Isoprenoid biosynthesis in the diatom Haslea ostrearia

Anastasia Athanasakoglou, Emilia Grypioti, Sofia Michailidou, Codruta Ignea, Antonios M Makris, Kriton Kalantidis, Guillaume Massé, Anagnostis Argiriou, Frederic Verret, Sotirios Kampranis

Isoprenoid biosynthesis in the diatom *Haslea ostrearia*

Anastasia Athanasakoglou[1], Emilia Grypioti[2], Sofia Michailidou[3], Codruta Ignea[1], Antonios M. Makris[3], Kriton Kalantidis[2,4], Guillaume Massé[5,6], Anagnostis Argiriou[3], Frederic Verret[2,*], Sotirios C. Kampranis[1,*]

[1]Department of Plant and Environmental Sciences, University of Copenhagen, Thorvaldsensvej 40, 1871 Frederiksberg C, Denmark [2]Department of Biology, University of Crete, P.O. Box 2208, Heraklion 71003, Greece [3]Institute of Applied Biosciences – Centre for Research and Technology Hellas (INAB-CERTH), P.O. Box 60361, Thermi 57001, Thessaloniki, Greece [4]Institute of Molecular Biology and Biotechnology - Foundation of research and technology Hellas (IMBB-FORTH), Heraklion, Greece [5]UMI 3376 TAKUVIK, Centre national de la recherche scientifique (CNRS), France [6]Département de Biologie, Université Laval, Québec, Canada

*Author for correspondence: Sotirios C. Kampranis, Department of Plant and Environmental Sciences, University of Copenhagen, Thorvaldsensvej 40, 1871 Frederiksberg C, Denmark. Phone: +45 35 33 29 89 Email: soka@plen.ku.dk; Frederic Verret, Department of Biology, University of Crete, Heraklion, Crete, Greece. Email: fverret@imbb.forth.gr.

Summary

- Diatoms are eukaryotic, unicellular algae that are responsible for about 20% of the Earth's primary production. Their dominance and success in contemporary oceans have prompted investigations on their distinctive metabolism and physiology. One metabolic pathway that remains largely unexplored in diatoms is isoprenoid biosynthesis, which is responsible for the production of numerous molecules with unique features.
- We selected the diatom species *Haslea ostrearia* because of its characteristic isoprenoid content and carried out a comprehensive transcriptomic analysis and functional characterization of the genes identified.
- We functionally characterized one farnesyl diphosphate synthase, two geranylgeranyl diphosphate synthases, one short-chain polyprenyl synthase, one bifunctional isopentenyl diphosphate isomerase - squalene synthase and one phytoene synthase. We inferred the phylogenetic origin of these genes and used a combination of functional analysis and subcellular localization predictions to propose their physiological roles.
- Our results provide insight into isoprenoid biosynthesis in *H. ostrearia* and propose a model of the central steps of the pathway. This model will facilitate the study of metabolic pathways of important isoprenoids in diatoms, including carotenoids, sterols and highly branched isoprenoids.

Keywords: Heterokonts, isoprenoids, phytoene synthase, prenyltransferase, squalene synthase

Total word count (excluding summary, references and legends): 6088

-Summary: 176

-Introduction: 683

-Materials and Methods: 918

-Results: 3005

-Discussion: 1482

- Acknowledgements: 84

Number of figures: 4 (all in colour)

Number of tables: 2

No of Supporting Information files: 1 (Methods S1, Table S1-Table S5, Fig. S1-Fig. S8, Notes S1-S2)

Introduction

Diatoms (phylum Heterokontophyta, class Bacillariophyceae) are one of the most diverse and ecologically important groups of phytoplankton. With more than 100,000 species that are widely distributed in aquatic environments, it is estimated that they contribute to about 20% of the global primary production. Consequently, they play central roles in aquatic food webs and in the biogeochemical cycling of nutrients (Nelson *et al.*, 1995; Field *et al.*, 1998; Falkowski, 2002). This profound ecological success has created great interest in distinctive physiological features of diatoms. Even though, several studies have already investigated unique facets of their metabolism (Kroth *et al.*, 2008; Ast *et al.*, 2009; Gruber *et al.*, 2009; Allen *et al.*, 2011; Fabris *et al.*, 2012; Smith *et al.*, 2012; Obata *et al.*, 2013), there are still important gaps in our understanding of key biochemical pathways. One such unexplored area is the biosynthesis of isoprenoids, a large class of metabolites that have vital biological functions in all domains of life (Holstein & Hohl, 2004).

Despite their structural diversity, all isoprenoids are assembled from the same five-carbon atom precursors, isopentenyl diphosphate (IPP) and its isomer dimethylallyl diphosphate (DMAPP). Two distinct biosynthetic routes are responsible for the synthesis of these five-carbon precursors: the mevalonate (MVA) pathway and the methyl-erythritol-phosphate (MEP) pathway. Beyond these early steps, IPP and DMAPP are condensed to form prenyl diphosphate molecules of various lengths that serve as substrates for the synthesis of the different isoprenoid classes (geranyl diphosphate (GPP) for C_{10} monoterpenoids; farnesyl diphosphate (FPP) for C_{15} sesquiterpenoids, C_{30} triterpenoids and sterols; geranylgeranyl diphosphate (GGPP) for C_{20} diterpenoids and C_{40} carotenoids; etc.). This set of condensation reactions comprise the central steps of the pathway and are catalyzed by prenyltransferase-type enzymes. After this central part, synthesis of carotenoids and sterols proceeds through the commitment of FPP and GGPP to squalene and phytoene, respectively, by squalene/phytoene synthase-type enzymes (Fig. S1) (Vranová *et al.*, 2012).

Initial work using feeding experiments with labelled precursors and specific pathway inhibitors has shown that both the MVA and MEP pathways are present in diatoms (Cvejić & Rohmer, 2000; Massé *et al.*, 2004). Following the genomic sequencing of model diatom species *Thalassiosira pseudonana* (Armbrust *et al.*, 2004) and *Phaeodactylum tricornutum* (Bowler *et al.*, 2008), the focus of investigation has moved towards the final steps of the pathway, i.e. the carotenoid biosynthetic branch that provides important light harvesting and photoprotective molecules (Coesel *et al.*, 2008; Dambek *et al.*, 2012; Eilers *et al.*, 2016) and the synthesis of sterols, which serve as membrane structural components (Fabris *et al.*, 2014). An additional branch of isoprenoid biosynthesis that has attracted special interest is that of the highly branched isoprenoids (HBI). HBIs are only synthesized by a limited number of diatom species. They are extensively used as geochemical and paleoenvironmental markers (Massé *et al.*, 2011; Belt & Müller, 2013) and show potential for use as pharmaceuticals and as an alternative form of fuels (Rowland *et al.*, 2001; Ferriols *et al.*, 2015, 2017).

Nevertheless, the central steps of the isoprenoid pathway are less well explored and there is limited understanding on the function, subcellular localization and regulation of prenyltransferases from diatoms. Considering that the synthesis of prenyl diphosphates is a key regulatory step in the pathway that determines the flux towards different branches, a thorough investigation of the central steps will significantly improve our overall understanding of isoprenoid biosynthesis in diatoms. To this end, we selected the diatom species *Haslea ostrearia* for further studies, as this species is also able to synthesize HBIs (Wraige *et al.*, 1999) (Fig. S2), thus providing a more comprehensive system to study isoprenoid biosynthesis. Through transcriptomic analysis, we *in silico* reconstructed the MVA and MEP pathways and confirmed the expression of five putative prenyltransferases and two putative phytoene/squalene synthases. We cloned six of these genes and characterized their function in a heterologous host or *in vitro*. By the combination of phylogenetic analysis, subcellular localization prediction and functional characterization, we proposed a model of the isoprenoid pathway in *H.*

ostrearia. This model will serve as a basis for the elucidation of the biosynthesis of HBIs or other useful isoprenoids from diatoms.

Materials and Methods

Diatom cultures

The *H. ostrearia* clone NCC 153.8 used in this study was a product of heterothallic reproduction between clones NCC 141 and NCC 171, both of which were isolated in 2003 from natural populations of the oyster pond Lainard (La Barre de Monts 46° 53' 33" N - 02° 07' 59" W). Clone NCC 153.8 was kindly provided by Dr. Vona Medeler (University of Nantes) and cultured at the University of Crete Greece, in f/2 medium (Guillard, 1975) at 20 °C under an irradiance of 50 µmol photons m^{-2} s^{-1} (12h:12h light:dark cycle) prior to analysis. The ability of the strain NCC 153.8 to synthesize HBIs was confirmed by extraction and gas chromarography-mass spectrometry GC-MS analysis (Methods S1, Fig. S3).

RNA extraction and transcriptome sequencing

Total RNA was extracted with Spectrum™ Plant Total RNA Kit (Sigma Aldrich, USA) and quantified using the Qubit™ RNA BR Assay Kit (Invitrogen, Carlsbad, CA, USA). Messenger RNA (mRNA) was isolated from total RNA using the NEBNext® Poly(A) mRNA Magnetic Isolation Module (New England Biolabs). Complementary DNA (cDNA) library was constructed using the NEBNext Ultra Directional RNA Library Kit for Illumina (New England Biolabs), according to the manufacturer's instructions. Library quantification was conducted with the KAPA Library Quantification kit for Illumina sequencing platforms (KAPA BIOSYSTEMS, U.S.A.) on a Rotor-Gene Q thermocycler (Qiagen). Sequencing was

performed at the Institute of Applied Biosciences of the Centre for Research and Technology Hellas, on an Illumina NextSeq500 platform (Illumina, USA) using the NextSeq™ 500/550 Mid Output Kit (2 x 150 cycles) (Illumina, USA).

Transcriptome Analysis

The overall bioinformatics strategy included the following steps: (i) Trim and clean-up of the sequencing reads using the trim galore wrapper (https://github.com/FelixKrueger/TrimGalore) with default parameters, except for --length 40 and the --fastqc option, so as to remove adaptors and low-quality sequences; (ii) de novo assembly of the read using the Trinity software suite (Grabherr *et al.*, 2011) with default parameters. All the analysis was implemented on a Linux/based HPC cluster assigning one node with 32 cores and 256 GB RAM.

Gene identification, sequence and phylogenetic analysis

Identification of candidate biosynthetic genes from *H. ostrearia* was based on homology with characterized and annotated homologues from red, green algae and heterokonts retrieved from the Genbank protein database (NCBI & Resource, 2017) and DiatomCyc database (Fabris *et al.*, 2012). For the phylogenetic analysis all diatom homologues were retrieved from the the Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP) (Keeling *et al.*, 2014) database using Blast search. The species names and corresponding MMETSP ID numbers are listed on Table S1. Sequences from other heterokonts, red algae, green algae, land plants, cyanobacteria, bacteria, archaea, fungi and metazoa were retrieved from GenBank protein database (NCBI & Resource, 2017). Sequences were aligned using ClustalW (Larkin *et al.*, 2007) and alignments were manually edited by exclusion of ambiguously aligned regions. Phylogenies were inferred using the maximum likelihood method

(Whelan & Goldman, 2001) and JTT matrix-based model in MEGA version 7 (Kumar *et al.*, 2016). All positions with less than 85% site coverage were eliminated. Branch support was generated using non-parametric bootstrap analysis with 100 replicates. Conserved motifs in the selected sequences were identified using NCBI Conserved Domain Search (Marchler-Bauer *et al.*, 2015). Prediction of subcellular localization was carried out through ASAFind (Gruber *et al.*, 2015) and Hectar (Gschloessl *et al.*, 2008) in combination with SignalP 3.0 (Bendtsen *et al.*, 2004) and TMHMM v. 2.0 (Krogh *et al.*, 2001) for signal peptide and transmembrane domain prediction, respectively.

Gene amplification and cloning

Basic sequence analysis and design of primers (Table S2) was performed using CLC workbench (QIAGEN). Total RNA was extracted using Trizol and cDNA was synthesized using SuperScript III RT (Thermo) and DNAseI (Roche) treatment. Both full length and truncated variants of the selected genes were PCR amplified using Phusion High-Fidelity DNA Polymerase (New England BioLabs) and cDNA as template. PCR products were gel purified, A-overhangs were added using MyTaq Polymerase (BIOLINE) and the products were subsequently cloned into vector pCRII-TOPO using TOPO TA cloning kit (Thermo). After digestion with the appropriate restriction enzymes, the digests were ligated to bacterial and yeast expression vectors. For bacterial expression, pRSET (N-terminal 6xHis tag) or pET102-His (C-terminal 6xHis tag) or pET102-Trx-His (N-terminal TRX and C-terminal 6xHis tag) plasmid backbones, digested with the same restriction enzymes were used. For yeast expression, vectors pUTDH3myc, pWTDHmyc and pHTDH3myc (Ignea *et al.*, 2012) were used. Final constructs were verified by sequencing. Gene expression in bacteria and yeast and protein purification protocols are described in Methods S1.

In vitro enzymatic assays

For the characterization of the prenyltransferases 200 µl enzymatic assays were carried out in glass vials. The reaction mixture contained 10 mM MOPS buffer (pH 7.0), 5 mM $MgCl_2$, 1 mM DTT, 1 mg/ml BSA, 100 µM prenyl diphosphate substrate. The substrates used were: dimethylallyl pyrophosphate (D4287, Sigma-Aldrich), isopentenyl diphosphate (I0503, Sigma-Aldrich), geranyl diphosphate (G6772, Sigma-Aldrich), farnesyl diphosphate (F6892, Sigma-Aldrich), geranylgeranyl diphosphate (G6025, Sigma-Aldrich). The reactions were initiated by addition of 50 ng of purified enzyme. After 16 h incubation at 25 °C, the reactions were terminated by addition of equal volume of 2N HCl in 83% EtOH and after 20 min incubation they were neutralized with 0.14 mL of 10% NaOH. The hydrolyzed diphosphates were extracted three times using 300 µl of hexane. The hexane extracts were concentrated to a final volume of 50 µl and 1 µl of each reaction was used for GC-MS analysis (Methods S1). Individual compounds were identified by comparing their GC retention indices and mass spectra with those of authentic standards.

Results

Transcriptomic analysis and identification of candidate biosynthetic genes

H. ostrearia NCC 153.8 strain was cultured and its ability to produce HBIs was confirmed (Fig. S3) prior to RNA extraction, library construction and sequencing on Illumina platform. After quality filtering and trimming a total of 38631556 pair-end reads were *de novo* assembled into contigs using the Trinity suite (Haas *et al.*, 2013). A total of 45508 contigs were obtained.

We screened the assembled transcriptome to identify genes that putatively encode enzymes of the MVA and MEP pathways. Exploiting the high degree of sequence conservation among genes of isoprenoid biosynthesis, we based our screening on other characterized or annotated algal sequences (See Materials and methods section). A pair-wise alignment between the queries and the contigs

obtained from the transcriptome assembly was performed with the Massblast tool (Veríssimo *et al.*, 2017). Through this analysis, we identified seven contigs corresponding to full-length protein sequences with high similarities to enzymes that catalyze the seven steps of the MEP pathway. We additionally identified six contigs with homology to genes involved in the conversion of acetyl-CoA to IPP, through the MVA route. The last step of the MVA pathway involves the isomerization of IPP to DMAPP by isopentenyl diphosphate isomerase (IDI) (Berthelot *et al.*, 2012). Searching for an IDI, we were only able to identify a contig that corresponds to a fusion of an IDI with a squalene synthase (*Ho*IDI-SQS). Squalene synthases catalyze dimerization of two farnesyl diphosphate molecules to form squalene, the precursor of all sterols (Spanova & Daum, 2011). It was not possible to identify independent transcripts of IDI or SQS in our sequencing data and we concluded that only the fusion of the two genes was expressed under the specific conditions.

We continued our analysis focusing on the central steps, which are catalyzed by prenyltransferase-type enzymes. Mining our assembled transcriptome, we were able to retrieve five contigs with similarity to annotated trans-prenyltransferases (PTS). These included one putative farnesyl pyrophosphate (FPP) synthase, sharing 57% sequence similarity at the amino acid level, with a functionally characterized homologue from *Rhizosolenia setigera* (Ferriols *et al.*, 2015) (from now on referred to as *Ho*PTS1) and four other putative polyprenyl diphosphate synthases (named *Ho*PTS2-5) that exhibited similarities to other diatom and algal prenyltransferases (Fig. 1a, Table 1). We investigated whether this set of the five trans-prenyltransferases is conserved among different diatom species by examining the publicly available diatom transcriptomes in the Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP) database (Keeling *et al.*, 2014). Based on homology searches, *Ho*PTS1, *Ho*PTS2 and *Ho*PTS4 orthologs are present in all diatom species investigated (a total of twenty-six species, representing both centric and pennate diatoms). An *Ho*PTS5 ortholog seems to be missing only in one species (*Skeletonema menzelii*), while a *Ho*PTS3 ortholog seems to be absent from seven diatom

species (Table S3, Fig. S4). Among the measured transcripts, *Ho*PTS5 showed significantly high expression levels (Fig. S5).

Sequence analysis of the identified prenyltranserases, led to identification of conserved polyprenyl synthase domains (Fig. 1b) and conserved motifs. Among them, the two aspartic acid-rich motifs DDxx(xx)D (First Aspartic acid-Rich Motif -FARM; and the Second Aspartic acid-Rich Motif -SARM), found in characterized prenyltransferases from all domains of life. These motifs are involved in the binding of magnesium ions and are essential for prenyltransferase activity. The identification of intact motifs in all five PTSs suggested that these genes likely encoded for active enzymes. Conserved amino acids were also observed at positions 4 and 5 upstream of FARM. It has previously been shown that these residues are involved in the regulation of the product chain length (Tarshis *et al.*, 1996; Wang & Ohnuma, 1999; Liang *et al.*, 2002). In *Ho*PTS1, like in other characterized FPP synthases, this region contains aromatic residues that are bulkier and block further chain elongation (Fig. S6a). By contrast, smaller residues were observed at this region in the remaining four PTSs, suggesting that these likely synthesize longer chain products (Fig. S6b).

The prenyl diphosphates synthesized in the central steps are allocated to different branches of the pathway for the synthesis of final isoprenoid products, with sterol and carotenoid pathways being the main ones. Specific enzymes commit FPP and GGPP to these pathways. As described above, we identified a squalene synthase, N-terminally fused to an IDI (*Ho*IDI-SQS). In the absence of additional SQS transcripts, this bi-functional protein could be involved in committing FPP to the synthesis of sterols. We also identified a single transcript, termed *Ho*PSY, which shares 72% similarity with a characterized phytoene synthase from *P. tricornutum* (Dambek *et al.*, 2012) and could catalyze the first committed step of carotenoid biosynthesis in *H. ostrearia*. Both *Ho*IDI-SQS and *Ho*PSY were found to contain the conserved phytoene/squalene synthase domain (Fig. 1b).

Phylogenetic analysis

Diatoms have a distinctive evolutionary history. Red and green algae, as well as glaucophytes evolved after a primary endosymbiotic event, when a heterotrophic eukaryote engulfed a cyanobacterial cell (McFadden, 2001; Rodríguez-Ezpeleta *et al.*, 2005). Diatoms arose from secondary endosymbiosis, during which a different heterotrophic eukaryote captured a cell of red algal origin (Yoon *et al.*, 2004; Prihoda *et al.*, 2012). After incorporation, the engulfed cell was transformed into the plastid organelle (Bhattacharya *et al.*, 2007), lost its mitochondria and nucleus and genetic information was transferred into the nucleus of the heterotrophic host, in a process termed endosymbiotic gene transfer (EGT) (Timmis *et al.*, 2004). The evolutionary linkage between diatoms and red algae is usually observed in phylogenetic surveys (Armbrust *et al.*, 2004; Bowler *et al.*, 2008; Frommolt *et al.*, 2008). However, such surveys have also revealed a green phylogenetic signal. It has been proposed that this is due to a cryptic endosymbiotic event that involved a green algal cell, which was later replaced by the red algal endosymbiont (Moustafa *et al.*, 2009; Chan *et al.*, 2011). Alternatively, the green related genes could have been acquired through repeated horizontal gene transfer (HGT) events from green algae, early after the evolution of the first diatoms (Oborník & Green, 2005; Dorrell & Smith, 2011). This multisourced genetic background has created unique, chimeric metabolic pathways that combine features from multiple lineages.

To investigate how this is reflected to isoprenoid biosynthesis we inferred the phylogenetic origin for each of the candidate biosynthetic genes. Three different possibilities were distinguished: (1) origin from the secondary heterotrophic host (2) origin from the red algal endosymbiont and (3) origin from a green algal cell, either through EGT or HGT.

It is evident that the MEP pathway genes can only have algal origin, as only the autotrophic cells involved in the evolution of diatoms had the plastidial route of isoprenoid biosynthesis. Taking this into account, we set out to investigate whether it was a red or a green algal cell that mediated each of the

MEP pathway gene transfers to diatoms (Fig. S7). According to our analysis, 1-deoxy-D-xylulose 5-phosphate synthase (DXS), 1-deoxy-D-xylulose 5-phosphate reductoisomerase (DXR) and 2-C-methyl-D-erythritol-4-phosphate-cytidylyltransferase (MCT) were transferred from the red algal endosymbiont. This is evident from the highly supported clustering of diatom and red algal proteins. By contrast, 4-diphosphocytidyl-2c-methyl-d-erythritol kinase (CMK) and (E)-4-hydroxy-3-methylbut-2-enyl diphosphate synthases (HDS) genes were transferred from a green algal related cell. For the remaining two, 2-C-methyl-D-erythritol 2,4-cyclodiphosphate synthase (MDS) and hydroxymethylbutenyl diphosphate reductase (HDR), diatom proteins form well separated clades, in distance from their algal homologues. This topology suggests that the corresponding genes have diverged so that their origin cannot be traced.

Prenyltransferases and phytoene/squalene synthases could have been acquired from any of the involved cells. The poorly supported branching patterns obtain for *Ho*PTS1 and its homologues did not allow us to trace its exact evolutionary origin. *Ho*PTS3 clusters closely to prenyltransferases from opisthokonts (fungi/metazoan), while other algal and cyanobacterial homologues are branching separately. This topology suggests heterotrophic host origin. By contrast, the topologies obtained for rest of the prenyltransferases, as well as the phytoene/squalene synthases are suggestive of algal origin. For *Ho*PTS5 and *Ho*PSY this is the red algal endosymbiont, as diatom and red algal proteins form well-supported branches together. The *Ho*IDI-SQS gene is only conserved among heterokonts, haptophytes and dinoflagellates, so it is likely that the fusion occurred in the common ancestor of these groups (Davies *et al.*, 2015). We investigated the evolutionary origin of each domain seperately and both for IDI and SQS inferred phylogenies indicate green algal origin. While for SQS the corresponding cluster is poorly supported, for IDI, branching is more robust (Fig. S8). *Ho*PTS2 and *Ho*PTS4 have also been acquired from an algal cell, however, due to divergence it was not possible to pinpoint their exact origin.

This divergence could alternatively be explained by gene duplication events. To resolve this, we examined three other diatom species (*P. tricornutum, T. pseudonana and Fragilariopsis cylindrus*) that have their genomes sequenced. We obtained the homologues of the *H. ostrearia* genes with non-resolved origin and we identified their localization in the respective genome. In all cases, those genes are localized in different genomic regions (Table S4), indicating that they have been likely independently acquired.

For genes of the MVA pathway see the extended discussion on the phylogenetic analysis that is included in the Supporting Information (Notes S1).

Prediction of subcellular localization

To facilitate analysis of the physiological roles of the identified genes, we investigated the presence of signal/target peptides and obtained insight into the subcellular localization of the encoded proteins. The different endosymbiotic events provided diatoms with secondary plastids surrounded by four membranes (Kroth & Strotmann, 1999). As a result, nuclear encoded diatom proteins targeted to the chloroplast contain a bipartite N-terminal pre-sequence, consisting of a signal peptide with a conserved motif at its cleavage site (ASAFAP motif), followed by a chloroplast transit peptide. These pre-sequences are essential for efficient transportation through all of the four membranes (Gruber *et al.*, 2007). Two independent computational tools, ASAFind (Gruber *et al.*, 2015) and HECTAR (Gschloessl *et al.*, 2008), developed for identification of such pre-sequences were used. A different tool that is used to predict transmembrane domains in protein sequences (TMHMM Server v. 2.0) (Krogh *et al.*, 2001) was included to further support this analysis. The results are summarized on Table S5. As expected, enzymes of the MEP pathway are predicted to be targeted to the chloroplast and those of the MVA to the cytosol. *Ho*PTS1 was found to contain a Type II signal anchor and analysis with TMHMM

311 suggested a strong possibility for the presence of transmembrane domain at the N-terminal region.
312 Taken together, these results suggest that *Ho*PTS1 is likely localized on the ER membrane. The ER
313 membrane is continuous with the outermost chloroplastic membrane, so the possibility that *Ho*PTS1 is
314 related to the periplastic space (the space between the second and third outermost membranes) cannot
315 be ruled out. For *Ho*PTS3 and *Ho*IDI-SQS, no target or signal peptides were predicted, indicating
316 cytosolic localization. *Ho*PTS2, *Ho*PTS5 and *Ho*PSY were predicted to be targeted to the chloroplast,
317 and *Ho*PTS4 to the mitochondria.

318

319 **Functional characterization of prenyltransferases**

320 Since the subcellular compartmentalization analysis revealed the presence of signal/target peptides for
321 the majority of the identified prenyltransferases, both the full length and different truncated variants of
322 the enzymes (Table 2) were cloned into bacterial vectors, in-frame with a C- or N- terminal His-tag.
323 This allowed the purification of the expressed proteins and the assessment of their activity in *in vitro*
324 enzymatic reactions with prenyl diphosphate substrates. IPP was always used as the homoallylic
325 substrate, while DMAPP, GPP, FPP and GGPP were used as the allylic substrates. Control reactions
326 with cells transformed with empty vectors and subjected to the same purification steps were ran in
327 parallel. Following acid hydrolysis that facilitated the removal of diphosphates from the substrates and
328 products, all reactions were extracted and analyzed by GC-MS. Compound identification was based on
329 combination of accurate mass and retention time comparisons between reaction products and acid
330 hydrolysis products of authentic standards.

331 To characterize *Ho*PTS1, the full-length enzyme was initially incubated with DMAPP or GPP and IPP.
332 As no activity was detected, two truncated variants, *Ho*PTS1(Ser73-end) and *Ho*PTS1(Val89-end),
333 were tested with the same substrates. Analysis of the reaction extracts resulted in the identification of

linalool (LOH) and nerolidol (NOH), the acid hydrolysis products of GPP and FPP, respectively. (Fig. 2a). In the presence of DMAPP and IPP both alcohols were detected, while in the presence of GPP and IPP, NOH was the only product identified. No other substrate combination resulted in new product formation, demonstrating that *Ho*PTS1 is an FPP synthase.

Initial expression of full length and truncated versions of *Ho*PTS2 resulted in formation of inclusion bodies. To obtain soluble proteins the variants were subcloned in frame with thioreodoxin at the N-terminus and a C-terminal 6xHis-tag. Soluble *Ho*PTS2 fusions were obtained and purified. The activities of the full-length enzyme and *Ho*PTS2(Gly140-end) variant were tested with all substrate combinations but no new product was formed. However, variant *Ho*PTS2(Arg45-end) was active and gave a range of C_{20}-C_{30} prenyl diphosphates in the presence of FPP and IPP and C_{25}-C_{30} prenyl diphosphates when GGPP and IPP were used as reaction substrates. We were able to identify these products only after hydrolysis treatment and formation of the corresponding prenyl alcohols (Fig. 2b). The fact that the variant *Ho*PTS2(Gly140-end) was inactive indicates the presence of a functionally important region between amino acids 45 and 140. Taken together, these results suggest that *Ho*PTS2 is likely a short chain polyprenyl synthase.

As there was no prediction for the presence of any signal/target peptide in *Ho*PTS3, the full-length enzyme was expressed, purified and tested with different substrates. In presence of DMAPP and IPP, there was no formation of new products. However, when GPP and IPP were used as reaction substrates, GGPP was produced, detected as its acid hydrolysis product geranyl linalool (GGOH). The same product was formed when the enzyme was incubated with FPP and IPP (Fig. 2c). These results, in combination with the lack of a signal peptide predicted for *Ho*PTS3, suggest that *Ho*PTS3 is possibly a cytosolic GGPP synthase in *H. ostrearia.*

Efforts to characterize *Ho*PTS4 were unsuccessful, as there was no activity detected in any reaction extract analyzed, either when using the full-length protein or the truncated variant *Ho*PTS4(Leu121-end). The prediction of a mitochondrial target peptide at the N-terminal region suggests that this, possibly, affects the catalytic activity of *Ho*PTS4 and more truncated variants should be tested to identify the optimal cleavage position for obtaining an active enzyme.

Finally, full length *Ho*PTS5 and one truncated variant, *Ho*PTS5(Ser36-end), were tested with different substrate combinations. Even though the full-length enzyme was inactive, the truncated variant *Ho*PTS5(Ser36-end) produced GGPP as the main product. This was detectable by the formation of GLOH after acid hydrolysis of the reaction products. Minor peaks of LOH and NOH were detected in when DMAPP and IPP were used as substrates while NOH could be detected when GPP and IPP were used instead, indicating that the formation of GGPP proceeds via GPP and FPP (Fig. 2d). Taken together with the prediction of a chloroplastic target peptide in its N-terminus, these results suggest that *Ho*PTS5 possibly encodes for a GGPP synthase, likely functioning in the chloroplast providing the substrates for carotenoid and phytol biosynthesis.

Functional characterization of the squalene/phytoene synthase family members

IDI and SQS are enzymes that catalyze two non-consecutive reactions. IDI isomerizes IPP to DMAPP, while SQS catalyzes the formation of squalene using two FPP molecules. In order to characterize the *Ho*IDISQS fusion, we studied the two domains separately. Initially, we evaluated the functionality of the SQS domain by introducing the full-length gene into a *Saccharomyces cerevisiae* strain engineered to produce high amounts of FPP (Ignea *et al.*, 2012). Analysis of the non-saponifiable lipid extract of yeast cells expressing the fusion showed production of higher amounts of squalene when *Ho*IDISQS was expressed, comparing to the levels of squalene produced by the endogenous yeast squalene

synthase (Fig. 3a). This suggests that the SQS domain of *Ho*IDISQS likely encodes for a functional synthase.

The functionality of the IDI domain was assessed both in the fusion (*Ho*IDISQS) and as a separately cloned IDI domain (*Ho*IDI). These two enzymes were coexpressed with a monoterpene synthase from *Salvia fruticosa*, 1,8-cineole synthase (*Sf*CinS1) (Kampranis *et al.*, 2007), in a yeast strain engineered to produce GPP (Ignea *et al.*, 2012). By catalyzing the isomerization of IPP to DMAPP, IDI provides substrates to the endogenous yeast enzyme, Erg20p, which produces GPP. It has been shown that functional IDI expression in this strain increased synthesis of 1,8-cineole by *Sf*CinS1 (Ignea *et al.*, 2011). We used the endogenous yeast IDI (*Sc*IDI) as positive control. We sampled the head space of all yeast cell cultures expressing the different combinations with solid-phase microextraction (SPME). Their analysis showed that when *Ho*IDISQS, *Ho*IDI or *Sc*IDI were expressed, significantly higher amounts of 1,8-cineole were produced (Fig. 3a), indicating that the IDI domain is functional both individually and in the fusion.

We also employed the yeast expression system for the characterization of the candidate phytoene synthase, *Ho*PSY. During our signal/target peptide analysis, *Ho*PSY was predicted to have a chloroplastic signal peptide. Taking into account the cleavage position predictions, we introduced the truncated variant *Ho*PSY(Ser78-end) into a yeast expression vector. In order to evaluate the functionality of the variant, we employed an *in vivo* chromogenic assay. In this assay, the candidate phytoene synthase is coexpressed with a GGPP synthase and a phytoene desaturase. The GGPP synthase provides the substrate for phytoene biosynthesis. If the examined enzyme is functional and produces phytoene, the desaturase will take it up to produce lycopene (Fig. 3c). As yeast colonies that accumulate lycopene become orange colored, this assay provides a quick and reliable means for the evaluation of phytoene synthase activity. We used the yeast strain AM94, which is engineered for the efficient production of isoprenoids (Ignea *et al.*, 2012), and introduced *Ho*PSY(Ser78-end) together

with Erg20p(Y95A), an engineered GGPP synthase (Ignea *et al.*, 2015), and crtI, a phytoene desaturase from *Xantho*phyllomyces *dendrorhous* (Verwaal *et al.*, 2007). While control cells that were transformed with an empty vector, instead of *Ho*PSY(Ser78-end), did not develop any color, coexpression of the three genes resulted in orange colored yeast colonies, confirming the ability of *Ho*PSY to synthesize phytoene (Fig. 3b).

Discussion

Aiming to shed light to isoprenoid biosynthesis in diatoms, we investigated different aspects of the pathway in the species *H. ostrearia*. By combining transcriptomic analysis with functional characterization of enzyme activities and predictions for the subcellular localization of the corresponding proteins, we can propose a model of isoprenoid biosynthesis in *H. ostrearia* that is summarized in Fig. 4 and discussed below.

H. ostrearia retains a functional cytosolic MVA pathway and a functional plastidial MEP pathway. The majority of the sequenced diatom species investigated to date, have both routes for IPP and DMAPP synthesis. However, these precursors seem to be differentially allocated towards the final isoprenoid products across different species. A plant-like dichotomy has been observed in *P. tricornutum, Nitzschia ovalis and R. setigera.* Accordingly, sterols are synthesized using precursors from the MVA route, while the biosynthesis of carotenoids and the diterpenoid phytol proceeds via MEP generated precursors in the chloroplast (Cvejić & Rohmer, 2000; Massé *et al.*, 2004). On the contrary, *H. ostrearia* uses chloroplast-derived precursors to synthesize its main sterol (24-ethylcholest-5-en-3-ol) (Massé *et al.*, 2004). Contribution of the MEP pathway to sterol (24-methyl-cholesta-5,24(28)-dien-3b-ol) biosynthesis is also observed in the centric species *T. pseudonana* under fast growing, nitrogen-replete culture conditions (Zhang *et al.*, 2009). Similar differentiations were shown for HBI

biosynthesis, with *R. setigera* incorporating C_5 precursors derived from the MVA route and *H. ostrearia* synthesizing HBIs with precursors generated via the MEP pathway (Massé *et al.*, 2004). These different patterns on precursor partitioning indicate different regulation mechanisms that in addition to being responsive to external conditions (for example nutrient availability), also possibly involve precursor transportation between cytosol and chloroplast. MVA and MEP pathway cross-talk has already been reported in some plants (Bick & Lange, 2003; Hemmerlin *et al.*, 2003a). Even though there are substantial differences between the organization of the primary plastids of plants and the secondary plastids of diatoms, the four membranes of diatoms' plastids were previously shown to be permeable through specific transporters (Ast *et al.*, 2009).

The first key step towards prenyl diphosphate synthesis is the isomerization of IPP to DMAPP by IDI. Photosynthetic heterokonts including diatoms and brown algae, haptophytes and dinoflagellates are characterized by the expression of a bifunctional protein fusion between IDI and SQS that catalyzes the dimerization of FPP towards sterol synthesis. In *H. ostrearia* this enzyme fusion does not appear to have any target peptide, while it is predicted to contain a pair of transmembrane helices at the C-terminal part, suggesting ER localization. No other contig encoding for an IDI could be identified in our transcriptomic data. Analysis of thirty three diatom species by Ferriols and coworkers (2017) (Ferriols *et al.*, 2017) showed that all but one specific strain of *R. setigera,* express an IDI-SQS fusion and 19 of them additionally express an independent IDI. Many of these fall within the same phylogenetic clade with the independent IDI from *R. setigera* that contains a putative chloroplast targeting peptide. A growing body of evidence recently supports the occurrence of alternative splicing in diatoms (Rastogi *et al.*, 2018). Thus, the possibility that diatoms that only contain the IDI-SQS fusion gene may also produce an alternatively spliced form producing only the IDI protein cannot be ruled out.

450 The intermediate step between the reactions catalyzed by IDI and SQS is FPP synthesis and according
451 to our functional characterization is likely catalyzed by *Ho*PTS1. Sequence analysis of *Ho*PTS1
452 suggested the presence of a type II signal anchor. This is responsible to anchor the enzyme to a
453 membrane. It is reasonable to assume that this is the ER membrane, which is continuous with the
454 outermost chloroplastic membrane in diatoms. Since *Ho*PTS1 and *Ho*IDISQS catalyze consecutive
455 reactions, we can speculate that they might also physically interact, forming an enzymatic complex that
456 is localized at the ER. Previous studies showed that *H. ostrearia*'s main sterol, 24-ethylcholest-5-en-3-
457 ol, incorporates precursors generated from the plastid localized MEP pathway (Massé *et al.*, 2004).
458 Taken together, these results indicate again precursor transportation from the chloroplastic stroma to
459 the cytosol. The mechanism and the exact regulation of such events are unknown. Recently, a novel
460 isoprenoid regulatory mechanism that possibly involves precursor transportation between subcellular
461 compartments has been described in plants. According to this, a cytosolic isopentenyl phosphate kinase
462 (IPK) and specific Nudix hydrolases regulate IPP supply to the pathway by active phosphorylation-
463 dephosphorylation of IP/IPP. Perturbation of IP supply in *Nicotiana tabacum* was shown to affect both
464 MVA and MEP derived isopenoids, highlighting a correlation with precursor transportation (Henry et
465 al., 2015, 2018). Mining the transcriptome of *H. ostrearia* for similar genes, we were able to identify
466 homologues of IPK and Nudix hydrolase (candidate sequences can be found in Supporting Information
467 Notes S2) that are also present in transcriptomes of other sequenced diatom species. Whether a similar
468 regulatory mechanism is present in diatoms and how this controls IP/IPP supply and/or transportation
469 are open questions that remain to be answered.

470 Our model suggests that *Ho*PTS5 acts in the plastids to generate the GGPP that is essential for
471 carotenoid synthesis. The high relative expression levels of HoPTS5 (Fig. S5) probably reflect the high
472 demand for plastidial isoprenoid biosynthesis (primarily GGPP-derived carotenoids and phytol) under
473 the specific growth conditions. The first committed step towards carotenoids is catalyzed by the

phytoene synthase *Ho*PSY. Both of these enzymes were acquired from the red algal endosymbiont, a fact that corroborates their predicted targeting to the chloroplast. *Ho*PTS2, which was also predicted to have chloroplastic localization, showed activity as a short-chain poly prenyl synthase. The lack of detailed information on the isoprenoid content of diatoms prevents us from assigning a role to *Ho*PTS2. It is possible that this enzyme synthesizes polyprenyl diphosphates or the corresponding alcohols *in vivo.* In plants these compounds have been shown to be incorporated in thylakoid membranes and modulate their fluidity, influencing the overall photosynthetic performance (Bajda *et al.*, 2009; Surmacz & Swiezewska, 2011; Akhtar *et al.*, 2017).

We propose that cytosolic isoprenoid biosynthesis is supported by precursors generated by the MVA pathway and prenyl diphosphates synthesized by *Ho*PTS3. Our analysis showed that this is probably the only enzyme that is not conserved among different diatom species. In *H. ostrearia*, the *Ho*PTS3 gene likely originates from the heterotrophic host of secondary endosymbiosis.

Although it was not possible to characterize the activity of *Ho*PTS4, we presume that this enzyme likely acts as a prenyl transferase in the mitochondria. As there was no other prenyltransferase predicted to be targeted to mitochondria, *Ho*PTS4 probably uses substrates transported from the cytosol or chloroplast. Crosstalk between mitochondria and chloroplast in diatoms has been previously proposed for other metabolic processes (Prihoda *et al.*, 2012).

Even though the majority of prenyltransferases catalyze the linear condensation of prenyl diphosphates, there are examples where these enzymes are involved in the synthesis of irregular, neither head-to-tail nor head-to-head, isoprenoids (Rivera *et al.*, 2001; Demissie *et al.*, 2013). A characteristic example is an FPPS in *Artemisia tridentata* that among other reactions catalyzes the head-to-middle linkage of two DMAPP molecules to produce the branched isoprenoid, lavandulyl diphosphate (Hemmerlin *et al.*, 2003b). It is thus likely that HBI biosynthesis also involves prenyltransferase-type enzymes. Furthermore, different HBI producing diatom species were shown to use different precursors for the

498 synthesis of these molecules (Massé *et al.*, 2004), suggesting that the pathway is regulated differently
499 among species. By providing here a model of the isoprenoid pathway in *H. ostrearia* that illustrates the
500 subcellular distribution of different prenyltransferases we believe that we provide a basis for future HBI
501 biosynthetic studies that will test the currently identified or different enzymes, substrates and/or
502 conditions.

503

504 **Conclusions**

505 In conclusion, our investigation has significantly contributed to our understanding of isoprenoid
506 biosynthesis in diatoms. At least five prenyltransferases mediate the prenyl diphosphate synthesis and
507 provide substrates to downstream branches of the pathway. It is likely that precursors are transported
508 from the plastids to the cytosol. Further studies that will confirm the localization of the enzymes *in vivo*
509 and explore the regulatory mechanisms and crosstalk between the MVA and MEP pathways are
510 essential for a thorough elucidation of the mechanisms involved. Our phylogenetic analysis provided an
511 insight into the multisourced genetic background of diatoms that is reflected clearly on isoprenoid
512 biosynthesis. While there is still a general debate on the extent of the contribution of each lineage (red
513 and green) to the genomes of diatoms (Dagan & Martin, 2009; Burki *et al.*, 2012; Deschamps &
514 Moreira, 2012), it is commonly accepted that these events have armed diatoms with the genetic
515 potential and the metabolic plasticity to succeed in contemporary oceans. Finally, our results provide
516 the blueprint for the elucidation of the biosynthetic pathways leading to unique diatom isoprenoids,
517 such as HBIs.

518 **Accession Numbers**

519 All sequences from *H. ostrearia* mentioned in this study have been submitted to the GenBank database
520 (www.ncbi.nlm.nih.gov) with accession numbers provided on Table 1. Transcriptomic data have been

submitted to the European Nucleotide Archive (ENA) Database under experiment accession ERX2834706 and Run accession ERR2827962.

Acknowledgements

The authors wish to thank Dr. David Ian Pattison for assistance in GC-MS sample running and analysis and Dr. Vona Medeler, the scientific curator of the Nantes Culture Collection (NCC), for providing the NCC 153.8 strain. We kindly acknowledge the financial support of the European Commission through projects PIRSES-GA-2010-269294-BIOVADIA and RISE 734708 – GHaNA (to SCK, KK and GM) and the Greek Secretariat of Research and Technology through project 11SYN_3_770 (to SCK, AMM and AAr) that was co-financed by the European Regional Development Fund.

Author contributions

AA, EG, SM, FV, CI carried out experiments. SCK, FV, KK, AMM, AAr, GM designed experiments. AA, EG, SM, CI, FV, SCK, AMM, GM, KK, AAr analyzed results. AA and SCK wrote the manuscript. All authors have read and commented on the manuscript.

References

Akhtar TA, Surowiecki P, Siekierska H, Kania M, Van Gelder K, Rea KA, Virta LKA, Vatta M, Gawarecka K, Wojcik J, *et al.* 2017. Polyprenols Are Synthesized by a Plastidial *cis* - Prenyltransferase and Influence Photosynthetic Performance. *The Plant Cell* **29**: 1709–1725.

Allen AE, Dupont CL, Oborník M, Horák A, Nunes-Nesi A, McCrow JP, Zheng H, Johnson DA, Hu H, Fernie AR, *et al.* 2011. Evolution and metabolic significance of the urea cycle in photosynthetic diatoms. *Nature* **473**: 203–207.

Armbrust EV, Berges JA, Bowler C, Green BR, Martinez D, Putnam NH, Zhou S, Allen AE, Apt KE, Bechner M, *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 HE, Kroth PG, Horn M, Haferkamp I. 2009. Diatom plastids depend on nucleotide import from the cytosol. *Proceedings of the National Academy of Sciences* **106**: 3621–3626.

Bajda A, Konopka-Postupolska D, Krzymowska M, Hennig J, Skorupinska-Tudek K, Surmacz L, Wójcik J, Matysiak Z, Chojnacki T, Skorzynska-Polit E, *et al.* 2009. Role of polyisoprenoids in tobacco resistance against biotic stresses. *Physiologia Plantarum* **135**: 351–364.

Belt ST, Müller J. 2013. The Arctic sea ice biomarker IP 25 : a review of current understanding , recommendations for future research and applications in palaeo sea ice reconstructions. *Quaternary Science Reviews*: 1–17.

Bendtsen JD, Nielsen H, Von Heijne G, Brunak S. 2004. Improved prediction of signal peptides: SignalP 3.0. *Journal of Molecular Biology* **340**: 783–795.

Berthelot K, Estevez Y, Deffieux A, Peruch F. 2012. Isopentenyl diphosphate isomerase: A

564 checkpoint to isoprenoid biosynthesis. *Biochimie* **94**: 1621–1634.

565 **Bhattacharya D, Archibald JM, Weber APM, Reyes-Prieto A**. **2007**. How do endosymbionts
566 become organelles? Understanding early events in plastid evolution. *BioEssays* **29**: 1239–1246.

567 **Bick JA, Lange BM**. **2003**. Metabolic cross talk between cytosolic and plastidial pathways of
568 isoprenoid biosynthesis: Unidirectional transport of intermediates across the chloroplast envelope
569 membrane. *Archives of Biochemistry and Biophysics* **415**: 146–154.

570 **Bowler C, Allen AE, Badger JH, Grimwood J, Jabbari K, Kuo A, Maheswari U, Martens C,**
571 **Maumus F, Otillar RP, *et al.* 2008**. The Phaeodactylum genome reveals the evolutionary history of
572 diatom genomes. *Nature* **456**: 239–244.

573 **Burki F, Flegontov P, Oborník M, Cihlář J, Pain A, Lukeš J, Keeling PJ**. **2012**. Re-evaluating the
574 green versus red signal in eukaryotes with secondary plastid of red algal origin. *Genome Biology and*
575 *Evolution* **4**: 626–635.

576 **Chan CX, Reyes-Prieto A, Bhattacharya D**. **2011**. Red and green algal origin of diatom membrane
577 transporters: Insights into environmental adaptation and cell evolution. *PLoS ONE* **6**: e29138.

578 **Coesel S, Oborník M, Varela J, Falciatore A, Bowler C**. **2008**. Evolutionary Origins and Functions
579 of the Carotenoid Biosynthetic Pathway in Marine Diatoms. *PLOS ONE* **3**: e2896.

580 **Cvejić JH, Rohmer M**. **2000**. CO2as main carbon source for isoprenoid biosynthesis via the
581 mevalonate-independent methylerythritol 4-phosphate route in the marine diatoms Phaeodactylum
582 tricornutum and Nitzschia ovalis. *Phytochemistry* **53**: 21–28.

583 **Dagan T, Martin W**. **2009**. Seeing green and red in diatom genomes. *Science* **324**: 1651–1652.

584 **Dambek M, Eilers U, Breitenbach J, Steiger S, Büchel C, Sandmann G**. **2012**. Biosynthesis of
585 fucoxanthin and diadinoxanthin and function of initial pathway genes in Phaeodactylum tricornutum.

586 *Journal of Experimental Botany* **63**: 5607–5612.

587 **Davies FK, Jinkerson RE, Posewitz MC**. **2015**. Toward a photosynthetic microbial platform for
588 terpenoid engineering. *Photosynthesis Research* **123**: 265–284.

589 **Demissie ZA, Erland LAE, Rheault MR, Mahmoud SS**. **2013**. The biosynthetic origin of irregular
590 monoterpenes in lavandula: Isolation and biochemical characterization of a novel cis-prenyl
591 diphosphate synthase gene, lavandulyl diphosphate synthase. *Journal of Biological Chemistry* **288**:
592 6333–6341.

593 **Deschamps P, Moreira D**. **2012**. Reevaluating the green contribution to diatom genomes. *Genome*
594 *Biology and Evolution* **4**: 683–688.

595 **Dorrell RG, Smith AG**. **2011**. Do red and green make brown?: Perspectives on plastid acquisitions
596 within chromalveolates. *Eukaryotic Cell* **10**: 856–868.

597 **Eilers U, Dietzel L, Breitenbach J, Büchel C, Sandmann G**. **2016**. Identification of genes coding for
598 functional zeaxanthin epoxidases in the diatom Phaeodactylum tricornutum. *Journal of Plant*
599 *Physiology* **192**: 64–70.

600 **Fabris M, Matthijs M, Carbonelle S, Moses T, Pollier J, Dasseville R, Baart GJE, Vyverman W,**
601 **Goossens A**. **2014**. Tracking the sterol biosynthesis pathway of the diatom Phaeodactylum tricornutum.
602 *New phytologist* **204**: 521–535.

603 **Fabris M, Matthijs M, Rombauts S, Vyverman W, Goossens A, Baart GJE**. **2012**. The metabolic
604 blueprint of Phaeodactylum tricornutum reveals a eukaryotic Entner-Doudoroff glycolytic pathway.
605 *Plant Journal* **70**: 1004–1014.

606 **Falkowski PG**. **2002**. The ocean's invisible forest. *Scientific American* **287**: 54–61.

607 **Ferriols VMEN, Yaginuma-Suzuki R, Fukunaga K, Kadono T, Adachi M, Matsunaga S, Okada**

S. **2017**. An exception among diatoms: unique organization of genes involved in isoprenoid biosynthesis in Rhizosolenia setigera CCMP 1694. *Plant Journal* **92**: 822–833.

Ferriols VMEN, Yaginuma R, Adachi M, Takada K, Matsunaga S, Okada S. **2015**. Cloning and characterization of farnesyl pyrophosphate synthase from the highly branched isoprenoid producing diatom Rhizosolenia setigera. *Scientific Reports* **5**: 10246.

Field CB, Behrenfeld MJ, Randerson JT, Falkowski P. **1998**. Primary production of the biosphere: Integrating terrestrial and oceanic components. *Science* **281**: 237–240.

Frommolt R, Werner S, Paulsen H, Goss R, Wilhelm C, Zauner S, Maier UG, Grossman AR, Bhattacharya D, Lohr M. **2008**. Ancient recruitment by chromists of green algal genes encoding enzymes for carotenoid biosynthesis. *Molecular Biology and Evolution* **25**: 2653–2667.

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, *et al.* 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology* **29**: 644–652.

Gruber A, Rocap G, Kroth PG, Armbrust EV, Mock T. **2015**. Plastid proteome prediction for diatoms and other algae with secondary plastids of the red lineage. *Plant Journal* **81**: 519–528.

Gruber A, Vugrinec S, Hempel F, Gould SB, Maier UG, Kroth PG. **2007**. Protein targeting into complex diatom plastids: Functional characterisation of a specific targeting motif. *Plant Molecular Biology* **64**: 519–530.

Gruber A, Weber T, Bartulos CR, Vugrinec S, Kroth PG. **2009**. Intracellular distribution of the reductive and oxidative pentose phosphate pathways in two diatoms. *Journal of Basic Microbiology* **49**: 58–72.

Gschloessl B, Guermeur Y, Cock JM. **2008**. HECTAR: A method to predict subcellular targeting in

heterokonts. *BMC Bioinformatics* **9**: 393.

Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger MB, Eccles D, Li B, Lieber M, *et al.* 2013. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature Protocols* **8**: 1494.

Hemmerlin A, Hoeffler JF, Meyer O, Tritsch D, Kagan IA, Grosdemange-Billiard C, Rohmer M, Bach TJ. **2003a**. Cross-talk between the cytosolic mevalonate and the plastidial methylerythritol phosphate pathways in tobacco bright yellow-2 cells. *Journal of Biological Chemistry* **278**: 26666–26676.

Hemmerlin A, Rivera SB, Erickson HK, Poulter CD. **2003b**. Enzymes encoded by the farnesyl diphosphate synthase gene family in the Big Sagebrush Artemisia tridentata ssp. spiciformis. *Journal of Biological Chemistry* **278**: 32132–32140.

Holstein SA, Hohl RJ. **2004**. Isoprenoids: Remarkable diversity of form and function. *Lipids* **39**: 293–309.

Ignea C, Cvetkovic I, Loupassaki S, Kefalas P, Johnson CB, Kampranis SC, Makris AM. **2011**. Improving yeast strains using recyclable integration cassettes, for the production of plant terpenoids. *Microbial Cell Factories* **10**: 4.

Ignea C, Trikka FA, Kourtzelis I, Argiriou A, Kanellis AK, Kampranis SC, Makris AM. **2012**. Positive genetic interactors of HMG2 identify a new set of genetic perturbations for improving sesquiterpene production in Saccharomyces cerevisiae. *Microbial cell factories* **11**: 162.

Ignea C, Trikka F a., Nikolaidis AK, Georgantea P, Ioannou E, Loupassaki S, Kefalas P, Kanellis AK, Roussis V, Makris AM, *et al.* 2015. Efficient diterpene production in yeast by engineering Erg20p into a geranylgeranyl diphosphate synthase. *Metabolic Engineering* **27**: 65–75.

Kampranis SC, Ioannidis D, Purvis A, Mahrez W, Ninga E, Katerelos NA, Anssour S, Dunwell JM, Degenhardt J, Makris AM, *et al.* 2007. Rational Conversion of Substrate and Product Specificity in a *Salvia* Monoterpene Synthase: Structural Insights into the Evolution of Terpene Synthase Function. *The Plant Cell* **19**: 1994–2005.

Keeling PJ, Burki F, Wilcox HM, Allam B, Allen EE, Amaral-Zettler LA, Armbrust EV, Archibald JM, Bharti AK, Bell CJ, *et al.* 2014. The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP): Illuminating the Functional Diversity of Eukaryotic Life in the Oceans through Transcriptome Sequencing. *PLoS Biology* **12**: e1001889.

Krogh A, Larsson B, von Heijne G, Sonnhammer ELL. **2001**. Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes. *Journal of molecular biology* **305**: 567–580.

Kroth PG, Chiovitti A, Gruber A, Martin-Jezequel V, Mock T, Parker MS, Stanley MS, Kaplan A, Caron L, Weber T, *et al.* 2008. A model for carbohydrate metabolism in the diatom Phaeodactylum tricornutum deduced from comparative whole genome analysis. *PLoS One* **3**: e1426.

Kroth P, Strotmann H. **1999**. Diatom plastids: Secondary endocytobiosis, plastid genome and protein import. *Physiologia Plantarum* **107**: 136–141.

Kumar S, Stecher G, Tamura K. **2016**. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Molecular biology and evolution* **33**: 1870–1874.

Larkin M, Blackshields G, Brown N, Chenna R, McGettigan P, McWilliam H, Valentin F, Wallace I, Wilm A, Lopez R, *et al.* 2007. ClustalW and ClustalX version 2. *Bioinformatics* **23**: 2947–2948.

Liang PH, Ko TP, Wang AHJ. **2002**. Structure, mechanism and function of prenyltransferases. *European Journal of Biochemistry* **269**: 3339–3354.

675 **Marchler-Bauer A, Derbyshire MK, Gonzales NR, Lu S, Chitsaz F, Geer LY, Geer RC, He J,**
676 **Gwadz M, Hurwitz DI, *et al.* 2015**. CDD: NCBI's conserved domain database. *Nucleic Acids*
677 *Research* **43**: D222–D226.

678 **Massé G, Belt ST, Crosta X, Schmidt S, Snape I, Thomas DN, Rowland SJ**. **2011**. Highly branched
679 isoprenoids as proxies for variable sea ice conditions in the Southern Ocean. *Antarctic Science* **23**:
680 487–498.

681 **Massé G, Belt ST, Rowland SJ, Rohmer M**. **2004**. Isoprenoid biosynthesis in the diatoms
682 Rhizosolenia setigera (Brightwell) and Haslea ostrearia (Simonsen). *Proceedings of the National*
683 *Academy of Sciences* **101**: 4413–4418.

684 **McFadden GI**. **2001**. Primary and secondary endosymbiosis and the origin of plastids. *Journal of*
685 *Phycology* **37**: 951–959.

686 **Moustafa A, Beszteri B, Maier UG, Bowler C, Valentin K, Bhattacharya D**. **2009**. Genomic
687 footprints of a cryptic plastid endosymbiosis in diatoms. *Science* **324**: 1724–1726.

688 **NCBI, Resource C**. **2017**. Database Resources of the National Center for Biotechnology Information.
689 *Nucleic acids research* **45**: D12–D17.

690 **Nelson DM, Tréguer P, Brzezinski MA, Leynaert A, Quéguiner B**. **1995**. Production and
691 dissolution of biogenic silica in the ocean: Revised global estimates, comparison with regional data and
692 relationship to biogenic sedimentation. *Global Biogeochemical Cycles* **9**: 359–372.

693 **Obata T, Fernie AR, Nunes-Nesi A**. **2013**. The central carbon and energy metabolism of marine
694 diatoms. *Metabolites* **3**: 325–346.

695 **Oborník M, Green BR**. **2005**. Mosaic origin of the heme biosynthesis pathway in photosynthetic
696 eukaryotes. *Molecular Biology and Evolution* **22**: 2343–2353.

Prihoda J, Tanaka A, De Paula WBM, Allen JF, Tirichine L, Bowler C. **2012**. Chloroplast-mitochondria cross-talk in diatoms. *Journal of Experimental Botany* **63**: 1543–1557.

Rastogi A, Maheswari U, Dorrell RG, Vieira FRJ, Maumus F, Kustka A, McCarthy J, Allen AE, Kersey P, Bowler C, *et al.* 2018. Integrative analysis of large scale transcriptome data draws a comprehensive landscape of Phaeodactylum tricornutum genome and evolutionary origin of diatoms. *Scientific Reports* **8**: 4834.

Rivera SB, Swedlund BD, King GJ, Bell RN, Hussey CE, Shattuck-Eidens DM, Wrobel WM, Peiser GD, Poulter CD. **2001**. Chrysanthemyl diphosphate synthase: isolation of the gene and characterization of the recombinant non-head-to-tail monoterpene synthase from Chrysanthemum cinerariaefolium. *Proceedings of the National Academy of Sciences of the United States of America* **98**: 4373–4378.

Rodríguez-Ezpeleta N, Brinkmann H, Burey SC, Roure B, Burger G, Löffelhardt W, Bohnert HJ, Philippe H, Lang BF. **2005**. Monophyly of primary photosynthetic eukaryotes: Green plants, red algae, and glaucophytes. *Current Biology* **15**: 1325–1330.

Rowland SJ, Belt ST, Wraige EJ, Massé G, Roussakis C, Robert JM. **2001**. Effects of temperature on polyunsaturation in cytostatic lipids of Haslea ostrearia. *Phytochemistry* **56**: 597–602.

Smith SR, Abbriano RM, Hildebrand M. **2012**. Comparative analysis of diatom genomes reveals substantial differences in the organization of carbon partitioning pathways. *Algal Research* **1**: 2–16.

Spanova M, Daum G. **2011**. Squalene - biochemistry, molecular biology, process biotechnology, and applications. *European Journal of Lipid Science and Technology* **113**: 1299–1320.

Surmacz L, Swiezewska E. **2011**. Polyisoprenoids - Secondary metabolites or physiologically important superlipids? *Biochemical and Biophysical Research Communications* **407**: 627–632.

719 **Tarshis LC, Proteau PJ, Kellogg B a, Sacchettini JC, Poulter CD**. **1996**. Regulation of product
720 chain length by isoprenyl diphosphate synthases. *Proceedings of the National Academy of Sciences of*
721 *the United States of America* **93**: 15018–15023.

722 **Timmis JN, Ayliff MA, Huang CY, Martin W**. **2004**. Endosymbiotic gene transfer: Organelle
723 genomes forge eukaryotic chromosomes. *Nature Reviews Genetics* **5**: 123–135.

724 **Veríssimo A, Bassard J-E, Julien-Laferrière A, Sagot M-F, Vinga S**. **2017**. MassBlast: A workflow
725 to accelerate RNA-seq and DNA database analysis. *bioRxiv*.

726 **Verwaal R, Wang J, Meijnen JP, Visser H, Sandmann G, Van Den Berg JA, Van Ooyen AJJ**.
727 **2007**. High-level production of beta-carotene in *Saccharomyces cerevisiae* by successive
728 transformation with carotenogenic genes from *Xanthophyllomyces dendrorhous*. *Applied and*
729 *Environmental Microbiology* **73**: 4342–4350.

730 **Vranová E, Coman D, Gruissem W**. **2012**. Structure and Dynamics of the Isoprenoid Pathway
731 Network. *Molecular Plant* **5**: 318–333.

732 **Wang K, Ohnuma S ichi**. **1999**. Chain-length determination mechanism of isoprenyl diphosphate
733 synthases and implications for molecular evolution. *Trends in Biochemical Sciences* **24**: 445–451.

734 **Whelan S, Goldman N**. **2001**. A General Empirical Model of Protein Evolution Derived from
735 Multiple Protein Families Using a Maximum-Likelihood Approach. *Molecular Biology and Evolution*
736 **18**: 691–699.

737 **Wraige EJ, Johns L, Belt ST, Mass√© G, Robert J-M, Rowland S**. **1999**. Highly branched C25
738 isoprenoids in axenic cultures ofHaslea ostrearia. *Phytochemistry* **51**: 69–73.

739 **Yoon HS, Hackett JD, Ciniglia C, Pinto G, Bhattacharya D**. **2004**. A Molecular Timeline for the
740 Origin of Photosynthetic Eukaryotes. *Molecular Biology and Evolution* **21**: 809–818.

Zhang Z, Sachs JP, Marchetti A. **2009**. Hydrogen isotope fractionation in freshwater and marine algae: II. Temperature and nitrogen limited growth rate effects. *Organic Geochemistry* **40**: 428–439.

760 **Supporting Information**

761 **Additional supporting information may be found in the online version of this article.**

762 **Methods S1** Supplementary methods

763 **Table S1** IDs of the transcriptomes from the MMETSP used in the phylogenetic analysis

764 **Table S2** List of primers used in this study

765 **Table S3** Identification of *Ho*PTS1-*Ho*PTS5 homologues in other diatom species

766 **Table S4** Genomic location of isoprenoid biosynthetic genes in the *P. tricornutum, T. pseudonana*
767 and *F. cylindrus*

768 **Table S5** Summarized results from predictions of protein subcellular localization

769 **Fig. S1** Overview of isoprenoid biosynthesis highlighting the different stages of the pathway (early,
770 central, final steps)

771 **Fig. S2** Chemical structures of some of the isoprenoids synthesized by *H. ostrearia*

772 **Fig. S3** HBI profile of *Haslea ostrearia* NCC 153.8 strain

773 **Fig. S4** Phylogenetic relationship of prenyltransferases from diatoms

774 **Fig. S5** Expression patterns of genes involved in isoprenoid biosynthesis as detected by RNA-seq

775 **Fig. S6** (a) Multiple sequence alignment of *Ho*PTS1 with known farnesyl diphosphate synthases
776 from other species (b) Multiple sequence alignment of *Ho*PTS2-5 with known polyprenyl
777 diphosphate synthases from other species

778 **Fig. S7** Phylogenetic trees of MEP pathway proteins

779 **Fig. S8** Phylogenetic trees of prenyltransferases and squalene/phytoene synthase family members

780 **Notes S1** Extended discussion on the phylogenetic analysis

781 **Notes S2** Diatom sequences

782

783

784

785

786

787

788

789

790

791

792

793

794

795

796

797

798

Tables

Table 1 Candidate isoprenoid biosynthetic genes, their closest homologues their evolutionary origin and subcellular localization as indicated by analysis in this study

	Gene Name	Genebank ID	Closest homologue (percentage similarity)	Evolutionary origin	Subcellular localization prediction
MEP PATHWAY	1-deoxy-D-xylulose 5-phosphate synthase (DXS)	MH731010	P*haeodactylum tricornutum* XP_002176386.1 (83%)	Red algae	Chloroplast
	1-deoxy-D-xylulose 5-phosphate reductoisomerase (DXR)	MH731011	Thalassiosira pseudonana XP_002295597.1 (84%)	Red algae	Chloroplast
	2-C-methyl-D-erythritol-4-phosphate-cytidylyltransferase (MCT)	MH731012	Fistulifera solaris GAX13480.1 (78%)	Red algae	Chloroplast
	4-diphosphocytidyl-2c-methyl-d-erythritol kinase (CMK)	MH731013	Phaeodactylum tricornutum XP_002178363.1 (77%)	Green algae	Chloroplast
	2-C-methyl-D-erythritol 2,4-cyclodiphosphate synthase (MDS)	MH731014	Phaeodactylum tricornutum XP_002180038.1 (76%)	Algae	Chloroplast
	(E)-4-hydroxy-3-methylbut-2-enyl diphosphate synthase (HDS)	MH731015	Fragilariopsis cylindrus OEU20628.1 (77%)	Green algae	Chloroplast
	Hydroxymethylbutenyl diphosphate reductase (HDR)	MH731016	Phaeodactylum tricornutum XP_002178617 (73%)	Algae	Chloroplast
MVA PATHWAY	Acetyl-coa c-acetyltransferase 1 (AACT)	MH731017	Phaeodactylum tricornutum XP_002185228.1 (69%)		Cytosol
	Hydroxy-methylglutaryl-CoA synthase (HMGS)	MH731018	Fragilariopsis cylindrus OEU16767.1 (73%)		Cytosol
	Hydroxyl-methylglutaryl-CoA reductase (HMGR)	MH731019	Fragilariopsis cylindrus OEU16221.1 (82%)		Cytosol
	Mevalonate kinase (MVK)	MH731020	Thalassiosira pseudonana XP_002287787.1 (73%)		Cytosol
	Phospho-mevalonate kinase (PMK)	MH731021	Fragilariopsis cylindrus OEU13939.1 (49%)		Cytosol
	Mevalonate disphosphate decarboxylase (MVD)	MH731022	Fragilariopsis cylindrus OEU17781.1 (69%)		Cytosol
	Isopentenyl-diphosphate delta-isomerase fused to squalene synthase (*Ho*IDISQS)	MH720297	Fistulifera solaris GAX27897.1 (63%)	Green algae	Cytosol
CENTRAL STEPS	Farnesyl diphosphate synthase (*Ho*PTS1)	MH720291	Rhizosolenia setigera AKH49589.1 (56%)	Not traced	ER or PPC
	Polyprenyl diphosphate synthase (*Ho*PTS2)	MH720292	Thalassiosira oceanica EJK71722.1 (75%)	Algae	Chloroplast
	Geranylgeranyl diphosphate synthase (*Ho*PTS3)	MH720293	Phaeodactylum tricornutum XP_002181666.1 (74%)	Heterotrophic host	Cytosol
	Putative polyprenyl synthase (*Ho*PTS4)	MH720294	Phaeodactylum tricornutum XP_002185039.1 (70%)	Algae	Mitochondria
	Geranylgeranyl diphosphate synthase (*Ho*PTS5)	MH720295	Phaeodactylum tricornutum XP_002178555.1 (75%)	Red algae	Chloroplast
	Phytoene synthase (*Ho*PSY)	MH720296	Phaeodactylum tricornutum XP_002178776.1 (69%)	Red algae	Chloroplast

Table 2 Functionally characterized enzymes in this study

	Full length ORF	Variants studied	Accepted substrates	Products
*Ho*PTS1	1299 bp	Full length	no activity	-
		Ser73-end	DMAPP+IPP	GPP, FPP
			GPP+IPP	FPP
		Val89-end	DMAPP+IPP	GPP, FPP
			GPP+IPP	FPP
*Ho*PTS2	1584 bp	Full length	no activity	-
		Arg45-end	FPP+IPP	C20-C30 PPP
			GGPP+IPP	C25-C30 PPP
		Gly140-end	no activity	no activity
*Ho*PTS3	1014 bp	Full length	GPP+IPP	FPP, GGPP
			FPP+IPP	GGPP
*Ho*PTS4	1416 bp	Full length	no activity	-
		Leu121-end	no activity	
*Ho*TPS5	1011 bp	Full length	no activity	-
		Ser36-end	DMAPP+IPP	GPP,FPP,GGPP
			GPP+IPP	FPP, GGPP
			FPP+IPP	GGPP
*Ho*PSY	1485 bp	Full length	Not tested	-
		Ser78-end	GGPP+GGPP	phytoene
*Ho*IDISQS	2277 bp	Full length	FPP+FPP	squalene

Figure legends

Fig. 1. (a) Neighbor joining phylogenetic tree of the selected sequences and their closest homologues and protein domain structures. Numbers on the branches indicate bootstrap support values from 1,000 trees; values under 70 were removed **(b) Protein domain structure of selected sequences.** BSP; Bipartite Signal Peptide, MSP; Mitochondrial Signal Peptide, TD; Transmembrane domain.

Fig. 2. GC-MS profile of (a) *Ho*PTS1(Ser73-end), (b) *Ho*PTS2(Arg45-end), (c) *Ho*PTS3, (d) *Ho*PTS5(Ser36-end) in *vitro* reaction products using different prenyl diphosphates as substrates. Identification of compounds was based on comparison of accurate mass and retention time between reaction substrates/products and acid hydrolysis products of authentic standards (bottom panel). Ions m/z 137, 205, 273, 341 and 409 are derived from precursor ions (m/z 154, 222, 290, 358 and 426 respectively) by loss of H_2O in positive ion mode

Fig. 3. Functional characterization of squalene/phytoene synthase family members. **(a)** Expression of *Ho*IDISQS in the yeast strain AM94 resulted in high accumulation of squalene (up). Coexpression of *Ho*IDISQS with SfCinS1 resulted in high accumulation of 1,8-cineole. *Ho=Haslea ostrearia*, *Sf=Salvia fruticosa*, *Sc=Saccharomyces cerevisiae.* **(b)** Yeast colonies coexpressing ERG20(Y95A), crtI and *Ho*PSY(Ser78-end) (up) in comparison with the control colonies that carry an empty vector instead of a phytoene synthase (down) and **(c)** Lycopene biosynthetic pathway.

Fig. 4 A model for isoprenoid biosynthesis in *H. ostrearia.* Isopentenyl diphosphate (IPP) and dimethyl allyl diphosphate (DMAPP) are synthesized via the mevalonate (MVA) and methylerythritol phosphate (MEP) pathways in the cytosol and plastid. Each of these precursor pools are used for the synthesis of prenyl diphosphates by prenyltransferases at different subcellular compartments. *Ho*PTS1

846 is responsible for farnesyl diphosphate (FPP) synthesis at the endoplasmic reticulum (ER), in close
847 proximity to *Ho*IDISQS. *Ho*PTS3 mediates synthesis of geranylgeranyl diphosphate (GGPP) in the
848 cytosol. In the chloroplast *H*oPTS2 and *H*oPTS5 synthesize precursors for polyprenol and carotenoid
849 synthesis. The first committed step of carotenoid biosynthesis is catalyzed by *H*oPSY. *H*oPTS4 likely
850 mediates prenyl diphosphate synthesis in mitochondria. Dashed arrows indicate possible precursor
851 transportation between cytosol and plastids.

852

853

854

855

856

857

858

859