicking the gray boxes in (d).

literature-based manual curation that had been applied to construct the diatom PGDB. EPA is an omega-3 polyunsaturated fatty acid (PUFA), and one of the valuable compounds synthesized by diatoms. In *P. tricornutum*, the accumulation of EPA is variable in time and can reach remarkable portions of the total fatty-acid fraction (Siron *et al.*, 1989; Zhukova, 1995; Patil *et al.*, 2006) . Fatty acid profiling showed that *P. tricornutum* accumulates large quantities of EPA, whereas intermediates of the pathway are present only in traces (Siron *et al.*, 1989; Zhukova, 1995; Patil *et al.*, 2006), suggesting that the pathway has been optimized for the specific accumulation of EPA. The pathway template provided by MetaCyc had been adapted according to the outcome of other studies (Spiekermann *et al.*, 2003; Wen and Chen, 2003) to reconstruct the putative pathway that leads to the EPA synthesis (Figure 3.2). The reconstructed pathway is directly connected to the previously described fatty acid biosynthesis pathways, from which stearoyl-ACP (C18:0) is desaturated to oleoyl-ACP by a Δ9 desaturase, encoded by *PHATRDRAFT_9316*. The set of genes involved in the subsequent desaturation and elongation steps consists of two Δ6 elongases

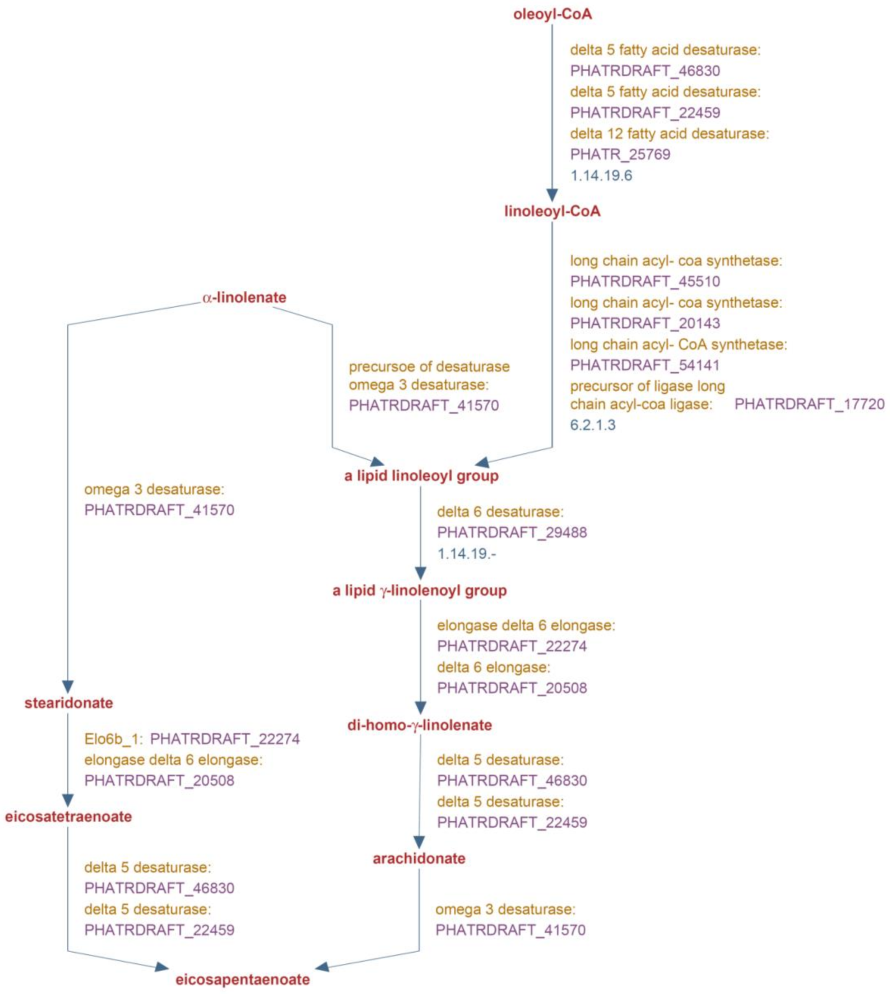


Figure 3.2. DiatomCyc screenshot of eicosapentaenoic acid biosynthesis in *Phaeodactylum tricornutum*

(encoded by *PHATRDRAFT_22274* and *PHATRDRAFT_20508*), two Δ5 desaturases (*PHATRDRAFT_22459* and *PHATR_46830*), a Δ12desaturase (*PHATR_25769*), a Δ6 desaturase (*PHATRDRAFT_ 29488*) and an omega-3 desaturase (*PHATRDRAFT_41570*) (Figure 3.2).

The 'Omics Viewer'

The web interface of DiatomCyc allows the user to import and picture several types of data, including microarray expression data, proteomics data, metabolomics data, reaction flux data or data from any other high-throughput 'omics' experiment, related to genes, proteins, reactions or compounds. With the 'Omics Viewer' tool implemented in DiatomCyc, data sets can be visualized on the cellular overview, and reactions and genes are marked with different color scales. As examples, the regulation of two isoprenoid precursor pathways will be illustrated: those of methyl-erythrol 4-phospate (MEP) and the mevalonate (MVA). Both pathways were mapped in DiatomCyc: the MEPP (http:// www.diatomcyc.org/DIATOM/NEW-IMAGE?type=PATHWAY& object=NONMEVIPP-PWY) and MVAP (http://www.diatomcyc. org/DIATOM/NEW-IMAGE?type=PATHWAY&object=PWY-922). Isoprenoids are important precursors for the biosynthesis of carotenoids and other pigments, such as b-carotene and the diatom-specific fucoxanthin (Bertrand, 2010), which have several applications in the food and pharmaceutical industries (Pangestuti and Kim, 2011). Previously*, P. tricornutum* had been found to use these two different pathways to synthesize isoprenoid precursors, just like plants (Cvejic *et al.*, 2000). It is commonly believed that the MEPP operates in the chloroplasts, providing precursors for carotenoids, and that the MVAP contributes to the biosynthesis of sterols in the cytosol (Massé *et al.*, 2004). Carbon-labeling experiments (Cvejic *et al.*, 2000; Massé *et al.*, 2004) demonstrated that the two pathways are triggered by different precursors. The main carbon source for the MEPP is CO_2, whereas the MVAP uses acetate-derived carbon.

Currently, the Diatom EST Database (Maheswari *et al.*, 2009) represents the only publicly available large-scale collection of gene expression data. Although this data set is still relatively small and does not allow the analysis of statistically significant changes in gene expression, it was used nonetheless to illustrate the utility of the DiatomCyc 'Omics Viewer' tools. As an example, the transcriptional reprogramming of the two isoprenoid precursor pathways under different culture conditions will be used. With the 'Omics

Viewer' utility, the imported EST data are mapped on the cellular pathway overview, and the reactions and genes involved are marked with different color scales (Figure 3.3). Specifically, as an example of the potential utility, we imported the EST data (Maheswari *et al.*, 2009) of the three culture conditions that affected the EST levels of

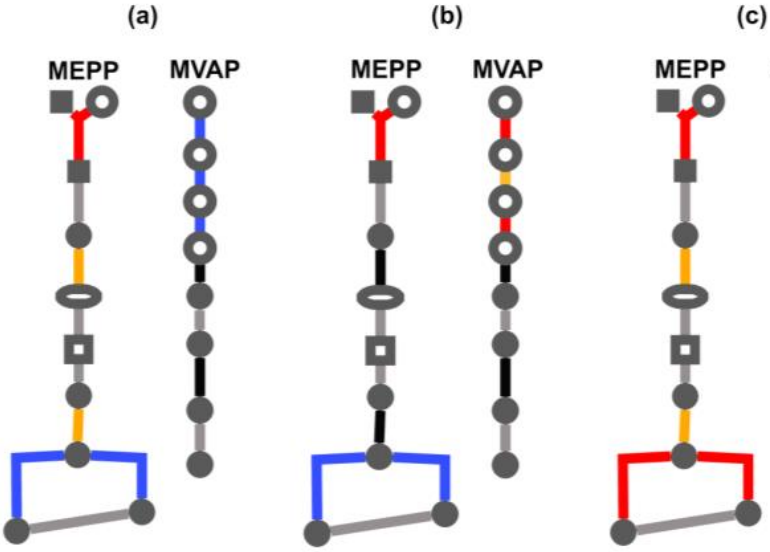


Figure 3.3. 'Omics Viewer' in DiatomCyc. Examples of visualization of expressed sequence tag (EST) data on the reconstructed methyl-erythrol-4-phosphate pathway (MEPP) and mevalonate pathway (MVAP). Each icon represents a single metabolite: squares, carbohydrates; ellipses, pyrimidines; circles, other compounds; and filled shapes indicate phosphorylated compounds. Different inductions of MEPP and MVAP under different culture conditions are shown, i.e. pseudo steady-state growth of *Phaeodactylum tricornutum* in the presence of: a high concentration of CO_2 (a); 2E,4E-decadienal (b); and under iron limitation (c) (Maheswari et al., 2009). Before importing EST data in DiatomCyc, we normalized the data by dividing the absolute number of ESTs for every gene in a given condition by the total number of ESTs for that condition. Subsequently, these normalized values were divided by the value relative to the condition used as a comparison (the standard condition). Colors indicate repression of a gene in a given condition (blue), no difference in expression relative to the standard condition (black) or induction of a gene in a given condition (red). Genes for which there are ESTs in the 'stress' condition but not in the standard condition are indicated in orange, and those for which no unambiguous EST data are available in any condition are indicated in light gray.

MVAP and MEPP most markedly when compared with those in control pseudo steady-state cultures, namely the presence of a high concentration of CO_2, the addition of a toxic aldehyde and iron limitation. This analysis indicated that contigs corresponding to genes involved in the MEPP were induced in cells exposed to a high concentration of CO_2, whereas genes involved in the MVAP were not (Figure 3.3a), according to their main carbon sources (Cvejic *et al.*, 2000). In contrast, in the presence of the toxic aldehyde 2E,4E-decadienal (DD), the MVAP genes were induced, but not those of the MEPP (Figure 3.3b). This observation is plausible, because an *S. cerevisiae* mutant lacking the *ERG6* gene (encoding the Δ(24) sterol C-methyltransferase involved in ergosterol biosynthesis) has been demonstrated to be significantly more sensitive to DD, which points to a direct link between the stress induced by the aldehyde and the synthesis of sterols (Adolph et al., 2004). Like many other toxic oxylipins, DD increases membrane permeability and programmed cell death in various organisms (Ribalet *et al.*, 2007; Adolph *et al.*, 2004). Sterols, fed by the MVAP, might represent an important cellular defense barrier against such toxic compounds. Finally, the lack of iron in culture media has been reported to trigger a rearrangement of the photosynthetic components, and to stimulate the expression of genes involved in the biosynthesis of chlorophyll, carotenoids and their precursors (Allen *et al.*, 2008). Accordingly, the EST data indicate that the MEPP genes, as well as those of the MVAP, are induced under iron restriction (Figure 3.3c). Although these data need to be interpreted with care because of the limited number of ESTs in the database, the above examples illustrate the usefulness of DiatomCyc as an interactive and illustrative laboratory tool in the study of diatom metabolism under different culture or stress conditions.

Completion of gaps in the current carbohydrate model

A genome-based overview of diatom metabolism provided the first insights into the acquisition of dissolved inorganic carbon, photorespiration and a detailed synthesis of carbohydrate metabolism (Kroth *et al.*, 2008). Our results confirmed the proposed carbohydrate model and allowed us to fill in several of the remaining gaps. For example, acetyl-CoA conversion to acetate appears to occur via a two-step reaction (Figure 3.4), involving phosphate acetyltransferase (EC 2.3.1.8, encoded by *PHATRDRAFT_50625* and *PHATRDRAFT_48580*) that converts acetyl-CoA to acetyl phosphate. The latter is dephosphorylated to acetate by acetate kinase (EC 2.7.2.1, encoded by PtACK, *PHATRDRAFT_36256*). The photorespiration pathway was further completed by identifying *PHATRDRAFT_56499* as the gene encoding glycerate kinase (EC 2.7.1.31; Figure 3.3). Contradictory to previous postulations (Kroth *et al.*, 2008), the presence of an N-terminal transit peptide that could be predicted *in silico* suggests that the enzyme might be mitochondrial rather than plastidial (Emanuelsson *et al.*, 2007). Notably, the identification of one ortholog (*THAPSDRAFT_261750*) in *Thalassiosira pseudonana* suggests that glycerate kinase might be conserved in diatoms. In agreement with a previous study (Kroth *et al.*, 2008), we also identified genes potentially involved in the biosynthesis of the polysaccharide storage compound chrysolaminaran, a β-1,3 glucan branched with β -1,6 linkages. A β-1,3-glucan synthase (encoded by *PHATRDRAFT_55327*) and three putative β-1,6 branching enzymes (encoded by *PHATRDRAFT_48300*, *PHATRDRAFT_56509* and *PHATRDRAFT_50238*) were predicted. The latter three genes share orthology relationships with two genes encoding b-1,6-glucan synthases of *S. cerevisiae* (i.e. *YPR159W* and *YGR143W*).

Conclusions

PGDBs and genome-scale metabolic networks represent relevant resources for the study of cellular metabolism. As shown by well-curated databases, such as EcoCyc (Keseler *et al.*, 2011), the level of information can have an impressive resolution and an important impact on the scientific community by becoming a common laboratory tool, in particular for comparative analysis and visualization of high-throughput experiments. DiatomCyc is a first step towards a comprehensive overview of diatom metabolism, and provides a user-friendly, interactive platform for current and future diatom research. The comparison between DiatomCyc and other PGDBs (Table 3.1) reflects the current status of knowledge of *P. tricornutum* metabolism, which still contains gaps in some metabolic pathways. Such gaps can sometimes be ascribed to missing or incomplete EC numbers in MetaCyc. Furthermore, isolated reactions without a gene association might involve diatom-specific genes or pathways. Proteins with obscure functions (POFs) that cannot be linked *in silico* to any cellular process yet, generally represent a significant portion in every eukaryotic genome, ranging from 18 to 38%) (Gollery *et al.*, 2006). In *P. tricornutum*, 44% of the genes code for such POFs and lack detectable functional domains (Maheswari *et al.*, 2010). Functional annotation and assembly of the *P. tricornutum* genome is an ongoing project; hence, continuous curation will be necessary to address new annotations and to complete the metabolic pathways. Therefore, DiatomCyc will be updated on a regular basis. The orthology-based approach that was used to construct DiatomCyc led to the characterization of the central metabolism of the model diatom *P. tricornutum*, with particular interest in pathways with a biotechnological relevance, such as carbohydrate, isoprenoid and fatty acid biosynthesis. Moreover, it allows us to propose hypothetical routes for unknown pathways, such as of PUFA biosynthesis. The utility and power of the methods used are illustrated by the identification of uncommon eukaryotic pathways as described in Chapter 4. DiatomCyc will become a powerful resource, both for fundamental research and the development of metabolic engineering strategies aiming at the industrial exploitation of diatoms.

Experimental procedures

Orthology prediction and database construction

The annotated genomes of *Phaeodactylum tricornutum* (Bowler *et al.*, 2008), 10 organisms with a published genome-scale metabolic model and curated annotation (*Arabidopsis thaliana, Homo sapiens, Saccharomyces cerevisiae, Escherichia coli, Helicobacter pilory, Neisseria meningiditis, Methanococcus jannaschii, Lactococcus lactis, Lactobacillus plantarum* and *Bacillus subtilis*) and 12 additional species (*Acaryochloris marina, Ectocarpus siliculosus, Prochlorococcus marinus, Ostreococcus tauri, Ostreococcus lucimarinus, Cyanidioschyzon merolae, Synechococcus* sp*.* JA-3-3Ab*, Synechococcus* sp. JA-2-3Ba*, Chlamydomonas reinhardtii, Trichodesmium erythraeum, Thalassiosira pseudonana*, and *Entamoeba histolytica*) were downloaded from the KEGG database (www.genome.jp/kegg , version February 2010; Kanehisa and Goto, 2000). Additionally the genome of *Aureococcus anophagefferens* was downloaded from the JGI website (http://genome.jgi.doe.gov/Auran1/Auran1.home.html).

Orthology prediction was carried out with Inparanoid 3.0 with the default cutoff value of 50 bits (Brien *et al.*, 2005; Remm *et al.*, 2001). Gene-to-function and function-to-reaction associations were transferred semi-automatically by means of the KEGG (Kanehisa and Goto, 2000) and MetaCyc (http://metacyc.org, Caspi *et al*., 2010) databases as input. In addition, FASTA headers of the reference genomes were screened and mined to improve the functional annotation of the *P. tricornutum* genome. Functions of the translated gene sequences with the highest score among the reference organisms were transferred to the corresponding *P. tricornutum* orthologs, yielding the primary genome-scale metabolic network. The genome-scale network was converted into a PathoLogic-specific data set according to the Pathway Tools documentation (Karp *et al.,* 2010), imported in Pathway Tools 15.0 (Karp *et al*., 2002) and, subsequently, refined and manually curated with literature references and bioinformatic tools, such as InterProScan (Hunter *et al.*, 2009), TargetP (Emanuelsson *et al.,* 2007) and TransportDB (Ren *et al.*, 2007). Metabolic pathways absent in the MetaCyc framework were added manually.

References

Adolph, S., Bach, S., Blondel, M., Cueff, A., Moreau, M., Pohnert, G., Poulet, S.A., Wichard, T. and Zuccaro, A. (2004) Cytotoxicity of diatom-derived oxylipins in organisms belonging to different phyla. *The Journal of experimental biology*, **207**, 2935–46.

Allen, A.E., Laroche, J., Maheswari, U., Lommer, M., Schauer, N., Lopez, P.J., Finazzi, G., Fernie, A.R. and Bowler, C. (2008) Whole-cell response of the pennate diatom Phaeodactylum tricornutum to iron starvation. *Proceedings of the National Academy of Sciences of the United States of America*, **105**, 10438–43.

Allen, A.E., Moustafa, A., Montsant, A., Eckert, A., Kroth, P.G. and Bowler, C. (2012) Evolution and functional diversification of fructose bisphosphate aldolase genes in photosynthetic marine diatoms. *Molecular biology and evolution*, **29**, 367–79.

Baart, G.J.E., Willemsen, M., Khatami, E., Haan, A. de, Zomer, B., Beuvery, E.C., Tramper, J. and Martens, D.E. (2008) Modeling Neisseria meningitidis B metabolism at different specific growth rates. *Biotechnology and bioengineering*, **101**, 1022–35.

Ball, C. a, Jin, H., Sherlock, G., *et al.* (2001) Saccharomyces Genome Database provides tools to survey gene expression and functional analysis data. *Nucleic acids research*, **29**, 80–1.

Bertrand, M. (2010) Carotenoid biosynthesis in diatoms. *Photosynthesis research*, **106**, 89–102.

Bowler, C., Allen, A.E., Badger, J.H., *et al.* (2008) The Phaeodactylum genome reveals the evolutionary history of diatom genomes. *Nature*, **456**, 239–44.

Brien, K.P.O., Remm, M., Sonnhammer, E.L.L. and O'Brien, K.P. (2005) Inparanoid: a comprehensive database of eukaryotic orthologs. *Nucleic acids research*, **33**, D476–80.

Caspi, R., Altman, T., Dale, J.M., *et al.* (2010) The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic acids research*, **38**, D473–9.

Chang, R.L., Ghamsari, L., Manichaikul, A., *et al.* (2011) Metabolic network reconstruction of Chlamydomonas offers insight into light-driven algal metabolism. *Molecular systems biology*, **7**, 518.

Chisti, Y. (2007) Biodiesel from microalgae. *Biotechnology advances*, **25**, 294–306.

Cvejic, J., Cvejić, J.H. and Rohmer, M. (2000) CO2 as main carbon source for isoprenoid biosynthesis via the mevalonate-independent methylerythritol 4-phosphate route in the marine diatoms Phaeodactylum tricornutum and Nitzschia ovalis. *Phytochemistry*, **53**, 21–8.

Domergue, F., Lerchl, J., Zähringer, U. and Heinz, E. (2002) Cloning and functional characterization of Phaeodactylum tricornutum front-end desaturases involved in eicosapentaenoic acid biosynthesis. *European Journal of Biochemistry*, **269**, 4105–4113.

Emanuelsson, O., Brunak, S., Heijne, G. von and Nielsen, H. (2007) Locating proteins in the cell using TargetP, SignalP and related tools. *Nature protocols*, **2**, 953–71.

Falkowski, P.G., Katz, M.E., Milligan, A.J., Fennel, K., Cramer, B.S., Aubry, M.P., Berner, R.A., Novacek, M.J. and Zapol, W.M. (2005) The Rise of Oxygen over the Past 205 Million Years and the Evolution of Large Placental Mammals. *Science*, **309**, 2202.

Field, C.B., Behrenfeld, M.J., Randerson, J.T. and Falkowski, P. (1998) Primary Production of the Biosphere: Integrating Terrestrial and Oceanic Components. *Science*, **281**, 237–240.

Fong, S.S., Burgard, A.P., Herring, C.D., Knight, E.M., Blattner, F.R., Maranas, C.D. and Palsson, B.O. (2005) In silico design and adaptive evolution of Escherichia coli for production of lactic acid. *Biotechnology and bioengineering*, **91**, 643–8.

Gino JE Baart (2007) Modeling Neisseria meningitidis metabolism: from genome to metabolic fluxes. *Genome Biology*.

Giordano, M., Beardall, J. and Raven, J. a (2005) CO2 concentrating mechanisms in algae: mechanisms, environmental modulation, and evolution. *Annual review of plant biology*, **56**, 99–131.

Gollery, M., Harper, J., Cushman, J., Mittler, T., Girke, T., Zhu, J.-K., Bailey-Serres, J. and Mittler, R. (2006) What makes species unique? The contribution of proteins with obscure features. *Genome biology*, **7**, R57.

Hua, Q., Joyce, A.R., Fong, S.S. and Palsson, B.Ø. (2006) Metabolic Analysis of Adaptive Evolution for In Silico-Designed Lactate-Producing Strains.

Hunter, S., Apweiler, R., Attwood, T.K., *et al.* (2009) InterPro: the integrative protein signature database. *Nucleic acids research*, **37**, D211–5.

Kanehisa, M. and Goto, S. (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, **28**, 27–30.

Karp, P.D., Paley, S. and Romero, P. (2002) The Pathway Tools software. *Bioinformatics (Oxford, England)*, **18 Suppl 1**, S225–32.

Keseler, I.M., Collado-Vides, J., Santos-Zavaleta, A., *et al.* (2011) EcoCyc: a comprehensive database of Escherichia coli biology. *Nucleic acids research*, **39**, D583–90.

Kroth, P.G., Chiovitti, A., Gruber, A., *et al.* (2008) A model for carbohydrate metabolism in the diatom Phaeodactylum tricornutum deduced from comparative whole genome analysis. *PloS one*, **3**, e1426.

Maheswari, U., Jabbari, K., Petit, J.-L., *et al.* (2010) Digital expression profiling of novel diatom transcripts provides insight into their biological functions. *Genome biology*, **11**, R85.

Maheswari, U., Mock, T., Armbrust, E.V. and Bowler, C. (2009) Update of the Diatom EST Database: a new tool for digital transcriptomics. *Nucleic acids research*, **37**, D1001–5.

Massé, G., Belt, S.T., Rowland, S.J. and Rohmer, M. (2004) Isoprenoid biosynthesis in the diatoms Rhizosolenia setigera (Brightwell) and Haslea ostrearia (Simonsen). *Proceedings of the National Academy of Sciences of the United States of America*, **101**, 4413–8.

May, P., Christian, J.-O., Kempa, S. and Walther, D. (2009) ChlamyCyc: an integrative systems biology database and web-portal for Chlamydomonas reinhardtii. *BMC genomics*, **10**, 209.

Notebaart, R. a, Enckevort, F.H.J. van, Francke, C., Siezen, R.J. and Teusink, B. (2006) Accelerating the reconstruction of genome-scale metabolic networks. *BMC bioinformatics*, **7**, 296.

Pangestuti, R. and Kim, S.-K. (2011) Biological activities and health benefit effects of natural pigments derived from marine algae. *Journal of Functional Foods*, **3**, 255–266.

Patil, V., Källqvist, T., Olsen, E., Vogt, G. and Gislerød, H.R. (2006) Fatty acid composition of 12 microalgae for possible use in aquaculture feed. *Aquaculture International*, **15**, 1–9.

Provost, a. and Bastin, G. (2004) Dynamic metabolic modelling under the balanced growth condition. *Journal of Process Control*, **14**, 717–728.

Radrich, K., Tsuruoka, Y., Dobson, P., Gevorgyan, A., Swainston, N., Baart, G. and Schwartz, J.-M. (2010) Integration of metabolic databases for the reconstruction of genome-scale metabolic networks. *BMC systems biology*, **4**, 114.

Remm, M., Storm, C.E. V and Sonnhammer, E.L.L. (2001) Automatic Clustering of Orthologs and In-paralogs from Pairwise Species Comparisons. , 1041–1052.

Ren, Q., Chen, K. and Paulsen, I.T. (2007) TransportDB: a comprehensive database resource for cytoplasmic membrane transport systems and outer membrane channels. *Nucleic acids research*, **35**, D274–9.

Ribalet, F., Wichard, T., Pohnert, G., Ianora, A., Miralto, A. and Casotti, R. (2007) Age and nutrient limitation enhance polyunsaturated aldehyde production in marine diatoms. *Phytochemistry*, **68**, 2059–67.

Romero, P., Wagg, J., Green, M.L., Kaiser, D., Krummenacker, M. and Karp, P.D. (2005) Computational prediction of human metabolic pathways from the complete human genome. *Genome biology*, **6**, R2.

Siron, R., Giusti, G. and Berland, B. (1989) Changes in the fatty acid composition of Phaeodactylum tricornutum and Dunaliella tertiolecta during growth and under phosphorus deficiency. *Marine Ecology Progress Series*, **55**, 95–100.

Smid, E.J., Molenaar, D., Hugenholtz, J., Vos, W.M. de and Teusink, B. (2005) Functional ingredient production: application of global metabolic models. *Current opinion in biotechnology*, **16**, 190–7.

Spiekermann, P., Lerchl, J., Beckmann, C., Kilian, O., Kroth, P.G., Boland, W., Za, U. and Heinz, E. (2003) New Insight into Phaeodactylum tricornutum Fatty Acid Metabolism . Cloning and Functional Characterization of Plastidial and Microsomal ○ 12-Fatty Acid Desaturases 1 [w]. , **131**, 1–13.

Teusink, B., Enckevort, F.H.J. Van, Francke, C., Wiersma, A., Wegkamp, A., Smid, E.J. and Siezen, R.J. (2005) In Silico Reconstruction of the Metabolic Pathways of Lactobacillus plantarum : Comparing Predictions of Nutrient Requirements with Those from Growth Experiments. *Society*, **71**, 7253–7262.

Tirichine, L. and Bowler, C. (2011) Decoding algal genomes: tracing back the history of photosynthetic life on Earth. *The Plant journal : for cell and molecular biology*, **66**, 45–57.

Wen, Z.-Y. and Chen, F. (2003) Heterotrophic production of eicosapentaenoic acid by microalgae. *Biotechnology Advances*, **21**, 273–294.

Whitaker, J.W., McConkey, G. a and Westhead, D.R. (2009) The transferome of metabolic genes explored: analysis of the horizontal transfer of enzyme encoding genes in unicellular eukaryotes. *Genome biology*, **10**, R36.

Zhang, P., Foerster, H., Tissier, C.P., *et al.* (2005) MetaCyc and AraCyc. Metabolic Pathway Databases for Plant Research. *Changes*, **138**, 27–37.

Zhukova, N. (1995) Fatty acid composition of 15 species of marine microalgae. *Phytochemistry*, **39**, 351–356.

Supplementary Data

Supplementary Table S1. Orthology prediction results.

Chapter 4

Discovery of a eukaryotic Entner-Doudoroff pathway

Adapted from the published article:
Fabris, M.; Matthijs, M.; Rombauts, S.; Vyverman, V.; Goossens, A.; Baart, G.J.E.
The metabolic blueprint of *Phaeodactylum tricornutum* reveals a eukaryotic Entner-Doudoroff
glycolytic pathway. *The Plant Journal* (2012), 70 (6):1004-14.

Author contribution

MF performed the experiments, analyzed the data and wrote the manuscript.

Abstract

Diatoms are unicellular photosynthetic microalgae that dominate phytoplankton in coastal upwelling regions, where nutrients and light availability fluctuates drastically. Diatoms genomes contain genes of different origins, which have direct implications on their metabolism, characterized by the presence of unusual features. These are the result of the combined effect of a highly variable habitat and numerous gene transfers, which conferred diatoms a remarkable ecological success. Through a genome-wide approach we reconstructed the metabolic network of the model diatom *Phaeodactylum tricornutum* and we surprisingly identified an Entner-Doudoroff pathway, the ancient alternative for the glycolytic Embden-Meyerhof-Parnas pathway, and a phosphoketolase pathway, both uncommon in eukaryotes.

Here, we report the functional characterization of the Entner-Doudoroff pathway of *P. tricornutum* both *in vivo*, *in vitro.* The presence of multiple glycolytic pathways in *P. tricornutum* implies a high level of metabolic flexibility and efficiency, suggesting that the conservation of these unusual pathways might be related to faster responses to sudden changes in nutrient availability.

Introduction

Glycolysis is the fundamental biochemical pathway that converts glucose into pyruvate. The free energy released in this process is used to form ATP and NADH. The Embden-Meyerhof-Parnas pathway (EMPP, Fig 4.1) is one of the most common and conserved glycolytic pathway in all domains of life. However, some microorganisms lack a complete EMPP and use alternative pathways as a consequence. One alternative glycolytic pathway is the Entner-Doudoroff pathway (EDP, Fig. 4.1). The EDP is very common in several prokaryotic organisms but, so far, had never been experimentally identified in eukaryotes. The EDP was first identified as a glycolytic alternative in 1952 by Nathan Entner and Michael Doudoroff, in *Pseudomonas putida* (Entner and Doudoroff, 1952) and it is considered to be the ancient glycolytic pathway (Romano and Conway, 1996). To enter this pathways, glucose-6-phosphate is first converted to 6-phosphogluconate, generating one NADPH molecule. The EDP enzymes consist in a 6-phosphogluconate dehydratase (EDD, EC 4.2.1.12) and a 2-keto-3-deoxyphosphogluconate aldolase (EDA, EC 4.2.1.14). EDD catalyzes the first reaction, in which the dehydratation of 6PG produces one molecule of 2-keto-3-deoxyphosphogluconate (2K3DPG). The second step of the pathway, catalyzed by EDA, cleaves 2K3DPG to one molecule of pyruvate and one of glyceraldehyde-3-phosphate (GAP). Here, the pathway connects to the lower glycolysis, which can then further degrade GAP to pyruvate. Compared to the EMPP, the energetic yield of the EDP is lower, as it produces only one net molecule of ATP, while the net yield of the EMPP consists in two ATP molecules (Fig 4.1).

The phosphoketolase pathway (PKP, Fig 4.1) is a catabolic variant of the pentose phosphate pathway (PPP) and phosphoketolase (XFP) is the key enzyme. To date, eukaryotic phosphoketolases have been indicated only in a few fungal species (Sánchez *et al*., 2010; Liu et al. 2012) and two types of XFP activities have been described: fructose-6-phosphate phosphoketolase (EC 4.1.2.22) and xylulose-5-phosphate phosphoketolase (EC 4.1.2.9) (Sánchez *et al.*, 2010). Most XFPs have dual substrate specificity (Sánchez *et al*., 2010) and organisms with both XFP activities are able to utilize a unique metabolic sugar pathway, also indicated as bifid shunt, which is generally considered a distinctive taxonomic mark for *Bifidobacteria*. XFP converts a molecule of xylulose-5-phosphate, originated from the first reactions of the PPP to GAP and acetyl-phosphate (AcP). Similarly to that produced by the EDP, phosphoketolase-derived GAP can enter the lower EMPP and be converted to pyruvate, yielding one

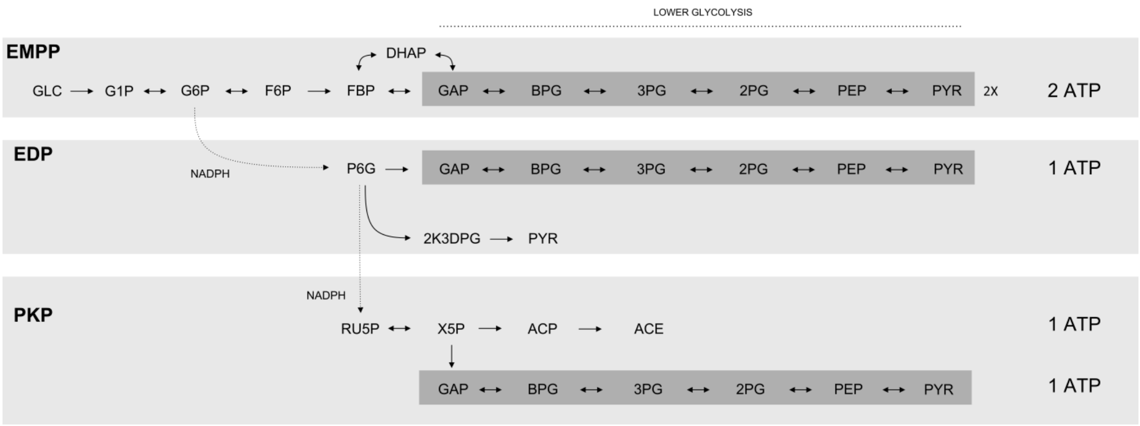


Figure 4.1. Schematic representation of the Embden–Meyerhof–Parnas pathway (EMPP), Entner-Doudoroff pathway (EDP) and phosphoketolase pathway (PKP). Abbreviations: GLC, glucose; G1P glucose-1-phosphate; G6P, glucose-6-phosphate; F6P, fructose-6-phosphate; FBP, fructose 2,6-bisphosphate; DHAP, di-hydroxy-acetone phosphate; GAP, glyceraldehyde-3-phosphate; BPG, 2,3-diphosphoglycerate; 3PG, 3-phosphoglycerate; 2PG, 2-phosphoglycerate; PEP, phosphoenolpyruvate; PYR, pyruvate, 6PG, 6-phospho gluconate; 2K3DPG, 2-keto-3-deoxyphosphogluconate; RU5P, ribulose-5-phopshate X5P, xylulose-5-phosphate; ACP, acetylphosphate; ACE, acetate.

molecule of ATP, whereas acetyl-phosphate is dephosphorylated by the enzyme acetate kinase (EC 2.7.2.1), yielding ATP and acetate (Fig 4.1).

This chapter describes the discovery of a eukaryotic EDP and the identification of PKP genes in *P. tricornutum*. Whereas the evidences for the presence of PK are computational, the occurence of the EDP in *P. tricornutum* has been demonstrated experimentally, both *in vivo* and *in vitro* and represents the first report of the occurrence of this pathway in a eukaryotic organism.

Results and discussion

In silico prediction of a phosphoketolase pathway

The reconstruction of the genome scale metabolic network made the *in silico* identification of a phosphoketolase pathway in *P. tricornutum* possible. Precisely, we report the presence of a xylulose-5-phosphate/fructose-6-phosphate phosphoketolase (PtXPK), encoded by the gene *PHATRDRAFT_36257*. PtXPK possibly catalyzes the cleavage of xylulose-5-phosphate to acetyl-phosphate and glyceraldehyde-3-phosphate (GAP). While the latter can be further converted to pyruvate by entering the lower part of the glycolysis, acetyl phosphate can be dephosphorylated to acetate by PtACK, hereby producing ATP. Hence, the presence of the enzymes PtXPK and PtAcK might confer to the PPP of diatoms the ability to generate ATP from the degradation of C5 sugars. Similar to the enzymes involved in the oxidative PPP (Kroth *et al.*, 2008), PtXPK and PtACK are putatively cytosolic enzymes (Emanuelsson *et al.*, 2007). Among the reference organisms used in the orthology prediction, *PtXPK* shared orthologs with Cyanobacteria (*Acaryochloris marina, Synechococcus* sp. and *Trichodesmium erythraeum*) and Lactobacillaceae (*Lactobacillus plantarum* and *Lactococcus lactis*) (Table S1). Interestingly, no orthologs were found in *T. pseudonana.* Similarly, through BLAST searches, it was not possible to identify genes showing similarities to *PtXFP* in any of the other available sequenced diatom genomes (*Fragilariopsis cylindrus, Pseudo-nitzschia multiseries* and *Thalassiosira oceanica*). In the genome of *P. tricornutum, PtXFP* is localized on chromosome 9 adjacent to the PtACK-encoding locus that catalyzes the next reaction in the pathway (Figure 4.2). The spatial association between phosphoketolase and acetate kinase open reading frames has been reported to occur relatively frequently in Proteobacteria and Cyanobacteria (Sánchez *et al.*, 2010) and might hint at a bacterial origin of the PKP in *P. tricornutum*, as indicated previously by computational analysis (Bowler *et al.*, 2008).

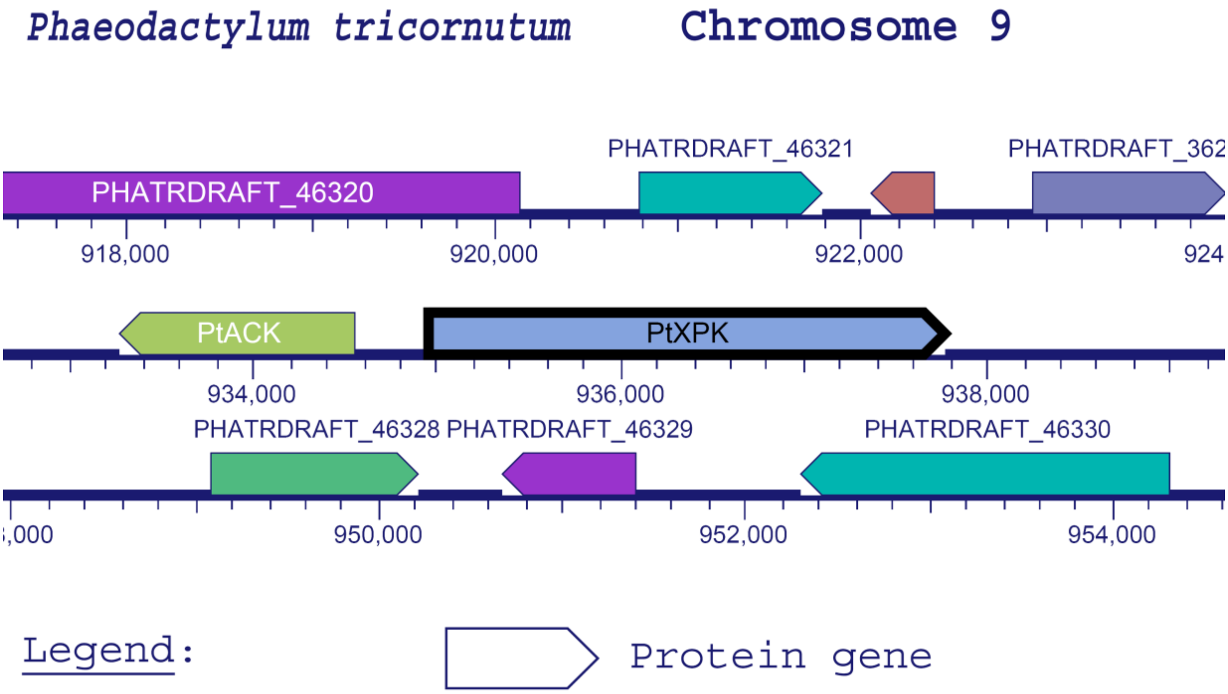


Figure 4.2. Genome browser of DiatomCyc. A particular region of chromosome 9 of *Phaeodactylum tricornutum*, where the genes encoding acetate kinase (PtACK) and phosphoketolase (PtXPK), key enzymes of the PKP, are spatially associated, is shown. Numbers indicate genomic positions.

In silico prediction of the Entner-Doudoroff Pathway

A striking finding that emerged from the reconstruction of DiatomCyc (Chapter 3) was the identification of the Entner-Doudoroff pathway (EDP) in *P. tricornutum*. Both EDP enzymes in *P. tricornutum* have a predicted mitochondrial signal peptide and seem to complete the lower glycolytic (EMPP) pathway present in the mitochondria (Kroth *et al.*, 2008). Whereas the EDD enzyme is widely conserved in nature, as a multifunctional protein involved in several cellular metabolic processes (Conway, 1998), EDA is considered the key and distinctive enzyme of the EDP. For instance, the *PtEDD* gene

(*PHATRDRAFT_20547*) has orthologs in 17 out of 21 reference organisms (Table S1). In contrast, the hits for *PtEDA* (*PHATRDRAFT_34120*) are predominantly restricted to prokaryotes, although orthologs had been identified in some eukaryotes as well, such as *Cyanidioschyzon merolae, Ostreococcus tauri* and *Ostreococcus lucimarinus* (Table S1). Notably, the two *EDP* genes of *P. tricornutum* are not part of the list of genes previously predicted to be of bacterial origin (Bowler *et al.* 2008). Orthologs of *PtEDD* and *PtEDA* have been identified also in *T. pseudonana, F. cylindrus, P. multistriata and T. oceanica,* suggesting that the EDP might be conserved in diatoms, in contrast to the PKP.

Experimental confirmation by genetic complementation in *Escherichia coli*

To confirm the function of the predicted EDP genes, we performed a genetic complementation experiment in the triple knockout mutant strain *ΔeddΔedaΔgnd* of *E. coli* and assayed growth on minimal medium containing gluconate as the sole carbon source (Figure 4.3). In this mutant*,* the native ED genes *eda* (*b4477*) and *edd* (*b3771*) were inactivated by knockout. Additionally, to avoid the gluconate flux to be channeled through the PPP, the gene encoding a 6-phosphogluconate dehydrogenase (*GND*, *b2029*) was deleted as well. The resulting triple KO strain was unable to grow on the gluconate-containing minimal medium (Figure 4.3.). Transformation of the *E. coli* mutant with a polycistronic plasmid carrying the *PtEDA* and *PtEDD* genes from *P. tricornutum* restored growth on minimal gluconate medium (Figure 4.3), thus confirming the functionality of the PtEDA and PtEDD proteins. The restored EDP allowed the metabolization of gluconate in the bacterial cell, after its phosphorylation to 6-phosphogluconate by endogenous gluconate kinase.

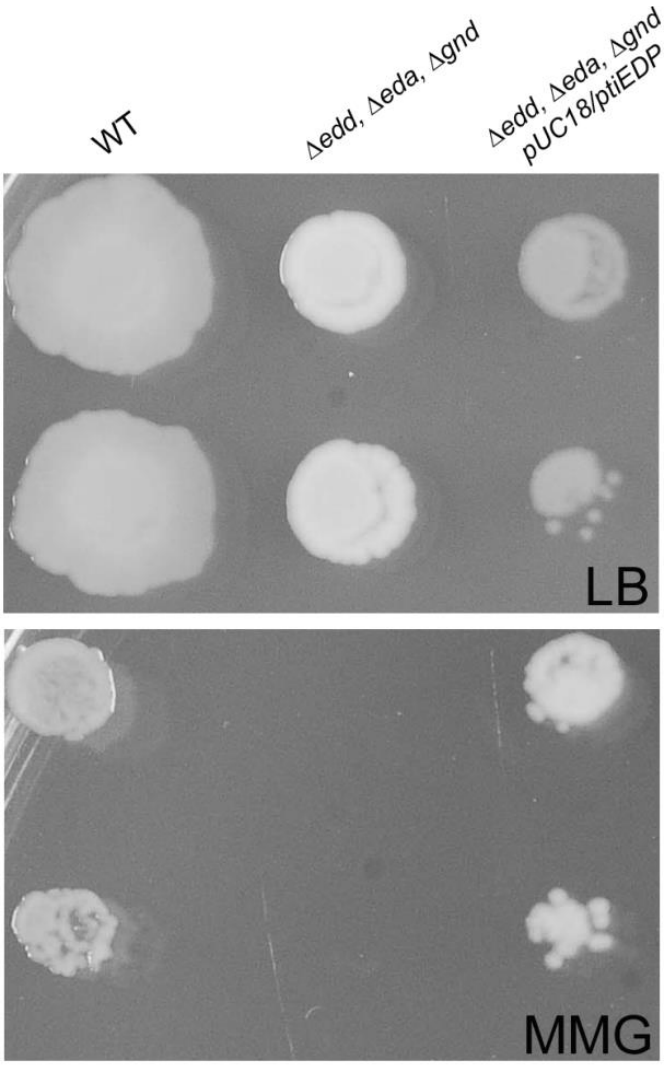


Figure 4.3. Genetic complementation of the Entner–Doudoroff pathway in *Escherichia coli*. Transformed wild-type and mutant *E. coli* strains (*ΔeddΔedaΔgnd*) were grown for 48 h at 37°C on LB medium or minimal medium supplemented with gluconate as the sole carbon source (MMG).

EDP genes expression analysis

The occurrence of the EDP was supported by the detection of *PtEDA* and *PtEDD* expression in cultured *P. tricornutum* cells. Cells grown under different light regimes showed differential transcript accumulation (Fig. 4.4). In particular, the transcript levels corresponding to PtEDA, the key enzyme of the EDP, showed a pronounced light modulated pattern, with a gradual decrease under continuous light and a sharp increase following the switch to a dark phase.

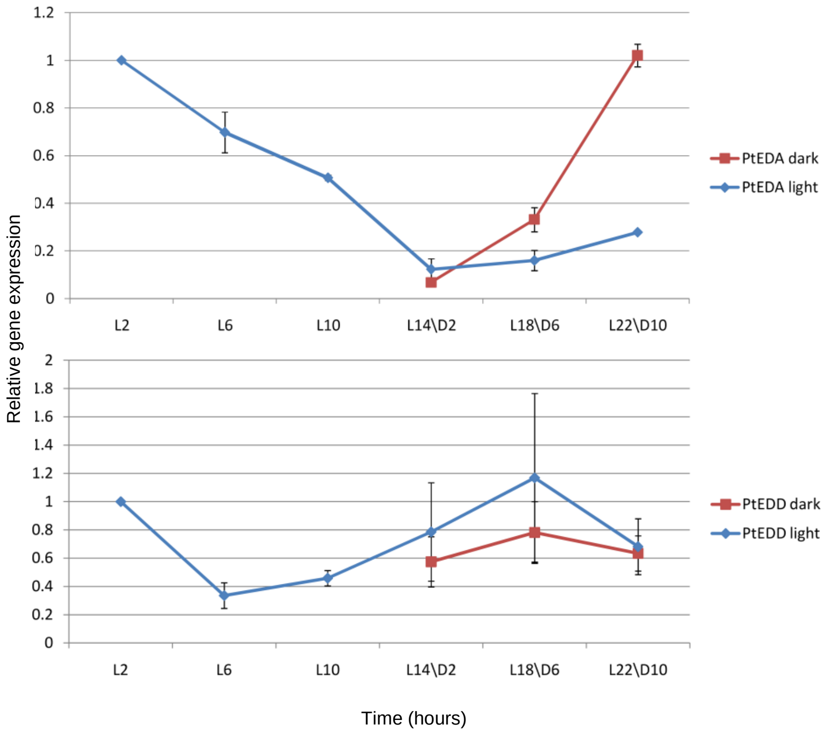


Figure 4.4. Expression of the *PtEDP* genes in different light regimes. Quantitative RT-PCR analysis for expression of *PtEDA* (upper panel) and *PtEDD* (lower panel) in *P. tricornutum* cultures grown under different light regimes. Cultures are either grown in continuous light (L) for 22 hours (blue line) or switched to the dark (D) after 12h of growth in the light (red line). The abscissa and the ordinate give the time points (in hours) and the normalized expression ratio with *Histone H4* and *Tubulin β chain* as the reference genes and the light 2h (L2) sample as the reference sample, respectively. Error bars represent standard errors of the mean of two biological replicates.

EDP enzymatic assay *in vitro*

The activity of the EDP enzymes was tested with an enzymatic assay with soluble protein extracts from *P. tricornutum* cells grown under different light regimes and the different *E. coli* strains that we generated. 6PG was used as substrate and its conversion to pyruvate was measured with a fluorometric assay (Table 4.1). Pyruvate was detected in cells of *P. tricornutum*, wild-type *E. coli* and mutant *E. coli ΔeddΔedaΔgnd* complemented with pUC18-PtiEDP, but not in cells of the uncomplemented *ΔeddΔedaΔgnd* mutant that lack an EDP. *E. coli* K12 *ΔeddΔedaΔgnd* transformed with pUC18-PtiEDP exhibits an exceptionally high EDP activity compared to *E.coli* K12 wild-type, possibly due to the presence of the elevated number of PtEDP genes transcripts in the bacterial cells. These observations provide direct evidence for the activity of the

	μg pyruvate/μg total soluble protein extract	Standard error
P. tricornutum (L)	0.155	0.026
P. tricornutum (D)	0.557	0.074
E. coli K12	0.439	0.054
E. coli K12 ΔeddΔedaΔgnd	0	0
E. coli K12 ΔeddΔedaΔgnd pUC18-PtiEDP	76.14	10.37

Table 4.1. Enzymatic activity of the PtEDP proteins. Activity of the PtEDP enzymes is expressed as μg pyruvate/μg total soluble protein and determined by a fluorometric assay after addition of 6PG to soluble enzyme fractions obtained from cells of *E. coli* K12 MG1655, *E. coli*K12 MG1655 *ΔeddΔedaΔgnd*, *E. coli* K12 MG1655 *ΔeddΔedaΔgnd* pUC18-PtiEDP, *P. tricornutum* grown either for 10 hours in the light (L) or for 10 hours in a prolonged dark phase (D). Standard errors of the mean of three biological replicates are indicated**.**

PtEDP enzymes in both endogenous and heterologous cells. Interestingly, the measured pyruvate concentration was higher in samples of *P. tricornutum* grown in a prolonged dark phase than in the samples obtained from diatoms grown in continuous light, which correlates with the *PtEDA* transcript levels (Fig.4.4).

Conclusions

The *in silico* analysis of the metabolic network of *P. tricornutum* highlighted a surprising redundancy of glycolytic pathways. We reported the identification of a predicted phosphoketolase pathway and, for the first time in eukaryotes, provided experimental evidence for the presence of a functional Entner-Doudoroff pathway. We focused our attention particularly on the latter, as it is commonly referred to as a bacterial metabolic pathway. Therefore, we functionally characterized the activity of EDP genes both *in vivo* and *in vitro*, providing transcriptional support to the occurrence of the pathway in *P. tricornutum* cultures.

The role of multiple glycolytic pathways in diatoms remains unclear. We hypothesize that the coordination of multiple central carbon pathways in *P. tricornutum* might be the consequence of cellular 'economical strategies'. Although the EDP produces less energy per glucose, it requires fewer resources to synthesize the enzymes than the EMPP (Carlson, 2007) and possibly enables a fast shunt to match the required production and demands for ATP/NADPH (Kramer and Evans, 2011). The shift to energetically inefficient metabolism originates from the trade-off between low investment cost in enzyme synthesis and a high operation cost for alternative catabolic pathways (Molenaar *et al.*, 2009). In *E. coli*, a low operation cost/high investment cost strategy (i.e.

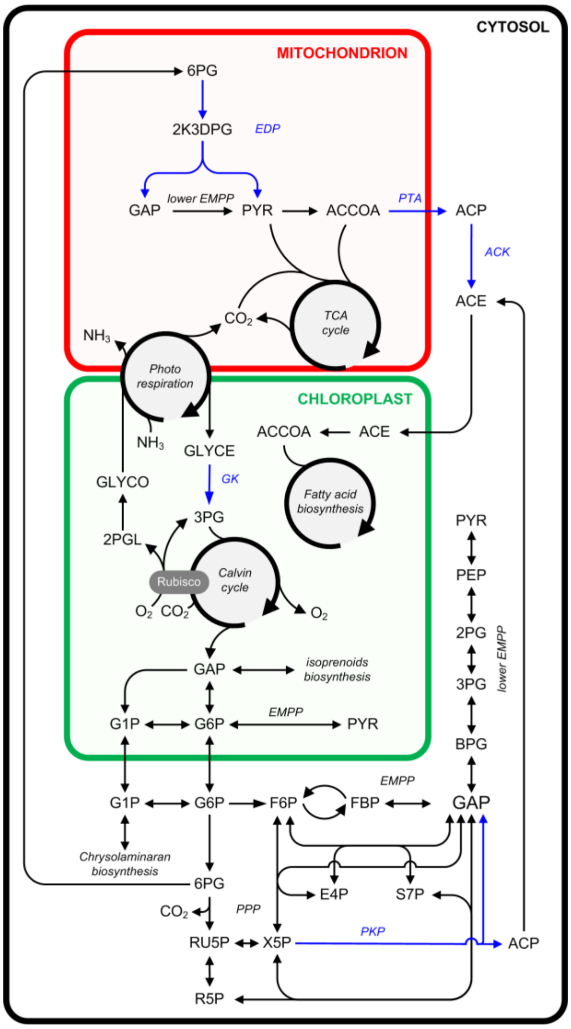


Figure 4.5. Simplified overview of carbohydrate metabolism and photorespiration in *Phaeodactylum tricornutum*. The Entner–Doudoroff pathway (EDP), phosphoketolase pathway (PKP, both described in Chapter 4) and completed gaps are indicated in blue. 2K3DPG, 2-keto-3-deoxyphosphogluconate; 2PGL, 2-phosphoglycolate; 3PG, 3-phosphoglycerate; 6PG, 6-phospho gluconate; ACCOA, acetyl-CoA; ACE, acetate; ACK, acetate kinase; ACP, acetylphosphate; BPG, 1,3-diphosphateglycerate; CO_2, carbon dioxide; EMPP, Embden–Meyerhof–Parnas pathway; E4P, erythrose-4-phosphate; FBP, fructose-1,6-bisphosphate; F6P, fructose-6-phosphate; GAP, glyceraldehyde-3-phosphate; GK, glycerate kinase; GLYCE, glycerate; GLYCO, glycolate; G1P, glucose-1-phosphate; G6P, glucose-6-phosphate; NH_3, ammonium; O_2, oxygen; PEP, phosphoenolpyruvate; 2PG, 2-phosphoglycerate; PPP, pentose phosphate pathway; PTA, phosphate acetyltransferase; PYR, pyruvate; R5P, ribulose-5-phosphate; RU5P, ribulose-5-phopshate; S7P, sedoeptulose-7-phosphate; X5P, xylulose-5-phosphate

maximizing energy yield), has been recognized as a competitive strategy during nutrient-limited chemostat growth (Schuetz *et al.*, 2007). The combined use of different pathways for glucose utilization might reflect a great metabolic flexibility: "expensive" pathways in terms of investment costs, such as the EMPP, are more energetically efficient and finely tuned by transcriptional regulation (Wessely *et al.*, 2011). In contrast, "cheaper" pathways, such as the EDP, might be less tightly regulated and be the predominant pathway in an organism subjected to frequent environmental changes, resulting in faster metabolic responses that provide an immediate selective advantage (Wessely *et al.*, 2011).

The implication that *P. tricornutum* has multiple pathways for glucose metabolism emphasizes its marked metabolic versatility. Their discovery highlights the efficacy of the orthology-based approach that we used in reconstructing *P. tricornutum's* metabolic network (Chapter 3).

Currently, the coordination of simultaneous and concurrent glycolytic pathways in the same organism has not been elucidated yet. This remains unclear even in bacteria, despite the extensive detailed knowledge available on their biology and metabolism and the relative "ease" of investigation that they allow, compared to other microorganisms. Questions about the conservation of the EDP in diatoms and their role prompt further efforts to elucidate the evolutionary and ecological advantage represented by the presence of such unusual pathways in eukaryotes.

Experimental procedures

Expression of *P. tricornutum* genes in *E. coli*

Wild-type *E. coli* strain K-12 MG1655 and the *ΔeddΔeda* mutant were kindly provided by Prof. Dr. Daniël Charlier (Vrije Universiteit Brussel, Belgium). A third gene deletion (*gnd*, 6-phosphogluconate dehydrogenase) was introduced into the latter strain to block the flux through the PPP by replacing the target gene by a kanamycin-resistant gene (Datsenko and Wanner, 2000). Luria Broth (LB) medium enriched with 50 µM kanamycin (Duchefa Biochemie; http://www.duchefa.com) was used to select the triple knockout *E. coli* mutants. The final mutant (*ΔeddΔedaΔgnd*) was checked with PCR.

Axenic cultures of the *P. tricornutum* CCAP 1055/1 were grown in f/2 medium without silica (Guillard and Ryther, 1962) at 21°C in a 12-h light/12-h dark regime (average 75 µmol•photons•m^{-2}•s^{-1}) and shaken at 100 rpm. Total RNA was extracted with Tri-Reagent (Molecular Research Center; http://www.mrcgene.com) according to the manufacturer's protocol. DNase treatment was done with RQ1 RNase-Free DNase (Promega; http://www.promega.com) and cDNA synthesized with the iScript cDNA synthesis kit (Bio-Rad; http://www.bio-rad.com). Predicted open reading frames of *EDD* and *EDA* (*PHATRDRAFT_34120 and PHATRDRAFT_20547*) were amplified with PrimeSTAR® HS DNA Polymerase (Takara Bio; http://www.takara-bio.com). For primer design, the JGI gene model of *PHATRDRAFT_20547* was manually adjusted. The full-lengths open reading frames of *EDD* and *EDA* were subsequentially cloned into the pUC18 vector as a polycistron (Ye *et al.*, 2010)to yield the pUC18-PtiEDP construct.

The *E. coli* mutants were transformed with the pUC18-PtiEDP vector by heat shock, selected on LB medium enriched with 100 µM ampicillin and verified by PCR. The minimal medium for the complementation experiments contained 15 g/l agar, 6.75 g/l NH4Cl, 1.25 g/l (NH4)2SO4, 1.15 g/l KH2PO4, 0.5 g/l NaCl, 0.5 g/l MgSO4.7H2O, 1 ml/l vitamin solution and 100 µl/l molybdic acid solution. The vitamin solution consisted of 4.89 g/l FeCl3.6H2O, 5 g/l CaCl2.2H2O, 1.3 g/l MnCl2.2H2O, 0.5 g/l CoCl2.6H2O, 0.94 g/l ZnCl2, 0.0311 g/l H3BO4, 0.4 g/l Na-EDTA.2H2O and 1.01 g/l thiamine-HCl. The molybdate solution contained 0.76 g/l molybdic acid. In the complementation assays, D-gluconic acid sodium salt (16.5 g/l) was used as the sole carbon source. Isopropyl β-D-1-thiogalactopyranoside (IPTG) was added to a final concentration of 500 µM to induce transgene expression.

Analysis of PtEDP gene expression

Complementary DNA (cDNA) obtained from RNA from axenic cultures of *P. tricornutum* CCAP 1055/1 grown in f/2 medium without silica (Guillard and Ryther, 1962) at 21°C either in continuous light (average 75 µmol•photons•m^{-2}•s-1) for 22 hours or switched to the dark after 12h in continuous light, was prepared and analyzed by quantitative Real Time PCR as described before (Huysman *et al.*, 2010). *Histone H4* (*H4*) and *Tubulin β chain* (*TubB*) were used as the internal control genes (Siaut *et al.*, 2007)for the normalization of the relative expression ratio of *PtEDA* and *PtEDD* transcripts. Primers for amplification of *PtEDA* (Fw 5'-CGCTACTTCGGATGATTGC-3', Rv-5'-GGAGTCGTGAGGGTGAAC-3') and *PtEDD* (Fw 5'-AGAAGCGAAGAACAGAATGG-3', Rv 5'-GGAGCGGCAATCACAATC-3') were designed with Beacon Designer (Premier Biosoft; www.premierbiosoft.com).

Analysis of PtEDP enzymatic activity

Axenic cultures of *P. tricornutum* CCAP 1055/1, were grown in ESAW medium (Harrison et al., 1980) at 21°C in a 16-h light/8-h dark regime (average 75 µmol•photons•m^{-2}•s^{-1}), harvested either after 10 hours of cultivation in the light or after 10 hours of prolonged dark phase. *E. coli* K-12 MG1655, *E.coli* K-12 MG1655 *ΔeddΔedaΔgnd* and *E.coli* K-12 MG1655 *ΔeddΔedaΔgnd* pUC18-PtiEDP were grown on LB medium, LB medium enriched with 50 µM kanamycin and 100 µM ampicillin respectively, for 12 hours at 37°C on an orbital shaker. Cells were harvested by centrifugation, washed in PBS and resuspended in 0.5 ml lysis buffer (50mM Tris-HCl pH7.5, 150mM NaCl, 1mM EDTA, 1% Triton X-100, Complete Protease Inhibitor) according to the manufacturer's protocol (Roche; www.roche.com). Resuspended samples were sonicated on ice for 5 minutes with 10 sec pulses with a Heat Systems Ultrasonics sonicator (Heat Systems Incorporated; www.misonix.com). Cell debris and non-solubilized material were removed by centrifugation (30 min at 14,000 g) and the supernatant was subsequently centrifuged (2 h at 40,000 g) in order to obtain the soluble enzyme fraction. The protein concentration was determined by the Bradford assay (Bio-Rad; www.bio-rad.com) according to the manufacturer's protocol and using Bovine Serum Albumin (BSA) as a standard.

The soluble protein concentration of *P. tricornutum* and *E.coli* lysates was equalized between samples of the same organism. The EDP in vitro reactions were carried out in 40 µl TRIS-HCl pH 7.5, 20 µl sodium arsenite 0.2 M, by adding 40 µl of soluble enzyme fraction and 40 µl of 0.2M 6PG as the substrate. The reaction mix was incubated for 90 min at 37°C and 21°C for *E. coli* and *P.tricornutum* samples, respectively. Pyruvate concentrations were determined fluorometrically with a Pyruvate Assay Kit (BioVision; http://www.biovision.com) according to manufacturer's protocol.

References

Bowler, C., Allen, A.E., Badger, J.H., *et al.* (2008) The Phaeodactylum genome reveals the evolutionary history of diatom genomes. *Nature*, **456**, 239–44.

Carlson, R.P. (2007) Metabolic systems cost-benefit analysis for interpreting network structure and regulation. *Bioinformatics (Oxford, England)*, **23**, 1258–64.

Conway, T. (1998) MINIREVIEW What ' s for Dinner ?: Entner-Doudoroff Metabolism in Escherichia coli. , **180**, 3495–3502.

Datsenko, K. a and Wanner, B.L. (2000) One-step inactivation of chromosomal genes in Escherichia coli K-12 using PCR products. *Proceedings of the National Academy of Sciences of the United States of America*, **97**, 6640–5.

Emanuelsson, O., Brunak, S., Heijne, G. von and Nielsen, H. (2007) Locating proteins in the cell using TargetP, SignalP and related tools. *Nature protocols*, **2**, 953–71.

Entner, N. and Doudoroff, M. (1952) GLUCOSE AND GLUCONIC ACID OXIDATION OF PSEUDOMONAS SACCHAROPHILA. *Journal of Biological Chemistry*, **196**, 853–862.

Huysman, M.J.J., Martens, C., Vandepoele, K., *et al.* (2010) Genome-wide analysis of the diatom cell cycle unveils a novel type of cyclins involved in environmental signaling. *Genome biology*, **11**, R17.

Kramer, D.M. and Evans, J.R. (2011) The importance of energy balance in improving photosynthetic productivity. *Plant physiology*, **155**, 70–8.

Kroth, P.G., Chiovitti, A., Gruber, A., *et al.* (2008) A model for carbohydrate metabolism in the diatom Phaeodactylum tricornutum deduced from comparative whole genome analysis. *PloS one*, **3**, e1426.

Molenaar, D., Berlo, R. van, Ridder, D. de and Teusink, B. (2009) Shifts in growth strategies reflect tradeoffs in cellular economics. *Molecular systems biology*, **5**, 323.

Sánchez, B., Zúñiga, M., González-Candelas, F., los Reyes-Gavilán, C.G. de and Margolles, A. (2010) Bacterial and eukaryotic phosphoketolases: phylogeny, distribution and evolution. *Journal of molecular microbiology and biotechnology*, **18**, 37–51.

Schuetz, R., Kuepfer, L. and Sauer, U. (2007) Systematic evaluation of objective functions for predicting intracellular fluxes in Escherichia coli. *Molecular systems biology*, **3**, 119.

Siaut, M., Heijde, M., Mangogna, M., Montsant, A., Coesel, S., Allen, A., Manfredonia, A., Falciatore, A. and Bowler, C. (2007) Molecular toolbox for studying diatom biology in Phaeodactylum tricornutum. *Gene*, **406**, 23–35.

Wessely, F., Bartl, M., Guthke, R., Li, P., Schuster, S. and Kaleta, C. (2011) Optimal regulatory strategies for metabolic pathways in Escherichia coli depending on protein costs. *Molecular systems biology*, **7**, 515.

Ye, Q., Cao, H., Yan, M., *et al.* (2010) Construction and co-expression of a polycistronic plasmid encoding carbonyl reductase and glucose dehydrogenase for production of ethyl (S)-4-chloro-3-hydroxybutanoate. *Bioresource technology*, **101**, 6761–7.

Chapter 5

Reconstruction of the mevalonate and the sterol biosynthetic pathways

Manuscript in preparation for publication
Fabris, M.; Matthijs, M.; Carbonelle, S.; Moses T.; Baart, G.J.E.; Vyverman, V.; Goossens, A.

Author contribution

MF performed the experiments, analyzed the data and wrote the manuscript.

Abstract

Diatoms are unicellular photosynthetic microalgae that dominate phytoplankton and have enormous impact on marine and terrestrial ecosystems. Endosymbiotic events and recurrent gene transfers uniquely shaped the genome and metabolism of diatoms, which contains features from all domains of life. The biosynthetic pathways of sterols, essential compounds in any eukaryotic cell, and many of the enzymes involved are evolutionarily conserved in eukaryotes. Although well characterized in most eukaryotes, the pathway leading to sterol synthesis in diatoms remained hitherto unidentified. Through DiatomCyc (www.diatomcyc.org), we reconstructed *in silico* the mevalonate and sterol biosynthetic pathways of the model diatom *Phaeodactylum tricornutum.* We experimentally verified the predicted pathways by using enzyme inhibitors, gene silencing and heterologous gene expression. Our analysis revealed the peculiar, chimeric, organization of the diatom sterol biosynthetic pathway, which possesses features of both plant and fungal pathways. Strikingly, it lacks a conventional squalene epoxidase and possesses an unusual oxidosqualene cyclase and a multifunctional isopentenyl diphosphate isomerase/squalene synthase enzyme. Finally, we detected a link between sterol biosynthesis and the accumulation of triacylglycerol.

The reconstruction of the *P. tricornutum* sterol pathway underscores the metabolic plasticity of diatoms and offers important insights for the engineering of diatoms for sustainable production of biofuels and high-value chemicals.

Introduction

Sterols are fundamental isoprenoid compounds in eukaryotes. They are important components of the plasma membrane and regulate its physical and chemical properties (Dufourc, 2008), play relevant roles in defense (Adolph *et al.*, 2004; Galea and Brown, 2009) and signaling phenomena (Benveniste, 2004). Often, and especially in plants and animals, steroid compounds are precursors of several secondary metabolites and important hormones (Tomazic *et al.*, 2011), oxysterols and vitamin D (Vinci *et al.*, 2008). Isoprenoids and sterols are considered interesting compounds also for their industrial applications. Isoprenoid molecules are attractive resources for the production of biofuels (Peralta-Yahya *et al.*, 2012) and have important potential applications in animal food, cosmetic and pharmaceutical industry (Kirby and Keasling, 2009). Similarly, several plant and algal sterols have promising applications (Cardozo *et al.*, 2007) and demonstrated beneficial effects on animals (Caroprese *et al.*, 2012) and on human health as cholesterol-lowering (EFSA, 2010). The ability to synthetize steroid compounds is a common feature of eukaryotes, with rare exceptions represented by insects, nematodes (Desmond and Gribaldo, 2009) and oomycete plant pathogens, such as *Phytophtora spp.* (Gaulin *et al.*, 2010). A few examples of bacteria with a minimal sterol pathway have been reported (Pearson *et al.*, 2003; Lamb *et al.*, 2007) although most prokaryotes synthesize hopanoids, structurally similar compounds that do not incorporate oxygen in position C-3. Therefore, being deeply rooted in the early history of eukaryotic life, sterol biosynthetic pathway is a prime example of metabolic conservation in evolution. It is believed that the Last Common Ancestor (LCA) of eukaryotes already possessed in the set of metabolic enzymes, some of the present sterol pathway (Desmond and Gribaldo, 2009). The origin of the sterol biosynthetic pathway is tightly connected to the advent of photosynthesis on Earth and to the rise of the level of molecular oxygen in the atmosphere (Summons *et al.*, 2006). This phenomenon, started in the oceans by cyanobacteria 200-300 million years before the Great Oxidation Event (GOE), is temporally collocated approximately 2.4 billion years ago (Lyons and Reinhard, 2011). The eukaryotic pathway of sterol biosynthesis might have evolved as a first protection against the increasingly oxidizing environment in which the excess of oxygen would have been channeled into more stable structured molecules (Desmond and Gribaldo, 2009). The presumed presence of a primitive form of this pathway in the LCA is reflected by the conservation of many aspects of the pathway in every sterol producing organism.

Generally, the sterol biosynthetic pathway is well studied in model organisms and is commonly divided into three main variants: that of animals yielding cholesterol, that of fungi, yielding ergosterol, and that of land plants, yielding a broad diversity of phytosterols. Although commonly accepted, this subdivision oversimplifies the actual organization of the pathway. In the last few years, the increasing number of sequenced genomes of non-model-organisms and technical advances in genomics and metabolomics started a re-evaluation of the organization of sterol biochemistry. This is the case of green algae (Miller *et al.*, 2012; Massé *et al.*, 2004), choanozoa (Kodner *et al.*, 2008), protozoans (Nes *et al.*, 2012), dinoflagellates (Leblond and Lasiter, 2012) and even plants, in which an alternative branch of the pathway has been discovered in *A. thaliana* (Ohyama *et al.*, 2009).

Despite the diversity of the final products of the pathway in the different taxonomical groups, several reactions are ubiquitously conserved, as well as many intermediate molecules, particularly in the first part of the pathway (Figure 5. 1). Land plants and many other photosynthetic organisms use two distinct parallel pathways for the synthesis of isoprenoids. The production of precursors for the sterol biosynthesis occurs through the mevalonate (MVA) pathway while precursors of carotenoids, pigments, tocopherols and other isoprenoid are synthetized through the plastidic methyl-erythrol phosphate pathway (MEPP). While MEPP is conserved among photosynthetic organisms, the MVA pathway was lost in green algae and in some members of the red algal group (i. e. *Cyanidioschyzon merolae*) (Lohr *et al.*, 2012). In these organisms the necessary building blocks for the sterol biosynthesis are provided by the plastidic MEPP (Lohr et al., 2012). Isopentenyl diphosphate (IPP) and dimethylallyl diphosphate (DMAPP) are the two important isomeric intermediates of both pathways. In the MVA pathway, IPP is formed from mevalonate phosphate (EC 4.1.1.33) and then further converted to DMAPP by the enzyme isopentenyl diphosphate isomerase (IDI, EC 5.3.3.2). IDI catalyzes also the reverse reaction, but the equilibrium is generally towards the formation of DMAPP (Sun *et al.*, 1998). IPP and DMAPP are then condensed to produce one molecule of geranyl diphosphate (GPP) that, with an additional molecule of IPP, forms farnesyl-diphosphate (FPP). FPP is the direct precursor of squalene, from which the sterol biosynthesis starts. While prokaryotes directly cyclize this molecule with squalene-hopanoid cyclases (SHCs) to form hopanoids, eukaryotes epoxidize squalene to 2, 3-epoxysqualene through the enzyme squalene epoxidase (SQE, referred also as squalene monooxygenase SQMO, EC 1.14.99.7) before the cyclization step. SQE catalyzes the first oxygen dependent step in the synthesis of sterols and is present in every sterol producing organism. Generally, SQEs are flavoproteins and catalyze the oxidation of squalene by using molecular oxygen and FAD as cofactor. FAD is subsequently restored by NAD(P)H, which is reduced by a cytochrome P450 reductase (Summons *et al.*, 2006). Oxidosqualene cyclase (OSC) catalyzes the third reaction of the conserved portion of the sterol biosynthesis. From 2,3-epoxysqualene, OSC forms the steroid scaffold, which will be further modified. The cyclization of 2,3-epoxysqualene, operated by OSC produces either lanosterol in animals and fungi or cycloartenol in plants and algae and this conversion is catalyzed by distinct OSCs, lanosterol synthases (LAS, EC 5.4.99.7) or cycloartenol synthases (CAS, EC 5.4.99.8), respectively. Although the two alternative products are similar, the formation of one or the other follows the taxonomical classification. The next conserved reaction is carried out by the cytochrome P450 sterol-14-demethylase enzyme (CYP51, EC 1.14.13.70) which is responsible of the oxidative removal of a methyl group from lanosterol in animals and fungi, and from obtusifoliol, a product of cycloartenol conversions in the green lineage. After this reaction, the synthesis of sterols becomes more specific and varies depending on phylogeny. In contrast to the detailed knowledge on the biochemistry of sterols of animals, plants and fungi, on organisms which do not fall in the above mentioned groups, such as algae, this is fragmentary. Although efforts to chemically characterize the sterol composition of several diatom species revealed a marked diversity of products (Volkman, 2003; Giner and Wikfors, 2011; Rampen *et al.*, 2010), the biosynthetic pathway is unknown. Some of diatom sterols derive from cycloartenol, such as common phytosterols, and others from lanosterol, making difficult the collocation of the pathway in the above mentioned tripartite subdivision (Rampen *et al.*, 2010). However, very little has been done to characterize the biosynthetic pathway in diatoms at the genetic and enzymatic level, despite its considerable scientific value, as diatoms are close relatives of the oomycete parasites, one of the isolated groups of eukaryotes that lost the ability to synthetize sterols (Gaulin *et al.*, 2010). In the present work, we report the reconstruction of the mevalonate and sterol biosynthetic pathway of the model pennate diatom *Phaeodactylum tricornutum*. By using DiatomCyc (Fabris *et al.*, 2012, Chapter 3) to

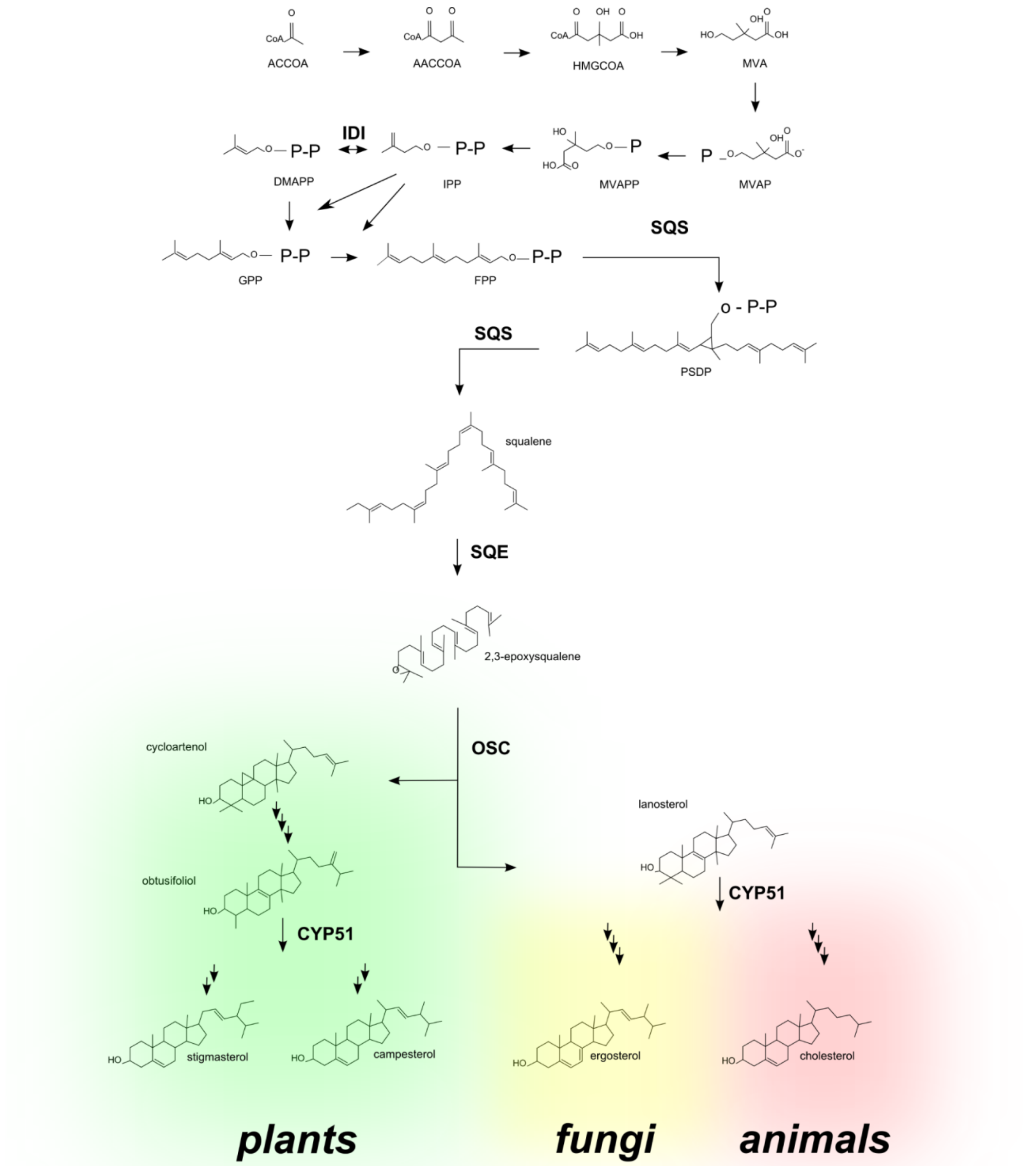


Figure 5.1 Conserved reactions of the MVA and sterol biosynthetic pathways. The enzymatic steps described in the text are highlighted in bold. Abbreviations: IDI, isopentenyl diphosphate isomerase; PMVK, phosphomevalonate kinase; SQS, squalene synthase; SQE squalene epoxidase; OSC oxidosqualene cyclase; CYP51, cytochrome P450 sterol-14-demethylase; ACCOA, acetyl-CoA; AACCOA, aceto-acetyl-CoA; HMGCOA, 3-hydroxy-3-methylglutaryl-coenzyme A; MVA, mevalonic acid, MVAP, mevalonic acid phosphate; MVAPP, mevalonic acid diphosphate; IPP isopentenyl diphosphate; DMAPP, dimethylallyl diphosphate; GPP, geranyl diphosphate; FPP, farnesyl diphosphate; PSDP, presqualene diphosphate.

identify the set of genes putatively involved in diatom sterol synthesis, we mapped and experimentally validated the main biochemical steps of the pathway. The proposed biochemical route is a hybrid pathway as it shares elements with both fungi and plants. We report that diatoms, as well as many other groups of marine organisms surprisingly lack the ubiquitously conserved enzyme SQE and we show the presence of uncommon enzymes, such as a particular OSC encoded by an unusually extended gene and an IDI-SQS multifunctional fusion enzyme. Our results imply that sterol biosynthesis in the marine environment might have evolved differently and highlight the need to reconsider the sterol pathway as a rigorously conserved pathway. In addition, we show that in diatoms the regulation of the sterol metabolism is tightly connected to the accumulation of triacylglycerol (TAG), opening an innovative view on the metabolic engineering of diatoms for sustainable production of biofuels.

Results

Sterol composition of *P. tricornutum*

We profiled the pool of sterol produced by three-day old *P. tricornutum* cultures by GC/MS analysis of trimethylsilyl (TMS) derivatized metabolites. *P. tricornutum* only accumulates C-28 sterols, mainly brassicasterol (24-methylcholest-5,22-dien-3β-ol) and campesterol (24-methylcholest-5-en-3 β -ol) (Figure 5.2). This is in agreement with earlier reports (Rampen *et al.*, 2010).

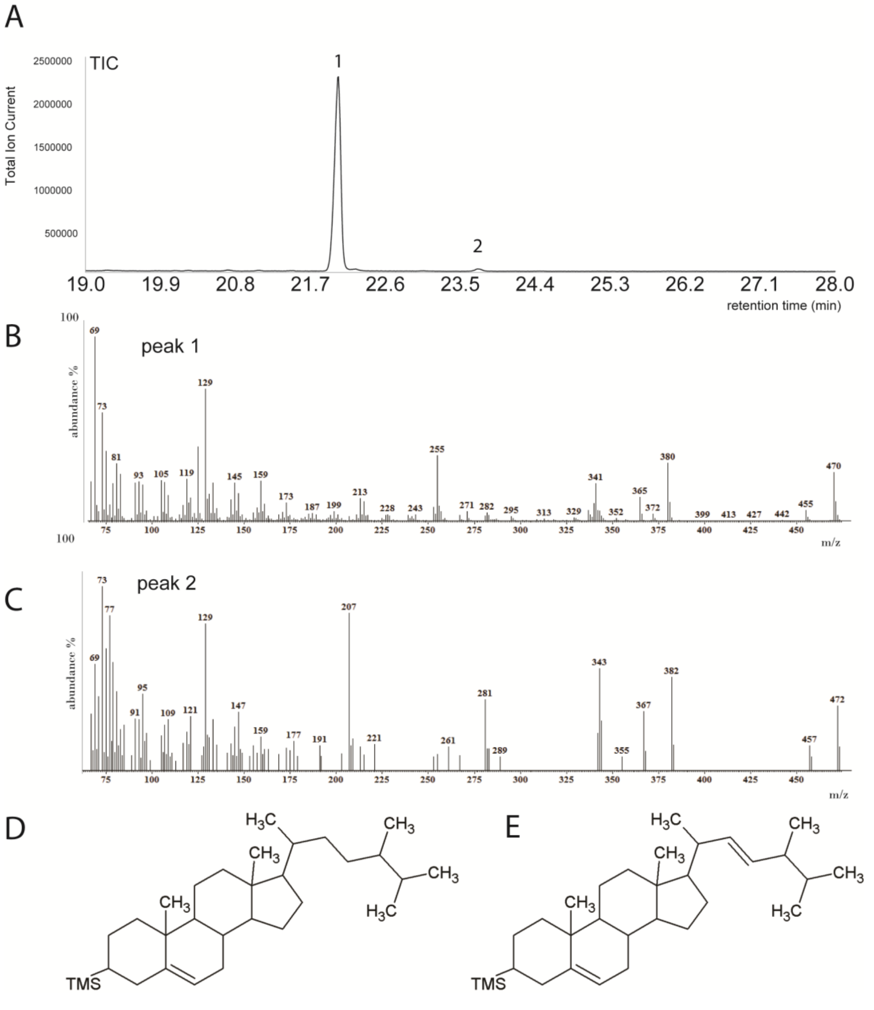


Figure 5.2. *P. tricornutum* sterol composition**.** A) GC/MS chromatograms of TMS-derivatized extracts from three-days old *P. tricornutum* cultures; B) EI-MS spectra of TMS-brassicasterol and (C) TMS-campesterol; D) chemical structures of campesterol and (E) brassicasterol.

In silico pathways analysis

We mined DiatomCyc (Fabris *et al.*, 2012) for the set of genes putatively encoding the biosynthetic enzymes involved in the MVA and sterol synthesis pathways in *P .tricornutum* . The identified genes, listed in Table 5.1, generally show homology with orthologs from plant, yeast, algae and animals, with substantially high orthology scores (Table S1). Notably, most genes lacked either a correct start or stop codon or both, thus their gene models were manually corrected with the aid of available RNA-Seq data and uploaded in DiatomCyc v2.0 (Matthijs et al., unpublished).

Mevalonate pathway - We identified the whole set of genes encoding the enzymes involved in the MVA pathway, which in diatoms, like in land plants but differently than in green algae, provides the immediate precursors of the biosynthesis of sterols and derived compounds (Massé *et al.*, 2004), whereas the same pathway is absent from green algae. Accordingly, many of the MVA pathway genes of *P. tricornutum* show pronounced similarity with plant orthologs (Table S1). Due to incorrect genome annotation, the preliminary set of genes involved MVA pathways lacked a phosphomevalonate kinase (PMVK, EC 2.7.4.2), while we could find corresponding orthologs in all the other sequenced diatom genomes. As reported in the next section, we manually identified and reconstructed *P. tricornutum*'s *PMVK* gene among a set of *de-novo* assembled gene models.*P. tricornutum* possesses two isoforms of the IDI enzyme type I (EC 5.3.3.2), encoded by *PHATRADRFT_12533* and by *PHATRDRAFT_12972*, localized on chromosome 8 and 10, respectively. While the first exists as an independent ORF, the latter is part of an elongated gene model, not predicted by the original genome annotation (Bowler *et al.*, 2008). Remarkably, the manual refinement of the gene model, revealed a fusion with the only locus encoding the SQS enzyme (PHATRADRFT_13160, EC 2.5.1.21).

Table 5.1

Gene ID	EC #	Predicted function	Closest ortholog in Orthology-prediction (DiatomCyc)* (score) GeneID [organism]	
			1st InParanoid hit	2nd InParanoid hit
PHATRDRAFT_23913	2.3.1.9	Acetyl-coa c-acetyltransferase	(423) AT5G47720 [A. thaliana]	(399) YPL028W [S. cerevisiae]
PHATRDRAFT_16649	2.3.3.10	Hydroxymethylglutaryl-CoA (HMG-CoA) synthase	(385) AT4G11820 [A. thaliana]	(355) 3157 [H. sapiens]
PHATRDRAFT_16870	1.1.1.34	Hydroxymethylglutaryl-CoA (HMG-CoA) reductase	(410) AT1G76490 [A. thaliana]	(270) MJ0705 [M. jannaschii]
PHATRDRAFT_53929	2.7.1.36	Mevalonate kinase	(164) YMR208W [S. cerevisiae]	(140) 4598 [H. sapiens]
PHATRDRAFT_44478	2.7.4.8	Phosphomevalonate kinase	-	
PHATRDRAFT_BD1325	4.1.1.33	Diphosphomevalonate decarboxylase	(318) AT2G38700 [A. thaliana]	(315) 4597 [H. sapiens]
PHATRDRAFT_12972	5.3.3.2	Isopentenyl-diphosphate δ-isomerase	(265) CHLREDRAFT_24471 [C. reinhardtii]	(237) OSTLU_13493 [O. lucimarinus]
PHATRDRAFT_12533	5.3.3.2	Isopentenyl-diphosphate δ-isomerase	(265) CHLREDRAFT_24471 [C. reinhardtii]	(156) 3422 [H. sapiens]
PHATRDRAFT_13160**	2.5.1.21	Squalene synthase	(279) OSTLU_31144 [O. lucimarinus]	(274) 2222 [H. sapiens]
PHATRADRFT_47271	2.5.1.1	Geranyl-diphosphate synthase	(309) 9453[H. sapiens]	(202) YPL069C [S. cerevisiae]
PHATRDRAFT_49325	2.5.1.1		(343) Ot03g02400 [O. tauri]	(340) YJL167W [S. cerevisiae]
PHATRDRAFT_49325	2.5.1.10		(343) Ot03g02400 [O. tauri]	(340) YJL167W [S. cerevisiae]
not found	1.14.99.7	Squalene epoxidase	-	-
PHATR_645**	5.4.99.8	Oxidosqualene cyclase	(602) Ot03g00850 [O. tauri]	(596) AT2G07050 [A. thaliana]
PHATRDRAFT_10824	2.1.1.41	24-methylenesterol C-methyltransferase	(301) AT5G13710 [A. thaliana]	(296) OSTLU_30710 [O. lucimarinus]
PHATRDRAFT_49447	5.5.1.9	Cycloeucalenol cycloisomerase	(250) Ot04g04910 [O. tauri]	(231) OSTLU_6401 [O. lucimarinus]
PHATRDRAFT_31339	1.14.13.70	14-α-demethylase	(435) CMS319C [C. merolae]	(427) OSTLU_43938 [O. lucimarinus]
PHATRDRAFT_10852	1.14.13.72	Methylsterol monooxygenase	(82) AT1G07420 [A. thaliana]	-
PHATRDRAFT_48864	1.1.1.170	3-β-hydroxysteroid-4-α-carboxylate-3-dehydrogenase	(318) Ot04g04390 [O. tauri]	(305) OSTLU_87094 [O. lucimarinus]
PHATRDRAFT_5780	1.1.1.270	putative SDR oxidoreductase	(103) AT5G65205 [A. thaliana]	(97) 3248 [H. sapiens]
PHATR_36801	5.3.3.5	3-Beta-hydroxysteroid-delta(8), delta(7)-isomerase	-	-
PHATRDRAFT_14208	1.14.21.6	δ-7-sterol δ-5-dehydrogenase	(193) AT3G02580 [A. thaliana]	(178) CHLREDRAFT_59933 [C. reinhardtii]
PHATRDRAFT_30461	1.3.1.21	Δ7-sterol reductase	(426) AT1G50430 [A. thaliana]	-
PHATRDRAFT_51757	1.3.1.-.	sterol C-22 desaturase	(230) CHLREDRAFT_196874 [C. reinhardtii]	(219) CMJ284C [C. merolae]

Table 5.1 (*previous page*). List of genes putatively involved in MVA pathway and sterol biosynthesis of *P. tricornutum* . Genes retrieved from DiatomCyc are listed with their predicted function and first and second InParanoid hits (Fabris *et al.*, 2012), (Table S1). *Stramenopiles are excluded. ** Reconstructed gene model are discussed in the text.

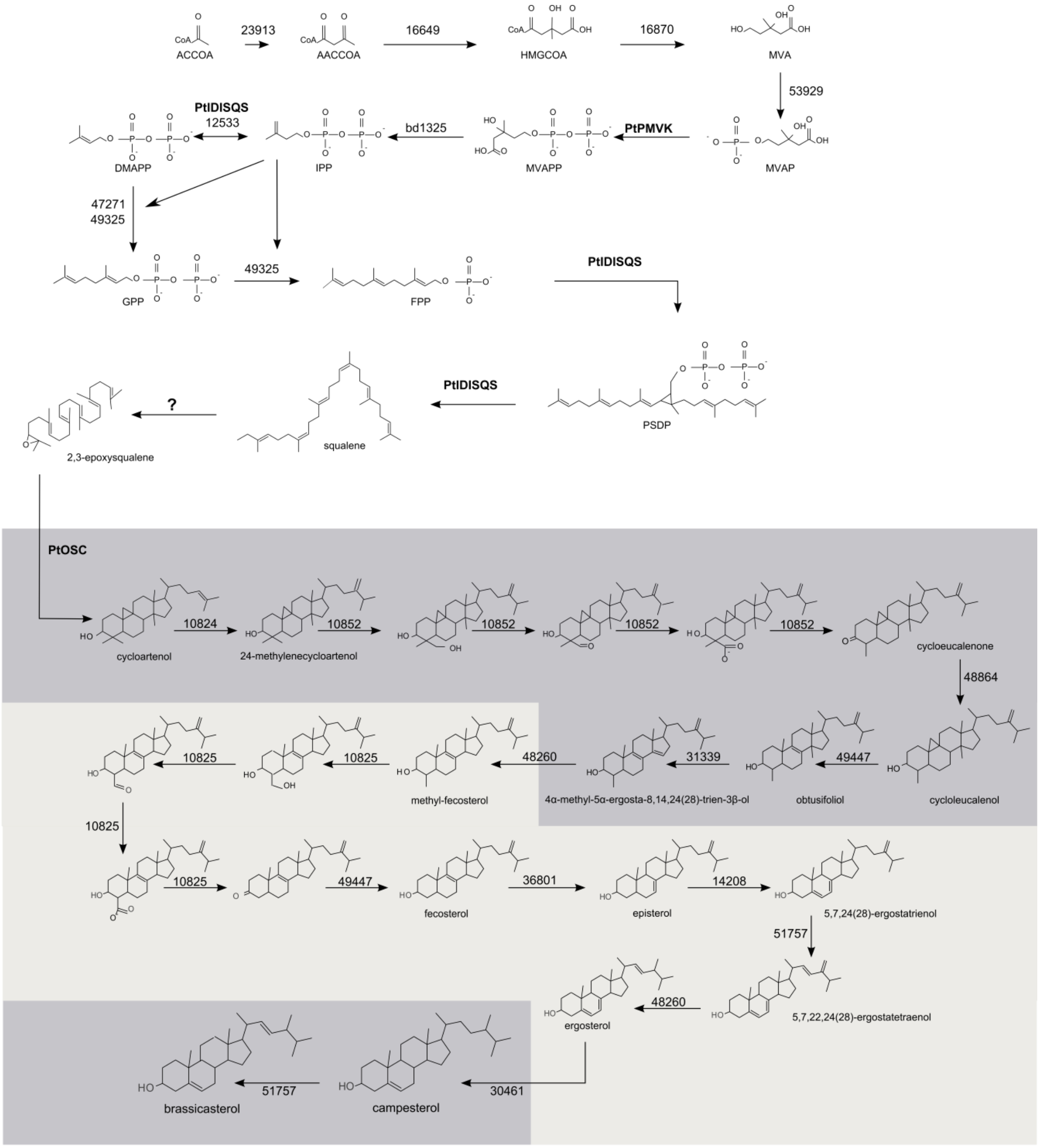
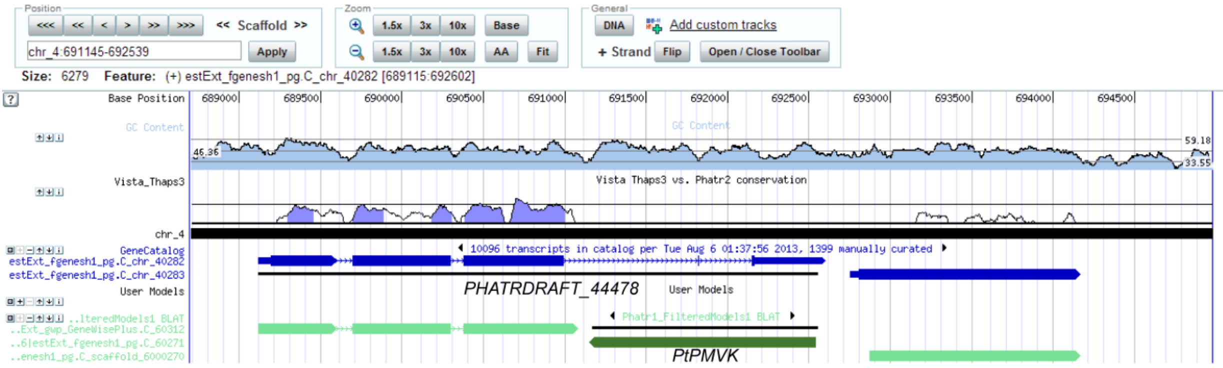


Figure 5.3. Hypothetical reconstruction of mevalonate and sterol biosynthetic pathways of *P. tricornutum*. Dark and light gray areas indicate similarities with the sterol biosynthesis of plants and fungi, respectively. The pathway is described in the text. Numbers on arrows indicate enzyme protein ID. PHATRDRAFT- suffix and co-factors are omitted. Abbreviations are listed in the description of Figure 5.1.

Sterol biosynthetic pathway - The sterol biosynthesis of *P. tricornutum* involves 11 genes (Table 1) that encode the enzymes necessary for the conversion of squalene into the final steroid products brassicasterol and campesterol. The same group of enzymes, encoded by the above mentioned genes, allows different hypothetical pathway reconstructions. In particular, both routes used by fungi and by plants are theoretically possible. Among the identified genes we found a single OSC (EC 5.4.99.7/8). Surprisingly, this gene in *P. tricornutum* is encoded by an unusually extended ORF, with a structure that recurs in all other diatom genomes sequenced so far. Another striking feature that emerged from the *in silico* analysis of the pathway is the lack of conventional SQE (EC 1.14.99.7).

Identification of the missing phospomevalonate kinase

Initially, a PMVK could not be identified in *P. tricornutum*, neither by the orthology prediction method (Fabris et al., 2012), nor by BLAST searches on both NCBI and JGI databases http://genome.jgi-psf.org/Phatr2/Phatr2.home.html). By performing BLAST searches using *A. thaliana*'s PMVK (*AT1G31910*) as query, on an in-house BLAST database constructed on a *de-novo* assembly of *P. tricornutum* (Matthijs et. al., unpublished), we identified a 1395 bp transcript encoding a phosphomevalonate kinase (*PtPMVK*). The novel *PtPMVK* transcript could be mapped, in reverse orientation, onto a specific region on chromosome 4 (chr_4:691145-692539), as part of the locus *PHATRADRFT_44478*. Specifically, the reconstructed *PtPMVK* ORF maps to a gene model that was predicted by a previous JGI assembly and supported by Expressed Sequence Tags (ESTs)in which the current PHATRDRAFT_*44478* locus, it erroneously corresponds to a large intron (Figure 5.4).



chr_4:691145-692539

Size: 6279 Feature: (+) estExt_fgenesh1_pg.C_chr_40282 [689115:692602]

10096 transcripts in catalog per Tue Aug 6 01:37:56 2013, 1399 manually curated

Figure 5.4. Screenshot of the JGI genome browser of *P. tricornutum* (http://genome.jgi-psf.org/Phatr2/Phatr2.home.html) relatively to the erroneously predicted locus *PHATRDRAFT_44478* (in dark blue) on chromosome 4. The identified *PtPMVK* ORF (dark green) corresponds to a previous version of the gene model, based on expressed sequences tags (ESTs).

Identification of a fusion *IDI-SQS* gene in *P. tricornutum*

In *P. tricornutum*'s genome we identified the presence of an unusual ORF that putatively encodes a metabolic fusion enzyme with predicted IDI and SQS activities.

The gene model was manually adjusted, resulting in a 2.335 bp long ORF. Originally predicted as independent adjacent ORFs, *PHATRDRAFT_12972* (*PtIDI*) and *PHATDRAFT_13160* (*PtSQS*) are actually spaced by a short coding sequence that does not include stop codons. The encoded protein consists of 763 amino acids subdivided in NUDIX hydroxylase/isopentenyl diphosphate delta-isomerase domain of type I at the N-terminus, and a squalene/phytoene synthase domain at the C-terminus (Hunter *et al.*, 2009), both sharing similarity with the respective IDI and SQS enzymes of mainly brown and green algae (Table S1). Interestingly, the same IDI-SQS fusion is present in all sequenced diatoms, as well as in *Aureococcus anophagefferens* and *Ectocarpus siliculosus* (Figure 5.5), suggesting that this postulated multifunctional enzyme might be conserved

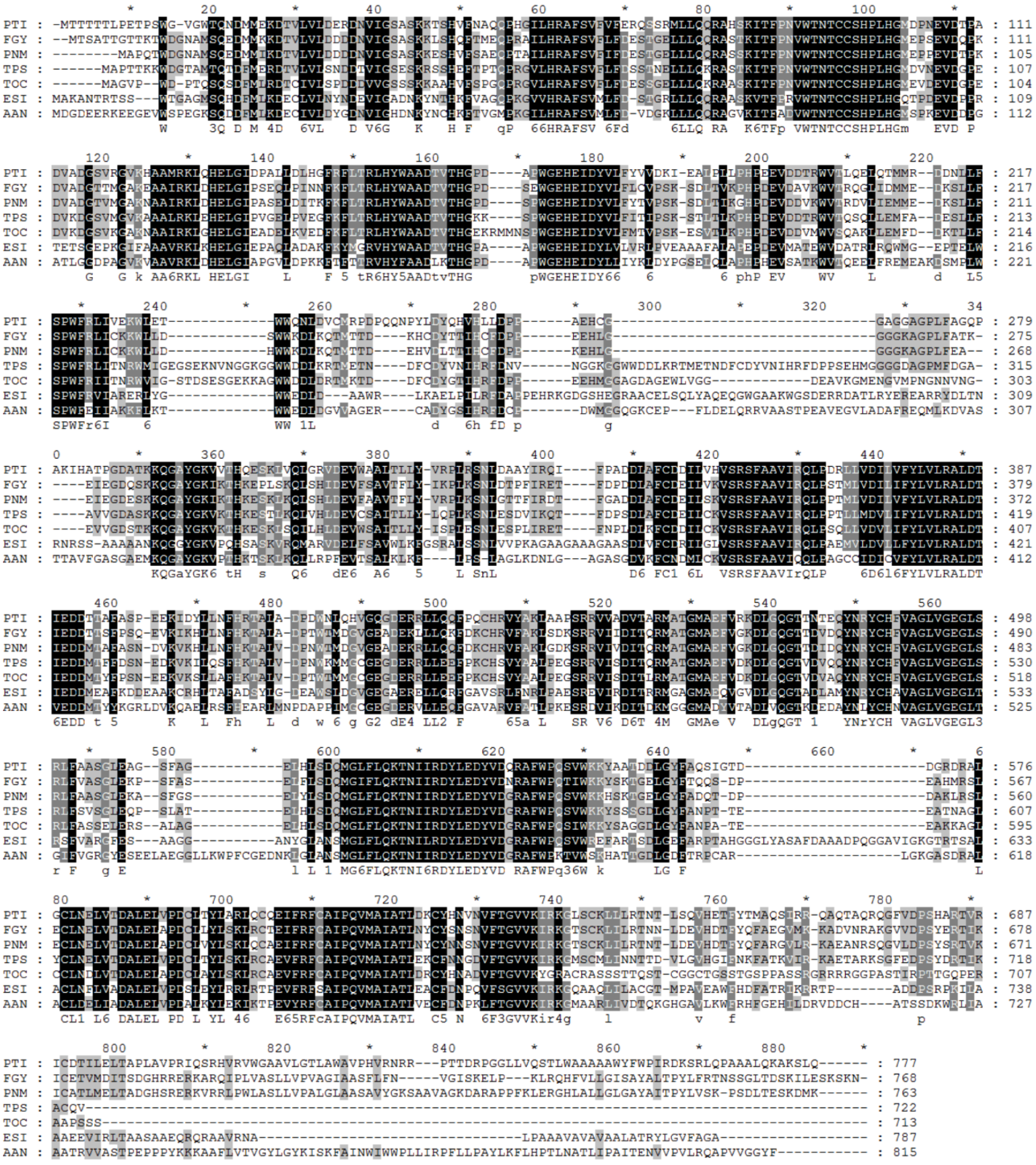


Figure 5.5. Alignment of the amino acid sequences of the reconstructed IDISQS fusion enzyme of diatoms with the corresponding orthologs of *Aureococcus anophagefferens* and *Ectocarpus siliculosus*. Abbreviations: PTI, *Phaeodactylum tricornutum*; FCY, *Fragilariopsis cylindrus*; TPS, *Thalassiosira pseudonana*; TOC, *Thalassiosira oceanica*; PNM, *Pseudo-nitzschia multistriata*; AAN, *Aureococcus anophagefferens*. ESI, *Ectocarpus siliculosus*;

among Stramenopiles. A notable exception in this regard is the heterokont oomycete *Phytophtora*, the species of which possess the IDI gene but lack the SQS gene as well as the whole sterol biosynthetic pathway (Gaulin et al., 2010). The *in silico* predicted gene structure was validated both at the transcript level by in-house RNA-Seq data (Matthijs *et al.,* unpublished), by PCR amplification on a cDNA library template and at the protein level, by cloning and expressing the *PtIDISQS* ORF in *E. coli*. (Figure 5.6) According to the *in silico* predictions, the protein should have a mass of 87.7 kDa, which was in agreement with the size of the product found when the *PtIDISQS* gene was expressed in *E. coli* (Figure 5.6).

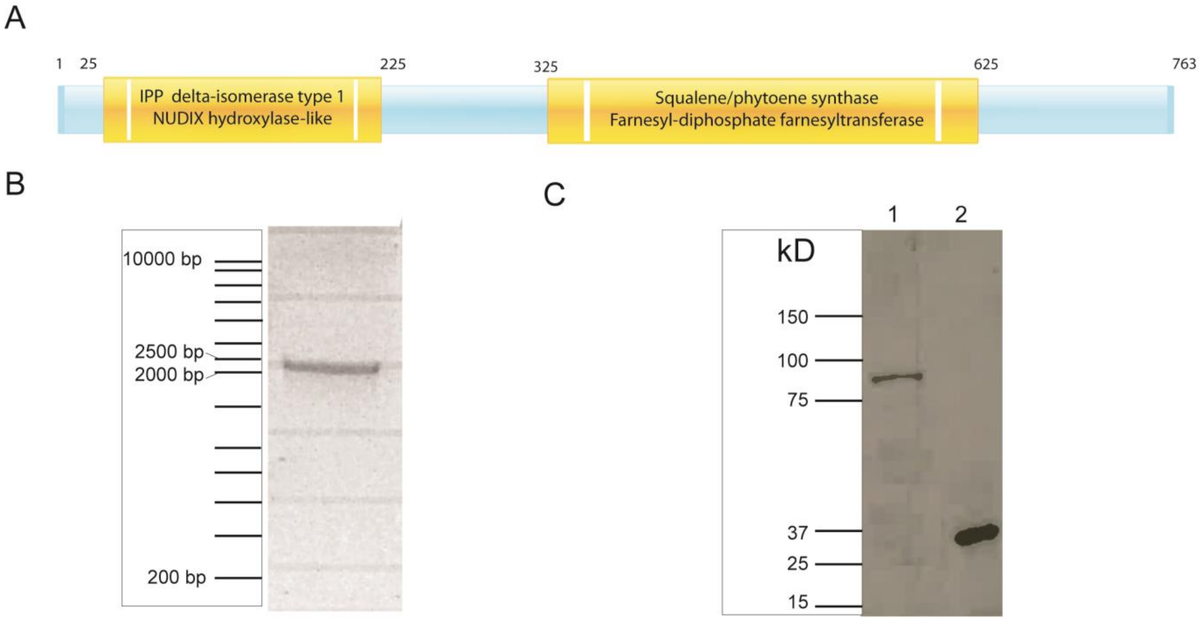

A
1
25
225
325
625
763
IPP delta-isomerase type 1
NUDIX hydroxylase-like
Squalene/phytoene synthase
Farnesyl-diphosphate farnesyltransferase
B
10000 bp
2500 bp
2000 bp
200 bp
C
1
2
kD
150
100
75
37
25
15

Figure 5.6. The endogenous *PtIDISQS* fusion gene. A) Schematic organization of the IDI and SQS domains in PtIDISQS according to InterProScan predictions (Hunter et al. 2009). Numbers refer to the approximate amino acid residue position; B) RT-PCR amplification of *PtIDISQS* from a cDNA library of *P. tricornutum*; C) Immunoblot analysis with anti-His antibodies of *E. coli* BL21 (DE3) samples expressing *6xHis-PtIDISQS* (lane 1) or *6xHis-PtEDA* as control (lane 2). Molecular weight (MW) of 6xHis is 1 kD. Predicted MW of PtEDA is 27.04 kD.

Functional characterization of *PtIDISQS*

IDI activity - The isomerization of IPP to DMAPP is a crucial step in the MVA pathway, but also in the methylerythritol phosphate pathway (MEPP), which is involved in the biosynthesis of isoprenoid precursors for the biosynthesis of carotenoids in photosynthetic organisms (Kajiwara *et al.*, 1997). Based on that, we evaluated the functionality of the IDI domain of *PtIDISQS* by co-expressing it in *E.coli* BL21 (DE3) with the plasmid pAC-LYC (Cunningham *et al.*, 1994). This plasmid carries a set of genes *of Erwinia herbicola* that allow the synthesis of lycopene in *E. coli* and belongs to a versatile collection of plasmids for the study of carotenogenic genes (Cunningham and Gantt, 2007). *E.coli* BL21 (DE3) cells transformed with pAC-LYC yields pink colonies, due to the accumulation of lycopene. The synthesis of this pigment is made possible by the availability of the precursors DMAPP and IPP, provided by the endogenous MVA pathway of the bacteria, further processed to lycopene by the genes carried by pAC-LYC (Cunningham *et al.*, 1994). Although IDI catalyzes the isomerization between IPP and DMAPP, the equilibrium favors the formation of the latter (Cunningham *et al.,* 1994). It has been observed that enhanced activity of the IDI enzyme generally results in an increased accumulation of carotenoid products in the pathway (Kajiwara *et al.* 1997, Sun *et al.,* 1998, Cunningham and Gantt, 2007), causing a visible change in the color of *E. coli* cultures (Cunningham and Gantt, 2007). In our experiment, we evaluated the differential accumulation of lycopene both qualitatively and quantitatively in *E.coli* BL21 (DE3) co-transformed with pAC-LYC and *PtIDISQS* cloned into a pDEST17 plasmid. As positive control *E. coli* cells co-transformed with the *IDI* gene of *E. herbicola* cloned into pAC-LYC plasmid (pAC LYCipi) (Cunningham et *al.,* 2007) and pDEST17_*AtJAZ1*, in order to meet the same antibiotic requirements of the other strains. *E. coli* transformed with pAC LYC and a pDEST17 plasmid carrying the carotenoid-unrelated gene *PtEDA* was used as negative control. As shown in Figure 5.7, the presence of multiple copies of genes encoding IDI enzyme and its overexpression caused an increased accumulation of lycopene and a consequent darker pink color in all the samples expressing *PtIDISQS* and *EhIPI*. Differently, 2 lines of the strain expressing *PtEDA*, which is not related to the isoprenoid biosynthesis, only accumulated basal levels of lycopene, resulting in light pink color. However, one *PtEDA* line shows a phenotype similar to *IDI* expressing strains. Although these results were preliminary and require the screening of a larger number of *E. coli* colonies, they provided good indications towards the functionality of the IDI domain of PtIDISQS.

SQS activity - To determine whether the *PtIDISQS* fusion gene produces a functional enzyme with squalene synthase activity, its ability to convert the isoprenoid FPP to squalene was tested in *E coli*. Wild-type *E. coli* does not synthetize squalene or sterols, but it naturally produces FPP thus represents a convenient test system for the functional characterization of this domain as well. Non-saponifiable lipid fractions extracted from the *E. coli* strains BL21 (DE3) transformed with a plasmid carrying either *PtIDISQS* or *PtEDA* and treated for 48 hours with 0.1 mM isopropyl β-D-1-thiogalactopyranoside (IPTG) were analyzed with GC/MS. The resulting chromatograms showed a recurrent small peak corresponding to squalene in strains expressing *PtIDISQS* (Figure 5.7), but not in control strains (Figure 5.7). This result confirmed the predicted SQS activity of *PtIDISQS* and further supported the occurrence of squalene as a sterol pathway intermediate in *P. tricornutum*.

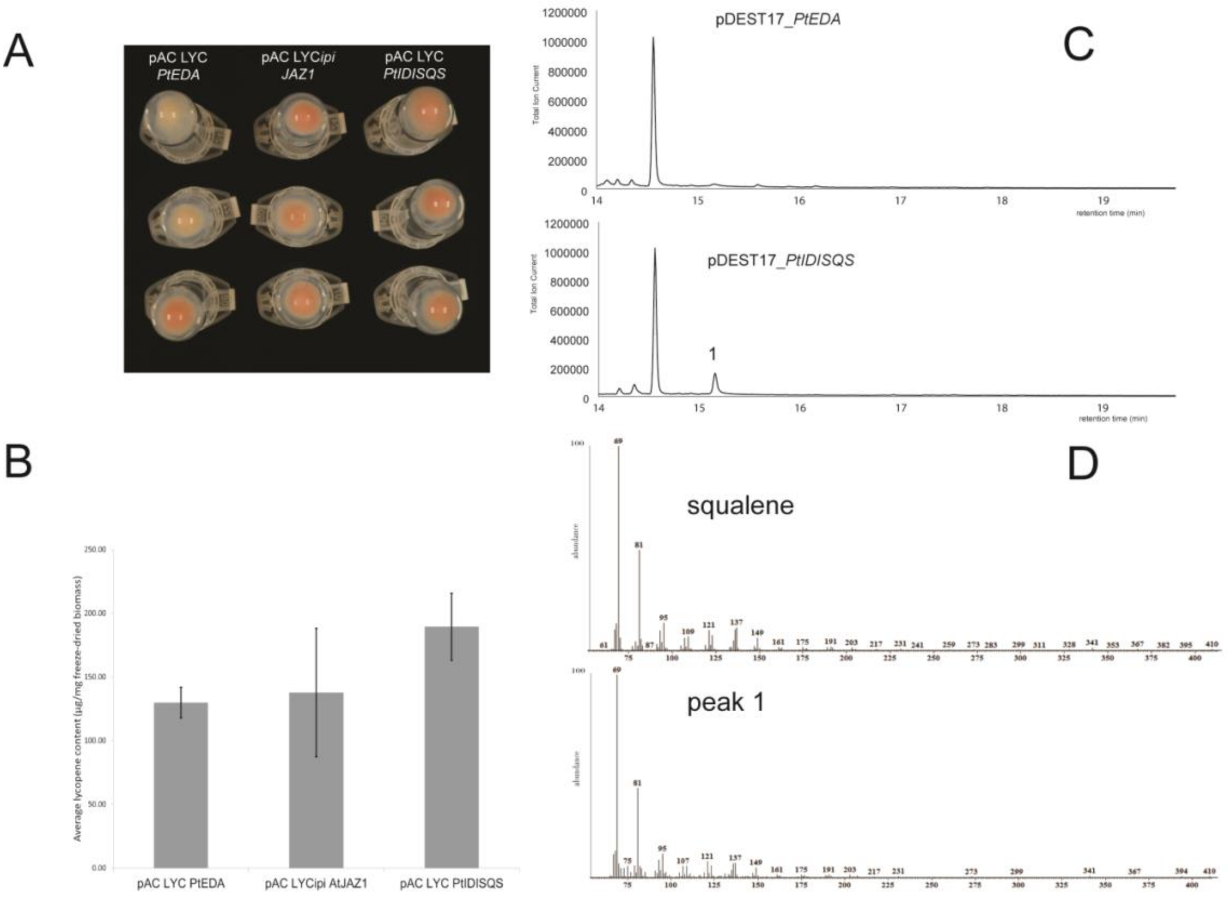


Figure 5.7. Enzymatic activity of PtIDISQS. A) IDI activity assay in *E. coli* cells transformed with the pAC LYC plasmid. Bacterial pellets showing a darker pink color indicate an increased IDI activity caused by overexpression of the *IDI* genes (pAC LYC/*PtIDISQS* and pAC LYCipi/*AtJAZ1*). The strain with pAC LYC/*PtEDA* indicates the basal levels of lycopene accumulation caused by the endogenous *idi* gene of *E. coli*. B) Average HPLC quantification of the lycopene content of *E. coli* colonies transformed with pAC LYC/*PtEDA,* pAC LYCipi/*AtJAZ1*C and pAC LYC/*PtIDISQS.* Error bars indicate standard deviations of the mean of three biological replicates. C) GC/MS chromatogram of *E. coli* BL21 (DE3) cells expressing *PtIDISQS* and showing accumulation of squalene (peak1), compared to *E.coli* cultures expressing *PtEDA* used as negative control. D) Comparison of mass spectrum of authentic squalene with that of peak 1.

Many Stramenopiles lack a conventional SQE gene

Since we could not identify *SQE* orthologs in the P. tricornutum genome in the dataset generated for the reconstruction of DiatomCyc (Fabris *et al.,* 2012), we performed an extensive BLASTP search in eukaryotes, excluding the group of fungi, animals, and land plants, and using the amino acid sequences of SQE from *Arabidopsis*, *Homo sapiens*, and *Saccharomyces cerevisae* as queries. Surprisingly, SQE seemed to be widely absent in the analysed groups, in particular in most of the analysed Stramenopiles, with the exception of E. siliculosus and the oomycete pathogens Aphanomyces eusteiches and Saprolegnia diclina. No hits were obtained in whole groups of Alveolata and Choanoflagellida either but SQE seemed to be conserved in the red and green algal lineages, with the exception of *Chlamydomonas reinhardtii* (Fig 5.8)

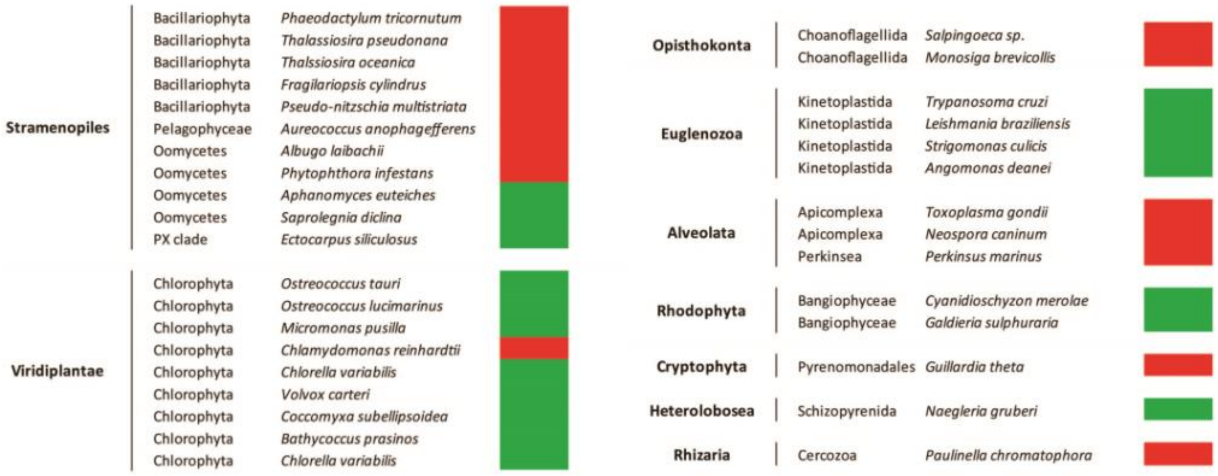

Stramenopiles
Bacillariophyta Phaeodactylum tricornutum
Bacillariophyta Thalassiosira pseudonana
Bacillariophyta Thalssiosira oceanica
Bacillariophyta Fragilariopsis cylindrus
Bacillariophyta Pseudo-nitzschia multistriata
Pelagophyceae Aureococcus anophagefferens
Oomycetes Albugo laibachii
Oomycetes Phytophthora infestans
Oomycetes Aphanomyces euteiches
Oomycetes Saprolegnia diclina
PX clade Ectocarpus siliculosus
Viridiplantae
Chlorophyta Ostreococcus tauri
Chlorophyta Ostreococcus lucimarinus
Chlorophyta Micromonas pusilla
Chlorophyta Chlamydomonas reinhardtii
Chlorophyta Chlorella variabilis
Chlorophyta Volvox carteri
Chlorophyta Coccomyxa subellipsoidea
Chlorophyta Bathycoccus prasinos
Chlorophyta Chlorella variabilis
Opisthokonta
Choanoflagellida Salpingoeca sp.
Choanoflagellida Monosiga brevicollis
Euglenozoa
Kinetoplastida Trypanosoma cruzi
Kinetoplastida Leishmania braziliensis
Kinetoplastida Strigomonas culicis
Kinetoplastida Angomonas deanei
Alveolata
Apicomplexa Toxoplasma gondii
Apicomplexa Neospora caninum
Perkinsea Perkinsus marinus
Rhodophyta
Bangiophyceae Cyanidioschyzon merolae
Bangiophyceae Galdieria sulphuraria
Cryptophyta
Pyrenomonadales Guillardia theta
Heterolobosea
Schizopyrenida Naegleria gruberi
Rhizaria
Cercozoa Paulinella chromatophora

Figure 5.8. Conservation (green) and loss (red) of the SQE gene in representative organisms belonging to the groups of Stramenopiles, green algae, Alveolata, Choanoflagellida, Kinetoplastids, Cryptophyta, Heterolobosea and Rizharia, according to BLASTP searches.

Oxidosqualene cyclase (OSC) gene model adjustment

We identified the locus *PHATR_645* as part of an ORF putatively encoding an OSC. The gene model provided by JGI assembly (Bowler *et al.*, 2008) was incomplete, as no correct start nor stop codon were present and the gene was wrongly annotated as "acetyl-coenzyme A synthase". The incorrect gene model was manually adjusted, which resulted to correspond to an ORF of 2.931 nt, spanning the coding regions of both the previous and the subsequent gene on chromosome 11, *PHATRDRAFT_46724* and *PHATRDRAFT_46726*, respectively (from here on referred to as *PtOSC*). The correctness of the manually reconstructed gene model and the existence of such gene at the transcript level were further confirmed by *in-house* generated RNA-Seq data from *P. tricornutum* samples (Matthijs *et al.,* unpublished) and by PCR amplification on a cDNA library template, producing a single band of the predicted size (Figure 5.9). Notably, as for the PtIDISQS gene, a similarly extended OSC locus -as well as the erroneous incomplete gene model prediction- could be identified in all annotated diatom genomes (Figure 5.10). In the predicted PtOSC protein, the sequence corresponding to the first 46 N-terminal amino acids is possibly part of an ER-targeting peptide (Emanuelsson *et al.*, 2007), the following 103 amino acids potentially correspond to three transmembrane domain repeats (Emanuallsson *et al.,* 2007 and the C-terminus shows similarity to common OSC enzymes. Although all diatom OSC sequences share a specific N-terminal extension, its sequence is not conserved among these species (Figure 5.10). To confirm the existence of the predicted gene product at the protein level, it was cloned and expressed in *E. coli*. The existence of the full-length (FL) protein could not be verified as the expression of FL *PtOSC* resulted to be extremely difficult, perhaps due to differences in codon usage, the considerable size of the polypeptide and/or to the presence of a large hydrophobic domain at the N- terminus. Therefore, a truncated version of *PtOSC* in

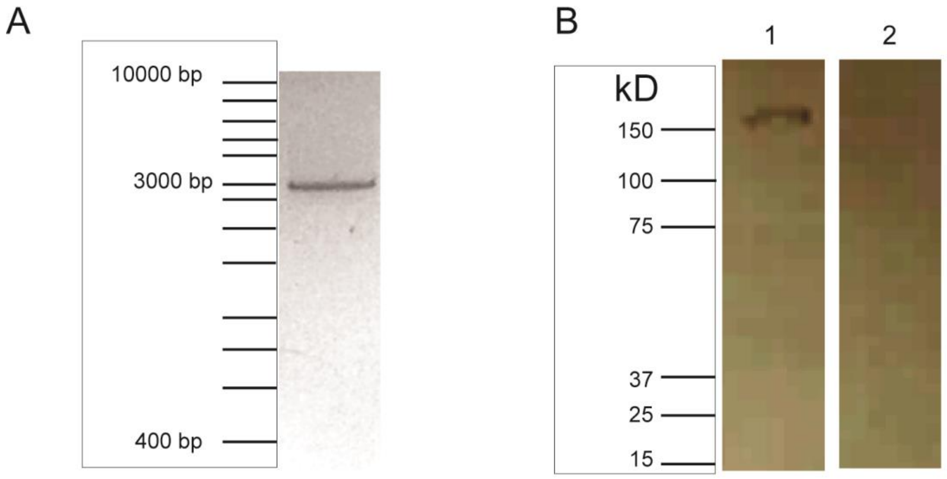
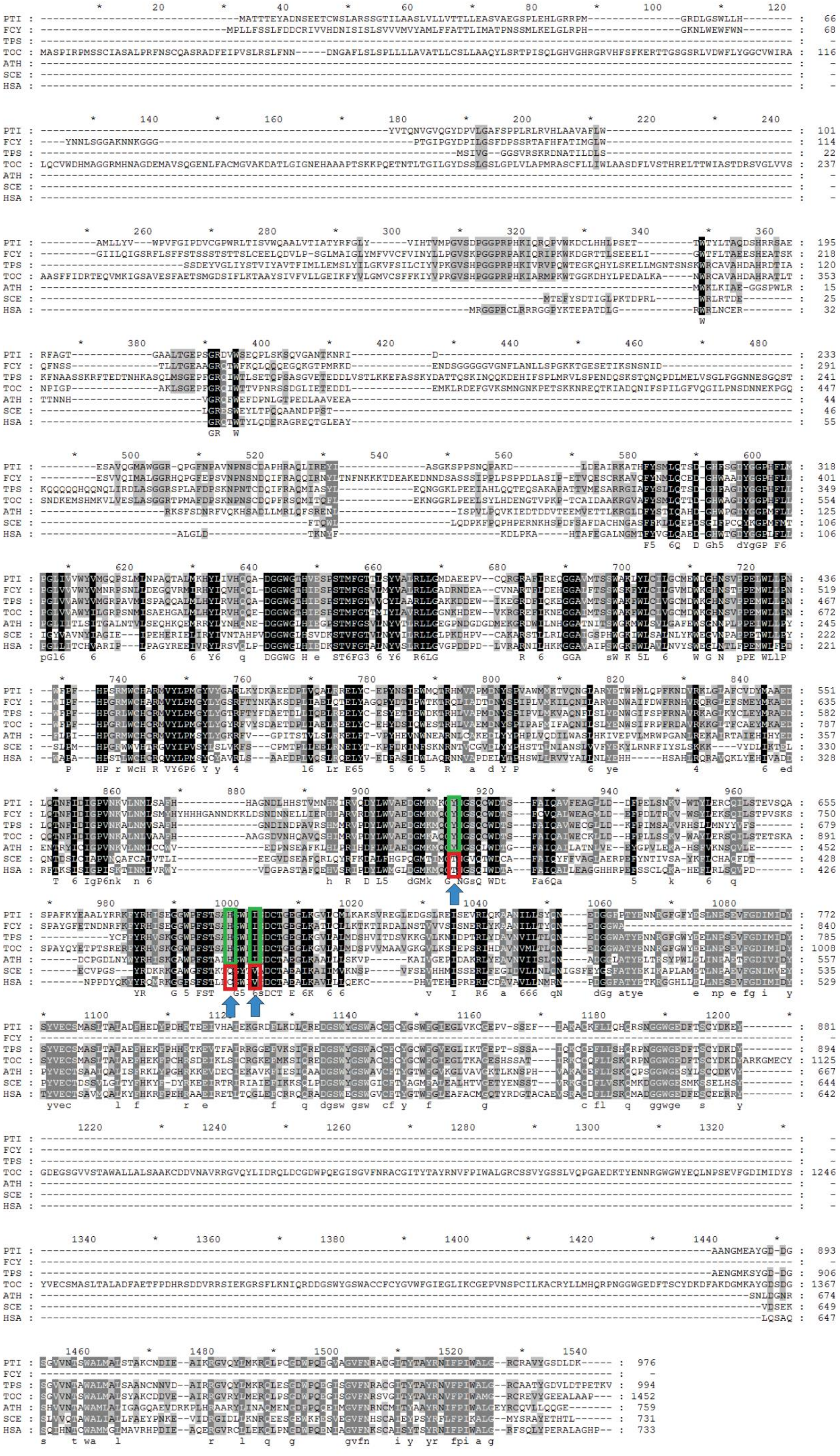


Figure 5.9. The *PtOSC* gene. A) RT-PCR amplification of *PtOSC* from a cDNA library of *P. tricornutum*, B) Immunoblot analysis with anti-His antibodies of *E. coli* BL21 (DE3) samples transformed with pDEST17-*6xHis-MBP-PtOSC,* induced for 24 hours with isopropyl β-D-1-thiogalactopyranoside (IPTG) (lane 1) and not induced (lane 2). The molecular weight MBP (Maltose Binding Protein) is 42.5 kD; of 6xHis is 1 kD.

Figure 5.10 (*next page*). Alignment of the amino acid sequences of the reconstructed OSC gene models of diatoms with those of plants (*A. thaliana*), animals (*H. sapiens*) and fungi (*S. cerevisiae*). To be noted is the extended N-terminal portion characterized by a low level of sequence conservation, present only in diatoms, while the remaining part of the proteins shows similarity with known OSCs. Arrowheads indicate positions in which conserved residues are specific for the formation of cycloartenol (green rectangle) or lanosterol (red rectangle), according to previous studies (Summons *et al.* 2006). Abbreviations: PTI, *Phaeodactylum tricornutum*; FCY, *Fragilariopsis cylindrus*; TPS, *Thalassiosira pseudonana*; TOC, *Thalassiosira oceanica*; ATH, *Arabidopsis thaliana*; SCE, *Saccharomyces cerevisiae*; HSA, *Homo sapiens*. Abbreviations: PTI, *Phaeodactylum tricornutum*; FCY, *Fragilariopsis cylindrus*; TPS, *Thalassiosira pseudonana*; TOC, *Thalassiosira oceanica*; ATH, *Arabidopsis thaliana*; SCE, *Saccharomyces cerevisiae*; HSA, *Homo sapiens*.


PTI
FCY
TPS
TOC
ATH
SCE
HSA

which the N-terminal 46 amino acids (corresponding to the putative signal peptide) had been removed was fused to a maltose binding protein (MBP) in an attempt to facilitate the synthesis. The western blot analysis revealed the existence of a large protein of approximately 104 KDa, matching the *in silico* predictions (Figure 5.9) and confirmed the protein model.

Treatments with chemical inhibitors

Ro 48-807 and terbinafine - The apparent absence of a conventional SQE and the unusual OSC enzyme raised the question whether diatoms would use 2,3-epoxysqualene as precursor for the cyclization step. To determine this, we treated diatom cultures with terbinafine (TB) and Ro 48-8071, specific inhibitors of conventional SQE and OSC enzymes, respectively. Whereas *P. tricornutum* cultures could be treated with 40 μM TB without major impairments, they were heavily affected by low concentrations of Ro 48-8071. In pilot experiments it was observed that Ro 48-8071 30 μM kills 3 days-old diatom cultures in less than 48 hours (data not shown). Therefore, timecourse experiments were set up, involving treatments with TB and Ro 48-8071, in order to detect intermediate compounds at the appropriate timing. Diatom cells treated with 10 μM Ro 48-8071 showed most significant effects between 8 and 12 hours after the treatment, in which diatoms accumulated both squalene and 2,3-epoxysqualene as a consequence of the blockage of the OSC enzyme (Figure 5.11). Also, as a consequence of Ro 48-8071 treatment, *P. tricornutum* cells accumulated intracellular lipid droplets, consisting mainly of TAG as they could be stained with Nile Red (Greenspan *et al.*, 1985)(Figure 5.11).

In accordance with the predicted absence of a conventional SQE, treatments with TB 40 μM (figure 5.11) had no effect on squalene accumulation in *P. tricornutum* cultures, not even at high (340 mM) concentrations (data not shown) whereas it clearly triggered accumulation of ergosterol (fig 5.11), possibly interfering with another enzyme in the pathway. Overall, these findings indicate that diatoms produce sterols through the conventional cyclisation of 2,3-epoxysqualene by a conventional OSC but using a non-conventional SQE enzyme to generate the 2,3-epoxysqualene precursor.

Fluconazole - The next step in the reconstruction of the pathway was to determine the specificity of PtOSC. The specificity of this enzyme in synthetizing either cycloartenol or lanosterol, allows the location of the early portion of the pathway within the photosynthetic or non-photosynthetic lineage. Being conserved enzymes, indications about the specificity of OSCs for one product or the other can be inferred by the analysis of the amino acid sequence of the protein, as residues in positions 381, 449 and 453 are specific for the formation of either lanosterol (T381/C,Q449/V453) or cycloartenol (Y381/H449/I453) (numbering relative to OSC of *H. sapiens*) (Summons *et al.*, 2006). The analysis of the active site of PtOSC, suggests that *P. tricornutum* and other diatoms cyclize 2, 3-epoxysqualene to cycloartenol as plants do (Figure 5.10).

Unfortunately, we were not able to detect any enzymatic activity with the recombinant PtOSC protein from *E. coli*. Therefore, we employed a pharmacological approach to verify the product specificity of PtOSC. We experimentally determined the accumulation of intermediate compounds in *P. tricornutum* cultures, by treating diatom cultures with different concentrations of fluconazole, a triazole specific inhibitor of CYP51 (EC 1.14.13.70, PHATRDRAFT_31339), another conserved step in the sterol pathway downstream of OSC. The optimal effect of fluconazole on 3 days old *P. tricornutum* cultures was observed to occur 48 hours after treatment, at a concentration of 200 µg/ml (653 mM), although clear effects were also visible at lower concentrations (data not shown). The GC/MS chromatograms of the TMS derivatized sterol fraction of

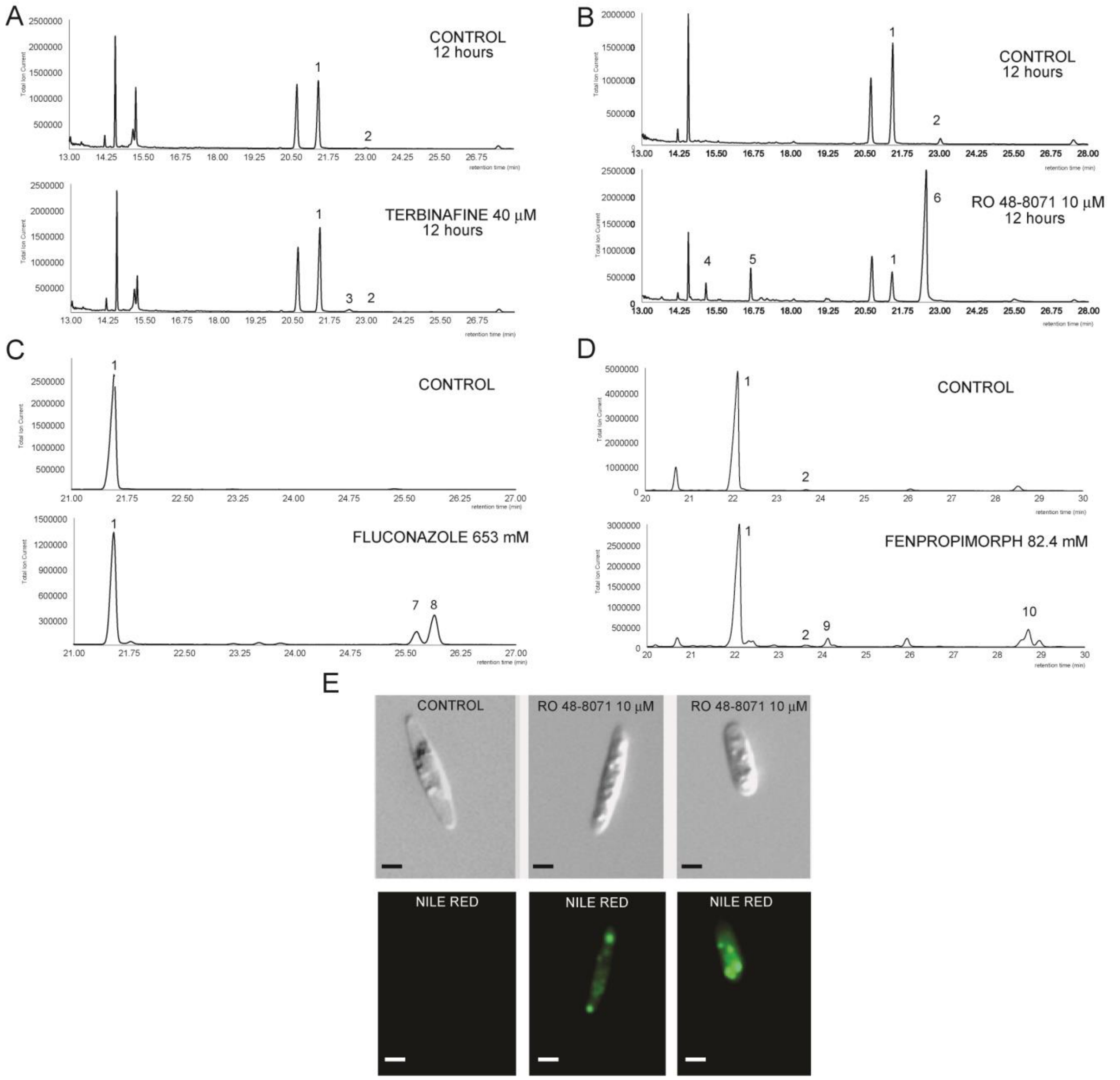


Figure 5.11. Chemical perturbation of *P. tricornutum* sterol synthesis. GC/MS chromatograms of TMS-derivatized non-saponifiable lipid extracts relative to *P. tricornutum* cultures treated for 12 hours with TB 40 μM (A) , Ro 48 8071 10μM (C), fluconazole 653 mM or (D) fenpropimorph 82.4 mM and compared with respective mock treatments. Peak numbering: 1, TMS-brassicasterol; 2, TMS-campesterol; 3, TMS-ergosterol; 4, squalene; 5, 2,3-epoxysqualene; 6, Ro 48-8071; 7 TMS-obtusifoliol; 8 TMS derivative of unknown steroid compound of FW 500; 9, putative TMS-fecosterol; 10, TMS-cycloartenol. EI-MS spectra are reported in Figure 5.12. (E) Effects of Ro 48-8071 on *P. tricornutum* cells compared to those of mock controls. Upper panels show differential interference contrast (DIC) images of *P. tricornutum* cells, while lower panels report epifluorescence images of the same cells upon staining with Nile Red, to visualize intracellular triacylglycerols (TAGs). Scale bars represent 3 μm.

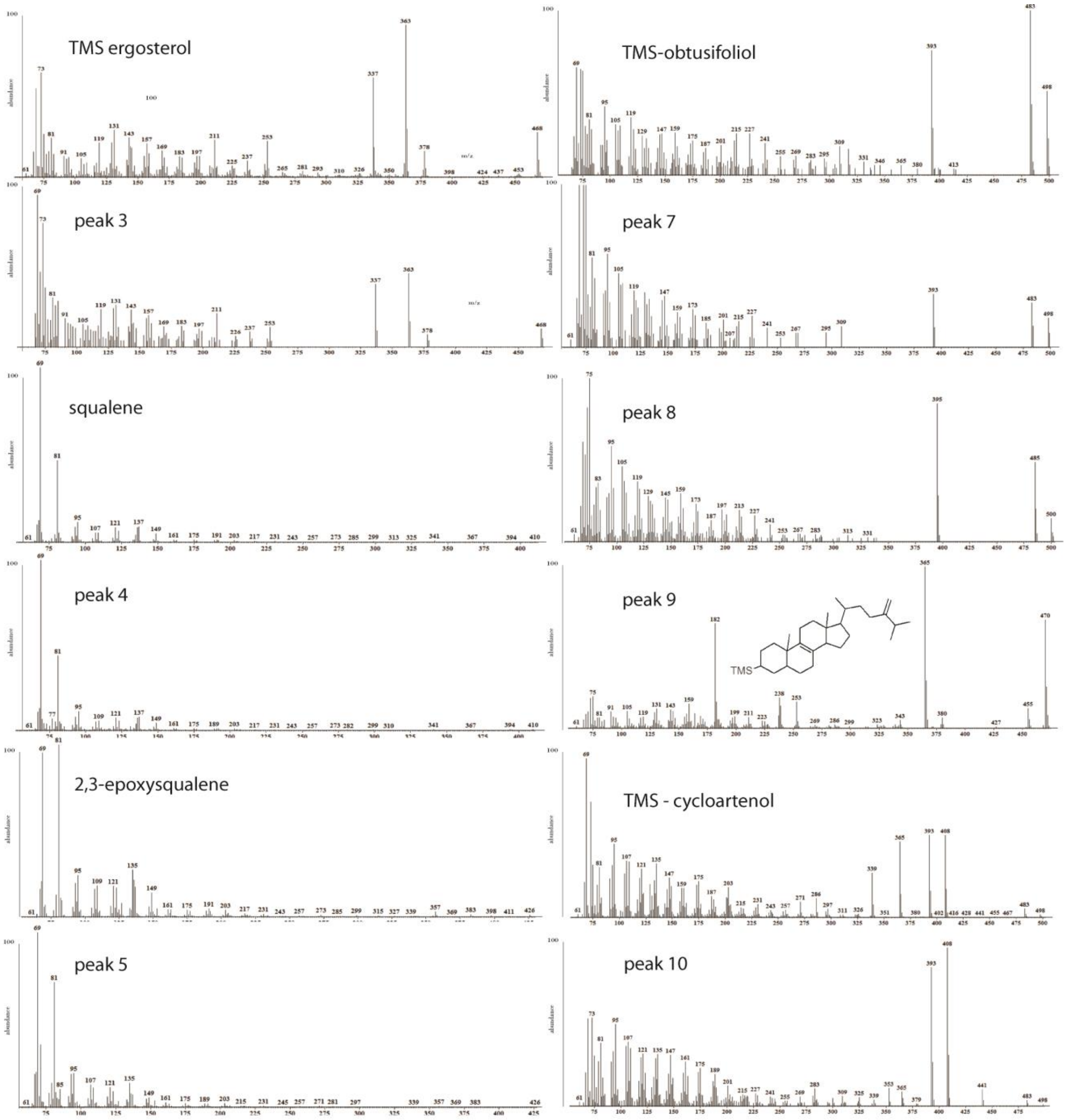


Figure 5.12. EI-MS spectra of the main peaks of the chromatograms showed in Figure 5.11 and that of relative standards. EI-MS spectrum of peak 9 possibly corresponds to TMS-fecosterol, although no standard is available.

treated diatoms demonstrated the presence of two distinct peaks, as compared to the control samples (Figure 5.11). The first corresponded to obtusifoliol (Figure 5.12), a known derivative of cycloartenol thus confirming the *in silico* predicted specificity of

PtOSC. The second peak was related to an additional unknown steroid compound of FW 428, which might possibly correspond to a reduced form of obtusifoliol (Figure 5.12).

Fenpropimorph - Fenpropimorph has strong effects and possibly multiple targets in the sterol pathway of *P. tricornutum*. While drug concentrations between 100 and 50 μg/ml rapidly killed diatoms (data not shown), treatments at concentrations of 25 μg/ml (82.4 mM) and 12.5 μg/ml (41.2 mM) caused a perturbed sterol profile. The GC/MS analysis of TMS derivatized sterols accumulated by diatom cultures after 2 days of treatment with fenpropimorph 41.2 mM revealed the presence of cycloartenol, and another peak, presumably fecosterol (Figure 5.11), a typical fungal sterol. The accumulation of cycloartenol, which further confirms the predicted activity of PtOSC, is likely caused by the inhibition of the methylsterol monooxygenase encoded by *PHATRDRAFT_10852*, as this enzyme type is also reported to be a known target of fenpropimorph (Burden *et al.*, 1989). While the identity of the peak relative to cycloartenol was confirmed by comparison with a commercial standard, that of fecosterol was inferred from the EI-MS spectra. The presence of this compound may be supported by the fact that the enzyme Δ7-sterol isomerase, that converts this compound to episterol, is known target of fenpropimorph (Campagnac *et al.*, 2009).

Screening for SQE activity

The presence of both squalene and 2,3-epoxysqualene in diatoms extracts implies that the epoxidation of squalene might occur through an alternative mechanism, presumably involving an unconventional enzyme. Therefore, we mined the genome of *P. tricornutum* for genes encoding enzymes possibly related to such reaction, identifying a list of candidate SQEs that included cytochrome P450s, FAD-dependent monooxygenases, carotenoid epoxidases and hydroxylases (Table 5.2). Given its unusual structure, *PtOSC* was incorporated in the list as possible multifunctional SQE-OSC enzyme. Several approaches were used to screen these genes for squalene epoxidase activity. For example, we tried to generate a *S. cerevisiae ERG1* mutant, which would have served as background for complementation experiments, but none of the attempts of deleting the gene were successful, given the lethality of the *ERG1*gene. Therefore, we screened a cDNA library of *P. tricornutum* cloned in yeast pAG423-GPD expression vectors by transforming it in *S. cerevisae* and subsequently plating it onto selective medium enriched with TB. TB prevents growth of WT yeast due to the inhibition of ERG1, but not of *P. tricornutum*, possibly allowing a complementation that might result in TB-resistant colonies. The sequencing of the plasmids retained by colonies grown on TB resulted in a group of *P. tricornutum* genes not related to epoxidation of squalene, with no correspondences in the list of candidate genes reported in Table 5.2. Thus, we evaluated the squalene epoxidase activity by setting up an *in vitro* assay, using both cell lysates and cell-free extracts of *E.coli* strain BL21 (DE3) transformed with pDEST17 plasmids carrying the genes reported in Table 5.2. *In vitro* reactions were carried out following several established protocols for in vitro SQE activity determination (Nagumo *et al.*, 1995; Laden *et al.*, 2000; Germann *et al.*, 2005), in presence or in absence of *E.coli* lysates containing *ATR1* (*AT4G24520*), since most CYP450 and FAD monooxygenase systems require a CYP450 NADPH-reductase for their functional reconstitution *in vitro*. Unfortunately, no 2,3-epoxysqualene was detected in any of the samples analyzed by GC/MS (data not shown). Due to practical reasons, a sub-group of genes was selected for the co-regulation of their expression level with genes involved in the sterol biosynthesis (*PHATRDRAFT_46438*, *PHATRDRAFT_26422* *PHATRDRAFT_45845*) (Matthijs *et al.,* unpublished), for their similarity to particularly flexible class of CYP450s

(*PHATRDRAFT_50101, PHATRDRAFT_26422*) (Syed *et al.*, 2013) and for their unusual structure (*PtOSC*).

Gene ID	Functional annotation
PHATRDRAFT_41947	FAD dependent oxidoreductase
PHATRDRAFT_24362	FAD dependent oxidoreductase
PHATRDRAFT_48932	monooxygenase activity
PHATRDRAFT_16283	monooxygenase activity
*PHATRDRAFT_46438**	monooxygenase activity (CYP450)
PHATRDRAFT_6940	monooxygenase activity (CYP450)
*PHATRDRAFT_50101**	LUT1-1 putative β-carotene hydroxylase (CYP450)
*PHATRDRAFT_26422**	LUT1-1 putative β-carotene hydroxylase (CYP450)
PHATRDRAFT_31339	CYP51
*PHATRDRAFT_45845**	zeaxanthin epoxidase
PHATRDRAFT_56492	similar to zeaxanthin epoxidase
PHATRDRAFT_56488	similar to zeaxanthin epoxidase
*PtOSC**	oxidosqualene cyclase

Table 5.2. List of candidate genes screened for SQE activity. Genes marked with asterisks have been included in the shortlist for *in vitro* SQE assays involving purified proteins and *in vivo* SQE assays in squalene-producing *E. coli* strains

These genes were optimized for expression in *E. coli* as described earlier (*PtOSC*) or by removing trans-membrane domains, where present (*PHATRDRAFT_46438, PHATRDRAFT_26422, and PHATRDRAFT_45845*). Squalene epoxidase activity was tested *in vitro* using both cell lysates and purified protein extracts, either in presence or in absence of *E.coli* lysates containing *ATR1* (*AT4G24520*). Given the difficult expression of *PtOSC* in *E. coli*, it was not possible to purify the protein. None of the samples was observed to convert squalene to 2,3-epoxysqualene (data not shown). Thus, given the ability to synthetize and accumulate little amounts of intracellular squalene demonstrated by *E. coli* cells when expressing *PtIDISQS* (Figure 5.), SQE activity was evaluated *in vivo* in strains co-transformed with *PtIDISQS* and *SQE* candidate genes belonging to the above mentioned short-list*,* either with or without *ATR1.* The accumulation of possible reaction products was checked by GC/MS after two and three days of induction; however, in none of the samples we could detect the presence of 2,3-epoxysqualene (data not shown).

Activity of Δ7-sterol reductase and sterol C-22 desaturase

As indicated in Table 5.1, genes *PHATRDRAFT_30461* and *PHATRDRAFT_51757* putatively encode a sterol Δ7 reductase and a sterol C-22 desaturase (of the CytP450 CYP710 subfamily), respectively. Both genes have high homology with similar genes of *A. thaliana* (Table S1). *PHATRDRAFT_30461* is the putative ortholog of the gene *DWF5* (*AT1G50430*) that encodes a sterol Δ7 reductase and catalyzes the reaction EC 1.3.1.21, whereas *PHATRDRAFT_51757* is the putative ortholog of *AT2G34490*, which encodes a cytochrome P450 with sterol 22-desaturase activity (Morikawa *et al.*, 2006). In our proposed pathway reconstruction (Figure 5.3), these enzymes were associated to the last two reactions, theoretically converting ergosterol to campesterol (PHATRDRAFT_30461) and subsequently to brassicasterol (PAHTRADRAFT_51757), respectively.

To confirm their predicted activity, both *P. tricornutum*'s genes were cloned and expressed in *S. cerevisiae*. Yeast naturally accumulates ergosterol as the main sterol compound and therefore it represents an ideal background for the purpose of the experiment. GC/MS analysis of yeast expressing *PHATRDRAFT_30461* confirmed its

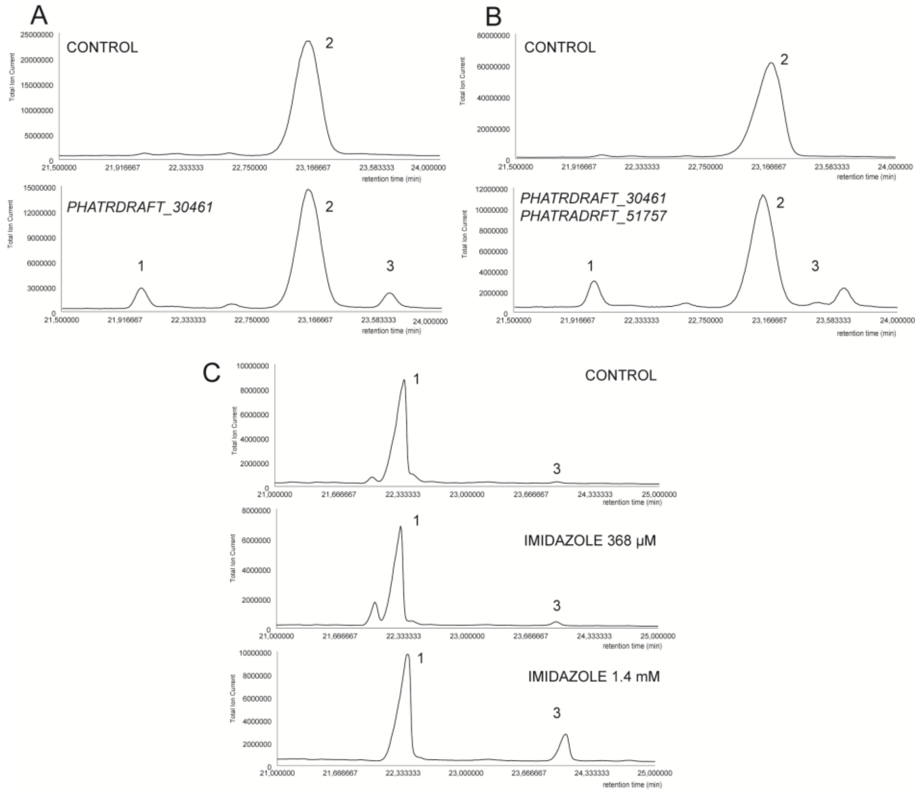


Figure 5.13. Functional characterization of *PHATRADRFT_30461* and *PHATRADRFT_51757*. (A) GC/MS chromatograms of TMS-derivatized non-saponifiable lipid extracts relative to *S. cerevisiae* W303 expressing *PHATRDRAFT_30461* independently (B) or in combination with *PHATRADRFT_51757* compared to controls transformed with corresponding empty vector(s). (C) GC/MS chromatograms showing the perturbed sterol composition of *P. tricornutum* cultures treated with imidazole 368 µM or 1.4 mM for 48 hours, compared to mock treatment. Peak numbering: 1, TMS-brassicasterol; 2, TMS-ergosterol; 3, TMS- campesterol.

Δ7-reductase activity in converting ergosterol to campesterol, by the detection of the latter compound in the sterol fraction of the transformed yeast (Figure 5.13). Additionally, in the same samples, brassicasterol was also detected (Figure 5.13), presumably as a result of a desaturation of campesterol in position C-22. Mass spectra of both compounds were compared to those obtained from GC/MS profiling of *P. tricornutum*, resulting identical. Similarly, both sterols were detected in samples which expressed both diatom genes. The expression of *PHATRDRAFT_30461* alone was sufficient to enable campesterol and brassicasterol synthesis in yeast cells, possibly due to the presence of the gene *ERG5*, the endogenous C-22 sterol reductase. In support of the possible activity of PHATRADRAFT_51757 in converting campesterol to brassicasterol, we observed that in the presence of increasing concentrations of imidazole, which inhibits CYP710s with minor specificity*, P. tricornutum* accumulates campesterol in a concentration-dependent manner (Figure 5.13). Unfortunately, none of our attempts to delete the *ERG5* gene from the *S. cerevisiae* genome to obtain a cleaner background for the functional analysis of PHATRDRAFT_51757 was successful.

A link between sterol and TAG accumulation

To further confirm the involvement of the putative *PtOSC* in diatom sterol biosynthesis, *P. tricornutum* cells were transformed with RNA interference (RNAi) constructs targeted to a portion of this gene (within the locus *PHATR_645*). Resulting *PtOSC* knock-down (KD) lines exhibited a striking phenotype (Figure 5.14), characterized by dramatically impaired growth (data not shown), overall reduced sterol content, abnormalities in the cellular membrane and, remarkably, significantly increased triacylglycerol (TAG) accumulation (Figure 5.14). These findings support the involvement of *PtOSC* in the sterol pathways and, in addition, demonstrate the existence of a link between sterol biosynthesis and TAG accumulation. Notably, none of the *KD* lines accumulated detectable amounts of precursors of the targeted enzymes, as otherwise observed in chemical inhibition experiments with the drug Ro 48-8071 (Figure 5.11).

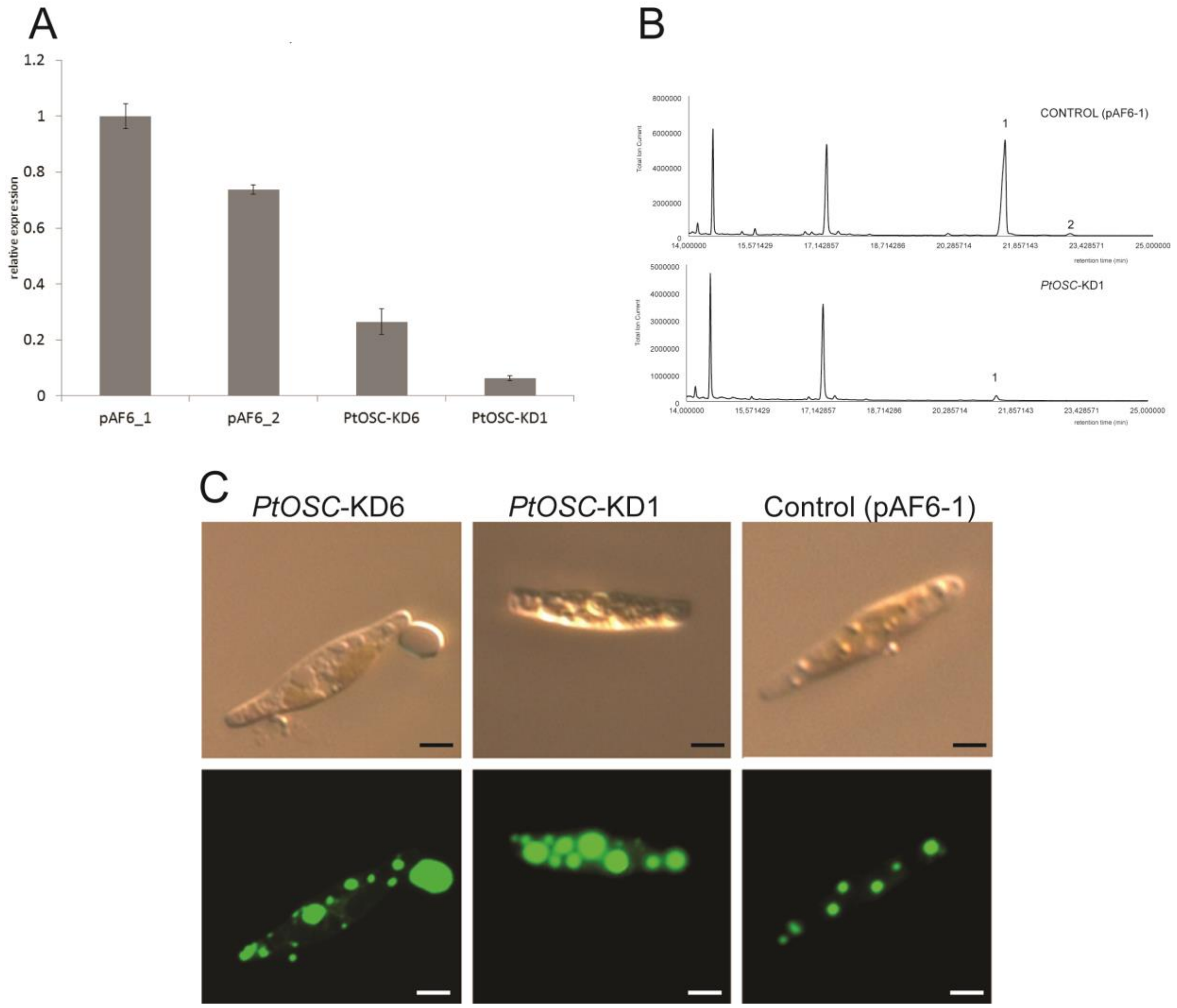


Figure 5.14 Analysis of *PtOSC*-KD lines. (A) RT-qPCR relative quantification *PtOSC* transcripts in *PtOSC*-KD lines compared to controls transformed with pAF6 empty vectors. Expression ratio was normalized with *histone H4* and *tubulin β chain* as reference genes and with pAF6_1 sample as reference sample, respectively. Error bars represent standard deviation of three technical replicates (B) GC/MS chromatograms of TMS-derivatized non-saponifiable lipid fraction of *PtOSC*-KD1 showing a dramatic decrease in the accumulation of the main steroid compound compared to pAF6. Peak numbering: 1, TMS-brassicasterol; 2, TMS-campesterol. (C) Effects of the silencing of *PtOSC* on three-day old *P. tricornutum* cells compared to a pAF6 control line. Upper panels show differential interference contrast (DIC) images of *P. tricornutum* cells, lower panels show epifluorence images of intracellular TAG accumulation, stained with Nile Red. Scale bars represent 3 μm.

Discussion

Reconstruction of hypothetical sterol pathway of *P. tricornutum*

The presented experiments and computational analysis, allowed a theoretical reconstruction of the pathway that *P. tricornutum* utilizes for the biosynthesis sterols. In accordance with previous analysis (Rampen *et al.,* 2010), we conclude that the main sterols produced by *P. tricornutum* are brassicasterol and campesterol (Figure 5.2). *P. tricornutum* possibly utilizes a hybrid pathway that has the initial and final part in common with the sterol biosynthesis of plants, while the central part appears to be similar to that of fungi (Figure 5.3). The lack of a conventional SQE, arisen from *in silico* pathway analysis, has been confirmed *in vivo* by diatom insensitivity to TB, which, differently, blocks SQE and triggers intracellular accumulation of squalene in fungal (Ta *et al.*, 2012), plant (Wentzinger and Bach, 2002) and at higher concentration, animal cells (Ryder, 1992). Despite the apparent absence of a conventional SQE, *P. tricornutum* uses squalene as precursor indicating that the epoxidation of squalene occurs but presumably involves a different mechanism/enzyme. As in plants and green algae, 2, 3-epoxysqualene is cyclized to cycloartenol by a conventional OSC. In *P. tricornutum* this is encoded by *PtOSC*, an unusually large OSC, characterized by an unknown domain at its N-terminus. Cycloartenol is then methylated in position C-24 by PHATRDRAFT_10824, a sterol methyltransferase (PtSMT), yielding 24-methylene-cycloartenol. One of the two methyl groups present at C-4 is removed by the subsequent action of the methylsterol monooxygenase encoded by PHATRDRAFT_10852 (PtMSMO, EC 1.14.13.72), which shows significant similarity only with plant orthologs, besides those of diatoms (Table S1), suggesting the occurrence a plant specific reaction. The result of this complex conversion is a hydroxysterol, which is decarboxylated on the C-4 hydroxy-methyl group and, at the same time, dehydrogenated at the hydroxyl group at the C-3 carbon by a NADPH-dependent reaction, catalyzed by a 3-β-hydroxysteroid-4-α-carboxylate-3-dehydrogenase (PtHSDD, EC 1.1.1.170) encoded by PHATRADRFT_48864. The resulting keto-sterol requires a reduction on the oxygen at C-3 in order to be further converted. This step involves a 3-keto-steroid reductase, which is currently unidentified in plants and algae. We postulate that a reductase involved in other reactions might have acquired specificity for keto-sterols. For example, using Pathologic (Karp *et al.*, 2002), we identified PHATRDRAFT_5870, encoding a short-chain dehydrogenase/reductase as best candidate for this reaction, potentially yielding a molecule of cycloleucalenol, the typical precursor of obtusifoliol in plants. Similarly to land plants, a cycloleucalenol cycloisomerase (EC 5.5.1.9) breaks the cyclopropane ring present between C-19 and C-9 with a consequent instauration of a double bond at position C-9/C-8. In *P. tricornutum* such enzyme is possibly encoded by the gene *PHATRADRFT_49447* (*PtCCI*), which shares best orthology relationships with organisms of the "green lineage". The result of this reaction is obtusifoliol, a common intermediate in the sterol biosynthesis of plants and green algae and specific substrate of the cytochrome P450 sterol 14-α-demethylase (PtCYP51, EC 1.14.13.70). PtCYP51 is encoded by *PHATRDAFT_31339* and its specificity for obtusifoliol was confirmed by the detection of obtusifoliol in inhibition experiments with fluconazole. In the same experiment, the presence of a similar compound is possibly due to a reduction of obtusifoliol as a response of its unnatural intracellular accumulation, or, less likely, to a parallel branch of pathway that from 2,3-epoxysqualene produces another substrate for PtCYP51. This enzyme is responsible for the removal of the methyl group in position C-14 of obtusifoliol and for the formation of the double bond between C-14 and C-15, yielding (4α)-methyl-(5α)-ergosta-8,14,24(28)-trien-3β-ol. The presence of a sterol C14-24 reductase (PtSC14-24R, EC

1.3.1.70), corresponding to the locus *PHATRDRAFT_48260*, allows the conversion of this compound to methyl-fecosterol that, through the re-iteration of the oxidative demethylation described above catalyzed by PtMSMO, is converted to fecosterol, a typical intermediate of ergosterol biosynthesis in yeast and fungi. *PHATRDRAFT_36801* encodes a Δ8-Δ7 sterol isomerase (PtEBP, EC 5.3.3.5) possibly responsible for the shift of the double bond from C-8/C-9 of fecosterol to C-8/C-7 of episterol. *PtEBP* does not have orthologs in any of the organisms used as reference in the pipeline used to construct DiatomCyc, not even in *T. pseudonana* (Table S1), although preliminary BLAST analysis indicated that the gene has similarity with animal and fungal enzymes. Episterol is desaturated by a Δ7-sterol 5-desaturase (EC 1.14.21.6, PtDES-5, PHATRDRAFT_14208) to form dehydro-episterol, subsequently converted to 5,7,22,24(28)-ergostatetraenol by a second desaturation reaction in position C-22, catalyzed by the cytochrome P450 C-22 sterol desaturase (PtCY61) encoded by *PHATRADRFT_51757*. Continuing this part of the pathway that closely resembles that of fungi, 5,7,22,24(28)-ergostatetraenol is reduced in position C-24 by the product of *PtSC14-24R* (EC 1.3.1.71) giving ergosterol as product. Despite being the main sterol of fungi, ergosterol is often found in algae (Miller *et al.,* 2012). We showed with the heterologous expression of the Δ7-sterol reductase gene of *P. tricornutum* in yeast that the corresponding diatom enzyme is capable of converting ergosterol in the phytosterols campesterol and brassicasterol, the final products of the sterol biosynthesis of *P. tricornutum*. Although in our experiments, the expression of *PHATRDRAFT_30461* alone was sufficient for the synthesis of both campesterol and brassicasterol; in *P. tricornutum* the biochemical conversion is likely to occur in two steps, involving PHATRADRFT_30461 in the reduction of ergosterol to campesterol and possibly PHATRADRAFT_51757 (CYP710) in the C-22 desaturation of campesterol to form brassicasterol. Interestingly, orthologs of *PHATRADRAFT_51757* have not been found in *T. pseudonana*, (Table S1) indicating a possible difference in the pathways of the two diatoms. In agreement with that, the sterols produced by *T. pseudonana*, are not desaturated in position C-22 (Rampen *et al.,* 2010).

IDI-SQS fusion in diatoms and lack of SQE

Besides its chimeric nature, the sterol biosynthetic pathway of P. tricornutum harbors several other peculiarities that make it unique among the known variants of this pathway across the kingdom of life. The MVA and the sterol pathways of *P. tricornutum* are characterized by the fusion of the IDI and SQS activities in a single multifunctional enzyme, by the lack of a conventional, yet unidentified, SQE, and the presence of an exotic OSC that synthetizes cycloartenol and that is composed of a conserved C-terminal domain and a large, less conserved N-terminal region. Interestingly, these features recur in all sequenced diatoms.

Joining of different enzymatic activities in a single fusion protein is a frequent event in protein evolution and often involves enzymes that are related, for example subjected to co-regulation or sharing the same pathway (Hwang *et al.*, 2011), as in the case of IDI and SQS. Since the IDI-SQS fusion observed in *P. tricornutum* is conserved in diatoms, and presumably in the Stramenopiles group as well, it suggests that the origin of the fusion might be considerably ancient and might have conferred a selective advantage. Although this is difficult to assess, we hypothesize that the advantage may be represented by a possible increased metabolic efficiency of the mevalonate and sterol pathways, reached by more economical and immediate transcriptional regulation (Morris *et al.*, 2009) and, at the same time, physical proximity of the enzymes (Yanai *et al.*, 2001). The latter phenomenon can sometimes result in metabolic channeling, in which the vicinity of the different catalytic sites is sufficiently tight to ensure that the substrate does not diffuse in the cell (Morris *et al.,* 2009). This prevents side reactions and dispersion of the compound, eventually increasing the efficiency of the metabolic process (Yanai *et al.* 2000) and lowering the resource-investment costs for its regulation and coordination (Morris *et al.* 2009 In diatoms, the presence of fusion enzymes that catalyse subsequent reactions appears relatively frequent. Indeed, examples are found in carbohydrate metabolism, where a triosephosphate-isomerase/glyceraldehyde-3-phosphate dehydrogenase (PHATRADRAFT_25308), a UDP-glucose-pyrophosporylase/phosphoglucomutase (PHATRADRFT_50444) and a glucose-6-phosphate-dehydrogenase/6-phosphogluconate-dehydrogenase (PHATRADRFT_54663) were predicted to exist as fusion proteins in *P. tricornutum* (Kroth et al. 2008).

In contrast to the former fusion enzymes, IDI and SQS activity do not occupy a consecutive position within the sterol pathway. DMAPP, the product of IDI, is converted to FPP before being converted by SQS to squalene. This intermediate conversion occurs through two additional reactions catalyzed by PHATRADRFT_49325 and PHATRDRAFT_47271, of which ORFs are localized on different chromosomes than *PtIDISQS*. Although the occurrence of protein-protein interactions with other enzymes of the pathway cannot be excluded, it is possible that the selective advantage of this genetic fusion is mainly represented by more efficient regulation, rather than by metabolic channeling.

Despite the lack of SQE and the presence of an unusual OSC, *P. tricornutum* utilizes both squalene and 2, 3-epoxysqualene as sterol precursors. So far, the epoxidation of squalene by SQE, requiring FAD as cofactor and molecular oxygen, was believed to be a ubiquitous reaction, occurring in aerobic conditions in every sterol producing organism through an identical mechanism. As for most of the other proteins working upstream of OSC, the degree of conservation of SQE is so high that no sterol-producing organisms with alternative SQE enzymes have been reported yet. However, our survey clearly indicates that the SQE gene seems to be lost in several groups. The existence of an alternative biochemical mechanism for the epoxidation of squalene thus seems plausible (Summons *et al.* 2006) and diatoms might use one of them.

For example, diatoms and other organisms lacking SQE might have evolved a particular CYP450 enzyme that acquired novel activity or specificity. CYP450s form a large family of enzymes involved in many metabolic reactions, including hydroxylations and epoxydations to which the evolution of the sterol pathway is intrinsically bound (Omura, 2013). In rat tumor cells it has been demonstrated that a cytochrome P450 17-alpha hydroxylase-17,20 lyase (CYP17) has a secondary squalene epoxidase activity (Liu *et al.*, 2005). Alternatively, among the consistent number of genes encoding proteins with obscure function (POF) (Fabris *et al.* 2012), diatoms might harbor an enzyme with unprecedented squalene epoxidase activity. Hypothetically, squalene epoxidation could occur anaerobically using water as source of oxygen instead of O_2 (Raymond and Blankenship, 2004), although the thermodynamics of such reaction would be significantly less favorable (Summons *et al.* 2006). Since the synthesis of sterols presumably evolved in anaerobic eukaryotes facing an increasingly oxidative environment, it is possible that the primordial SQE was anaerobic (Raymond and Blankenship, 2004) and substantially different. Before being replaced by the 'modern' SQE, this hypothetical primitive enzyme might have persisted in some groups, while others evolved the 'modern' SQE enzyme. Based on these suppositions, we compiled a list of possible alternative enzymes that might have replaced SQE in diatoms (Table 5.2) and we tested their activity in a series of preliminary assays, both *in vivo* and *in vitro*. Although these assays resulted unsuccessful, it cannot be excluded that the group of candidate enzymes includes the alternative SQE. CYP450 and FAD-dependent monooxygenase systems are known to be difficult to be functionally reconstituted in *E. coli*, due to the lack of appropriate NADPH reductases (Cunningham *et al.,* 2007) and to the usually insoluble character of these enzymes, which are often ER membrane-bound proteins.

Link between sterol biosynthesis and TAG accumulation

The phenotype exhibited by *PtOSC-KD* lines shows traits that are typical for unicellular organisms with impaired sterol biosynthesis (Wentiznger *et al.* 2002, Ta *et al.,* 2012) and strongly suggest the enzyme plays a key role in the sterol pathway. The silencing of *PtOSC* produces effects on growth, intracellular TAG accumulation and membrane defects that are comparable to those caused by Ro 48-8071, and supports the presence of major impairments in the sterol biosynthesis. In plant (Wentiznger *et al.* 2002) and yeast (Ta *et al.,* 2012) cells, the perturbation of the sterol metabolism through chemical inhibitors causes the appearance of lipid droplets, while in human cells, the blockage of HMGR, catalyzing the rate limiting step of MVA pathway, triggers accumulation of polyunsaturated fatty acids (PUFAs) and up-regulation of the fatty acid biosynthetic genes (Plée-Gautier *et al.*, 2012). The increased accumulation of lipid droplets in *PtOSC-KD* lines indicates the existence of a link between the regulation of sterol biosynthesis and TAG accumulation. To some extents this might be due to the reduced growth rate observed in diatom mutant lines. When growth rate slows down, cells need less chemical energy and building blocks for division, therefore carbon is in excess and it is stored in lipid droplets. However, the blockage of the sterol biosynthesis represents the interruption of a major constant flux of carbon in the cell, which may trigger cellular metabolic re-organizations and cause carbon to be channeled to alternative pathways, such as fatty acid biosynthesis. Additionally, sterols are crucial for cellular membrane fluidity and dynamism. Since the shortage of sterols in diatoms cells causes membrane abnormalities that can result in cell damage (Figure 5.14), it is likely that the cell increases the synthesis of lipid molecules that have similar roles in the membrane, to replace sterols and restore the membrane integrity and fluidity.

Conclusions

The reconstruction of two important pathways such as MVA and sterol biosynthesis of *P. tricornutum* presented in this chapter represents the first study of this type in diatoms, combining computational and biochemical analysis with functional characterization of metabolic enzymes. The experimental validation of the reconstructed pathways reported earlier (Fabris *et al.* 2012), highlights the value and reliability of DiatomCyc as a research tool for metabolic studies on diatoms.

The proposed biosynthetic pathway follows a hybrid fungal/plant route, involving a novel multifunctional enzyme and novel alternatives to conserved biochemical solutions. This underscores the prominent metabolic flexibility of diatoms and requires a general reconsideration of the sterol biosynthetic pathway, currently considered as a highly conserved pathway and subdivided into rigid phylogenetic variants.

The absence of SQE in diatoms and seemingly in other groups of organisms raises intriguing questions on the selective pressure that might have caused the loss and/or the replacement of such an enzyme. Further efforts will be required to identify the alternative enzyme that diatoms and all the organisms that lost SQE use. The reconstruction of the MVA pathway of *P. tricornutum* is an important step towards the understanding of the isoprenoid biosynthesis in diatoms, which is very attractive for the sustainable production biofuels and high valuable compounds. A better knowledge of this pathway and its regulation will ultimately drive metabolic engineering strategies aimed at the production of specific compounds for important industrial applications. Finally, we showed a link between sterol biosynthesis and TAG accumulation in diatoms by generating sterol mutant strains that showed increased lipid accumulation. Understanding the mechanisms underlying the balance and regulation between sterol biosynthesis and lipid accumulation might lead to the identification of useful genetic targets for the generation of metabolically optimized diatoms for increased oil productivity.

Experimental procedures

Chemicals

Squalene, 2,3-epoxysqualene, lanosterol, ergosterol, fenpropimorph, fluconazole, imidazole and terbinafine were purchased from Sigma-Aldrich (USA), cycloartenol from Extrasynthese (Germany) and Ro 48-8071 from Cayman Chemicals (USA).

Cloning and bacterial and yeast culturing

All genes were amplified from a cDNA library of *P. tricornutum* (Huysman *et al.*, 2013) with PrimeSTAR® HS DNA Polymerase (Takara Bio, Japan), were cloned into pDONR221 entry vectors using Gateway™ cloning technology (Invitrogen, USA) and were sequence-verified. PCR primers (SI Table S1) were designed with Vector NTI (Invitrogen). Destination vectors were all generated with the Gateway™ cloning technology, unless specified otherwise. The plasmid pDEST17 (Invitrogen) was used for the heterologous expression of *P. tricornutum* genes in *E. coli*. *PtOSC* was fused to a Maltose Binding Protein (MBP) at its N-terminus, by cloning it into a modified version of pDEST17 (pDEST17-MBP). For the co-expression of *PtIDISQS* and *SQE* candidate genes, the cassette including *PtIDISQS* flanked by T7 promoter and terminator was amplified with primers introducing *Sma*I and *Spe*I restriction sites and digested with the appropriate enzymes. The resulting insert was cloned into the backbone of pDONR223, *de-novo* amplified by PCR to introduce restriction sites compatible with the *PtIDISQS* insert and remove the attB1/2 cassette.

E. coli DH5α and BL21 (DE3) (Invitrogen, Life Technologies Corporation, USA) cells were used for cloning and heterologous gene expression, respectively. Bacterial cultures for heterologous gene expression were grown on Luria Broth (LB) medium enriched with appropriate antibiotics at 37°C until A_{600nm} of 0.6. Protein synthesis was induced with addition of 1mM isopropyl β-D-1-thiogalactopyranoside (IPTG) followed by 48 hours incubation at 20°C in darkness.

S. cerevisiae strain W303 was used for the heterologous expression of *PHATRDRAFT_51757* and *PHATRADRFT_30461*, which were cloned into pAG426-GAL1 and pAG423-GAL1 plasmids (Alberti *et al.*, 2007), respectively. Yeast transformants were selected on minimal SD base medium (Clontech Laboratories, Inc, USA) with appropriate amino acid drop out supplements (Clontech Laboratories). Transformed yeast cultures were grown overnight in liquid minimal SD base medium with appropriate amino acid drop out supplements at 30° on an orbital shaker at 280 rpm. Cells were harvested by a brief centrifugation at 3000×g, washed twice with sterile water and resuspended in minimal SD base GAL/RAF medium (Clontech Laboratories) with appropriate drop out supplements. Induced cultures were incubated 2 days at 30°C on an orbital shaker at 280 rpm.

Protein extraction and immunoblot analysis

200 ml *E. coli* culture was centrifuged at 6000×g for 10 minutes. Cell pellets were resuspended in phosphate buffer saline (PBS) and sonicated for 1 minute with 10-s pulses with a Heat Systems Ultrasonics sonicator (Heat Systems Incorporated). Cell lysates were either directly used for *in vitro* enzymatic assays or separated from cell debris with an additional centrifugation step at 10.000×g for 20 minutes. Protein purification was performed using Ni-NTA Superflow nickel-charged resin (Qiagen, The Netherlands) according to manufacturer instruction. Immunoblot analysis was carried out wih Mini-PROTEAN® Precast Gels and related equipment (Bio-Rad Laboratories,

USA), anti·His Antibody Selector Kit ECL Mouse IgG and HRP-linked whole Ab (Ge Healthcare, UK) as primary and secondary antibodies, respectively, and the Clarity™ Western ECL Substrate (Bio-Rad Laboratories, USA).

Treatments with chemical inhibitors

Three-days old *P. tricornutum* CCAP 1055/1 cultures grown in ESAW medium (Harrison *et al.*, 1980) at 21°C in continuous light regime (average 75 µmol•photons•m^{-2}•s^{-1}) were treated in triplicate with terbinafine, fluconazole, Ro48-8071, fenpropimorph, or imidazole for 48 hours. Samples were harvested 1, 2, 4, 6, 8, 12, 24 and/or 48 hours after treatment. Mock treatments were performed with the corresponding solvents: dimethyl sulfoxide (DMSO) for terbinafine and fenpropimorph, methyl-acetate for Ro48-8071, 10% ethanol for fluconazole, and water for imidazole.

Metabolite extraction, trimethylsilyl (TMS) derivatization and Gas Chromatography Mass Spectrometry (GC/MS) analysis

50 ml of *P. tricornutum*, 1 ml of *S. cerevisiae* or 2 ml of *E. coli* cultures were harvested by centrifugation for 15 min at 4500×g and cell pellets snap-frozen. Cells were lysed by the addition of equal volumes (250 µL) 40% KOH and 50% ethanol followed by incubation for 10 minutes at 95°C. The extraction was performed by adding 900 µl hexane to the lysate. The phase separation was obtained by centrifugation at 10000×g for 1 minute and the organic upper phase was transferred in a fresh tube. This procedure was repeated two times on the cell lysates. The pooled organic fractions were evaporated and derivatised by adding 100 µl (trimethylsilyl)trifluoroacetamide (TMS) (Sigma-Aldrich) and 20 µl pyridine (Sigma-Aldrich). Standards were dissolved in hexane, evaporated and derivatised similarly.

GC-MS analysis was performed with the GC model 6890 and MS model 5973 (Agilent). One µL of sample aliquot was injected in splitless mode into a VF-5ms capillary column (Varian CP9013, Agilent). Helium carrier gas was set at a constant flow of 1 mL/min. The injector temperature was set to 280°C and the oven temperature was programmed as follows: temperature was held at 80°C for 1 min post injection; subsequently ramped to 280°C at 20°C/min, held at 280°C for 45 min, ramped to 320°C at 20°C/min, held at 320°C for 1 min and cooled to 80°C at 50°C/min at the end of the run. The MS transfer line was set to 250°C, the MS ion source to 230°C, and the quadrupole to 150°C, throughout. Metabolites were identified from full EI-MS spectra, generated by scanning the m/z range of 60-800 with a solvent delay of 7.8 min.

Extraction, HPLC analysis and quantification of lycopene

Pre-cultures of *E. coli* BL21 (DE3) co-transformed with pAC LYC, pAC LYCipi and pDEST17 plasmids for the functional characterization of *PtIDISQS* were grown at 37°C overnight and used to inoculate 200 ml of LB enriched with chloramphenicol 50 mg/l and carbenicillin 100 mg/l. Cultures were grown at 37°C until A_{600nm} of 0.6 was reached. Gene expression was induced with addition of 1mM isopropyl β-D-1-thiogalactopyranoside (IPTG) followed by 72 hours incubation at 20°C in darkness. Cells were harvested by centrifugation (10 minutes at 6000×g) and the obtained biomass was lyophilized.

Aliquots of lyophilized samples were weighed (0.1 mg accuracy) and extracted with acetone/H_2O, 90:10 (v:v) and sonicated with a tip sonicator at 40 W for 30s, 2s pulses. Extracts were filtered over a 0.2 µm Alltech® nylon syringe filter (Fisher Scientifics, USA) to remove cell debris and injected into a Agilent 1100 series HPLC system equipped with a Grace® reverse phase Eclipse XDB C_8 column (150 mm x 4.6mm;

3.5μm). Lycopene was analysed as described (Van Heukelem and Thomas, 2001) using 2 solvents: solvent A, 70:30 (v:v) methanol, 28 mM aqueous tert-butylacrylamide (TBAA)^2, pH 6.5; solvent B, methanol. Lycopene was identified by comparing retention times and absorption spectra and quantified by calculating response factors, using pure lycopene standards (DHI, Denmark).

Gene silencing in *P. tricornutum*

The *PtOSC*-RNAi vector was prepared as described elsewhere (De Riso *et al.*, 2009) using the primers listed in Table S1. *P tricornutum* cells were transformed with *PtOSC*-RNAi or empty pAF6 vectors by biolistic transformation as reported (Falciatore *et al.*, 1999) with a PDS-1000/He™ System (Bio-Rad Laboratories). Transformants were selected and continuously cultivated on ESAW medium containing 100 μ/ml phleomycin.

Nile Red staining and fluorescence microscopy

One ml of *P. tricornutum* culture was stained with 5 μl Nile Red (9-(Diethylamino)-5H benzo [α] phenoxazin-5-one) solution and incubated for 15 minutes in the dark. Fluorescence microscopy images were acquired with a Zeiss Axio Imager.M2m microscopy (Carl Zeiss, Germany) as described previously (Greenspan *et al.*, 1985).

Quantitative Real Time PCR (qRT-PCR) Analysis

cDNA was generated with iScript kit (Bio-Rad) from RNA extracted with RNeasy (Qiagen) from *P. tricornutum* cultures.

qRT-PCR was carried out with a Lightcycler 480 (Roche) and SYBR Green QPCR Master Mix (Stratagene). *Histone H4* (*H4*) and *Tubulin β chain* (*TubB*) were used as the reference genes for normalization (Siaut *et al.*, 2007). Primers for amplification of *PtOSC* (Table S1) were designed with Beacon Designer (Premier Biosoft; www.premierbiosoft.com).

In vitro SQE assays

Biochemical assays to test SQE activity of candidate genes were carried following protocols previously reported (Nagumo *et al.,* 1995, Laden *et al.,* 2000; Germann et al., 2005), using non-labeled squalene. Samples were incubated 3 hours and further extracted and analyzed by GC/MS as described above.

SQE cDNA library screening

A cDNA library of *P. tricornutum* kindly provided by Dr. Marie Huysman (VIB, Ghent University, Belgium) was cloned into pAG423-GPD yeast expression vectors (Alberti *et al.*, 2007) using Gateway™ cloning technology (Invitrogen, Life Technologies Corporation, USA) according to manufacturer instructions **.** 10 µg of plasmid DNA were used to transform *S. cerevisiae* strain 303W as described above. Transformants were selected on minimal SD base medium (Clontech Laboratories, Inc, USA) with appropriate amino acid drop out supplements (Clontech Laboratories, Inc, USA) containing agar 20 g/L and TB 40 µM. Resistant colonies (n=24) were checked by PCR for the presence of the insert with primers GIV119 and GIV120 (Table 5.3). The PCR product was sequenced.

nr	ID	Sequence
1	qPCR_645_F	GTTCGTGGTATGGCTCCTG
2	qPCR_645_R	TGTCATAGCAAGATGTGAAGTC
3	PHATR_645_RNAi_F	tatataGAATTCGCTCAAAGCAAAGTCTGTCAGGG
4	PHATR_645_RNAi_1R	tatatatctagaGCTGGAAACGGAACGAATAAATATG
5	PHATR_645_RNAi_2R	tatatatctagaAATCGGCCAAAGCTGTCAAGCT
6	PtOSC_F	GGGGACAAGTTTGTACAAAAAAGCAGGCTCCATGGCGACAACTACGGAGTACGC
7	PtOSC_R	GGGGACCACTTTGTACAAGAAAGCTGGGTCCTACTTATCCAGATCGCTTCCATAAACAG
8	PtIDISQS_F	GGGGACAAGTTTGTACAAAAAAGCAGGCTCCatgacgacgacgacgacgttacccga
9	PtIDISQS__R	GGGGACCACTTTGTACAAGAAAGCTGGGTCTCMCTGTAGACTCTTGGCTTTTTGCA
10	pDEST17_ T7 cassette _F_SmaI	ATATATCCCGGGtaatacgactcactataggg
11	pDEST17 cassette _R_SpeI	ATATATACTAGTATCCGGATATAGTTCCTCCTTT
12	pDONR223 full plasmid amp_SmaI_F	ATATATCCCGGGAGCTTAAGACTGGCCGTCGTTTTAC
13	pDONR223 full plasmid amp_SpeI_R	ATATATACTAGTGCCCGTGTCTCAAAATCTCTGATG
14	PHATRDRAFT_30461	GGGGACAAGTTTGTACAAAAAAGCAGGCTCCATGACCTCCTCTAGCTCCAGCAA
15	PHATRDRAFT_30461	GGGGACCACTTTGTACAAGAAAGCTGGGTCTCMGACGACACCGGGTACAATCTTG
16	PHATRDRAFT_51757	GGGGACAAGTTTGTACAAAAAAGCAGGCTCCatgtcgttgtcgaaacagcg
17	PHATRDRAFT_51757	GGGGACCACTTTGTACAAGAAAGCTGGGTCTCMCTCGGTCGCCTTGACCCGTTT
18	PHATRDRAFT_24362_F	GGGGACAAGTTTGTACAAAAAAGCAGGCTCCatggatgacgagaagtcgggat
19	PHATRDRAFT_24362_R	GGGGACCACTTTGTACAAGAAAGCTGGGTCCTATGGATATAAAGGGAAAAAATGCTCG
20	PHATRDRAFT_48932_F	GGGGACAAGTTTGTACAAAAAAGCAGGCTCCATGCCCTATAATGATACAGCTACCGAAC
21	PHATRDRAFT_48932_R	GGGGACCACTTTGTACAAGAAAGCTGGGTCTCMTACAATGGGAAAACACGTAATTTAAGATTGC
22	PHATRDRAFT_16283_F	GGGGACAAGTTTGTACAAAAAAGCAGGCTCCATGTTGTTCACTTCCTGGCGATC
23	PHATRDRAFT_16283__R	GGGGACCACTTTGTACAAGAAAGCTGGGTCTCAATCCTTCAACACTCCCTGACAG
24	PHATRDRAFT_46438_F	GGGGACAAGTTTGTACAAAAAAGCAGGCTCCATGAACATCCTTTTTTCAAAGCCTGAT
25	PHATRDRAFT_46438_R	GGGGACCACTTTGTACAAGAAAGCTGGGTCTTACTTCCTTTGAACGGTAACTTTGA
26	PHATRDRAFT_6940_F	GGGGACAAGTTTGTACAAAAAAGCAGGCTCATGAACGATTACAAAGCCCAGCTCT
27	PHATRDRAFT_6940_R	GGGGACCACTTTGTACAAGAAAGCTGGGTCTCAAATCTCTTGCACCGCACG
28	PHATRDRAFT_50101_F	GGGGACAAGTTTGTACAAAAAAGCAGGCTCCATGCAAGTTGGCAAGTCAGGTGA
29	PHATRDRAFT_50101_R	GGGGACCACTTTGTACAAGAAAGCTGGGTCTTATTTCGGTACCCCACGCTTC
30	PHATRDRAFT_26422_F	GGGGACAAGTTTGTACAAAAAAGCAGGCTCCATGCGCTCGTCGGATTACAGTC
31	PHATRDRAFT_26422_R	GGGGACCACTTTGTACAAGAAAGCTGGGTCTTACATCTTGTGCATTGGACATCCG
32	PHATRDRAFT_31339_F	GGGGACAAGTTTGTACAAAAAAGCAGGCTCCATGATTGTGGCACTTGTGGTAGTTACG
33	PHATRDRAFT_31339_R	GGGGACCACTTTGTACAAGAAAGCTGGGTCTCACGATTGGACGCGCTTACG
34	PHATDRAFT_45845 _F	GGGGACAAGTTTGTACAAAAAAGCAGGCTCCatgaaaagatcttgcagtatagtcacaatc
35	PHATDRAFT_45845 _R	GGGGACCACTTTGTACAAGAAAGCTGGGTCTCMTACAAACCGGCTGCGCCA
36	PHATRDRAFT_56492_F	GGGGACAAGTTTGTACAAAAAAGCAGGCTCCATGAAAAGATCTTGCAGTATAGTCACAATCC
37	PHATRDRAFT_56492_R	GGGGACCACTTTGTACAAGAAAGCTGGGTCTCMTACAAACCGGCTGCGCCACCTC
38	PHATRDRAFT_56488_F	GGGGACAAGTTTGTACAAAAAAGCAGGCTCCATGGGTCTTTCGTTTCTATCATTATGC
39	PHATRDRAFT_56488_R	GGGGACCACTTTGTACAAGAAAGCTGGGTCTCMTAGCTTAGTTCTTCTTTCGTAGCTGC

Table 5.3 Primers used

References

Adolph, S., Bach, S., Blondel, M., Cueff, A., Moreau, M., Pohnert, G., Poulet, S.A., Wichard, T. and Zuccaro, A. (2004) Cytotoxicity of diatom-derived oxylipins in organisms belonging to different phyla. *J. Exp. Biol.*, **207**, 2935–46.

Alberti, S., Gitler, A.D. and Lindquist, S. (2007) A suite of Gateway cloning vectors for high-throughput genetic analysis in Saccharomyces cerevisiae. , 913–919.

Benveniste, P. (2004) Biosynthesis and accumulation of sterols. *Annu. Rev. Plant Biol.*, **55**, 429–57.

Bowler, C., Allen, A.E., Badger, J.H., *et al.* (2008) The Phaeodactylum genome reveals the evolutionary history of diatom genomes. *Nature*, **456**, 239–44.

Burden, R.S., Cooke, D.T. and Carter, G.A. (1989) Inhibitors of sterol biosynthesis and growth in plants and fungi. *Phytochemistry*, **28**, 1791–1804.

Campagnac, E., Fontaine, J., Lounès-Hadj Sahraoui, a, Laruelle, F., Durand, R. and Grandmougin-Ferjani, a (2009) Fenpropimorph slows down the sterol pathway and the development of the arbuscular mycorrhizal fungus Glomus intraradices. *Mycorrhiza*, **19**, 365–74.

Cardozo, K.H.M., Guaratini, T., Barros, M.P., *et al.* (2007) Metabolites from algae with economical impact. *Comp. Biochem. Physiol. C. Toxicol. Pharmacol.*, **146**, 60–78.

Caroprese, M., Albenzio, M., Ciliberti, M.G., Francavilla, M. and Sevi, A. (2012) A mixture of phytosterols from Dunaliella tertiolecta affects proliferation of peripheral blood mononuclear cells and cytokine production in sheep. *Vet. Immunol. Immunopathol.*, **150**, 27–35.

Cunningham, F.X. and Gantt, E. (2007) A portfolio of plasmids for identification and analysis of carotenoid pathway enzymes: Adonis aestivalis as a case study. *Photosynth. Res.*, **92**, 245–59.

Cunningham, F.X., Sun, Z., Chamovitz, D., Hirschberg, J. and Gantt, E. (1994) Molecular Structure and Enzymatic Function of Lycopene Cyclase from the Cyanobacterium. *Plant Cell*, **6**, 1107 – 1121.

Desmond, E. and Gribaldo, S. (2009) Phylogenomics of sterol synthesis: insights into the origin, evolution, and diversity of a key eukaryotic feature. *Genome Biol. Evol.*, **1**, 364–81.

Dufourc, E.J. (2008) Sterols and membrane dynamics. *J. Chem. Biol.*, **1**, 63–77.

EFSA (2010) Scientific Opinion on the substantiation of health claims related to plant sterols and plant stanols and maintenance of normal blood cholesterol 3140), and maintenance of normal prostate size and normal urination (ID 714 , 1467 , 1635) pursuant to Arti. *EFSA J.*, **8**, 1–22.

Emanuelsson, O., Brunak, S., Heijne, G. von and Nielsen, H. (2007) Locating proteins in the cell using TargetP, SignalP and related tools. *Nat. Protoc.*, **2**, 953–71.

Fabris, M., Matthijs, M., Rombauts, S., Vyverman, W., Goossens, A. and Baart, G.J.E. (2012) The metabolic blueprint of Phaeodactylum tricornutum reveals a eukaryotic Entner-Doudoroff glycolytic pathway. *Plant J.*, **70**, 1004–14.

Falciatore, a, Casotti, R., Leblanc, C., Abrescia, C. and Bowler, C. (1999) Transformation of Nonselectable Reporter Genes in Marine Diatoms. *Mar. Biotechnol. (NY).*, **1**, 239–251.

Galea, A.M. and Brown, A.J. (2009) Special relationship between sterols and oxygen: were sterols an adaptation to aerobic life? *Free Radic. Biol. Med.*, **47**, 880–9.

Gaulin, E., Bottin, A. and Dumas, B. (2010) Sterol biosynthesis in oomycete pathogens. , **5**, 258–260.

Germann, M., Gallo, C., Donahue, T., ***et al.*** (2005) Characterizing sterol defect suppressors uncovers a novel transcriptional signaling pathway regulating zymosterol biosynthesis. *J. Biol. Chem.*, **280**, 35904–13.

Giner, J.-L. and Wikfors, G.H. (2011) "Dinoflagellate Sterols" in marine diatoms. *Phytochemistry*, **72**, 1896–901.

Greenspan, P., Mayer, E.P. and Fowler, S.D. (1985) Nile red: a selective fluorescent stain for intracellular lipid droplets. *J. Cell Biol.*, **100**, 965–73.

Harrison, P.J., Waters, R.E. and Taylor, F.J.R. (1980) A broad spectrum artificial seawater medium for coastal and open ocean phytoplankton. *J. Phycol.*, **16**, 28–35.

Heukelem, L. Van and Thomas, C.S. (2001) Computer-assisted high-performance liquid chromatography method development with applications to the isolation and analysis of phytoplankton pigments. *J. Chromatogr. A*, **910**, 31–49.

Hunter, S., Apweiler, R., Attwood, T.K., ***et al.*** (2009) InterPro: the integrative protein signature database. *Nucleic Acids Res.*, **37**, D211–5.

Huysman, M.J.J., Fortunato, A.E., Matthijs, M., ***et al.*** (2013) AUREOCHROME1a-mediated induction of the diatom-specific cyclin dsCYC2 controls the onset of cell division in diatoms (Phaeodactylum tricornutum). *Plant Cell*, **25**, 215–28.

Hwang, S., Rhee, S.Y., Marcotte, E.M. and Lee, I. (2011) Systematic prediction of gene function in Arabidopsis thaliana using a probabilistic functional gene network. *Nat. Protoc.*, **6**, 1429–42.

Kajiwara, S., Fraser, P.D., Kondo, K. and Misawa, N. (1997) Expression of an exogenous isopentenyl diphosphate isomerase gene enhances isoprenoid biosynthesis in Escherichia coli. *J. Biochem.*, **426**, 421–426.

Karp, P.D., Paley, S. and Romero, P. (2002) The Pathway Tools software. *Bioinformatics*, **18 Suppl 1**, S225–32.

Kirby, J. and Keasling, J.D. (2009) Biosynthesis of plant isoprenoids: perspectives for microbial engineering. *Annu. Rev. Plant Biol.*, **60**, 335–55.

Kodner, R.B., Summons, R.E., Pearson, A., King, N. and Knoll, A.H. (2008) Sterols in a unicellular relative of the metazoans EVOLUTION.

Laden, B.P., Tang, Y. and Porter, T.D. (2000) Cloning, heterologous expression, and enzymological characterization of human squalene monooxygenase. *Arch. Biochem. Biophys.*, **374**, 381–8.

Lamb, D.C., Jackson, C.J., Warrilow, A.G.S., Manning, N.J., Kelly, D.E. and Kelly, S.L. (2007) Lanosterol biosynthesis in the prokaryote Methylococcus capsulatus: insight into the evolution of sterol biosynthesis. *Mol. Biol. Evol.*, **24**, 1714–21.

Leblond, J.D. and Lasiter, A.D. (2012) Sterols of the green-pigmented, aberrant plastid dinoflagellate, Lepidodinium chlorophorum (Dinophyceae). *Protist*, **163**, 38–46.

Liu, Y., Yao, Z.-X. and Papadopoulos, V. (2005) Cytochrome P450 17alpha hydroxylase/17,20 lyase (CYP17) function in cholesterol biosynthesis: identification of squalene monooxygenase (epoxidase) activity associated with CYP17 in Leydig cells. *Mol. Endocrinol.*, **19**, 1918–31.

Lohr, M., Schwender, J. and Polle, J.E.W. (2012) Isoprenoid biosynthesis in eukaryotic phototrophs: a spotlight on algae. *Plant Sci.*, **185-186**, 9–22.

Lyons, T.W. and Reinhard, C.T. (2011) Earth science: Sea change for the rise of oxygen. *Nature*, **478**, 194–195.

Massé, G., Belt, S.T., Rowland, S.J. and Rohmer, M. (2004) Isoprenoid biosynthesis in the diatoms Rhizosolenia setigera (Brightwell) and Haslea ostrearia (Simonsen). *Proc. Natl. Acad. Sci. U. S. A.*, **101**, 4413–8.

Matthijs, M. Nitrogen deprivation alters the primary metabolism of Phaeodactylum tricornutum. *In preparation*

Miller, M.B., Haubrich, B. a, Wang, Q., Snell, W.J. and Nes, W.D. (2012) Evolutionarily conserved Delta(25(27))-olefin ergosterol biosynthesis pathway in the alga Chlamydomonas reinhardtii. *J. Lipid Res.*, **53**, 1636–45.

Morikawa, T., Mizutani, M., Aoki, N., *et al.* (2006) Cytochrome P450 CYP710A Encodes the Sterol C-22 Desaturase in Arabidopsis and Tomato. , **18**, 1008–1022.

Morris, P.F., Schlosser, L.R., Onasch, K.D., Wittenschlaeger, T., Austin, R. and Provart, N. (2009) Multiple horizontal gene transfer events and domain fusions have created novel regulatory and metabolic networks in the oomycete genome. *PLoS One*, **4**, e6133.

Nagumo, a, Kamei, T., Sakakibara, J. and Ono, T. (1995) Purification and characterization of recombinant squalene epoxidase. *J. Lipid Res.*, **36**, 1489–97.

Nes, C.R., Singha, U.K., Liu, J., Ganapathy, K., Villalta, F., Waterman, M.R., Lepesheva, G.I., Chaudhuri, M. and Nes, W.D. (2012) Novel sterol metabolic network of Trypanosoma brucei procyclic and bloodstream forms. *Biochem. J.*, **443**, 267–77.

Ohyama, K., Suzuki, M., Kikuchi, J., Saito, K. and Muranaka, T. (2009) Dual biosynthetic pathways to phytosterol via cycloartenol and lanosterol in Arabidopsis. *Proc. Natl. Acad. Sci. U. S. A.*, **106**, 725–30.

Omura, T. (2013) Contribution of cytochrome P450 to the diversification of eukaryotic organisms. *Biotechnol. Appl. Biochem.*, **60**, 4–8.

Pearson, A., Budin, M. and Brocks, J.J. (2003) Phylogenetic and biochemical evidence for sterol synthesis in the bacterium Gemmata obscuriglobus. *Proc. Natl. Acad. Sci. U. S. A.*, **100**, 15352–7.

Peralta-Yahya, P.P., Zhang, F., Cardayre, S.B. del and Keasling, J.D. (2012) Microbial engineering for the production of advanced biofuels. *Nature*, **488**, 320–8.

Plée-Gautier, E., Antoun, J., Goulitquer, S., Jossic-Corcos, C. Le, Simon, B., Amet, Y., Salaün, J.-P. and Corcos, L. (2012) Statins increase cytochrome P450 4F3-mediated eicosanoids production in human liver cells: a PXR dependent mechanism. *Biochem. Pharmacol.*, **84**, 571–9.

Rampen, S.W., Abbas, B.A., Schouten, S. and Damste, J.S.S. (2010) A comprehensive study of sterols in marine diatoms (Bacillariophyta): Implications for their use as tracers for diatom productivity. , **55**, 91–105.

Raymond, Jason Blankenship, R.E. (2004) Biosynthetic pathways, gene replacement and the antiquity of life. *Geobiology*, **2**, 199–203.

Riso, V. De, Raniello, R., Maumus, F., Rogato, A., Bowler, C. and Falciatore, A. (2009) Gene silencing in the marine diatom Phaeodactylum tricornutum. *Nucleic Acids Res.*, **37**, e96.

Ryder, N.S. (1992) Terbinafine: mode of action and properties of the squalene epoxidase inhibition. *Br. J. Dermatol.*, **126 Suppl** , 2–7.

Siaut, M., Heijde, M., Mangogna, M., Montsant, A., Coesel, S., Allen, A., Manfredonia, A., Falciatore, A. and Bowler, C. (2007) Molecular toolbox for studying diatom biology in Phaeodactylum tricornutum. *Gene*, **406**, 23–35.

Summons, R.E., Bradley, A.S., Jahnke, L.L. and Waldbauer, J.R. (2006) Steroids, triterpenoids and molecular oxygen. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.*, **361**, 951–68.

Sun, Z., Cunningham, F.X. and Gantt, E. (1998) Differential expression of two isopentenyl pyrophosphate isomerases and enhanced carotenoid accumulation in a unicellular chlorophyte. *Proc. Natl. Acad. Sci. U. S. A.*, **95**, 11482–8.

Syed, K., Porollo, A., Lam, Y.W., Grimmett, P.E. and Yadav, J.S. (2013) A catalytically versatile fungal P450 monooxygenase (CYP63A2) capable of oxidizing higher polycyclic aromatic hydrocarbons, alkylphenols, and alkanes. *Appl. Environ. Microbiol.*

Ta, M.T., Kapterian, T.S., Fei, W., Du, X., Brown, A.J., Dawes, I.W. and Yang, H. (2012) Accumulation of squalene is associated with the clustering of lipid droplets. *FEBS J.*, **279**, 4231–44.

Tomazic, M.L., Najle, S.R., Nusblat, A.D., Uttaro, A.D. and Nudel, C.B. (2011) A novel sterol desaturase-like protein promoting dealkylation of phytosterols in Tetrahymena thermophila. *Eukaryot. Cell*, **10**, 423–34.

Vinci, G., Xia, X. and Veitia, R. a (2008) Preservation of genes involved in sterol metabolism in cholesterol auxotrophs: facts and hypotheses. *PLoS One*, **3**, e2883.

Volkman, J.K. (2003) Sterols in microorganisms. *Appl. Microbiol. Biotechnol.*, **60**, 495–506.

Wentzinger, L.F. and Bach, T.J. (2002) Inhibition of Squalene Synthase and Squalene Epoxidase in Tobacco Cells Triggers an Up-Regulation of 3-Hydroxy-3-Methylglutaryl Coenzyme A Reductase. , **130**, 334–346.

Yanai, I., Derti, A. and Delisi, C. (2001) Genes linked by fusion events are generally of the same functional category : A systematic analysis of 30 microbial genomes.

Chapter 6

General conclusions and perspectives

Author contribution
MF wrote the chapter.

The work presented in this thesis aimed to decipher the metabolic capabilities encrypted in the genome of the model diatom *Phaeodactylum tricornutum* using a systems biology approach and, subsequently, to experimentally characterize relevant metabolic pathways related to the carbohydrate and isoprenoid metabolism.

The main deliverables of this work are (I) a substantial improvement of the knowledge on the metabolism of *P. tricornutum*, (II) the development of a solid pipeline for the semi-automated generation of Pathway/Genome Databases (PGDB) suitable for any organism and (III) the creation of DiatomCyc, an important research resource. Moreover, we provided evidences of (IV) the occurrence of novel glycolytic pathways and (V) of the peculiar organization of others, such as the mevalonate and the sterol biosynthetic pathways, which showed striking peculiarities. Finally, (VI) we anticipated hints for possible targets for the metabolic engineering of diatoms.

A solid pipeline for the generation of metabolic databases

The methodology developed and used for the reconstruction of genome-scale metabolic network of *P. tricornutum*, significantly leveraged the amount of information obtainable from its genome, resulting in a notable level of resolution and coverage of its central metabolism (Chapter 3). The robustness of our approach was validated by the identification of novel unexpected pathways and by the experimental functional characterization of the metabolic genes (Chapter 4 and Chapter 5). Such reconstruction pipeline is readily applicable to any organism with available genetic information, either genomic or transcriptomic. For example, it was successfully implemented in the reconstruction of the metabolic network of the medicinal plant *Catharanthus roseus* (www.cathacyc.org; Chapter 8, Van Moerkercke *et al.*, 2013), constructed from RNA sequencing (RNA-Seq) data. Differently from the case of *P. tricornutum*, the amount of initial functional annotation of *C. roseus* transcripts was null, therefore our method entirely generated the amount of information necessary for the reconstruction of the complete cellular metabolism and specific important secondary metabolic pathways (Van Moerckerke *et al.,* 2013). Additionally, both metabolic networks of *C. roseus* and *P. tricornutum* have been made publicly available through freely accessible online metabolic databases. This format, (I) offers logical and interactive organization of the metabolic and genomic information, (II) greatly enhances the usability of such dataset with a suite of web tools (Chapter 3 , Karp *et al.*, 2002; Fabris *et al.*, 2012; Van Moerkercke *et al.*, 2013; Caspi *et al.*, 2010)) and (III) easily allows cooperation and collaborations within the scientific community.

Towards metabolic modeling of diatoms

A common application of metabolic networks is metabolic modeling. Genome-scale metabolic models are used to reproduce and simulate *in silico* different cellular and metabolic phenomena and to predict the effect of physiological or genetic modifications on metabolic fluxes. Also, metabolic models can be used to determine essential metabolic steps and to design metabolic engineering strategies. Modeling can be performed on metabolic networks by translating them into mathematical terms, such as stoichiometric matrixes, in which the network is simplified by grouping and reducing the number of reactions and lowering their connectivity.

To date, algal metabolism has been object of metabolic modeling only concerning *Chlamydomonas reinhardtii, Ostreococcus tauri* and *Ostreococcus lucimarinus* (Boyle & Morgan, 2009; Chang et al., 2011; Davis et al., 2013; Kliphuis et al., 2012; Kliphuis et al., 2011; Krumholz et al., 2012). Main achievements in modeling algal metabolism involved minimal extension network analysis that allowed to match the *in silico* production of experimentally detected metabolites, overcoming missing reactions and pathways and contributing in network completeness (May *et al.*, 2008). More recently, by using Metabolic Flux Analysis (MFA) and Flux Balance Analysis (FBA) it has been possible to successfully predict intracellular metabolic fluxes defining specific functions under fixed constrains, for instance under different light (Kliphuis *et al.,* 2012, Chang et al. 2011) and gas ratio regimes (Kliphuis *et al.* 2011) or either in autotrophic, mixotrophic or heterotrophic growth conditions, providing insights for optimizing algal biomass yield (Boyle and Morgan, 2009; Dal'Molin *et al.*, 2011; Manichaikul *et al.*, 2009), hydrogen production (Dal Molin *et al.* 2011; Manichaikul *et al.* 2009) and to estimate the impact that photorespiration has on cell growth (Kliphuis *et al.* 2012). However, while predictions obtained from metabolic models of bacterial organisms are often consistent with experimental results, similar achievements are far from being obtained in the case of eukaryotic photosynthetic organisms. This is manly imputable to the excessive simplification that characterize the reconstruction of metabolic models of such organisms. This does not match the high genetic, metabolic and biological complexity of those organisms, resulting in highly underdetermined models, which significantly diverge from the reality. Nevertheless, although the biological significance of model-inferred predictions of cellular performances is often limited, FBA on genome-scale metabolic models of microalgae represents a useful resource for the evaluation of possible effects of specific gene knock-outs *in silico*, providing the basis for model-driven metabolic engineering of microalgae and indications about gene essentiality within the metabolic network.

So far, no diatom genome-scale metabolic models have been published. We are currently involved in the realization and validation of the first genome-scale metabolic model of *P. tricornutum*, based on DiatomCyc (Kim *et al.,* unpublished). Within this project, in collaboration with the groups of Prof. Paul Falkowski (Rutgers University, USA), Prof. Imogen Foubert (KU Leuven, Belgium) and Prof. Rene Wijffels (Wagenigen University, The Netherlands) the biomass composition of *P. tricornutum* chemostat cultures, grown in fully controlled, custom designed photobioreactors (PBRs) is being used to validate the *in silico* model (Kim *et al.,* unpublished).

Novel metabolic features in *P. tricornutum*

Important results of this work consisted in the identification of several metabolic peculiarities in the metabolism of *P. tricornutum*. For the first time, we proved the existence of a functional eukaryotic Entner-Doudoroff pathway (EDP) in a eukaryotic organism, and we computationally predicted the existence in *P. tricornutum* of a phosphoketolase pathway (PKP). These findings, together with the fact that a notable redundancy of genes encoding enzymes involved in the Embden-Meyerhof-Parnas pathway (EMPP) exists in *P .tricornutum* (Kroth *et al.*, 2008), suggest that the carbohydrate metabolism of diatoms is exceptionally efficient and flexible. At the same time, the presence of multiple glycolytic pathways raises a number of intriguing questions about role, coordination and regulation of EMPP, EDP and PKP. In Chapter 4 we hypothesized that the presence of EDP in *P. tricornutum* is linked to cellular "economic strategies", as trade-off between ATP yield and resource investment costs for the synthesis of the metabolic enzyme and the transcriptional regulatory machinery, resulting in more rapid and cost-effective adaptation strategies in response to changing environmental conditions. On this regard, a recent study showed that the EMPP is thermodynamically less favorable compared to the EDP when ATP yield is related to overall protein synthesis costs (Flamholz *et al.*, 2013), further supporting this hypothesis. However, the EDP in diatoms might have also other roles. For example, in

Pseudomonas putida the presence of a functional EDP has been put in direct relation to the ability of the bacterium to scavenge reactive oxygen species (ROS) and tolerate oxidative stress (Chavarría *et al.*, 2012). In fact, EDP produces one NAPDH in the conversion of glucose-6-phosphate to D-glucono-d-lactone-6-phosphate (EC 1.1.1.49) and it may be a fast source of reducing equivalents. Diatoms, as many other marine photosynthetic organisms populating upwelling coastal waters, are faced not only to sudden changes in terms of nutrients, but also to other harsh and dynamic conditions such as intense illumination, thermal stress and pollution. These particular stresses are the main causes of ROS formation in algae (Ernani et al., 2003; Latham, 2008), with deleterious effects on lipids, proteins and DNA (Lesser, 2006). To scavenge ROS burst, algae adopt methods that are similar to those used by plants (Sabatini *et al.*, 2009), which mostly use NADPH as reducing agent.

To investigate the role of EDP genes in *P. tricornutum*, we generated several mutant lines in which they were either overexpressed or silenced. Unfortunately, in none of the lines the transformation resulted stable and the induced overexpression or silencing effect was reverted to wild-type levels over a time-span of two-three weeks (data not shown). To understand the role of these pathways, it would be interesting to determine the distribution of carbon fluxes in *P. tricornutum* in different growth conditions. This could be done using CO_2 as a radio-labeled substrate. Alternatively, labeled glucose could be fed to transgenic *P. tricornutum* strains that allow hetero- and mixotrophy (Zaslavskaia *et al.*, 2001). Either the measurement of the carbon fluxes through the EMPP, EDP and PKP or the characterization of stable silencing mutants will help to clarify the contribution of each glycolytic pathway in the cell and their coordination in response to environmental changes and, possibly, elucidate the evolutionary advantage that determined their conservation.

The absence of a gene encoding the conserved squalene epoxidase (SQE) from diatom genomes is one of the most fascinating aspects emerged from this work and it remains an unanswered question (Chapter 5). To synthetize sterols diatoms use the typical substrate and product, squalene and 2, 3-epoxysqualene, respectively. We proposed that diatoms might use a yet to be discovered alternative enzyme to catalyze the same epoxidation reaction. Our preliminary survey suggested that SQE does not lack only in diatoms and further analysis will help in reconstructing the evolutionary scenario in which the SQE divergence originated. Identifying the enzyme that replaced or pre-existed the conventional SQE will have a great impact, since so far, alternative for this conserved enzyme were never reported. This information will shed light in the reconstruction of the evolution of the sterol biosynthetic pathway, which is still uncertain in some aspects. Origin, evolutionary advantage of the different products and occasional disappearance of the sterol pathway are still debated (Summons *et al.*, 2006; Desmond and Gribaldo, 2009; Raymond and Blankenship, 2004), as well as its rare occurrence in prokaryotes, which possibly involved unusual Horizontal Gene Transfer (HGT) events from eukaryotes to prokaryotes (Desmond and Gribaldo, 2009). Interestingly, in the genome of sterol synthetizing bacteria such as *Methylococcus capsulatum* and *Gemmata obscuriblobus*, the genes encoding SQE and OSC are physically associated, constituting a sort of "sterol operon" (Lamb *et al.*, 2007; Pearson *et al.*, 2003). The identification of an unusually extended portion at the N-terminus of PtOSC led us to hypothesize that this unknown domain might serve as second catalytic domain, possibly SQE. Therefore, further investigations are essential on this regard. As showed by the presence of an IDI-SQS multifunctional enzyme (Chapter 5) and by other examples previously identified in the carbohydrate metabolism (Kroth *et al.*, 2008), fusions of proteins with related activity are relatively frequent in *P. tricornutum*, presumably contributing to enhance the efficiency of certain metabolic processes.

The reconstructed sterol pathway of *P. tricornutum* proposed in Chapter 5 resembles an unprecedented mix of the fungal and plant pathways. According to previous biochemical investigation on sterols produced by 106 species of diatoms (Rampen *et al.*, 2010), several variants of this pathway might exist in the Bacillariophyta group. In fact, not only the sterols identified showed a marked diversity and none of them seemed to be constant among all the analyzed species, but also many of them are derived from lanosterol (Rampen *et al.*, 2010), anticipating that OSC has different specificity in some species. The reconstruction of this pathway and its comparative analysis with other species of diatoms and related organisms will surely give interesting information on the level of conservation, on the occurrence of specific variations and possibly on the selective advantage represented by specific sterol molecules.

An efficient metabolism

Optimal resource allocation and investment seem to be the point of convergence of the metabolic peculiarities emerged from this work. Faster and simpler transcriptional regulation, rapid and efficient substrate conversion, optimization of cellular resource-investment costs might be the selective advantage conferred by the conservation of alternative metabolic pathways composed by fewer enzymes (EDP), the fusion of related enzymatic activities in multifunctional enzymes (*PtIDISQS*) and by the genomic clustering of genes involved in the same pathway (PKP). We suggest that these strategies are undoubtedly linked to the ecological success of unicellular non-motile organisms that thrive in a highly unstable environment. The comprehensive picture emerged from this and the other works aimed to characterize specific aspects of the biochemistry of diatoms seem to point towards an extraordinarily optimized, flexible and rapidly adaptable metabolism. Such metabolic configuration presumably enables diatoms to take advantage of sudden changes in light and nutrient availability, during which they possibly respond faster and more efficiently than other phytoplanktonic species.

Towards metabolic engineering of diatoms

The availability of DiatomCyc, a metabolic database in which detailed gene-reaction associations cover all the main pathways of the central metabolism of *P. tricornutum* is, *per se*, a valuable resource for designing metabolic engineering strategies and potential genetic targets and it represents a significant contribution towards the industrial exploitation of diatoms. In addition, it provides the basis for the development of resources such as metabolic models, which will further improve the rational optimization of strains and cultivation technology. The work presented in this thesis offered relevant insights on potential strategies for the metabolic engineering of diatom strains for improved lipid accumulation. The characterization of the sterol biosynthesis in *P. tricornutum* highlighted the existence of a relationship between the sterol biosynthetic pathway and the accumulation of triacylglycerol (TAG) (Chapter 5). Although similar effects have been observed in plants (Wentzinger and Bach, 2002), fungi (Ta *et al.*, 2012) and on different extents in animal cells (Plée-Gautier *et al.*, 2012), to our knowledge the specific mechanisms that triggers it has not been elucidated. The knock-down cell lines generated in this thesis, despite showing a dramatic accumulation, were characterized by severely impaired growth which made the further characterization of the diatom strains impossible. Therefore, a possible approach to study this metabolic cross-talk and to eventually improve the lipid productivity *P.*

tricornutum would be represented by the association of RNA interference (RNAi) constructs to inducible promoters, in order to trigger the silencing of sterol biosynthetic genes or their regulators, once the algal culture has reached the desired cell density.

On the other hand, the mevalonate (MVA) pathway is considered an attractive target for many metabolic engineering strategies for the production of several high valuable compounds and non-fatty acid derived biofuels (Peralta-Yahya *et al.*, 2012). The accurate reconstruction of the pathway described in Chapter 5 will possibly help in designing genetic modifications aimed to increase the production of isoprenoid precursors, which could be ultimately combined with the heterologous expression of specific enzymes for the direct production of biofuel precursors (Peralta-Yahya *et al.*, 2012) and of transporters for the secretion of the precursors directly into the culture medium (Doshi *et al.*, 2013).

Future perspectives

Although the study of diatom metabolism is still in its infancy, compared to other organisms, its characterization is advancing rapidly. Tools and methods for the molecular study of diatoms until a few years ago were scarce and ineffective. Recently, they have been greatly refined and many new ones have been developed or adapted from other fields, such as yeast-2-hybrid (Y2H), yeast-1-hybrid (Y1H) and transient reporter assays to cite some examples (Huysman *et al.*, 2013). Recent breakthroughs allowed targeted gene knock-out (KO) in *P. tricornutum* (Cellectis Press Release) by using the innovative TALEN™ nucleases (Beurdeley *et al.*, 2013), which will certainly have a principal role for the future metabolic engineering of diatoms. However, molecular research on diatoms still presents substantial technological limitations, compared to other fields of biotechnology, such as those of plants and yeasts. The often problematic codon usage, the limited availability of vectors and promoters and the unfeasibility of (low cost) targeted gene KO are factors that make the study of the molecular biology of diatoms a challenging task. On the other hand, the availability of modern resources such as Next Generation Sequencing (NGS) technologies and impressive computational power at competitive costs, recently determined a remarkable increase in the number of (meta-) genomic and (meta-) transcriptomic projects, aimed to characterize *non*-model marine organisms and to discover novel biochemical traits from phytoplanktonic species. However, the sequencing of the first diatom genomes (Bowler *et al.*, 2008; Armbrust *et al.*, 2004) highlighted the significant limits represented by the use of computational approaches originally developed for land species on diatom datasets, resulting in gene models which are often incomplete (*PtEDD*, Chapter 3; *PtOSC* and *PtIDISQS,* Chapter 5) or even not predicted (*PtPMVK*, Chapter 5). Thus, the increasing volume of genomic data originated from phytoplankton will require the development and optimization of specific pipelines for gene model construction and functional annotation. RNA sequencing is becoming an affordable lab tool and it will be key for the prediction of gene models and provide understanding on the regulation of important metabolic pathways, which in diatoms appear to be different than in plants or green algae (Hockin *et al.*, 2012). The further development and improvement of resources such as DiatomCyc will be possible with the addition of databases relative to other species, the implementation of transcriptomics, metabolomics and flux balance analysis data. As such, it will provide the diatom community with a platform similar to the Plant Metabolic Network (PMN, www.plantcyc.org), which includes metabolic data of more than 350 plant species. The future generation of more genome-scale metabolic models of diatoms and other microalgal species will surely benefit from the reconstruction of DiatomCyc.

Reciprocally, the availability of additional metabolic networks will result in an increasing accuracy and resolution of future versions of *P. tricornutum* metabolic network, which will take advantage from comparative studies with phylogenetically closely related organisms.

In conclusion, this work contributed to gain extensive knowledge on the central metabolism of diatoms, providing the basis for further metabolic studies and raising a number of intriguing questions on the role, origin and evolution of the peculiarities of diatom metabolism. Future efforts will be aimed to address them and to further understand the contribution of these metabolic strategies and of others, yet to be discovered, to the ecological success of diatoms. Moreover, the wealth of information that will be gained from future studies and from the sequencing of genomes and transcriptomes of other diatom species will certainly result in important breakthroughs that will favor the industrial exploitation of diatoms.

References

Armbrust, E.V., Berges, J. a, Bowler, C., *et al.* (2004) The genome of the diatom Thalassiosira pseudonana: ecology, evolution, and metabolism. *Science*, **306**, 79–86.

Beurdeley, M., Bietz, F., Li, J., *et al.* (2013) Compact designer TALENs for efficient genome engineering. *Nat. Commun.*, **4**, 1762.

Bowler, C., Allen, A.E., Badger, J.H., *et al.* (2008) The Phaeodactylum genome reveals the evolutionary history of diatom genomes. *Nature*, **456**, 239–44.

Boyle, N.R. and Morgan, J. a (2009) Flux balance analysis of primary metabolism in Chlamydomonas reinhardtii. *BMC Syst. Biol.*, **3**, 4.

Caspi, R., Altman, T., Dale, J.M., *et al.* (2010) The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res.*, **38**, D473–9.

Cellectis Press Release Cellectis has successfully engineered the genome of photosynthetic algae with a view to biofuel production. Available at: http://www.cellectis.com/media/press-release/2013/cellectis-has-successfully-engineered-genome-photosynthetic-algae-view-biof.

Chang, R.L., Ghamsari, L., Manichaikul, A., *et al.* (2011) Metabolic network reconstruction of Chlamydomonas offers insight into light-driven algal metabolism. *Mol. Syst. Biol.*, **7**, 518.

Chavarría, M., Nikel, P.I., Pérez-Pantoja, D. and Lorenzo, V. de (2012) The Entner-Doudoroff pathway empowers Pseudomonas putida KT2440 with a high tolerance to oxidative stress. *Environ. Microbiol.*

Dal'Molin, C.G.D.O., Quek, L.-E., Palfreyman, R.W. and Nielsen, L.K. (2011) AlgaGEM--a genome-scale metabolic reconstruction of algae based on the Chlamydomonas reinhardtii genome. *BMC Genomics*, **12 Suppl 4**, S5.

Davis, M.C., Fiehn, O. and Durnford, D.G. (2013) Metabolic acclimation to excess light intensity in Chlamydomonas reinhardtii. *Plant. Cell Environ.*, 1391–1405.

Desmond, E. and Gribaldo, S. (2009) Phylogenomics of sterol synthesis: insights into the origin, evolution, and diversity of a key eukaryotic feature. *Genome Biol. Evol.*, **1**, 364–81.

Doshi, R., Nguyen, T. and Chang, G. (2013) Transporter-mediated biofuel secretion. *Proc. Natl. Acad. Sci.*, 1–6.

Ernani Pinto Teresa C . S . Sigaud-Kutner, M.A.. S.. L. (2003) Heavy metal induced oxidative stress in algae. *J. Phycol.*, **1018**, 1008–1018.

Fabris, M., Matthijs, M., Rombauts, S., Vyverman, W., Goossens, A. and Baart, G.J.E. (2012) The metabolic blueprint of Phaeodactylum tricornutum reveals a eukaryotic Entner-Doudoroff glycolytic pathway. *Plant J.*, **70**, 1004–14.

Flamholz, a., Noor, E., Bar-Even, a., Liebermeister, W. and Milo, R. (2013) Glycolytic strategy as a tradeoff between energy yield and protein cost. *Proc. Natl. Acad. Sci.*, 2–7.

Hockin, N.L., Mock, T., Mulholland, F., Kopriva, S. and Malin, G. (2012) The response of diatom central carbon metabolism to nitrogen starvation is different from that of green algae and higher plants. *Plant Physiol.*, **158**, 299–312.

Huysman, M.J.J., Fortunato, A.E., Matthijs, M., *et al.* (2013) AUREOCHROME1a-mediated induction of the diatom-specific cyclin dsCYC2 controls the onset of cell division in diatoms (Phaeodactylum tricornutum). *Plant Cell*, **25**, 215–28.

Karp, P.D., Paley, S. and Romero, P. (2002) The Pathway Tools software. *Bioinformatics*, **18 Suppl 1**, S225–32.

Kliphuis, A.M.J., Klok, A.J., Martens, D.E., Lamers, P.P., Janssen, M. and Wijffels, R.H. (2012) Metabolic modeling of Chlamydomonas reinhardtii: energy requirements for photoautotrophic growth and maintenance. *J. Appl. Phycol.*, **24**, 253–266.

Kliphuis, A.M.J., Martens, D.E., Janssen, M. and Wijffels, R.H. (2011) Effect of O(2) : CO(2) ratio on the primary metabolism of Chlamydomonas reinhardtii. *Biotechnol. Bioeng.*, **108**, 2390–2402.

Kroth, P.G., Chiovitti, A., Gruber, A., *et al.* (2008) A model for carbohydrate metabolism in the diatom Phaeodactylum tricornutum deduced from comparative whole genome analysis. *PLoS One*, **3**, e1426.

Krumholz, E.W., Yang, H., Weisenhorn, P., Henry, C.S. and Libourel, I.G.L. (2012) Genome-wide metabolic network reconstruction of the picoalga Ostreococcus. *J. Exp. Bot.*, **63**, 2353–62.

Lamb, D.C., Jackson, C.J., Warrilow, A.G.S., Manning, N.J., Kelly, D.E. and Kelly, S.L. (2007) Lanosterol biosynthesis in the prokaryote Methylococcus capsulatus: insight into the evolution of sterol biosynthesis. *Mol. Biol. Evol.*, **24**, 1714–21.

Latham, H. (2008) Temperature stress-induced bleaching of the coralline alga Corallina officinalis: a role for the enzyme bromoperoxidase. *Biosci. Horizons*, **1**, 104–113.

Lesser, M.P. (2006) Oxidative stress in marine environments: biochemistry and physiological ecology. *Annu. Rev. Physiol.*, **68**, 253–78.

Manichaikul, A., Ghamsari, L., Hom, E.F.Y., *et al.* (2009) Metabolic network analysis integrated with transcript verification for sequenced genomes. *Nat. Methods*, **6**, 589–92.

May, P., Wienkoop, S., Kempa, S., *et al.* (2008) Metabolomics- and proteomics-assisted genome annotation and analysis of the draft metabolic network of Chlamydomonas reinhardtii. *Genetics*, **179**, 157–66..

Peralta-Yahya, P.P., Zhang, F., Cardayre, S.B. del and Keasling, J.D. (2012) Microbial engineering for the production of advanced biofuels. *Nature*, **488**, 320–8.

Plée-Gautier, E., Antoun, J., Goulitquer, S., Jossic-Corcos, C. Le, Simon, B., Amet, Y., Salaün, J.-P. and Corcos, L. (2012) Statins increase cytochrome P450 4F3-mediated eicosanoids production in human liver cells: a PXR dependent mechanism. *Biochem. Pharmacol.*, **84**, 571–9.

Rampen, S.W., Abbas, B.A., Schouten, S. and Damste, J.S.S. (2010) A comprehensive study of sterols in marine diatoms (Bacillariophyta): Implications for their use as tracers for diatom productivity. , **55**, 91–105.

Raymond, Jason Blankenship, R.E. (2004) Biosynthetic pathways, gene replacement and the antiquity of life. *Geobiology*, **2**, 199–203.

Sabatini, S.E., Juárez, A.B., Eppis, M.R., Bianchi, L., Luquet, C.M. and Ríos de Molina, M.D.C. (2009) Oxidative stress and antioxidant defenses in two green microalgae exposed to copper. *Ecotoxicol. Environ. Saf.*, **72**, 1200–6.

Summons, R.E., Bradley, A.S., Jahnke, L.L. and Waldbauer, J.R. (2006) Steroids, triterpenoids and molecular oxygen. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.*, **361**, 951–68.

Ta, M.T., Kapterian, T.S., Fei, W., Du, X., Brown, A.J., Dawes, I.W. and Yang, H. (2012) Accumulation of squalene is associated with the clustering of lipid droplets. *FEBS J.*, **279**, 4231–44.

Van Moerkercke, A., Fabris, M., Pollier, J., *et al.* (2013) CathaCyc, a Metabolic Pathway Database Built **Pearson, A., Budin, M. and Brocks, J.J.** (2003) Phylogenetic and biochemical evidence for sterol synthesis in the bacterium Gemmata obscuriglobus. *Proc. Natl. Acad. Sci. U. S. A.*, **100**, 15352–7

Wentzinger, L.F. and Bach, T.J. (2002) Inhibition of Squalene Synthase and Squalene Epoxidase in Tobacco Cells Triggers an Up-Regulation of 3-Hydroxy-3-Methylglutaryl Coenzyme A Reductase. , **130**, 334–346.

Zaslavskaia, L. a, Lippmeier, J.C., Shih, C., Ehrhardt, D., Grossman, a R. and Apt, K.E. (2001) Trophic conversion of an obligate photoautotrophic organism through metabolic engineering. *Science*, **292**, 2073–5.

Chapter 7

Summary

Diatoms represent a heterogeneous group of Stramenopile microalgae, estimated to comprise approximately 200.000 species, which dominates phytoplankton in coastal regions (Armbrust, 2009). Diatoms have a relevant ecological impact on global scale. In the oceans, they sustain the food web and play a central role in the geochemical cycling of silica and carbon. On larger scale, diatom produce the 20-25% of the oxygen present in the atmosphere (Falkowski *et al.*, 2004) and adsorb one third of the carbon dioxide emission related to anthropogenic human activities (Sabine *et al.*, 2004). The sequencing of diatom genomes revealed the surprising presence of genes of different origin and a considerable number of unknown diatom-specific genes, highlighting the contribution of two distinct endosymbiotic events, frequent horizontal gene transfers (HGTs) in their evolution path (Armbrust *et al.*, 2004; Bowler *et al.*, 2008). This had a major reflection on the metabolism of diatoms, which differs significantly from that of other microalgae and plants. Diatom metabolism includes features from all domains of life and possibly represents one of the key factors in their ecological success. Mainly due to their metabolic capabilities, diatoms have a great industrial and biotechnological potential, in particular for the sustainable production of bioenergy and high-value bio-products.

Although metabolic systems biology had entered the microalgal field with the reconstruction of genome-scale metabolic networks of a few green algal model species, this approach had never been used on diatoms. This dissertation presented the first attempt to map diatom metabolism through a genome-wide approach, which unveiled novel metabolic arrangements and pathways.

In Chapter 3 we described the reconstruction of the genome-scale metabolic network of *P. tricornutum* through a semi-automated method in which the original functional annotation of the single genes was substantially improved. The resulting metabolic network covered in details the central metabolism of *P. tricornutum*, and made possible the filling of pathway-gaps emerged from previous studies (Kroth *et al.*, 2008). *P. tricornutum*'s metabolic network was converted in a freely accessible online database, DiatomCyc, based on the BioCyc collection of Pathway/Genome Databases (Caspi *et al.*, 2010), which offers a logical organization of the generated data and numerous useful tools for comparative studies and for the upload of experimental high-throughput datasets. The reconstruction of the metabolic network led to the identification of unexpected pathways in the carbohydrate metabolism of *P. tricornutum*. Chapter 4 describes the discovery of a putative phosphoketolase pathway (PKP) and functional Entner-Doudoroff pathway (EDP). We characterized the functionality of the EDP by complementation experiments in *Escherichia coli* mutant strains and by *in vitro* enzymatic assays with cell free extracts. Furthermore, the occurrence of the pathway *in alga* was supported by gene expression analysis. Although the roles of the EDP and PKP in *P. tricornutum* remain uncertain, we suggested that their advantage, compared to the Embden-Meyerhof-Parnas (EMP), might be represented by higher thermodynamic efficiency and faster responses in facing sudden changes in nutrient availability.

In Chapter 5 we characterized the mevalonate and the sterol biosynthetic pathway of *P. tricornutum*. From the preliminary reconstruction offered by DiatomCyc, these two pathways resulted peculiarly organized. The reconstruction of both pathways was possible by combining computational analysis with experimental evidences obtained by using specific chemical inhibitors, gene silencing and gene functional characterization in yeast and *E. coli*. Our results highlighted striking peculiarities, such as a multifunctional isopentenyl diphosphate isomerase (IDI) – squalene synthase (SQS) enzyme in the mevalonate pathway; the lack of the key enzyme squalene epoxidase (SQE) and the presence of a peculiar oxidosqualene cyclase (OSC) in the sterol biosynthesis. We proposed that *P. tricornutum* synthetizes brassicasterol and campesterol through an unusual pathway that resemble a mix between those occurring in plants and in fungi, with the formation of cycloartenol and ergosterol as intermediate compounds. Moreover, the phenotype of mutant diatom lines in which an important sterol gene such as OSC was silenced, indicated the existence of a link between the sterol biosynthesis and the accumulation of triacylglycerol (TAG) in *P. tricornutum*, offering important hints for the metabolic engineering of diatoms.

In summary the application of metabolic system biology for the study of the metabolism of diatoms resulted successful and provided a solid base for the future metabolic studies. The reconstruction of the metabolism of *P. tricornutum* revealed important novelties in the central metabolism of diatoms, directly linked to their peculiar evolution and their exceptional metabolic efficiency and flexibility, which significantly contributed to their ecological success.

References

Armbrust, E.V. (2009) The life of diatoms in the world's oceans. *Nature*, **459**, 185–92.

Armbrust, E.V., Berges, J. a, Bowler, C., *et al.* (2004) The genome of the diatom Thalassiosira pseudonana: ecology, evolution, and metabolism. *Science*, **306**, 79–86.

Bowler, C., Allen, A.E., Badger, J.H., *et al.* (2008) The Phaeodactylum genome reveals the evolutionary history of diatom genomes. *Nature*, **456**, 239–44.

Caspi, R., Altman, T., Dale, J.M., *et al.* (2010) The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res.*, **38**, D473–9.

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

Kroth, P.G., Chiovitti, A., Gruber, A., *et al.* (2008) A model for carbohydrate metabolism in the diatom Phaeodactylum tricornutum deduced from comparative whole genome analysis. *PLoS One*, **3**, e1426.

Sabine, C.L., Feely, R. a, Gruber, N., *et al.* (2004) The oceanic sink for anthropogenic CO2. *Science*, **305**, 367–71.

Chapter 8

CathaCyc, a Metabolic Pathway Database Built from *Catharanthus roseus* RNA-Seq Data

Published article

Van Moerkercke A*, Fabris M*, Pollier J*, Baart GJE, Rombauts S, Hasnain G, Rischer H, Memelink J, Oksman-Caldentey KM and Goossens A.

Plant and Cell Physiology (2013) 54(5): 673-685

(*These authors contributed equally to this work)

Author contribution

MF performed orthology-based transcript annotation, metabolic network reconstruction, manual curation, developed website and database and wrote part of the manuscript.

Abstract

The medicinal plant Madagascar periwinkle (*Catharanthus roseus*) synthesises numerous terpenoid indole alkaloids (TIAs), such as the anticancer drugs vinblastine and vincristine. The TIA pathway operates in a complex metabolic network that steers plant growth and survival. Pathway databases and metabolic networks reconstructed from 'omics' sequence data can help to discover missing enzymes, study metabolic pathway evolution and, ultimately, engineer metabolic pathways. To date, such databases have mainly been built for model plant species with sequenced genomes. Although genome sequence data are not available for most medicinal plant species, next-generation sequencing is now extensively employed to create comprehensive medicinal plant transcriptome sequence resources.

Here we report on the construction of CathaCyc, a detailed metabolic pathway database, from *C. roseus* RNA-Seq data sets. CathaCyc (version 1.0) contains 390 pathways with 1347 assigned enzymes and spans primary and secondary metabolism. Curation of the pathways linked with the synthesis of TIAs and triterpenoids, their primary metabolic precursors, and their elicitors, the jasmonate hormones, demonstrated that RNA-Seq resources are suitable for the construction of pathway databases. CathaCyc is accessible online (http://www.cathacyc.org) and offers a range of tools for the visualisation and analysis of metabolic networks and 'omics' data. Overlay with expression data from publicly available RNA-Seq resources demonstrated that two well-characterised *C. roseus* terpenoid pathways, those of TIAs and triterpenoids, are subject to distinct regulation by both developmental and environmental cues. We anticipate that databases such as CathaCyc will become key to the study and exploitation of the metabolism of medicinal plants.

Introduction

The medicinal plant *Catharanthus roseus* (Madagascar periwinkle) synthesises over 150 different terpenoid indole alkaloids (TIAs), including the pharmaceutically important molecules ajmalicine and serpentine. In addition, it is the sole source of the commercial anticancer TIA compounds vinblastine and vincristine (van der Heijden et al. 2004, Verma et al. 2012). All TIA compounds are synthesized in a highly branched and complex pathway from the central compound strictosidine, a condensation product of the monoterpenoid compound secologanin and the indole compound tryptamine (Facchini and De Luca 2008). The biosynthesis of TIAs occurs in a jasmonate (JA)-responsive manner and involves at least three different cell types and several intracellular compartments (Facchini and St-Pierre 2005, Guirimand et al. 2011, St-Pierre et al. 1999). Because TIAs accumulate in very low amounts, they are difficult to extract, leading to a high commercial production cost. Efforts to increase or alter the production of TIAs in *C. roseus* plant, cell culture or hairy root systems have only been partly successful, due to the complex cellular organisation of this pathway and our fragmented knowledge of the enzymes acting in the different branches (Hughes et al. 2004, Zhou et al. 2009). Indeed, most enzymes involved in the production of TIA compounds from strictosidine onwards have not been identified, with the exception of the 6-step conversion of tabersonine to vindoline (Facchini and De Luca 2008, Loyola-Vargas et al. 2007). *C. roseus* serves as one of the model systems of choice for TIA production, but it also produces several other classes of natural products, including phenolics and triterpenoids (Ferreres et al. 2011, Huang et al. 2012, Mustafa and Verpoorte 2007, Yu et al. 2012).

Inherent to the species-specific character of plant secondary metabolism, most of these molecules and/or pathways are absent in model systems such as *Arabidopsis thaliana* (Arabidopsis). Therefore, further exploration of secondary metabolism in medicinal plants will be needed to develop metabolic engineering strategies to alter the production of these compounds in plants and other systems. Many of the enzymes involved in secondary metabolic pathways in medicinal plants such as *C. roseus* await discovery and characterisation. Combined with our limited understanding of the regulation of these pathways, this impedes the development of efficient metabolic engineering strategies to increase yields of the high-value compounds they produce. As a consequence, the metabolic potential of medicinal plants in general is far from being fully explored. Therefore, categorisation and characterisation of all relevant pathways in these plants would benefit medicinal compound discovery and metabolic engineering. This can be streamlined by constructing metabolic databases from annotated sequence information.

Pathway databases (PDBs) constructed from annotated genomes are available for model systems like *Escherichia coli* (EcoCyc), *Saccharomyces cerevisiae* (YeastCyc), Arabidopsis (AraCyc), *Oryza sativa* (RiceCyc), *Populus trichocarpa* (PoplarCyc), *Chlamydomonas reinhardtii* (ChlamyCyc), *Homo sapiens* (HumanCyc) and over 100 bacteria (Jaiswal et al. 2006, Karp et al. 1997, May et al. 2009, Mueller et al. 2003, Romero et al. 2005, Zhang et al. 2010, Zhang et al. 2005). Recently, a PDB of the model diatom *Phaeodactylum tricornutum* (DiatomCyc) has been constructed, leading to the identification of novel glycolytic pathways in eukaryotes (Fabris et al. 2012). Furthermore, a PDB of *Medicago truncatula* (MedicCyc) was constructed from the draft genome sequence, complemented with EST sequence information (Urbanczyk-

Wochniak and Sumner 2007). The Solanaceae community also created metabolic PDBs (SolCyc) for some of its members (LycoCyc, PetCyc, CoffeaCyc, CapCyc and PotatoCyc) from EST collections (http://solcyc.solgenomics.net/). Finally, the MetaCyc PDB aims to collect every experimentally determined biochemical pathway for small molecule metabolism (Caspi et al. 2012, Krieger et al. 2004) and PlantCyc aims to catalogue all plant-specific molecules, enzymes and pathways (Zhang et al. 2010). To date PlantCyc (http://www.plantcyc.org) comprises 879 pathways and 3,455 compounds. Particularly the latter seems low given the large number of compounds present in plants. For instance, over 12,000 alkaloid compounds, of which 2,000 are TIAs, have already been identified in the plant kingdom (Ziegler and Facchini 2008). Many of these compounds and pathways are not represented in the current databases. Although some literature-curated pathways from *C. roseus* have been included in PlantCyc and MetaCyc (e.g. vinblastine biosynthesis), a comprehensive overview or database of *C. roseus* metabolism has not been constructed to date.

Even though the genome of *C. roseus* and other medicinal plants has not been sequenced yet, for many of them large EST or RNA-Seq collections are available (Desgagné-Penix et al. 2012, He et al. 2011, Wenping et al. 2011). Here, we show that such resources have great potential, not only for gene discovery but also for the establishment of metabolic PDBs. First, we have conducted an elaborate RNA sequencing experiment of the *C. roseus* transcriptome, spanning different organs and growth conditions. From this, we have reconstructed the *C. roseus* metabolic map, emphasising important compound classes, using AraCyc and MetaCyc as initial templates in conjunction with the Pathway Tools prediction software (Karp et al. 2010). Optimisation of the PDB by manual curation resulted in the first metabolic PDB of *C. roseus*, called CathaCyc.

Results

Illumina HiSeq2000 RNA sequencing

RNA-Seq is the most powerful transcript profiling method available to date and, unlike microarray technology, is applicable to species without existing genomic sequence (Wang et al. 2009), including *C. roseus*. Therefore we designed an Illumina HiSeq2000-based RNA sequencing strategy such that both *de novo* sequence assembly and transcript counting was possible. The former facilitates PDB assembly, cloning of full-length open reading frames (FL-ORFs) for gene discovery projects and proteomics analysis, amongst others. The latter enables comparative mining of gene expression, which in turn allows selecting candidate genes for gene discovery programs, for instance to fill the current gaps in TIA biosynthesis. Here, we focus on the generation and use of our RNA-Seq data set for the assembly of the first *C. roseus* PDB, CathaCyc. The *C. roseus* explants material that has been used encompasses suspension cells and shoots treated or not with methyl jasmonate (MeJA).

In total about 44.2 x10^9 bases were sequenced. The cell and shoot samples were run in separate batches and initially assembled as separate RNA-Seq sets (Supplementary Table S8.1). Of the cell-derived libraries, a total of 141,031,789 reads were sequenced corresponding to 28,206,357,800 bases. *De novo* assembly was performed with the VELVET transcriptome assembler (Zerbino and Birney 2008) and its module OASES (Schulz et al. 2012), which generated a total of 31,015 contigs with an average length of 840 nt. The contig maximum length was 8,252 nt. Of the shoot-derived libraries, a total of 80,127,936 reads were sequenced corresponding to 16,025,587,200 bases. A total of 36,363 contigs was generated from this set with an average length of 1,084 nt. The contig maximum length was 11,904 nt.

Pilot BLAST screens revealed that all known TIA genes could be retrieved either in the cell or shoot contig collections, or both (Supplementary Table S8.2). Importantly, all of them were nearly (minimum 98%) or 100% identical to the sequences in The National Center for Biotechnology Information (NCBI) database and in the assemblies from both explant sets. Furthermore, for 25 out of the 36 known TIA genes, the full-length (FL) sequence could be retrieved. This analysis supports the quality and potential utility of our sequence data set for gene discovery and the construction of CathaCyc.

Assembly and annotation of the 'reference transcriptome

To construct a reference transcriptome for *C. roseus*, the generated raw RNA-Seq reads of the suspension cell and plant samples were assembled into two distinct sets of contigs by the sequencing service provider. However, to serve as a basis for expression analysis, an exhaustive reference unigene set, representing the whole *C. roseus* transcriptome, was needed. To inspect the provided contigs and to ensure completeness of the sequence set, we made *de novo* assemblies from a subset of the publicly available data set from the Medicinal Plant Genomics Resource (MPGR) consortium (http://medicinalplantgenomics.msu.edu/), that comprises RNA-Seq data from more than 20 different *C. roseus* tissues and cultures grown in different conditions (Góngora-Castillo et al., 2012), and compared these with our assemblies. This evaluation did not yield any longer transcripts. Furthermore, allelic differences made it more difficult to build a strict unigene and would render future downstream analyses more complex. It was therefore decided to restrict and generate the reference *C. roseus* unigene set by combining and joining both our sets only. This assembly resulted in a unigene set of 31,450, in the majority full-length (FL), transcripts, designated with the prefix Caros, on which we predicted ORFs. Due to indels in the assemblies of some transcripts, more than one ORF was returned for a number of transcripts. For highly relevant *C. roseus* pathways, like the TIA and triterpenoid pathways, as well as their precursor and the MeJA biosynthesis pathways, we edited and curated all sequences that showed indels or frameshifts.

We assigned a functional description to each Caros transcript through a guilt-by-homology approach based on BLASTX analysis (see Materials and Methods). Subsequently, the annotation of the predicted translated ORFs was further refined by using an established orthology-based method (Fabris et al. 2012), in which translated genomic sequences of 17 annotated organisms were used as reference (see Materials and Methods). Thereby, gene functions and gene-reaction associations were transferred from the reference genomes to the query *C. roseus* transcriptome by means of a score-driven semi-automated pipeline (Fabris et al. 2012).

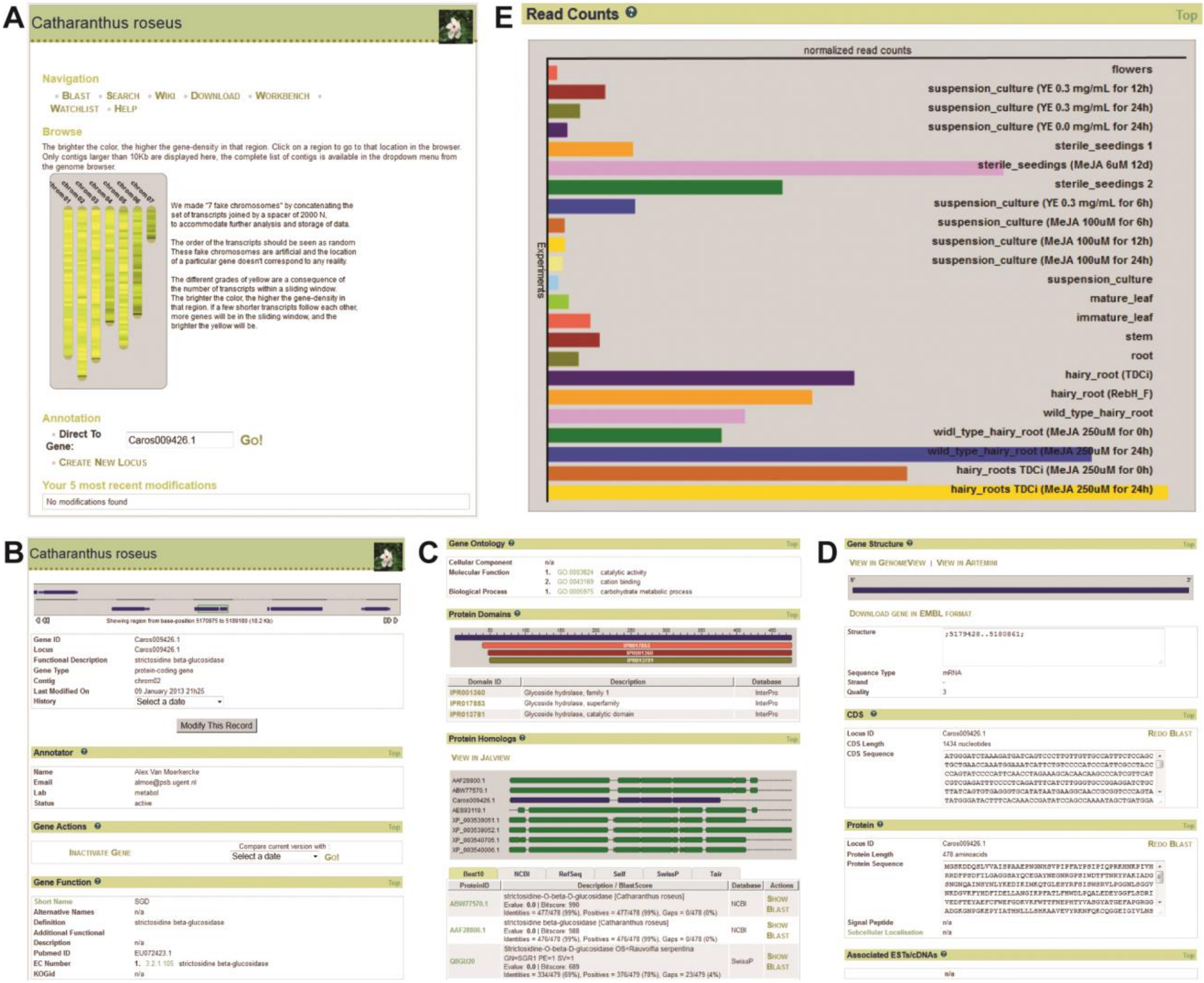
To accommodate analyses and storage in the ORCAE database with web interface (Sterck et al. 2012), we built 'fake chromosomes' by concatenating the set of Caros transcripts joined by a spacer of 2,000 N. This resulted in seven chromosomes, of which the first six contain 5,000 transcripts each. This platform, accessible at http://bioinformatics.psb.ugent.be/orcae/ offers the possibility to edit and curate the predicted Caros ORFs, append functional annotations where needed and display pre-computed analyses such as protein domains, BLAST alignments and expression data (e.g. FPKM values) depicted as bar diagrams. Besides displaying gene related data, blast and search options are at the disposal of the user (see Fig. 8.1 for an example).

To assess the quality of our reference transcriptome, we verified for the presence and completeness of publicly known *C. roseus* sequences. To this end we downloaded the

406 *C. roseus* protein sequences that were publicly available at NCBI (date October 24, 2012) and filtered out the 168 unique FL entries. A TBLASTN search with these 168 protein sequences against our reference transcriptome showed that 164/168 (98%) of the sequences were represented in our assembled transcriptome (i.e. the 'fake chromosomes'). Of the four proteins that were not present in our dataset, only one was present in the MPGR dataset. Verification of the sequences revealed that the FL sequence of 133 of the 164 proteins (81%) was present in our dataset (Supplementary Table S8.3).

Creation and manual curation of CathaCyc v1.0

The information obtained from the orthology-based functional annotation of transcripts was imported into Pathway Tools 16 (Karp et al. 2010, Paley et al. 2012) and used as input for Pathologic (Karp et al. 2010, Paley et al. 2012) for the raw construction of CathaCyc. The resulting draft metabolic network lacked several important pathways common to all plants and, at the same time, included many unnecessary ones, specific to prokaryotes, for instance. Therefore, extensive manual curation was carried out by importing missing pathways from the MetaCyc (Caspi et al. 2010, Caspi et al. 2012), AraCyc (Zhang et al. 2005) and PlantCyc (Zhang et al. 2010) databases and by using the literature as a reference to add missing gene-reaction associations, as well as to remove false-predicted pathways. After this refining procedure, the database still contained 910 pathway gaps, related to reactions, which lacked the correct link to a gene. By using the implemented BLAST-based Pathway Hole Filler algorithm we could find gene connections for 498 of pathway holes (see Materials and Methods). Currently, CathaCyc v1.0 consists of 1,802 reactions organized in 390 pathways, 1,347 enzymes and 1,322 compounds, which is similar to other plant PDBs (see www.plantcyc.org for an overview). CathaCyc v1.0 is accessible at www.cathacyc.org through an intuitive and user-friendly web- interface based on the MetaCyc family format.


A
B
C
D
E
chrom01
chrom02
chrom03
chrom04
chrom05
chrom06
chrom07
Read Counts
Top
normalized read counts
Experiments
flowers
suspension_culture (YE 0.3 mg/mL for 12h)
suspension_culture (YE 0.3 mg/mL for 24h)
suspension_culture (YE 0.0 mg/mL for 24h)
sterile_seedings 1
sterile_seedings (MeJA 6uM 12d)
sterile_seedings 2
suspension_culture (YE 0.3 mg/mL for 6h)
suspension_culture (MeJA 100uM for 6h)
suspension_culture (MeJA 100uM for 12h)
suspension_culture (MeJA 100uM for 24h)
suspension_culture
mature_leaf
immature_leaf
stem
root
hairy_root (TDCi)
hairy_root (RebH_F)
wild_type_hairy_root
widl_type_hairy_root (MeJA 250uM for 0h)
wild_type_hairy_root (MeJA 250uM for 24h)
hairy_roots TDCi (MeJA 250uM for 0h)
hairy_roots TDCi (MeJA 250uM for 24h)
Showing region from base-position 5170975 to 5189180 (18.2 Kb)
50
100
150
200
250
300
350
400
450
IPR017853
IPR001360
IPR013781
AAF28800.1
ABW77570.1
Caros009426.1
AES93119.1
XP_003539051.1
XP_003539052.1
XP_003540705.1
XP_003540006.1
5'
3'

Catharanthus roseus

Navigation

- BLAST - SEARCH - WIKI - DOWNLOAD - WORKBENCH - WATCHLIST - HELP

Browse

The brighter the color, the higher the gene-density in that region. Click on a region to go to that location in the browser. Only contigs larger than 10Kb are displayed here, the complete list of contigs is available in the dropdown menu from the genome browser.

We made "7 fake chromosomes" by concatenating the set of transcripts joined by a spacer of 2000 N, to accommodate further analysis and storage of data.

The order of the transcripts should be seen as random These fake chromosomes are artificial and the location of a particular gene doesn't correspond to any reality.

The different grades of yellow are a consequence of the number of transcripts within a sliding window. The brighter the color, the higher the gene-density in that region. If a few shorter transcripts follow each other, more genes will be in the sliding window, and the brighter the yellow will be.

Annotation

- Direct To Gene: Caros009426.1 Go!
- CREATE NEW LOCUS

Your 5 most recent modifications

No modifications found

Catharanthus roseus

Gene ID	Caros009426.1
Locus	Caros009426.1
Functional Description	strictosidine beta-glucosidase
Gene Type	protein-coding gene
Contig	chrom02
Last Modified On	09 January 2013 21h25
History	Select a date

Modify This Record

Annotator Top

Name	Alex Van Moerkercke
Email	almoe@psb.ugent.nl
Lab	metabol
Status	active

Gene Actions Top

INACTIVATE GENE — Compare current version with : Select a date Go!

Gene Function Top

Short Name	SGD
Alternative Names	n/a
Definition	strictosidine beta-glucosidase
Additional Functional Description	n/a
Pubmed ID	EU072423.1
EC Number	1. 3.2.1.105 strictosidine beta-glucosidase
KOGid	n/a

Gene Ontology Top

Cellular Component	n/a		
Molecular Function	1.	GO:0003824	catalytic activity
	2.	GO:0043169	cation binding
Biological Process	1.	GO:0005975	carbohydrate metabolic process

Protein Domains Top

Domain ID	Description	Database
IPR001360	Glycoside hydrolase, family 1	InterPro
IPR017853	Glycoside hydrolase, superfamily	InterPro
IPR013781	Glycoside hydrolase, catalytic domain	InterPro

Protein Homologs Top

VIEW IN JALVIEW

Best10 | NCBI | RefSeq | Self | SwissP | Tair

ProteinID	Description / BlastScore	Database	Actions
ABW77570.1	strictosidine-O-beta-D-glucosidase [Catharanthus roseus] Evalue: 0.0 | Bitscore: 990 Identities = 477/478 (99%), Positives = 477/478 (99%), Gaps = 0/478 (0%)	NCBI	SHOW BLAST
AAF28800.1	strictosidine beta-glucosidase [Catharanthus roseus] Evalue: 0.0 | Bitscore: 988 Identities = 476/478 (99%), Positives = 476/478 (99%), Gaps = 0/478 (0%)	NCBI	SHOW BLAST
Q8GU20	Strictosidine-O-beta-D-glucosidase OS=Rauvolfia serpentina GN=SGR1 PE=1 SV=1 Evalue: 0.0 | Bitscore: 689 Identities = 334/479 (69%), Positives = 376/479 (78%), Gaps = 23/479 (4%)	SwissP	SHOW BLAST

Gene Structure Top

VIEW IN GENOMEVIEW | VIEW IN ARTEMINI

DOWNLOAD GENE IN EMBL FORMAT

Structure	;5179428..5180861;
Sequence Type	mRNA
Strand	-
Quality	3

CDS Top

Locus ID	Caros009426.1
CDS Length	1434 nucleotides
CDS Sequence	ATGGGATCTAAAGATGATCAGTCCCTTGTTGTTGCCATTTCTCCAGC TGCTGAACCAAATGGAAATCATTCTGTCCCCATCCCATTCGCCTACC CCAGTATCCCCATTCAACCTAGAAAGCACAACAAGCCCATCGTTCAT CGTCGAGATTTCCCCTCAGATTTCATCTTGGGTGCCGGAGGATCTGC TTATCAGTGTGAGGGTGCATATAATGAAGGCAACCGCGGTCCCAGTA TATGGGATACTTTCACAAACCGATATCCAGCCAAAATAGCTGATGGA

REDO BLAST

Protein Top

Locus ID	Caros009426.1
Protein Length	478 aminoacids
Protein Sequence	MGSKDDQSLVVAISPAAEPNGNHSVPIPFAYPSIPIQPRKHNKPIVH RRDFPSDFILGAGGSAYQCEGAYNEGNRGPSIWDTFTNRYPAKIADG SNGNQAINSYNLYKEDIKIMKQTGLESYRFSISWSRVLPGGNLSGGV NKDGVKFYHDFIDELLANGIKPFATLFHWDLPQALEDEYGGFLSDRI VEDFTEYAEFCFWEFGDKVKFWTTFNEPHTYVASGYATGEFAPGRGG ADGKGNPGKEPYIATHNLLLSHKAAVEVYRKNFQKCQGGEIGIVLNS
Signal Peptide	n/a
Subcellular Localisation	n/a

REDO BLAST

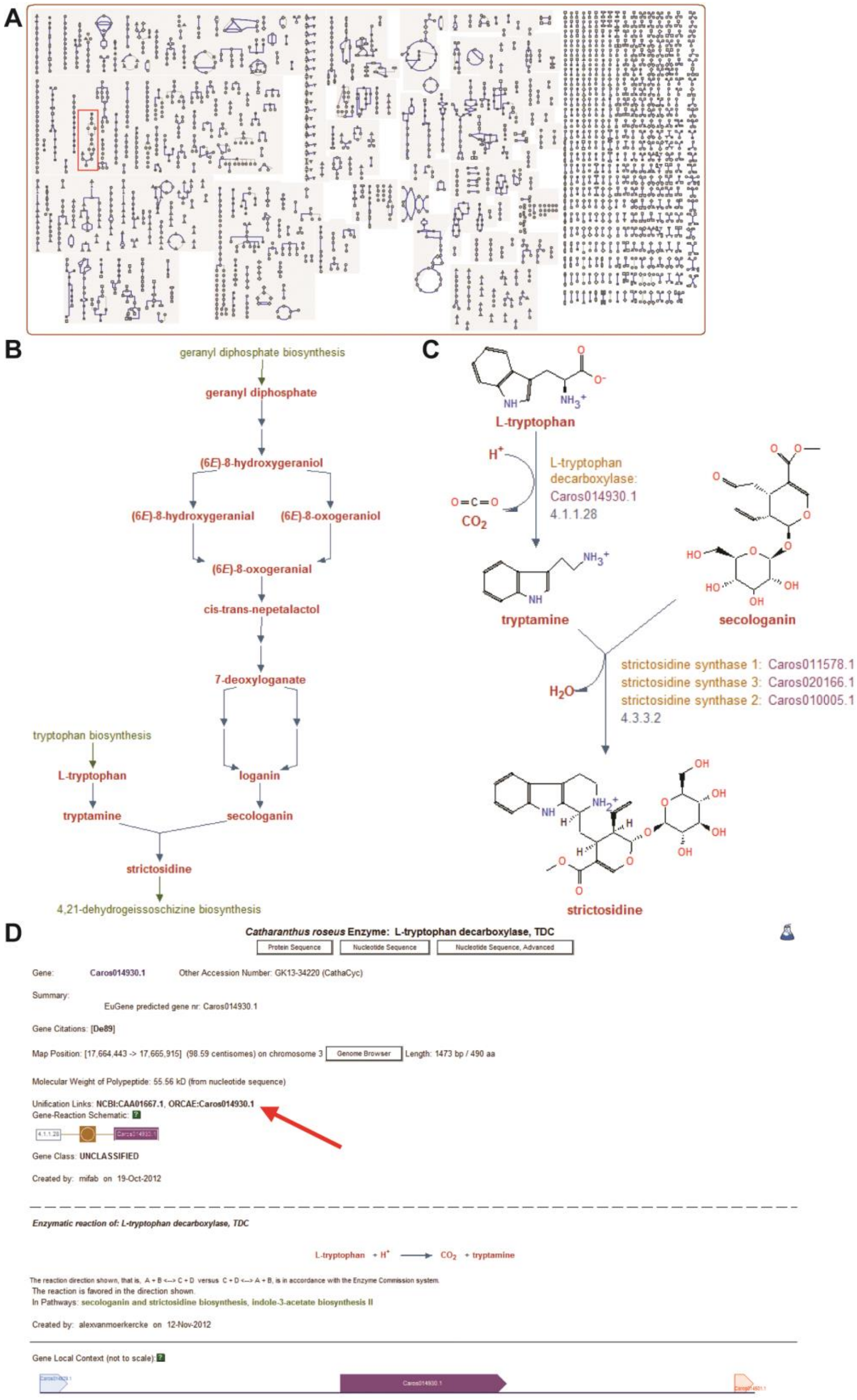
Associated ESTs/cDNAs Top

n/a

Figure 8.1 The *C. roseus* 'fake chromosomes' in the ORCAE database. (A) The ORCAE home page (http://bioinformatics.psb.ugent.be/orcae/) for the *C. roseus* transcriptome with visual representations of the 'fake chromosomes' and a search box for genes. (B) Caros009426.1, corresponding to SGD, has been given as an example. Clicking 'Go!' leads to a gene-specific webpage that displays Gene IDs and functional annotation. (C) Gene Ontologies and homologous genes from NCBI. (D) sequence data. (E) Atlas of expression data from publicly available *C. roseus* RNA-Seq analyses.

Users can explore *C. roseus* metabolism at different levels, by browsing specific information pages that provide graphical overviews of the whole metabolic network, single pathways and single reactions (see Fig. 8.2 for an example). Furthermore, specific pages are available for genes, proteins and metabolites, complete with links to external online databases and to a dedicated gene expression database (http://bioinformatics.psb.ugent.be/orcae/) (see Fig. 8.1 for an example). By using the set of tools provided by the website, users can query CathaCyc through keywords and searches by specific parameters. Furthermore, CathaCyc allows comparative analysis between external PDBs and the upload and graphical visualisation of high-throughput experimental data.

Figure 8.2 ***Next page***. Screenshots from CathaCyc illustrating the different levels. (A) Cellular overview of the complete metabolism of *C. roseus*. The 'Secologanin and strictosidine biosynthesis' pathway, shown in more detail on the information page (B) is boxed in red. (B) and (C) Parts of the 'Secologanin and strictosidine biosynthesis' pathway, from the least to the most detailed view. The latter provides details of the *Caros* genes and proteins associated with the corresponding reactions. (D) Selection and visualization of a single reaction on a gene information page that, in turn, includes links to external databases and literature references, and information relative to gene length and protein size. The genomic localization of the protein is graphically represented and the genomic coordinates are indicated. The link to ORCAE is indicated by a red arrow.



***Catharanthus roseus* Enzyme: L-tryptophan decarboxylase, TDC**

Gene: **Caros014930.1** Other Accession Number: GK13-34220 (CathaCyc)

Summary:

EuGene predicted gene nr: Caros014930.1

Gene Citations: **[De89]**

Map Position: [17,664,443 -> 17,665,915] (98.59 centisomes) on chromosome 3 Length: 1473 bp / 490 aa

Molecular Weight of Polypeptide: 55.56 kD (from nucleotide sequence)

Unification Links: **NCBI:CAA01667.1**, **ORCAE:Caros014930.1**

Gene-Reaction Schematic: ?

Gene Class: **UNCLASSIFIED**

Created by: mifab on 19-Oct-2012

Enzymatic reaction of: L-tryptophan decarboxylase, TDC

L-tryptophan + H^+ ⟶ CO_2 + tryptamine

The reaction direction shown, that is, A + B <--> C + D versus C + D <--> A + B, is in accordance with the Enzyme Commission system.

The reaction is favored in the direction shown.

In Pathways: **secologanin and strictosidine biosynthesis, indole-3-acetate biosynthesis II**

Created by: alexvanmoerkercke on 12-Nov-2012

Gene Local Context (not to scale): ?

Figure 8.2. ***Description on the previous page.***

Manually curated *C. roseus* pathways

CathaCyc significantly covers the central metabolism of *C. roseus*, however, here we focused particularly on its secondary metabolism. We extensively curated the reconstruction of biochemical pathways that are relevant for the biotechnological and pharmaceutical exploitation of this species, i.e. the biosynthesis of TIAs and their precursors, as well as of the triterpenoids, another reasonably well studied class of *C. roseus* metabolites.

The pathways that produce the precursors for TIA biosynthesis, the chorismate, 2-*C*-methyl-D-erythritol 4-phosphate (MEP), mevalonate (MVA), indole and tryptamine pathways (El-Sayed and Verpoorte 2007) have been well described in multiple organisms and are consequently present in PDBs like MetaCyc, PlantCyc and AraCyc. Our RNA-Seq data contained all transcripts of the above-mentioned pathways and all enzymes they encode have therefore been included in CathaCyc v1.0. For 1-deoxy-D-xylulose-5-phosphate synthase (DXS), acetoacetyl-CoA thiolase (AACT), anthranilate phosphoribosyltransferase and D-3-phosphoglycerate dehydrogenase, multiple Caros gene copies have been annotated to the corresponding reactions.

The monoterpenoid compound secologanin is produced from geranyl pyrophosphate (GPP) involving at least eight enzymatic steps, of which five have been characterized in *C. roseus* (El-Sayed and Verpoorte 2007, Geu-Flores et al. 2012, Simkin et al., 2013, Verma et al. 2012) (Fig. 8.2B). Since only three were included in MetaCyc or PlantCyc, we have added the recently discovered genes encoding geraniol synthase (GES) (Simkin et al., 2013) and iridoid synthase (IS) (Geu-Flores et al. 2012) in CathaCyc v1.0 (Fig. 8.3A-B). Secologanin is further condensed with tryptamine by strictosidine synthase (STR) resulting in the production of strictosidine (Fig. 8.2C), which is further metabolized by strictosidine beta-glucosidase (SGD) (Geerlings et al. 2000) to yield strictosidine aglycone, the precursor of all TIAs in *C. roseus* (Verma et al., 2012). For the vindoline branch of the TIA pathway, which starts from tabersonine, one transcript, corresponding to the putative enzyme catalysing the conversion of 16-methoxytabersonine to3-hydroxy-16-methoxy 2,3-dihydrotabersonine was not

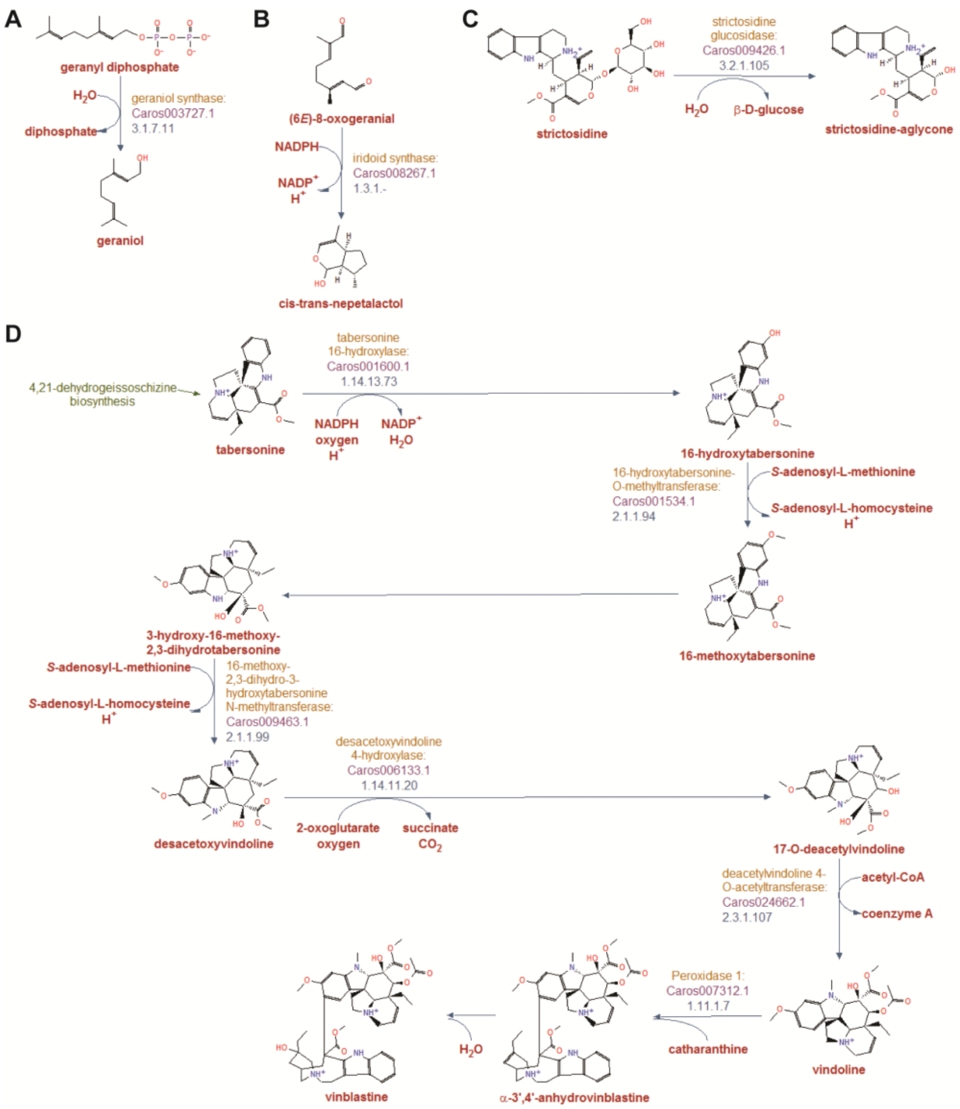


Figure 8.3 TIA pathways presented in CathaCyc. (A-C) Reactions from the 'Secologanin and strictosidine biosynthesis' pathway: (A) geraniol synthase, (B) iridoid synthase, (C) strictosidine glucosidase. (D) Vindoline and vinblastine biosynthesis.

included in CathaCyc v1.0, since the pathway has not been further characterised at the gene level so far. Likewise, the conversion of α-3'-4'-anhydrovinblastine to vinblastine is not included in CathaCyc. Contrary, Peroxidase 1, which catalyses the condensation of catharanthine with vindoline (Costa et al. 2008) and is not included in MetaCyc, is presented in CathaCyc (Fig. 8.3D).

The pentacyclic triterpenoids ursolic acid and oleanolic acid have been found in considerable amounts in the cuticular wax layer of *C. roseus* leaves (Murata et al. 2008). Synthesis of these triterpenoids depends on the same precursor pathways as sterol synthesis, i.e. the MVA pathway, which delivers the IPP necessary for synthesis of 2,3-oxidosqualene, the common substrate for the oxidosqualene cyclases that catalyse the first committed steps towards the different branching triterpenoid pathways (Pollier et al. 2011). As for the TIA precursor pathways, all of these triterpenoid precursor pathways have been well described and included in the PDBs. Similarly, our RNA-Seq data contained all the corresponding transcripts and all enzymes they encode have therefore been included in CathaCyc v1.0. Recently the genes encoding the amyrin synthase (AAS) and the amyrin C-28 oxidase (CYP716AL1/AO) have been isolated (Huang et al. 2012, Yu et al. 2012). AAS was characterized as a novel multifunctional oxidosqualene cyclase producing α- and β-amyrin, whereas AO was a multifunctional C-28 oxidase converting α-amyrin and β-amyrin to ursolic and oleanolic acid, respectively. Both genes are present in our RNA-Seq collection and have been added to CathaCyc v1.0 (Fig. 8.4).

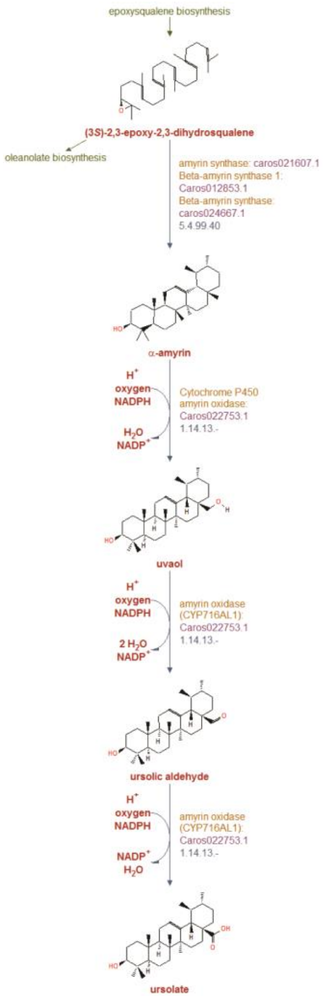


Figure 8.4 The ursolate pathway presented in CathaCyc.

The JA pathway

Furthermore, we curated some pathways involved in the metabolism of plant hormones, in particular in the biosynthesis of JAs, one of the main drivers of TIA synthesis in *C. roseus* and plant secondary metabolism in general (De Geyter et al. 2012, Rischer et al. 2006). The substrate for the synthesis of JAs is α-linolenic acid, which is released by lipase activity on chloroplast membranes. JAs comprise jasmonic acid (JA), methyl jasmonate (MeJA), JA amino acid conjugates, such as (+)-7-iso-jasmonoyl-L-Ile (JA-Ile) that is now accepted as the endogenous bioactive JA, and further JA metabolites (Pauwels and Goossens 2011, Wasternack 2007). The corresponding biosynthetic pathways have been thoroughly characterised in *Arabidopsis thaliana* mainly. For all known Arabidopsis JA and JA-Ile synthesis genes, the corresponding ortholog or homologs were detected in the CathaCyc database (Fig. 8.5A-B). We also encountered two homologs of the gene encoding S-adenosyl-L-methionine:jasmonic acid carboxyl methyltransferase (JMT), that catalyses the conversion of JA to MeJA. Since the corresponding pathway was missing from the MetaCyc databases, we created it in CathaCyc (Fig. 8.5C). No candidate ortholog for the *Arabidopsis* gene shown to be involved in hydroxyjasmonate sulfate biosynthesis was found in the *C. roseus* transcriptome.

Overall, the level of redundancy in the JA synthesis genes from *C. roseus* was similar to that of Arabidopsis. For instance, for allene oxide synthase (AOS), 3-oxo-2(2'-[Z]-pentenyl)cyclopentane-1-octanoic acid CoA ligase (OPCL1) and the JA-Ile synthetase JASMONATE RESISTANT 1 (JAR1) only one hit was encountered in the CathaCyc database, whereas for all other JA synthesis genes two or more copies were detected (Fig. 8.5), as is also the case in the Arabidopsis genome.

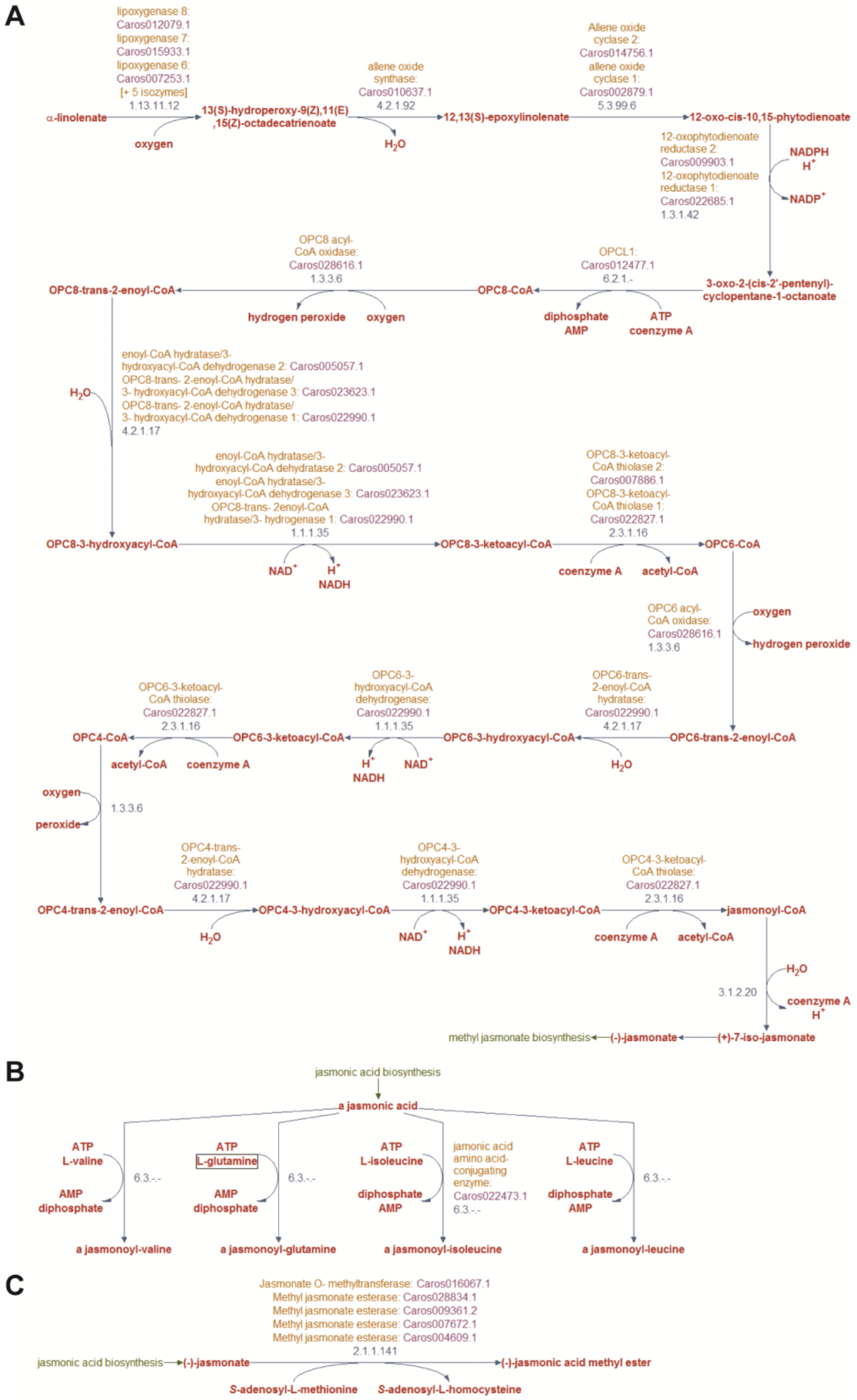

A
lipoxygenase 8: Caros012079.1
lipoxygenase 7: Caros015933.1
lipoxygenase 6: Caros007253.1
[+ 5 isozymes]
1.13.11.12
α-linolenate
oxygen
13(S)-hydroperoxy-9(Z),11(E),15(Z)-octadecatrienoate
allene oxide synthase: Caros010637.1
4.2.1.92
H_2O
12,13(S)-epoxylinolenate
Allene oxide cyclase 2: Caros014756.1
allene oxide cyclase 1: Caros002879.1
5.3.99.6
12-oxo-cis-10,15-phytodienoate
12-oxophytodienoate reductase 2: Caros009903.1
12-oxophytodienoate reductase 1: Caros022685.1
1.3.1.42
NADPH
H^+
$NADP^+$
3-oxo-2-(cis-2'-pentenyl)-cyclopentane-1-octanoate
OPCL1: Caros012477.1
6.2.1.-
diphosphate
AMP
ATP
coenzyme A
OPC8-CoA
OPC8 acyl-CoA oxidase: Caros028616.1
1.3.3.6
hydrogen peroxide
oxygen
OPC8-trans-2-enoyl-CoA
H_2O
enoyl-CoA hydratase/3-hydroxyacyl-CoA dehydrogenase 2: Caros005057.1
OPC8-trans- 2-enoyl-CoA hydratase/3- hydroxyacyl-CoA dehydrogenase 3: Caros023623.1
OPC8-trans- 2-enoyl-CoA hydratase/3- hydroxyacyl-CoA dehydrogenase 1: Caros022990.1
4.2.1.17
OPC8-3-hydroxyacyl-CoA
enoyl-CoA hydratase/3-hydroxyacyl-CoA dehydratase 2: Caros005057.1
enoyl-CoA hydratase/3-hydroxyacyl-CoA dehydrogenase 3: Caros023623.1
OPC8-trans- 2enoyl-CoA hydratase/3- hydrogenase 1: Caros022990.1
1.1.1.35
NAD^+
H^+
NADH
OPC8-3-ketoacyl-CoA
OPC8-3-ketoacyl-CoA thiolase 2: Caros007886.1
OPC8-3-ketoacyl-CoA thiolase 1: Caros022827.1
2.3.1.16
coenzyme A
acetyl-CoA
OPC6-CoA
OPC6 acyl-CoA oxidase: Caros028616.1
1.3.3.6
oxygen
hydrogen peroxide
OPC6-trans-2-enoyl-CoA
OPC6-trans-2-enoyl-CoA hydratase: Caros022990.1
4.2.1.17
H_2O
OPC6-3-hydroxyacyl-CoA
OPC6-3-hydroxyacyl-CoA dehydrogenase: Caros022990.1
1.1.1.35
H^+
NADH
NAD^+
OPC6-3-ketoacyl-CoA
OPC6-3-ketoacyl-CoA thiolase: Caros022827.1
2.3.1.16
acetyl-CoA
coenzyme A
OPC4-CoA
oxygen
peroxide
1.3.3.6
OPC4-trans-2-enoyl-CoA
OPC4-trans-2-enoyl-CoA hydratase: Caros022990.1
4.2.1.17
H_2O
OPC4-3-hydroxyacyl-CoA
OPC4-3-hydroxyacyl-CoA dehydrogenase: Caros022990.1
1.1.1.35
NAD^+
H^+
NADH
OPC4-3-ketoacyl-CoA
OPC4-3-ketoacyl-CoA thiolase: Caros022827.1
2.3.1.16
coenzyme A
acetyl-CoA
jasmonoyl-CoA
H_2O
3.1.2.20
coenzyme A
H^+
(+)-7-iso-jasmonate
(-)-jasmonate
methyl jasmonate biosynthesis
B
jasmonic acid biosynthesis
a jasmonic acid
ATP
L-valine
6.3.-.-
AMP
diphosphate
a jasmonoyl-valine
ATP
L-glutamine
6.3.-.-
AMP
diphosphate
a jasmonoyl-glutamine
ATP
L-isoleucine
diphosphate
AMP
jamonic acid amino acid-conjugating enzyme: Caros022473.1
6.3.-.-
a jasmonoyl-isoleucine
ATP
L-leucine
diphosphate
AMP
6.3.-.-
a jasmonoyl-leucine
C
Jasmonate O- methyltransferase: Caros016067.1
Methyl jasmonate esterase: Caros028834.1
Methyl jasmonate esterase: Caros009361.2
Methyl jasmonate esterase: Caros007672.1
Methyl jasmonate esterase: Caros004609.1
2.1.1.141
jasmonic acid biosynthesis
(-)-jasmonate
(-)-jasmonic acid methyl ester
S-adenosyl-L-methionine
S-adenosyl-L-homocysteine

Figure 8.5. Jasmonate synthesis pathways presented in CathaCyc. (A) JA pathway. (B) JA-Ile pathway. (C) MeJA pathway.

The *C. roseus* RNA-Seq atlas

Several *C. roseus* RNA-Seq libraries are already publicly available, for instance from the Medicinal Plant Genomics Resource (MPGR) consortium (http://medicinalplantgenomics.msu.edu/). However, to make full use of the information they contain, researchers have to download and process the vast amounts of data, which is currently beyond the capacity of many labs. Therefore, we created a *C. roseus* RNA-Seq atlas that archives all publicly available *C. roseus* expression data derived from RNA-Seq experiments. To generate the *C. roseus* RNA-Seq atlas, the MPGR reads were mapped to the artificial genome and counted. To allow the comparison of the expression of the genes in the different samples, the counts were normalized by transcript length and the total number of fragments to "fragments per kilobase of transcript per million fragments mapped" (FPKM) values (Trapnell et al. 2010). As a part of CathaCyc/ORCAE, the *C. roseus* RNA-Seq atlas allows retrieving sequences of genes of interest and visualising the expression pattern of the genes in the different experimental conditions. Currently, the *C. roseus* RNA-Seq atlas holds the expression data from the MPGR consortium that comprises gene expression profiles of different *C. roseus* plant organs and plant, suspension cell and hairy root cultures grown under different conditions and totals 23 different samples (Fig. 8.1E).

Triterpenoid and TIA biosynthesis is differentially regulated in *C. roseus* tissues

The *C. roseus* RNA-Seq atlas was used to analyse the expression of all CathaCyc genes from the curated TIA and triterpenoid pathways, as well as of their precursor and elicitor pathways. Average linkage hierarchical cluster analysis was performed with the full sample set from the MPGR collection or with the hairy root samples only (Fig. 8.6).

Most of the genes were expressed in all explant types except for deacetylvindoline-4-O-acetyltransferase (*DAT*), minovincinine-19-O-acetyltransferase (*MAT*), *AO*, *JMT1* and *JMT2* that were not expressed in non-elicited suspension cells (Magnotta et al. 2007), and tabersonine 16-hydroxylase (*T16H*) that was not expressed in non-elicited hairy roots (Schröder et al. 1999), respectively.

When using the full MPGR sample set, distinct regulation between the TIA and triterpenoid genes was apparent. Genes from both pathways clustered together in a group with lower expression in suspension cells. However, within this group, most TIA genes assembled together in a subcluster with relatively high expression in hairy root cultures (Fig. 8.6A), whereas nearly all triterpenoid genes, including *AAS*, *AO*, all those of the oxidosqualene precursor pathway and most of those of the MVA pathway, assembled in a subcluster that showed higher expression in seedlings and/or other whole-plant organs (Fig. 8.6B). The latter is in agreement with the reported accumulation profile of triterpenoid compounds *in planta* (Murata et al. 2008).

Assessing the expression within the hairy root sample set only, revealed another marked differential regulation between both pathways. In agreement with the existing literature, expression of all TIA pathway genes was MeJA inducible (Fig. 8.6C). The only exception was the N-methyltransferase *NMT* (Liscombe et al. 2010), which was repressed following MeJA treatment. Conversely, the expression of *AAS*, *AO* and some genes encoding rate-limiting enzymes from the triterpenoid precursor pathways, such as 3-hydroxy-3-methylglutaryl-Coa reductase (*HMGR*) and squalene epoxidase (*SQE*), was repressed in hairy roots by MeJA treatment (Fig. 8.6D). The latter observation is remarkable considering that triterpenoid biosynthesis has been reported to be JA-inducible in many different plant species (Pauwels et al. 2009, Yendo et al. 2010).

Interestingly, BLAST searches of the CathaCyc sequences disclosed the presence of clear homologs of two of the *Rauvolfia serpentina* genes that encode enzymes catalysing steps in the conversion of strictosidine to vinorine, i.e. Caros023628.1 for vinorine synthase (VS; Bayer et al. 2004) and Caros017545.1 for polyneuridine aldehyde esterase (PNAE; Dogru et al. 2000), respectively. *C. roseus* is not known to produce vinorine or derivatives thereof but expression analysis shows coregulation of both the *VS* and *PNAE* homolog with the known TIA genes (Fig. 8.6A, C), suggesting that both genes might be involved in TIA synthesis in *C. roseus* as well.

Discussion

C. roseus is the single biological source of the anti-cancer compounds vinblastine and vincristine (El-Sayed and Verpoorte 2007). Efforts to increase the production of these compounds have been moderately successful, partly because a significant number of steps of the pathways that produce these compounds have remained uncharacterised in *C. roseus*. The identification of missing steps can be simplified using a metabolic database of *C. roseus* that catalogues its metabolic potential based on annotated sequence data. In addition, metabolic databases allow for the visualisation and interpretation of large-scale -omics data. An increasing number of plant metabolic databases is being constructed (Fabris et al. 2012, May et al. 2009, Mazourek et al. 2009, Urbanczyk-Wochniak and Sumner 2007, Zhang et al. 2010). These species-specific databases are generally created from reference databases like MetaCyc, which contains many non-plant-specific pathways, and therefore require extensive literature-based curation (Zhang et al. 2010). In addition, highly significant pathways from medicinal plants are mostly absent or only partially included in these reference databases. For this reason, there is a need to create metabolic databases of medicinal plants, which are characterised by their highly specialised metabolism.

The genomes of many pharmaceutically important medicinal plants like *C. roseus* have not been sequenced. The use of transcriptomes rather than genomes to reconstruct pathway databases can circumvent the necessity for fully assembled and annotated genomes. By doing so, PDBs reconstructed from large EST collections have been built for a limited number of plant species such as *M. truncatula* (Urbanczyk-Wochniak and Sumner 2007) and some Solanaceae (SolCyc; SGN). To our knowledge, the use of RNA-Seq data for the creation of plant pathway databases has not been demonstrated.

We assembled and annotated reads of an RNA-Seq experiment of the *C. roseus* transcriptome spanning different plant and cell material. With this information we created CathaCyc v1.0, which significantly covers the central metabolism of *C. roseus*. Here, however, we focused particularly on its secondary metabolism. We extensively curated the reconstruction of biochemical pathways that are relevant for the biotechnological and pharmaceutical exploitation of this species, i.e. the biosynthesis of TIAs and their precursors, as well as of oleanane- and ursane-type triterpenoids. Furthermore, we also concentrated on the reconstruction of pathways involved in the metabolism of plant hormones, in particular in the biosynthesis of JAs, one of the main drivers of TIA synthesis in *C. roseus* and plant secondary metabolism in general (De Geyter et al. 2012). Starting from the transcriptome data, we were able to identify transcripts for all known steps in these pathways, illustrating the quality of the RNA-Seq data set, as well as its utility to reconstruct a metabolic PDB.

As with most metabolic network reconstructions, CathaCyc v1.0 still contains some pathway holes. Many of these holes might correspond to reactions with an incomplete or missing EC number, which is due to the absence of sufficient knowledge to accurately link them to a specific enzyme, or to reactions for which the predicted candidate genes show a confidence score that was lower than the minimal threshold. Some pathway holes, however, might also originate from the discrepancy between the pathway templates stored in MetaCyc that are often inferred from model organisms, and the real pathway organisation that occurs in *C. roseus*, which can be specifically re-arranged or slightly modified. Next-generation sequencing technologies demand increasing efforts to assemble and annotate sequencing projects. The enzymes annotated in CathaCyc v1.0 are linked to the online genome annotation web interface ORCAE (Sterck et al. 2012), which allows viewing and editing of the initial automatic predictions. Users are thus able to contribute to the completeness and accuracy of the annotation. In addition, it enables the integration and visualisation of large-scale omics data. As an example of the latter, we have mapped the publicly accessible RNA-Seq data of the MPGR consortium, spanning 23 different tissues and treatments (Góngora-Castillo et al., 2012), to our annotated *C. roseus* transcriptome. By doing so, expression of the genes of individual pathways can be viewed in the different conditions. Analysis of the expression of the genes from the curated CathaCyc v1.0 pathways demonstrated that the two well-characterised *C. roseus* terpenoid pathways, i.e. of TIAs and triterpenoids, are subject to distinct regulation by both developmental and environmental cues. Furthermore, BLAST searches of the CathaCyc sequences identified clear homologs of *R. serpentina* genes that encode enzymes catalysing steps in the conversion of strictosidine to vinorine and that show coregulation with the known *C. roseus* TIA genes. Since *C. roseus* is not known to produce vinorine or derivatives thereof, these genes might correspond to missing steps in the TIA pathway. These lead findings warrant further exploitation and underscore the value of CathaCyc for the study and exploitation of the metabolism of the Madagascar periwinkle.

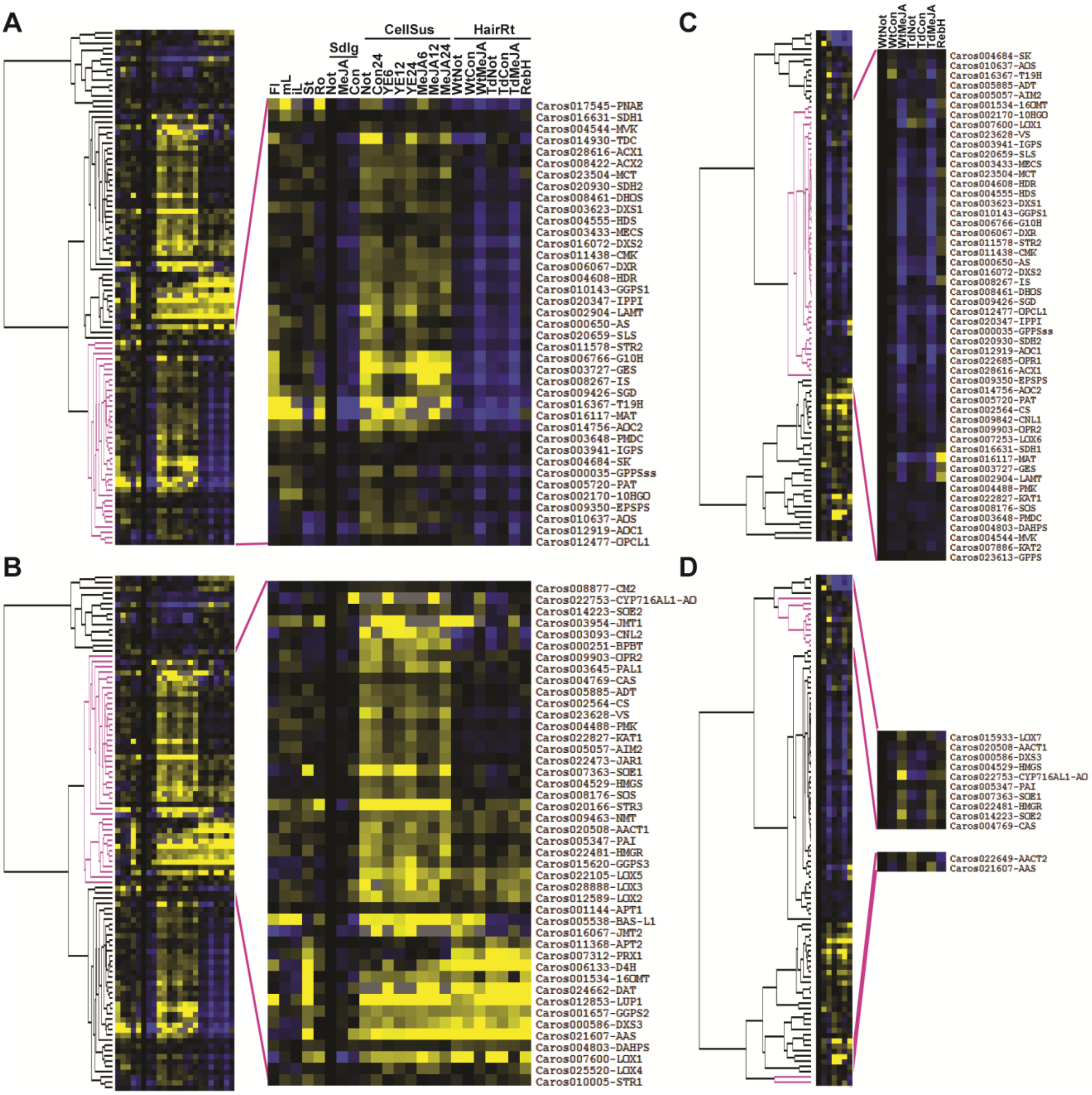


Figure 8.6 (*continues on the next page*) Average linkage hierarchical clustering of secondary metabolism related transcriptomes from *C. roseus*. (A) Clustering of the TIA genes based on all MPGR expression data. (B) Clustering of the triterpenoid genes based on all MPGR expression data. (C) Clustering of the TIA genes based on hairy root expression data. (D) Clustering of the triterpenoid genes

Figure 8.6 (*continued from previous page*)

based on hairy root expression data. Each panel shows the full cluster and a subcluster at the left and right, respectively. Blue and yellow boxes reflect transcriptional activation and repression, respectively, relative to the expression level of the non-treated seedlings (Not, for panels A and B) and non-treated wildtype hairy roots (WtNot, for panels C and D). Gray boxes correspond to no expression for a particular gene in a particular explant. Average linkage hierarchical cluster analysis was performed with the CLUSTER and TREEVIEW software (Eisen et al, 1998) and using the Log10 transformed values of the normalised FPKM values as input for CLUSTER. Caros numbers and enzyme names are indicated at the right (for enzyme abbreviations, see Supplementary Table S8.4). The explants and/or treatments and the time points (in hours) are indicated at the top. Abbreviations: Fl, flower; mL, mature leaves; iL, immature leaves; St, stem; Ro, root; Sdlg, seedling; CellSus, suspension cells; HairRt, hairy roots; Not, non-treated; MeJA, MeJA elicited; Con, mock-treated; YE, yeast-elicitor treated; Wt, wildtype; Td, TDCi hairy roots; RebH, RebH_F hairy roots (see http://medicinalplantgenomics.msu.edu/ for sample details).

Material and Methods

Illumina HiSeq2000 RNA sequencing

C. roseus cell lines were treated with 10 µM MeJA (Bedoukian Research Inc.) for 24 hours. Control treatments were with the solvent DMSO at 0.1% final concentration. Cell line and RNA extraction methods were described in Pauw et al. (2004).

C. roseus plants were germinated from seeds originally received from the Botanical Garden Wuerzburg (Germany). One clone was maintained and vegetatively propagated in a growth chamber in pots containing a substrate mixture (soil (Karkea Ruukutusseos, Kekkila, Finland):Vermiculite 50:50) at 27°C, 16 h photoperiod, 110 µmol $s^{-1}m^{-2}$ and 66.4 % humidity. For the elicitation experiment, shoots (ca. 2 g fresh weight) from flowering plants were cut with sterile scissors and individually placed in cups containing 20 ml of water. The shoots were left to recover from the cutting for 10 days in the growth chamber. Evaporated water was replaced during this time. Then the cups with the shoots were individually placed in zip-lock bags. The elicitation was started by adding MeJA solution to a final concentration of 1 mM or an equivalent amount of the solvent DMSO (800 µl) as a control. The bags were tightly closed and incubated at the same conditions for 0, 6 and 24 hours, respectively. Five replicates per treatment were used. Finally, samples were shock-frozen in liquid nitrogen and stored at -20°C until extraction. Total RNA from *C. roseus* shoots was prepared with the RNeasy Mini Kit (Qiagen).

RNA samples were sent to Fasteris Life Sciences SA (Plan-les-Ouates, Switzerland) for mRNA purification, cDNA library construction and Illumina HiSeq2000-based RNA sequencing with the Solexa technology. For both the cell and plant samples, non-normalised and duplex-specific nuclease (DSN) normalised libraries were prepared and sequenced. In brief, processing included DNase treatment, transcript purification, transcripts breaking, double-stranded cDNA synthesis using random primers and RNase H, ends repair, 3' A addition, ligation of adapters, gel purification to isolate fragments of an insert size of 150-250 bp, PCR amplification to generate the DNA Colonies Template Library, and library purification. A subset of the libraries was additionally normalised with the DSN protocol. The Hi-Seq 2000 instrument and the TruSeq™ SBS v5 kit were used for paired-end (2x100bp) and indexed sequencing of the RNA-Seq libraries.

Assembly and annotation of the 'reference transcriptome'

To construct a reference transcriptome for *C. roseus*, the generated raw RNA-Seq reads were assembled *de novo* by the sequencing service provider using VELVET (v1.1.04) (Zerbino and Birney 2008) and its module OASES (v0.1.21) (Schulz et al. 2012) into two sets of contigs, corresponding to the cell and plant samples, respectively. For VELVET, the pair-read insert average size was set at 200 bases with a standard deviation of 10%. No kmer average coverage and coverage threshold for nodes were used because of the high dynamic range of gene expression. Validation mapping indicated that the highest representativity was found for the assembly of hash 87, which was selected for the complete mapping for both sets of contigs.

The *C. roseus* unigene set was created by combining and joining both sets using the CAP3 software with default parameters (Huang and Madan 1999). On this unigene set we predicted open reading frames (ORFs) using the FrameDP (Gouzy et al. 2009) software. Each transcript was assigned a functional description through a guilt-by-homology approach. Therefore, BLASTX was run with the transcript sequences and was preferred to BLASTP with the predicted ORFs as the occasional indels would not affect hits too much for BLASTX. The BLASTX (v2.2.24) parameters were: -evalue 0.001 -num_descriptions 100 -num_alignments 100 -seg yes. The protein blast database was the non-redundant protein database supplemented with Tomato proteins including the human readable descriptions. The BLASTX output was then parsed using perl programs and hits scored according to their bitscore and number of occurring descriptions using comparable choice of words. To accommodate further analysis and storage of data, we made 7 ‘fake chromosomes’ by concatenating the set of transcripts joined by a spacer of 2000 N. These data were all loaded in the database with web interface of the ORCAE (Sterck et al. 2012) platform. The integrated data from ORCAE was extracted and appropriately formatted into genbank files to upload the data in a BioCyc system dedicated to *C. roseus*.

Creation of CathaCyc (orthology prediction and database reconstruction)

Translated annotated genomic sequences were downloaded from PLAZA (http://bioinformatics.psb.ugent.be/plaza/, April 2012; Van Bel et al. (2012)) for *Medicago truncatula, Zea mays*, *Ricinus communis*, *Populus trichocarpa*, *Vitis vinifera*, *Malus domestica, Brachypodium distachyon, Physcomitrella patens*, *Carica papaya, Glycine max* and *Oryza sativa*, or from the KEGG database (http://www.genome.jp/ keg, February 2010; Kanehisa and Goto (2000)) for *Escherichia coli*, *Saccharomyces cerevisiae*, *Homo sapiens*, *Phaeodactylum tricornutum, Chlamydomonas reinhardtii* and *Arabidopsis thaliana*, respectively, and were used as the reference dataset for the orthology prediction. Orthology prediction, semi-automated functional annotation, BioCyc database construction and manual curation have been done as described previously (Fabris et al. 2012). Pathway-gap filling was performed with the Pathway Hole Filler utility of Pathologic (Caspi et al. 2012) in the user-driven mode. Available gene-reaction associations of *C. roseus* were used as a training dataset. We confirmed the correctness of a random group of 20 candidate genes (with a confidence score of 0.1 or higher) by manual BLAST, InterProscan (Hunter et al. 2009) and Inparanoid (O'Brien et al., 2005) searches. Therefore, the minimal score for candidate genes to be automatically selected as Hole Fillers was set to 0.1 and any gene with a lower score was excluded.

Mapping of the transcriptome data

The FASTQ files containing the read sequences and quality scores of the MPGR consortium were extracted from the NCBI Short Read Archive accessions (accession number SRA030483) using the NCBI SRA Toolkit version 2.1.7. Processing of the extracted reads, mapping of the reads on the 'artificial genome' with TOPHAT version 2.0.3 (Trapnell et al. 2009), and counting of the uniquely mapped reads and calculation of the FPKM values with CUFFLINKS version 1.3.0 (Trapnell et al. 2010) were performed using default parameters as described (Pollier et al. 2013).

Supplementary data

Supplementary data are available at PCP online and on the CD attached.

Funding

This work was supported by the European Union Seventh Framework Programme FP7/2007-2013 [222716 – SMARTCELL]; the European Cooperation in Science and Technology Action [FA1006 – PlantEngine]; the Agency for Innovation by Science and Technology in Flanders ['Strategisch Basisonderzoek' 80031 – SUNLIGHT]; the European Molecular Biology Organization [Long-Term Fellowship to A.V.M.] and European Commission support from Marie Curie Actions [EMBOCOFUND2010, GA-2010-267154 to A.V.M].

Acknowledgments

We thank Annick Bleys for help in preparing the manuscript and Mari Lämsä, Robin Vanden Bossche, Thomas Van Parys and Dany Cuyt for technical assistance.

References

Bayer, A., Ma, X. and Stöckigt, J. (2004) Acetyltransfer in natural product biosynthesis - functional cloning and molecular analysis of vinorine synthase. *Bioorg. Med. Chem.* 12: 2787-2795.

Caspi, R., Altman, T., Dale, J.M., Dreher, K., Fulcher, C.A., Gilham, F. et al. (2010) The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res.* 38: D473-479.

Caspi, R., Altman, T., Dreher, K., Fulcher, C.A., Subhraveti, P., Keseler, I.M. et al. (2012) The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res.* 40: D742-753.

Costa, M.M.R., Hilliou, F., Duarte, P., Pereira, L.G., Almeida, I., Leech, M. et al. (2008) Molecular cloning and characterization of a vacuolar class III peroxidase involved in the metabolism of anticancer alkaloids in *Catharanthus roseus*. *Plant Phsyiol.* 146: 403-417.

De Geyter, N., Gholami, A., Goormachtig, S. and Goossens, A. (2012) Transcriptional machineries in jasmonate-elicited plant secondary metabolism. *Trends Plant Sci.* 17: 349-359.

Desgagné-Penix, I., Farrow, S.C., Cram, D., Nowak, J. and Facchini, P.J. (2012) Integration of deep transcript and targeted metabolite profiles for eight cultivars of opium poppy. *Plant Mol. Biol.* 79: 295-313.

Dogru, E., Warzecha, H., Seibel, F., Haebel, S., Lottspeich, F. and Stöckigt, J. (2000) The gene encoding polyneuridine aldehyde esterase of monoterpenoid indole alkaloid biosynthesis in plants is an ortholog of the α/β hydrolase super family. *Eur. J. Biochem.* 267: 1397-1406.

Eisen, M.B., Spellman, P.T., Brown, P.O. and Botstein, D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA* 95: 14863-14868.

El-Sayed, M. and Verpoorte, R. (2007) Catharanthus terpenoid indole alkaloids: biosynthesis and regulation. *Phytochem. Rev.* 6: 277-305.

Fabris, M., Matthijs, M., Rombauts, S., Vyverman, W., Goossens, A. and Baart, G.J.E. (2012) The metabolic blueprint of *Phaeodactylum tricornutum* reveals a eukaryotic Entner-Doudoroff glycolytic pathway. *Plant J.* 70: 1004-1014.

Facchini, P.J. and De Luca, V. (2008) Opium poppy and Madagascar periwinkle: model non-model systems to investigate alkaloid biosynthesis in plants. *Plant J.* 54: 763-784.

Facchini, P.J. and St-Pierre, B. (2005) Synthesis and trafficking of alkaloid biosynthetic enzymes. *Curr. Opin. Plant Biol.* 8: 657-666.

Ferreres, F., Figueiredo, R., Bettencourt, S., Carqueijeiro, I., Oliveira, J., Gil-Izquierdo, A. et al. (2011) Identification of phenolic compounds in isolated vacuoles of the medicinal plant *Catharanthus roseus* and their interaction with vacuolar class III peroxidase: an H_2O_2 affair? *J. Exp. Bot.* 62: 2841-2854.

Geerlings, A., Ibañez, M.M.-L., Memelink, J., van der Heijden, R. and Verpoorte, R. (2000) Molecular cloning and analysis of strictosidine β-D-glucosidase, an enzyme in terpenoid indole alkaloid biosynthesis in *Catharanthus roseus*. *J. Biol. Chem.* 275: 3051-3056.

Geu-Flores, F., Sherden, N.H., Courdavault, V., Burlat, V., Glenn, W.S., Wu, C. et al. (2012) An alternative route to cyclic terpenes by reductive cyclization in iridoid biosynthesis. *Nature* 492: 138-142.

Góngora-Castillo, E., Childs, K.L., Fedewa, G., Hamilton, J.P., Liscombe, D.K., Magallanes-Lundback, M. et al. (2012) Development of transcriptomic resources for interrogating the biosynthesis of monoterpene indole alkaloids in medicinal plant species. *PLoS ONE* 7: e52506.

Gouzy, J., Carrere, S. and Schiex, T. (2009) FrameDP: sensitive peptide detection on noisy matured sequences. *Bioinformatics* 25: 670-671.

Guirimand, G., Guihur, A., Poutrain, P., Héricourt, F., Mahroug, S., St-Pierre, B. et al. (2011) Spatial organization of the vindoline biosynthetic pathway in *Catharanthus roseus*. *J. Plant Physiol.* 168: 549-557.

He, M., Wang, Y., Hua, W., Zhang, Y. and Wang, Z. (2011) *De novo* sequencing of *Hypericum perforatum* transcriptome to identify potential genes involved in the biosynthesis of active metabolites. *PLoS ONE* 7: e42081.

Huang, L., Li, J., Ye, H., Li, C., Wang, H., Liu, B. et al. (2012) Molecular characterization of the pentacyclic triterpenoid biosynthetic pathway in *Catharanthus roseus*. *Planta* 236: 1571-1581.

Huang, X. and Madan, A. (1999) CAP3: A DNA sequence assembly program. *Genome Res.* 9: 868-877.

Hughes, E.H., Hong, S.-B., Gibson, S.I., Shanks, J.V. and San, K.-Y. (2004) Metabolic engineering of the indole pathway in *Catharanthus roseus* hairy roots and increased accumulation of tryptamine and serpentine. *Metab. Eng.* 6: 268-276.

Hunter, S., Apweiler, R., Attwood, T.K., Bairoch, A., Bateman A., Binns D. et al. (2009) InterPro: the integrative protein signature database. *Nucleic Acids Res.* 37: D211-D215.

Jaiswal, P., Ni, J., Yap, I., Ware, D., Spooner, W., Youens-Clark, K. et al. (2006) Gramene: a bird's eye view of cereal genomes. *Nucleic Acids Res.* 34: D717-723.

Kanehisa, M. and Goto, S. (2000) KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* 28: 27-30.

Karp, P.D., Paley, S.M., Krummenacker, M., Latendresse, M., Dale, J.M., Lee, T.J. et al. (2010) Pathway Tools version 13.0: integrated software for pathway/genome informatics and systems biology. *Briefings Bioinf.* 11: 40-79.

Karp, P.D., Riley, M., Paley, S.M., Pellegrini-Toole, A. and Krummenacker, M. (1997) EcoCyc: enyclopedia of *Escherichia coli* genes and metabolism. *Nucleic Acids Res.* 25: 43-50.

Krieger, C.J., Zhang, P., Mueller, L.A., Wang, A., Paley, S., Arnaud, M. et al. (2004) MetaCyc: a multiorganism database of metabolic pathways and enzymes. *Nucleic Acids Res.* 32: D438-442.

Liscombe, D.K., Usera, A.R. and O'Connor, S.E. (2010) Homolog of tocopherol C methyltransferases catalyzes *N* methylation in anticancer alkaloid biosynthesis. *Proc. Natl. Acad. Sci. USA* 107: 18793-18798.

Loyola-Vargas, V.M., Galaz-Ávalos, R.M. and Kú-Cauich, R. (2007) *Catharanthus* biosynthetic enzymes: the road ahead. *Phytochem. Rev.* 6: 307-339.

Magnotta, M., Murata, J., Chen, J. and De Luca, V. (2007) Expression of deacetylvindoline-4-*O*-acetyltransferase in *Catharanthus roseus* hairy roots. *Phytochemistry* 68: 1922-1931.

May, P., Christian, J.-O., Kempa, S. and Walther, D. (2009) ChlamyCyc: an integrative systems biology database and web-portal for *Chlamydomonas reinhardtii*. *BMC Genomics* 10: 209.

Mazourek, M., Pujar, A., Borovsky, Y., Paran, I., Mueller, L. and Jahn, M.M. (2009) A dynamic interface for capsaicinoid systems biology. *Plant Phsyiol.* 150: 1806-1821.

Mueller, L.A., Zhang, P. and Rhee, S.Y. (2003) AraCyc: a biochemical pathway database for Arabidopsis. *Plant Phsyiol.* 132: 453-460.

Murata, J., Roepke, J., Gordon, H. and De Luca, V. (2008) The leaf epidermome of *Catharanthus roseus* reveals its biochemical specialization. *Plant Cell* 20: 524-542.

Mustafa, N.R. and Verpoorte, R. (2007) Phenolic compounds in *Catharanthus roseus*. *Phytochem. Rev.* 6: 243-258.

O'Brien, K.P., Remm, M. and Sonnhammer, E.L.L. (2005) Inparanoid: a comprehensive database of eukaryotic orthologs. *Nucleic Acids Res.* 33: D476-D480.

Paley, S.M., Latendresse, M. and Karp, P.D. (2012) Regulatory network operations in the Pathway Tools software. *BMC Bioinformatics* 13: 243.

Pauw, B., van Duijn, B., Kijne, J.W. and Memelink, J. (2004) Activation of the oxidative burst by yeast elicitor in *Catharanthus roseus* cells occurs independently of the activation of genes involved in alkaloid biosynthesis. *Plant Mol. Biol.* 55: 797-805.

Pauwels, L. and Goossens, A. (2011) The JAZ proteins: a crucial interface in the jasmonate signaling cascade. *Plant Cell* 23: 3089-3100.

Pauwels, L., Inzé, D. and Goossens, A. (2009) Jasmonate-inducible gene: what does it mean? *Trends Plant Sci.* 14: 87-91.

Pollier, J., Moses, T. and Goossens, A. (2011) Combinatorial biosynthesis in plants: a (p)review on its potential and future exploitation. *Nat. Prod. Rep.* 28: 1897-1916.

Pollier, J., Rombauts, S. and Goossens, A. (2013) Analysis of RNA-Seq data with TOPHAT and CUFFLINKS for genome-wide expression analysis of jasmonate-modulated plant transcriptomes. In Jasmonate Signaling: Methods and Protocols. *Methods Mol. Biol.* 1011, in press.

Rischer, H., Orešič, M., Seppänen-Laakso, T., Katajamaa, M., Lammertyn, F., Ardiles-Diaz, W. et al. (2006) Gene-to-metabolite networks for terpenoid indole alkaloid biosynthesis in *Catharanthus roseus* cells. *Proc. Natl. Acad. Sci. USA* 103: 5614-5619.

Romero, P., Wagg, J., Green, M.L., Kaiser, D., Krummenacker, M. and Karp, P.D. (2005) Computational prediction of human metabolic pathways from the complete human genome. *Genome Biol.* 6: R2.

Schröder, G., Unterbusch, E., Kaltenbach, M., Schmidt, J., Strack, D., De Luca, V. et al. (1999) Light-induced cytochrome P450-dependent enzyme in indole alkaloid biosynthesis: tabersonine 16-hydroxylase. *FEBS Lett.* 458: 97-102.

Schulz, M.H., Zerbino, D.R., Vingron, M. and Birney, E. (2012) *Oases*: robust *de novo* RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics* 28: 1086-1092.

Simkin, A.J., Miettinen, K., Claudel, P., Burlat, V., Guirimand, G., Courdavault, V. et al. (2013) Characterization of the plastidial geraniol synthase from Madagascar periwinkle which initiates the monoterpenoid branch of the alkaloid pathway in internal phloem associated parenchyma. *Phytochemistry* 85: 36-43.

St-Pierre, B., Vazquez-Flota, F.A. and De Luca, V. (1999) Multicellular compartmentation of *Catharanthus roseus* alkaloid biosynthesis predicts intercellular translocation of a pathway intermediate. *Plant Cell* 11: 887-900.

Sterck, L., Billiau, K., Abeel, T., Rouzé, P. and Van de Peer, Y. (2012) ORCAE: online resource for community annotation of eukaryotes. *Nat. Methods* 9: 1041.

Trapnell, C., Pachter, L. and Salzberg, S.L. (2009) TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* 25: 1105-1111.

Trapnell, C., Williams, B.A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M.J. et al. (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* 28: 511-515.

Urbanczyk-Wochniak, E. and Sumner, L.W. (2007) MedicCyc: a biochemical pathway database for *Medicago truncatula*. *Bioinformatics* 23: 1418-1423.

Van Bel, M., Proost, S., Wischnitzki, E., Movahedi, S., Scheerlinck, C., Van de Peer, Y. et al. (2012) Dissecting plant genomes with the PLAZA comparative genomics platform. *Plant Phsyiol.* 158: 590-600.

van der Heijden, R., Jacobs, D.I., Snoeijer, W., Hallard, D. and Verpoorte, R. (2004) The Catharanthus alkaloids: pharmacognosy and biotechnology. *Curr. Med. Chem.* 11: 607-628.

Verma, P., Mathur, A.K., Srivastava, A. and Mathur, A. (2012) Emerging trends in research on spatial and temporal organization of terpenoid indole alkaloid pathway in *Catharanthus roseus*: a literature update. *Protoplasma* 249: 255-268.

Wang, Z., Gerstein, M. and Snyder, M. (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.* 10: 57-63.

Wasternack, C. (2007) Jasmonates: an update on biosynthesis, signal transduction and action in plant stress response, growth and development. *Ann. Bot.* 100: 681-697.

Wenping, H., Yuan, Z., Jie, S., Lijun, Z. and Zhezhi, W. (2011) *De novo* transcriptome sequencing in *Salvia miltiorrhiza* to identify genes involved in the biosynthesis of active ingredients. *Genomics* 98: 272-279.

Yendo, A.C.A., de Costa, F., Gosmann, G. and Fett-Neto, A.G. (2010) Production of plant bioactive triterpenoid saponins: elicitation strategies and target genes to improve yields. *Mol. Biotechnol.* 46: 94-104.

Yu, F., Thamm, A.M.K., Reed, D., Villa-Ruano, N., Quesada, A.L., Gloria, E.L. et al. (2012) Functional characterization of amyrin synthase involved in ursolic acid biosynthesis in *Catharanthus roseus* leaf epidermis. *Phytochemistry* in press.

Zerbino, D.R. and Birney, E. (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.* 18: 821-829.

Zhang, P., Dreher, K., Karthikeyan, A., Chi, A., Pujar, A., Caspi, R. et al. (2010) Creation of a genome-wide metabolic pathway database for *Populus trichocarpa* using a new approach for reconstruction and curation of metabolic pathways for plants. *Plant Physiol.* 153: 1479-1491.

Zhang, P., Foerster, H., Tissier, C.P., Mueller, L., Paley, S., Karp, P.D. et al. (2005) MetaCyc and AraCyc. Metabolic pathway databases for plant research. *Plant Phsyiol.* 138: 27-37.

Zhou, M.-L., Shao, J.-R. and Tang, Y.-X. (2009) Production and metabolic engineering of terpenoid indole alkaloids in cell cultures of the medicinal plant *Catharanthus roseus* (L.) G. Don (Madagascar periwinkle). *Biotechnol. Appl. Biochem.* 52: 313-323.

Ziegler, J. and Facchini, P.J. (2008) Alkaloid biosynthesis: metabolism and trafficking. *Annu. Rev. Plant Biol.* 59: 735-769.

Supplementary Data

Supplementary Table S8.1. *C. roseus* RNA-Seq libraries and number of reads from each library

Supplementary Table S8.2. Presence of known TIA genes in the RNA-Seq sets from *C. roseus* suspension cells and seedlings.

Supplementary Table S8.3. Overview of the BLAST results with the *C. roseus* NCBI sequences.

Supplementary Table S8.4. Enzyme abbreviations.

Acknowledgements

Finalizing this work would have been impossible without the precious help of many people, to which I am deeply grateful.

I would like to thank the members of the Evaluation Board for their availability, for their useful suggestions and for the stimulating and fruitful discussion about my work.

A great, sincere special thank goes to Alain and Wim. When I started my PhD, I could not have hoped for better guidance. I feel extremely lucky to have had the opportunity to carry out my PhD in their labs. I enjoyed working with them, both from professional and human perspectives.

I want to thank Wim for his guidance, his support, his suggestions and for the genuine enthusiasm that he showed since the first moments for every single little progress of my work. This enthusiasm lasted even during those long frustrating months without any good result. This was extremely important; it helped me to stay on track and to stay positive. I thank him for his trust, for providing extra funds to cover the last part of my research and for the great time we had in several occasions, at conferences, meetings, drinks and receptions.

I want to thank Alain, who had been a solid referral since the first day I arrived in his group. I learned much from him, both as great scientist and as great person. I thank him for having been always positive, supportive and available, always ready with an alternative solutions to any problem. Many times I walked into his office with frustration and anger but after a few minutes (or hours) I could walk out with a smile on my face. I am very grateful for everything he had done for me during my stay in his group.

I thank Alain also for the amazing team he built. This group has an enormous strength that comes from the cohesion and the feeling of being part of a team, of a group of friends. This is something really precious that I was glad to be part of and I will miss it. Hard working days, often turned into great parties, this is the Metabol-style. Therefore I thank all the Metabollers, current and past members, for such great time, professionally and beyond. I thank Gino, who contributed enormously to this work, both with hard work and with constant support in the lab and during long lunch-beer meetings about science and rock'n'roll. I thank Robin, the Mayor of the group and the solid cornerstone, always ready to help me and currently the best coffee maker of the lab; Amparo, with whom I shared joy and tears of this PhD, who made me feel home for being so Spanish and so Italian; Lander, and our daily "fake" fights; Michiel, the other side of my desk, my cigarette counter, the one and only who could fully understand my frustrations ("the beauty of science") of working with diatoms; Sophie, for the very important help provided and for taking good care of us and our messy corner; Tessa, for the constant patience shown in answering my (too many) questions, always with a smile (even when I would have deserved something else); Jacob for helping me so many times and for the incredible passion and professionalism that he puts in his tasks, which can turn into something dangerous when the task is partying or having "a few beers"; Astrid, always available for a break, for a party or for letting me move her washing-machines; Laurens, always prone in giving good advices and ideas; Rebecca for her smiles every morning and for her help whenever I needed it; Jan, for the funny jokes and for thinking that I could have been a good wingman; Azra, for being always friendly and for sharing Persian music with me in the lab on weekends; Sandra, for being always present in the group, despite the distance; Janine for being so kind and having helped me many times (last time to print this thesis!); Nathan for keeping the metabol reputation high, for loving my grappa so much and for flying into fountains and cupboards; Jonas for keeping Duvel in business, for making me swearing on the squash courts and for running as many OCR's as possible; Philipp for being so German, getting all the jokes about it and for cooking breakfast only at my place, which makes me feel special; Andrès for turning into the Animal that had a major role in my hangovers, especially during summer barbecue's; Sir Alex for having been so noisy, so late and so party-oriented: now I know how postdoc life is. I had such a great time with you guys, I'll miss you all.

I would like to thank Dirk Inzé and VIB management Committee for the additional funds that covered the last part of my PhD; Diane for the kind help provided during all these years; Kristof, Dirk, Nancy and Jackie; the IT group, especially Raf and Tim who fixed my laptop and desktop so many times and Dany, who sacrificed several Friday afternoons updating or fixing DiatomCyc and CathaCyc; Wilson for his fundamental help and for his constant smile; Nino for solving my qPCR problems and for the good time at drinks and parties; Jelle for crashing all our drinks, Stephane for the help provided in many occasions and for the nice collaborations; Marie for taking Michiel and me into the unknown world of diatoms and for all the nice discussions and suggestions; Tommaso for being my first guide in Ghent and in the lab and a good friend; Camilla e Simone for the Italian coffee breaks that made me feel home.

I wish to thank the whole PAE group at de Sterre, even if I did not spend much time there. In particular, I thank Renaat, for his help, always ready even after last-minute calls.

My stay in Ghent did not only include hard work in the lab. Luckily, I found a very nice group of friends that let me play the sport I love. I wish to thank all the players and the administrative board of the Ijshockey Club Eeklo Yeti Bears (aka Yeti Beers), for welcoming me in their team for four great ice-hockey seasons, during which I enjoyed good ice hockey and great time off-ice. I will always carry an indelible sign that will remind me of these guys.

I did not forget about my far away friends. Despite the distance, they have been constantly present in my life, also during my stay here in Belgium. I thank Marino, Massimo, Palmino, Valerio, Marco, Simo, and Davide. I thank who made it here today but also who could not make it. I thank them for the frequent visits in the past years and weeks, and for letting me destroy their plans whenever I showed up for one or two days. Therefore, I wish to thank even more Claudia, Chiara, Karolina, Giulia, Fede e Stefania. Of course, this very important group of friends also includes the tough guys of the Hockey Club Bologna Mummies. In particular I thank Hermann, Ricky, Cavini, Gas, Tommy, Matteo, Giamma, Nicola, Carlo, Claude and Dagheo for the great time on- and off- ice, in Bologna, in Prague and anywhere where ice, sticks, skates and beers were waiting for us. On this concern, I wish to thank the OIB, which, luckily or not, has never left me.

Ringrazio chi mi ha dato la forza e il supporto più importante nel portare a termine questo impegnativo capitolo della mia vita. Un grandissimo ringraziamento va alla mia famiglia: a mamma e papà che mi hanno sempre supportato e aiutato nonostante le mie scelte mi abbiano portato a essere distante da loro; a Lisa che mi ha sempre incoraggiato e spinto a dare il massimo, a crederci. La consapevolezza che continueranno a farlo mi da la forza per affrontare ogni sfida. In particolare voglio ringraziare Lisa e Mattia, perché hanno iniziato la loro famiglia dandomi un nipotino meraviglioso, Leonardo. Inoltre ringrazio i nonni, vicini e lontani, per avermi sempre sostenuto. Ringrazio Ruggero, Enrico e Patrizia, per avermi incoraggiato e per essermi stati così vicini; così come ringrazio Guglielmo, sempre disponibile e pronto ad aiutarmi.

Infine ringrazio Sarah, compagna di vita da sempre, che con amore mi ha affiancato, sostenuto, spronato, sopportato e motivato durante tutta quest'avventura. Non mi ha mai lasciato solo un momento, a volte con grandi sacrifici e difficoltà. E' sempre stata presente e importantissima nei momenti felici e in quelli difficili. E sempre ci sarà, nella splendida avventura della nostra vita. Senza di lei non sarei mai arrivato a questo traguardo, che alla fine è anche un po' suo. Per questo motivo, questo lavoro è dedicato a lei.

Curriculum vitae

MICHELE FABRIS
Verlorenkost 17
9000 Gent, Belgium
(+32) 0471 302986
(+39) 349 1992841
fabrismichele@gmail.com

Birth date: 23/06/1982
Birth place: Bolzano, Italy
Nationality: Italian

LANGUAGES

Italian	Level: mother tongue (written/spoken)
English	Level: proficient (written/spoken)
German	Level: good (written/spoken)

EDUCATION

PhD Biotechnology and Biochemistry, Flanders Institute of Biotechnology (VIB), 2013 Ghent University, Ghent, Belgium
Dissertation: " The reconstruction of *Phaeodactylum tricornutum*'s metabolism: unveiling biochemical peculiarities towards the biotechnological exploitation of diatoms"
Promoters: Prof. Dr. Alain Goossens and Prof. Dr. Wim Vyverman

M. Sc Cellular and Molecular Biology, March 2008, University of Bologna. Bologna, Italy.
Dissertation: "Organogenic proprieties, responses to methyl-jasmonate and polyamine production in *Nicotiana tabacum* overexpressing *AUX/IAA* genes."
Promoter: Prof. Stefania Biondi

B. Sc Biological Sciences, July 2005, University of Bologna, Bologna, Italy.
Research project: "Cross-talk between primary, secondary metabolism and alkaloid biosynthesis: polyamines and methyl-putrescine analysis in transgenic cells and roots of *Nicotiana tabacum* and *Hyoscyamus muticus"*
Supervisor: Prof. Stefania Biondi

RESEARCH EXPERIENCE

PhD student **Feb 2009 – Nov 2013**

Flanders Institute of Biotechnology (VIB), Department of Plant Systems Biology (PSB) Ghent University, Laboratory of Protistology and Aquatic Ecology (PAE), Ghent Belgium.
Labs of Prof. Dr. Alain Goossens and Prof. Dr. Wim Vyverman

Graduate student grant **Mar 2008 - Dec 2008**

University of Bologna, Department of Evolutionary Experimental Biology, Bologna Italy.
Labs of Prof. Stefania Biondi and Prof. Ferruccio Poli

PUBLICATIONS

<u>**Fabris M**</u>, Matthijs M, Rombauts S, Vyverman W, Goossens A, Baart GJE. **The metabolic blueprint of *Phaeodactylum tricornutum* reveals a eukaryotic Entner-Doudoroff glycolytic pathway** *The Plant Journa*l. 2012 Jun;70(6):1004-14. doi: 10.1111/j.1365-313X.2012.04941.x

Van Moerkercke* A, <u>**Fabris* M**</u>, Pollier* J, Baart GJE, Rombauts S, Hasnain G, Rischer H, Memelink J, Oksman-Caldentey KM and Goossens A. **CathaCyc, a Metabolic Pathway Database Built from *Catharanthus roseus* RNA-Seq Data**. *Plant and Cell Physiology.* 2013; 54 (5): 673-685.
doi: 10.1093/pcp/pct039 (* shared co-authorship)

In preparation

<u>**Fabris M**</u>**,** Matthijs M, Carbonelle S, Moses T, Baart GJE, Vyverman W and Goossens A. **The sterol biosynthetic pathway of diatoms shows features of fungi as well as plants.** *Article in preparation*

Websites/Databases

www.diatomcyc.org – First comprehensive database of diatom metabolism
www.cathacyc.org – Metabolic database of the medicinal plant *Catharanthus roseus*

INTERNATIONAL CONFERENCES CONTRIBUTIONS

"How to reconstruct a Pathway/Genome Database" <u>Michele Fabris</u> and Gino J.E. Baart **(Workshop)** August 2012, 22nd International Diatom Symposium. Ghent, Belgium

"Sterol biosynthesis in *Phaeodactylum tricornutum*" **(Oral presentation)** August 2012, 22nd International Diatom Symposium. Ghent, Belgium

"The metabolic blueprint of *Phaeodactylum tricornutum* reveals a eukaryotic Entner-Doudoroff pathway"**(Poster presentation)** June-July 2012, Banff Conference on Plant Metabolism 2012. Banff, Canada. Winner of poster presentation award (Noble Foundation).

"Metabolic blueprint for sustainable production of bioenergy" **(Poster presentation)** March 2011, Specific Light-Driven Reactions in Unicellular Model Algae Meeting. Jena, Germany.

ATTENDED COURSES AND TRAININGS

"Advanced Academic English: Writing Skills" by Catherine Verguts, Ghent University 2012
"Effective scientific communication" by Jean-Luc Dumond, Ghent University 2011
“Tech Transfer Course", by VIB Tech Transfer Team, 2011
“Workshop on Specific Light-Driven Reactions in Unicellular Model Algae“ Friedrich-Schiller-Universität Jena (Germany) 2010
"Perl introductory training” by Guy Bottu, VIB Bioinformatics Training & Service (BITS) 2009
“Introduction to Bioinformatics with BEN” VIB Bioinformatics Training & Service (BITS) 2009

Supervised students

Simoen Jeroen – Master Project (AY 2010-2011)
Huysmans Marlies – Master Project (AY 2011-2012)