



OPEN

SUBJECT AREAS:

BIOFUELS

GENE REGULATORY NETWORKS

Received
25 March 2014Accepted
9 June 2014Published
26 June 2014

Correspondence and requests for materials should be addressed to J.X. (xujian@qibebt.ac.cn) or K.N. (ningkang@qibebt.ac.cn)

Genome-wide identification of transcription factors and transcription-factor binding sites in oleaginous microalgae *Nannochloropsis*

Jianqiang Hu^{1,2}, Dongmei Wang¹, Jing Li^{1,2}, Gongchao Jing¹, Kang Ning¹ & Jian Xu¹

¹Single-Cell Center, CAS Key Laboratory of Biofuels and Shandong Key Laboratory of Energy Genetics, Qingdao Institute of BioEnergy and Bioprocess Technology, Chinese Academy of Sciences, Qingdao, Shandong 266101, China, ²University of Chinese Academy of Sciences, Beijing 100049, China.

Nannochloropsis spp. are a group of oleaginous microalgae that harbor an expanded array of lipid-synthesis related genes, yet how they are transcriptionally regulated remains unknown. Here a phylogenomic approach was employed to identify and functionally annotate the transcriptional factors (TFs) and TF binding-sites (TFBSs) in *N. oceanica* IMET1. Among 36 microalgae and higher plants genomes, a two-fold reduction in the number of TF families plus a seven-fold decrease of average family-size in *Nannochloropsis*, *Rhodophyta* and *Chlorophyta* were observed. The degree of similarity in TF-family profiles is indicative of the phylogenetic relationship among the species, suggesting co-evolution of TF-family profiles and species. Furthermore, comparative analysis of six *Nannochloropsis* genomes revealed 68 “most-conserved” TFBS motifs, with 11 of which predicted to be related to lipid accumulation or photosynthesis. Mapping the IMET1 TFs and TFBS motifs to the reference plant TF-“TFBS motif” relationships in TRANSFAC enabled the prediction of 78 TF-“TFBS motif” interaction pairs, which consisted of 34 TFs (with 11 TFs potentially involved in the TAG biosynthesis pathway), 30 TFBS motifs and 2,368 regulatory connections between TFs and target genes. Our results form the basis of further experiments to validate and engineer the regulatory network of *Nannochloropsis* spp. for enhanced biofuel production.

Microalgae are functionally diverse and phylogenetically heterogeneous groups of organisms that are primarily unicellular, aquatic and photosynthetic eukaryotes. They are responsible for over 45% of our planet’s annual net primary biomass¹. On the other hand, they represent a promising source of biomass feedstock for fuels and chemicals, as many species possess the ability to grow rapidly and synthesize large amounts of storage neutral lipids in a form of triacylglycerol (TAG) from sunlight and carbon dioxide and moreover can be cultivated on non-arable land with non-potable water and waste streams (e.g., flue gases and wastewaters) and thus pose little competition to food crops while providing environmental benefits². As vigorous growth and TAG accumulation are usually mutually exclusive in microalgae, genetic engineering of microalgae for improved growth while stimulated TAG production has been a key goal^{2,3}. However, the molecular and cellular mechanisms underlying lipid metabolism in microalgae are still elusive, which has hampered rational approaches to screen for or genetically engineer superior production strains^{3–5}.

Transcriptional factor (TF)-encoding genes have been recognized as one important source of the diversity and change that underlie the evolution of plants^{6,7}. Moreover, identification of the transcriptional factors (TF; *trans* elements) and their cognate transcriptional factor binding-sites (TFBS; *cis* elements) is one of the first steps in dissecting and engineering the regulatory network for enhanced productivity of the target molecules^{8–10}. Both experimental and computational methods have been developed for the identification of TFBS motifs in the promoters of TF target genes on a genome-wide scale. Experimental methods such as DNase footprinting¹¹ and electrophoretic mobility shift assays¹² have unfortunately fallen far behind the rapid accumulation of genome sequences. Moreover, high-throughput experiments such as Chip-Seq¹³ can be costly and time-consuming. Therefore *in silico* evidences between TFs and TFBSs can be valuable.

In microalgae, however, computational identifications of transcriptional factors (TFs) genome-wide were reported only for green algae such as *Chlamydomonas reinhardtii* and *Volvox carteri*^{14,15}, red algae such as



Galdieria sulphuraria^{14,15} and an Eustigmatophyceae strain (*Nannochloropsis oceanica* CCMP1779¹⁶). On the other hand, a global *in silico* prediction of *cis*-regulatory element (CRE) was reported in *C. reinhardtii*¹⁷. However, it remained largely unclear how the genomic profiles of TFs in microalgae is linked to or different from higher plants, and how such relationships is implicated in the evolution of these unicellular organisms and their multicellular higher plant siblings (all modern higher plants were derived from green algae¹). Furthermore, as potential model organisms of oleaginous microalgae have started to emerge only recently^{16,18,19}, few attempts have been made to model the links between TF and their cognate TFBS (i.e., the targeted genes) on a genome-wide scale in oleaginous microalgae.

Nannochloropsis spp. are a group of microalgae in the Eustigmatophyceae class, and are widely distributed in the marine environment as well as in fresh and brackish waters^{18,20}. These algae are of industrial interest due to their ability to grow rapidly, synthesize large amounts of TAG and high-value polyunsaturated fatty acids (e.g. eicosapentaenoic acid), and tolerate broad environmental and culture conditions^{21,22}. As a result, these organisms attracted particular attention and have emerged as a research model for microalgal oleaginousness^{16,18–20}. We have recently adopted a phylogenomic approach to unravel the genome-wide diversity and divergence of the oleaginous loci in this microalgal genus^{18,20}. A comparative analysis of six genomes of oleaginous *Nannochloropsis* spp. that includes two *N. oceanica* strains (IMET1 and CCMP531) and one strain from each of four other recognized species: *N. salina* (CCMP537), *N. gaditana* (CCMP526, which was previously reported¹⁹), *N. oculata* (CCMP525) and *N. granulata* (CCMP529) revealed a core genome of ca. 2,700 genes and a large pan-genome of >38,000 genes¹⁸. Moreover, the six genomes share key oleaginous traits such as the enrichment of selected lipid biosynthesis genes¹⁸. This genus-wide set of oleaginous genomes thus provides an opportunity to identify the diversity and evolution of TF families as well as TFBSs in *Nannochloropsis*. Furthermore, we have generated large-scale, highly reproducible transcript profiles from *N. oceanica* strain IMET1 as a function of time (i.e., over the six time points of 3, 4, 6, 12, 24, 48 h) under both N-replete (N+) and N-depleted (N-) conditions via mRNA-Seq²³. This time-series transcriptomic dataset in *Nannochloropsis* thus laid a foundation for unraveling the links between TFs and TFBSs via gene co-expression analysis.

Here we present a genome-wide *in silico* map of TFs and TFBSs and a computationally predicted, preliminary regulatory network that link TFs and target genes in *Nannochloropsis*. First, the TF-encoding genes in the genomes of *N. oceanica* IMET1, *N. oceanica* CCMP1779 and *N. gaditana* CCMP526 were identified. A two-fold reduction in the number of TF families plus a seven-fold decrease of average family-size in *Nannochloropsis*, *Rhodophyta* and *Chlorophyta* were apparent, as compared to those of the surveyed higher plants. The degree of similarity in TF-family profiles was found to be indicative of the phylogenetic relationship among the species, suggesting that the co-evolution of species and TF profiles occurred largely at the level of TF-family. Furthermore, an improved computational pipeline based on the MERCED algorithm¹⁷ was developed for TFBS identification via comparative analysis of six sequenced *Nannochloropsis* genomes. This analysis revealed 68 “most-conserved” TFBSs, with 11 of which predicted to be related to lipid accumulation or photosynthesis. Comparison of the IMET1 TFs and TFBSs to the reference plant TF-TFBS motif pairs in TRANSFAC²⁴ enabled us to predict 78 interaction pairs between a TF and a TFBS motif, which consisted of 34 TFs (with 11 TFs potentially involved in the TAG biosynthesis pathway), 30 TFBSs and 950 target genes. These results form the basis of further experiments to validate and engineer the regulatory network of *Nannochloropsis* spp. for enhanced biofuel production.

Results

Genome-wide identification of TFs in *Nannochloropsis* spp. and comparative analysis of TF-family profiles among 36 plant species.

A systematic identification of TFs in Nannochloropsis spp. The presence or absence of the defining features of TFs (e.g., their DNA-binding domains, auxiliary domains and forbidden domains²⁵) was typically employed as the major criteria for identification of TFs and moreover, for classification of the TFs into individual TF families. As a result, several public databases of plant TFs have been established, such as PlantTFDB (v. 2.0, <http://plantfdb.cbi.pku.edu.cn>)¹⁴ and PlnTFBS (v. 3.0, <http://plntfdb.bio.uni-potsdam.de/v3.0/>)¹⁵. These databases have cataloged the predicted TFs of over 50 species from the main lineages of the plant kingdom, including green algae, moss, fern, gymnosperm and angiosperm.

We first performed a genome-wide identification of TFs in *N. oceanica* IMET1¹⁸, *N. oceanica* CCMP1779¹⁶ and *N. gaditana* CCMP526¹⁹ via the characteristic domains of plant TFs using the methods of PlantTFDB¹⁴. The results revealed 125, 119 and 85 TFs in the three *Nannochloropsis* strains respectively (1.26%, 0.99% and 0.94% of their proteomes; Supplemental Table S1, Supplemental Dataset S1). Each predicted TF was then assigned into a specific TF-family based on its DNA-binding domain (based on the criteria of PlantTFDB). In the three strains, 26 TF families were found collectively, among which 19 are shared by all three strains. MYB, bZIP, MYB-related and NF-YC are the four largest TFs families, together accounting for 48 ~ 56% of all TFs in each of the strains.

The MYB group of TFs, found in the MYB family and the MYB-related family, is the largest in each of the three strains. *N. oceanica* IMET1 harbored 35 TF genes from this group, including 15 R2R3-MYB genes, 8 R1R2R3-MYB genes and 12 MYB-related genes (Supplemental Table S2). Numerous MYB genes have been characterized by genetic approaches and found to be involved in the control of plant-specific processes in higher plants including (i) primary and secondary metabolism, (ii) cell fate and identity, (iii) developmental processes and (iv) responses to biotic and abiotic stresses²⁶. It was also observed that the ratio (i.e., relative abundance of a group of TFs among all TFs in the genome) of MYB genes in higher plants are remarkably higher than those in fungi or animals⁷. Intriguingly, the ratios of the MYB group of TFs in *Nannochloropsis* spp., 28%, 30% and 36% in IMET1, CCMP1779 and CCMP526 respectively, are even higher than those of the higher plants such as *Arabidopsis thaliana* (12%), *Glycine max* (13%) and *Zea mays* (11%), which is perhaps indicative of its significant and broad roles in transcriptional regulation in microalgae. In higher plants, most MYB genes encode proteins of the R2R3-MYB class (e.g., *Arabidopsis thaliana* harbors 131 R2R3-MYB genes yet only five R1R2R3-MYB genes⁷), while R1R2R3-MYB proteins are the norm in animals. The plant-specific R2R3 organization is usually thought to have evolved from an R1R2R3-type ancestral gene by the loss of the first repeat²⁷; however, the evolution of 3R-MYB genes from R2R3-MYB genes by the gain of the sequences encoding the R1 repeat through an ancient intragenic duplication has also been proposed²⁸. Intriguingly, the proportions of R1R2R3-MYB genes are significantly higher in the three *Nannochloropsis* strains (e.g. 15 R2R3-MYB genes and 8 R1R2R3-MYB genes in IMET1), which appears to support the hypothesis of “loss” instead of “gain” in higher plants.

Comparison of TF-family profiles among 36 plant genomes. It has been proposed that alterations in the expression of TF-encoding genes serve as one major source of the diversity and changes that underlie higher plant evolution^{6,7}, however the potential link between genome-wide TF-family profiles and plant evolution remains elusive. To probe such a putative link, we compared the genome-wide TF-family profiles among 36 plants, including three *Nannochloropsis* strains, four red algae strains, nine green algae strains and 20 higher



plants in PlantTFDB. These 36 plants were classified into four phylogenetic lineages: *Nannochloropsis* (three strains in two species), green algae (*Chlorophyta*; nine species), red algae (*Rhodophyta*; four species) and higher plants (*Bryophyta*, *Lycopodiophyta*, *Dicotyledon* and *Monocotyledon*; totally 20 species). In total, 58 TF families were present in these 36 plants (Supplemental Dataset S2). However, only 17 TF families were present in all four lineages (Figure 1). Members of such “core plant TF-families” represent over 80% of all the TFs in *Nannochloropsis* and *Rhodophyta*, over 65% in *Chlorophyta* and about 50% in higher plants. These “core TFs” might play key roles in gene expression regulation of plants as they appear early in the evolution of plants. On the other hand, 16 TFs families (ZF-HD, NAC, GRAS, MIKC, EIL, RAV, TCP, HRT-like, LBD etc.; Figure 1) were found only in higher plants. It is possible that these “higher-plant specific TF-families” have emerged independently or diverged from other TF-families along the evolution of plant genomes.

Many fewer TF families were found in *Nannochloropsis* (26), *Rhodophyta* (26) and *Chlorophyta* (35) than in higher plants (ranging from 53 to 58; 57 on average). Moreover, no microalgae-specific TF-families were found. This likely reflects the unicellular life style and the aquatic environment of microalgae as compared to the much more intricate land environment of the multicellular land plants, although one cannot rule out the possibility of high false-negative rate of TF recognition in these microalgae (e.g., the HMM models in PlantTFDB were all derived from higher plants). On the other hand, within microalgae, most of the TF-families were shared by all the three lineages, except only a few TF families (which exist in one or two microalgal lineages) such as STAT, LFY and BBR-BPC (found in neither *Rhodophyta* nor *Chlorophyta*), and AP2, ERF, DBB, B3 and Whirly (absent in *Rhodophyta*) (Figure 1). These TFs that are specific to each of the microalgal lineages might contribute to differentiation of these lineages.

Principal component analysis (PCA) for these plant species based on their profiles of TF-families (with the ratios of 58 TF families as variables and the 36 plants as samples; Figure 2) revealed that: (i) The TF-family profiles of the four lineages (*Nannochloropsis*, green algae, red algae and higher plants) are all quite distinct from each other,

whereas those of the species within each of the lineage are more similar. In fact, the three microalgae lineages (*Nannochloropsis*, green algae and red algae) can be separated with higher plants on PC1 level (accounting for 42.5% of cumulative variance), while *Nannochloropsis*, green algae and red algae can be distinguished from each other on PC2 (with PC1 and PC2 together accounting for 54.5% of cumulative variance). (ii) The TF-family profiles of the three *Nannochloropsis* strains are most similar to those of red algae, which were known to be phylogenetically close to Eustigmatophyceae, yet least similar to those of higher plants. (iii) At PC1, the top five TF-families that are able to distinguish the algal species from higher plants are MYB, MYB_related, bZIP, C2H2 and C3H. At PC2, the top five drivers that separate the three microalgal lineages from one another are the TF-families of bHLH, NAC, ERF, C3H and WRKY. Therefore these TF-families appear to play particularly prominent roles within microalgae and between microalgae and higher plants, respectively.

The observations above suggested a potential link between TF-family profiles and organismal evolution. To test this hypothesis, for these 36 plant species, an organismal tree based on hierarchical clustering of their TF-family profiles was compared with a phylogenetic tree that was constructed based on the multiple-alignment of their 18S sequences (Figure 3A, 3B; Methods)^{29–32}. The two trees were correlated, as the 36 plants were divided into four main branches (*Nannochloropsis*, green microalgae, red microalgae and higher plants) in both trees; moreover, between the two trees, the topology of their green microalgae branches are quite similar (Figure 3A, 3B). Therefore, TF-family profiles can potentially be indicative of organismal phylogeny, suggesting that the co-evolution of species and their TF-profiles occurred largely at the level of TF-family.

TFs involved in lipid-related pathways. Recent genome sequencing and gene annotation studies of *Nannochloropsis* spp. have revealed genes involved in lipid production in these algae^{16,18,19}, however no attempts have been carried out to investigate regulators of lipid accumulation process. Therefore, a computational strategy was devised here to identify the TFs involved in lipid-related pathways. *Firstly*, for the three *Nannochloropsis* strains, the orthologs of TFs that were

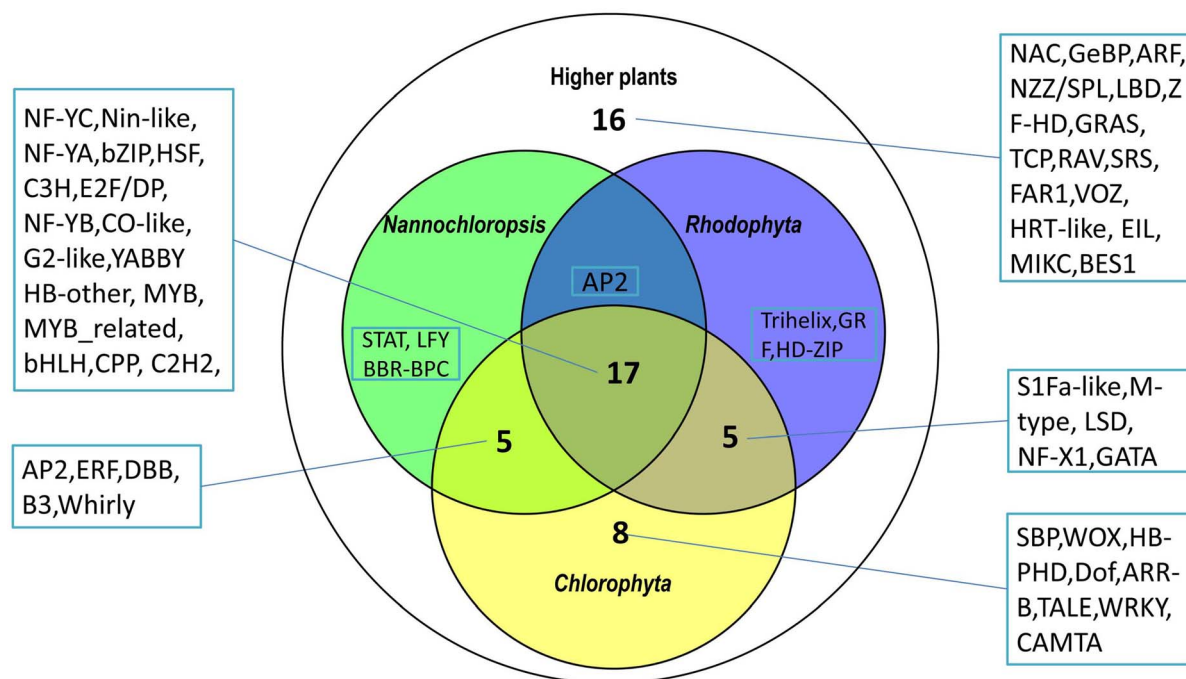


Figure 1 | TF-families composition in four species-lineages (*Nannochloropsis*, *Rhodophyta*, *Chlorophyta* and higher plants). Detailed background information about these TF-families can be accessed in the database PlantTFDB (<http://planttfdb.cbi.pku.edu.cn/>).

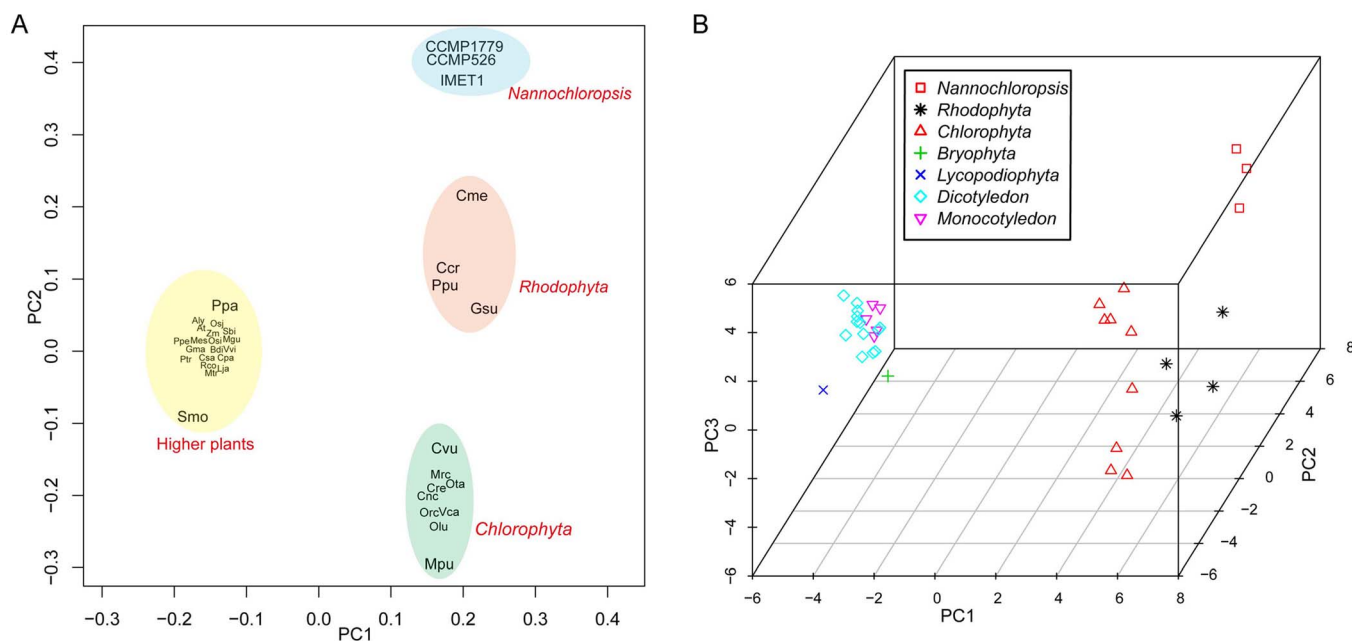


Figure 2 | PCA analysis of TF-family profiles in 36 plant species. Principal component analysis (PCA) was performed based on TF-family profiles, setting 58 TF families as variables and 32 plants as samples. (A) Two-dimensional PCA results based on PC1 and PC2. Each black name represents a plant species. (B) Three-dimensional PCA results based on PC1, PC2 and PC3. Abbreviations: (i) *Nannochloropsis* (IMET1, *N. oceanica* IMET1; CCMP1779, *N. oceanica* CCMP1779; CCMP526, *N. gaditana* CCMP526) (ii) *Rhodophyta* (Ppu, *Porphyridium purpureum*; Gsu, *Galdieria sulphuraria*; Cme, *Cyanidioschyzon merolae*; Ccr, *Chondrus crispus*) (iii) *Chlorophyta* (Cre, *Chlamydomonas reinhardtii*; Vca, *Volvox carterii*; Cnc, *Chlorella* sp. NC64A; Cvu, *Coccomyxa* sp. C-169; Mpu, *Micromonas pusilla* CCMP1545; Mrc, *Micromonas* sp. RCC299; Olu, *Ostreococcus lucimarinus* CCE9901; Orc, *Ostreococcus* sp. RCC809; Ota, *Ostreococcus tauri*) (iv) higher plants (Ppa, *Physcomitrella patens* subsp. *patens*; Smo, *Selaginella moellendorffii*; Aly, *Arabidopsis lyrata*; At, *Arabidopsis thaliana*; Cpa, *Carica papaya*; Csa, *Cucumis sativus*; Gma, *Glycine max*; Lja, *Lotus japonicas*; Mes, *Manihot esculenta*; Mtr, *Medicago truncatula*; Mgu, *Mimulus guttatus*; Ptr, *Populus trichocarpa*; Ppe, *Prunus persica*; Rco, *Ricinus communis*; Vvi, *Vitis vinifera*; Bdi, *Brachypodium distachyon*; Osi, *Oryza sativa* subsp. *indica*; Osj, *Oryza sativa* subsp. *japonica*; Sbi, *Sorghum bicolor*; Zm, *Zea mays*).

experimentally shown to be related to lipid accumulation in higher plants (including WRINKLED1 (WRI1) and GmDofc^{33,34}) were identified. The WRI1 family of TFs, which is a member of a plant-specific family of TFs (AP2/EREBP) that share either one or two copies of AP2 DNA-binding domain, serves as an important regulator of oil accumulation in maturing *Arabidopsis* seeds^{34–36}. Three, two and one genes were identified as putative WRI1 orthologs in *N. oceanica* IMET1 and *N. oceanica* CCMP1779 and *N. gaditana* CCMP526 respectively (via PSI-blast with the E-value cutoff at 1E-5) (Supplemental Table S3).

Secondly, putative lipid-synthesis-related TFs in IMET1 were identified via co-expression analysis between TFs and 118 lipid-synthesis-related genes based on our time-series transcriptome dataset that tracked the TAG accumulation process for 48 hours upon nitrogen depletion²³ (Supplemental Dataset S3, Methods). In the end, 27 putative lipid-related TF genes were identified, which were from 11 TF families that include NF-YC, bZIP, HB-other, HSF, C3H, E2F/DP, AP2, MYB_related, CPP, MYB and LFY (Supplemental Table S4). Three of these TF families (NF-YC, C3H and E2F/DP) were found enriched in lipid-related pathways (p -value ≤ 0.05).

Together, our analyses uncovered 30 lipid-related transcriptional factors in *N. oceanica* IMET1, which were found in 11 TF-families (Table 1). Among them, MYB_related (five), NF-YC (five), AP2 (four) and C3H (four) are the dominating families. The functional role of these TFs, which are presented to be involved in lipid metabolism, requires experimental characterization.

Genome-wide identification and functional analysis of TFBS in *N. oceanica* IMET1 via phylogenetic footprinting. An improved pipeline for whole-genome prediction of TFBS motifs via comparative

genomics. TFBS motifs are short genomic DNA segments that play important roles in gene regulation by modulating gene activities through their interaction with TFs. A widely used computational strategy for TFBS motifs identification is to detect over-represented and conserved patterns that might be good candidates for being TFBSs from promoter regions of co-regulated or co-expressed genes of a single genome^{37,38}. Such methods, however, are not suitable for non-model organisms such as *Nannochloropsis* spp., due to the paucity of transcriptomic or Chip-Seq resources. Alternative methods, such as phylogenetic footprinting, consider conserved patterns in the promoter regions of orthologous genes among several orthologous species as putative TFBS motifs, as functional sequences in promoter regions usually evolve slower than non-functional sequences due to the selective pressure^{9,10,39,40}. Such methods have been proved efficient in detecting TFBS motifs with biological significance in several studies^{9,10,17,41–43}. The major advantage of phylogenetic footprinting over the co-regulated genes approach is that it's possible to identify motifs on a genome-wide level based on orthologous sequences of considered genomes, while the latter requires a reliable method for identifying co-regulated genes⁴⁴. The recent availability of seven *Nannochloropsis* strains in five species^{16,18,19} therefore provides an opportunity for TFBS-identification via phylogenetic footprinting.

To identify the TFBSs in *Nannochloropsis* spp., we devised an improved pipeline for TFBSs identification based on phylogenomic footprinting, which represented an improvement based on the algorithms of MERCED¹⁷, by extending the comparative genomics method of two species into multi-species and applying parallelization programming to improve the computing performance. The genome sequences of six strains (from five *Nannochloropsis* species) that

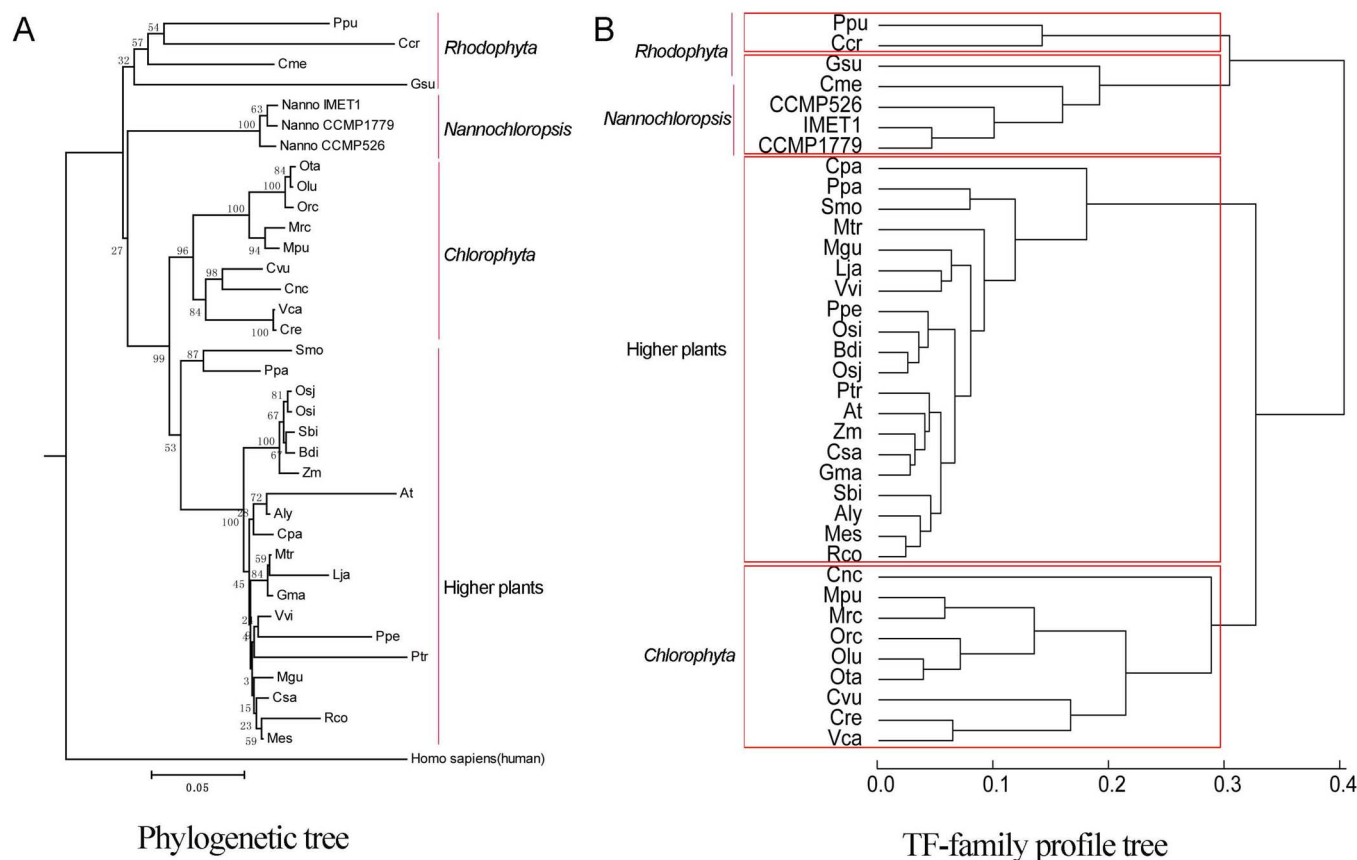


Figure 3 | Phylogenetic tree and TF-families profile tree of 36 plants. (A) The phylogenetic tree was constructed, using neighbor-joining method with bootstrap test (100 replicates) in MEGA5.2, based on the multiple-alignment of 18S sequences of these 36 selected plants using MUSCLE version 3.8 with default parameter, setting *Homo sapiens* as the compared species. The scale bar in the bottom represents the genetic changes of such amount (nucleotide substitutions per site). (B) The hierarchical cluster tree was constructed based on the ratio of each TF-family in these 36 species, using the “hclust” (average method) function for R language. The scale bar in the bottom represents the euclidean distance of such amount counted by the TF family profiles. The abbreviations are listed in Figure 2 legend.

included *N. oceanica* IMET1, *N. oculata* CCMP525, *N. gaditana* CCMP526, *N. granulata* CCMP529, *N. oceanica* CCMP531 and *N. salina* CCMP537 were used for TFBSs identification, employing the IMET1 genome as the “reference genome” and the other five as “query genomes”. This pipeline consists of five steps: (i) Orthologous gene groups among these six *Nannochloropsis* strains were first identified, by applying PSI-BLAST⁴⁵ (E-value cutoff at 1E-5) to all protein sequences of IMET1 and those of each “query genome”. (ii) A substitution matrix was constructed to model the neutral evolution rate of nucleotide substitution between IMET1 and each “query genome”⁴⁶. (iii) Conserved *k*-mers in the promoter sequences of orthologous genes were obtained between IMET1 and each “query genome”; (iv) Conserved *k*-mers were clustered using hierarchical clustering with average linkage⁴⁷. (v) The clustered TFBSs patterns were converted into a series of Position Frequency Matrix (PFM), each of which characterized a TFBS motif (See Methods for details).

The predicted TFBSs are consistent with experimentally verified TFBSs in databases. Analysis of the promoter regions of the six *Nannochloropsis* genomes thus revealed 68 TFBS motifs (8-mer) that were shared by all the strains (which were called “most-conserved” TFBS motifs), whereas 382 TFBS (8-mer) motifs shared by at least five strains (Supplemental Dataset S4). To test the specificity of our predicted TFBSs, our computationally determined TFBS motifs were then compared to the experimentally verified motifs in TRANSFAC (<http://www.gene-regulation.com/index2>)²⁴ and PLACE

(<http://www.dna.affrc.go.jp/PLACE/>)⁴⁸ using STAMP⁴⁹. TRANSFAC provides one of the most comprehensive collections of experimentally determined TFBSs and positional weight matrices, where 2173 TFBS motifs including 199 plant TFBSs were cataloged²⁴. PLACE compiled 469 experimentally verified TFBS motifs in plants, which also serves as an important reference for the studies of plant TFBSs⁴⁸. STAMP, which evaluated motif similarities, aligns input motifs against the chosen database (or alternatively against a user-provided dataset), and returns a lists of the highest-scoring matches⁴⁹.

Among our 68 predicted “most-conserved” TFBSs motifs, 46 (67%) were similar to TRANSFAC motifs (STAMP E-value cutoff as 1E-5). Among these 46 TFBSs motifs, 22 were related to experimentally verified motifs of plants, while the other 24 to motifs of vertebrata, insect, fungus and nematode (Supplemental Dataset S5). For example, for the predicted motif Nanno_M53, the consensus sequence of its reverse complement (GGCACGKG) is similar to TRANSFAC motif bHLH66_M01054 GCACGTGB (E-value = 1.22E-09), which is a TFBS motif found in *Arabidopsis* that regulates root hair elongation⁵⁰. On the other hand, 37 (54%) of the 68 “most-conserved” motifs were similar to the PLACE motifs. Taking predicted motif Nanno_M27 for example, the consensus sequence of its reverse complement (CCACGTMC) is similar to PLACE motif ABRE3HVA22 (E-value of 4.3E-11), which is the *cis*-acting element of an abscisic acid(ABA)-inducible gene⁵¹. Collectively, these evidences suggested that our predicted TFBS motifs appeared to be largely consistent with experimentally verified TFBS motifs in higher plants.


Table 1 | Lipid-related transcriptional factors predicted in *N. oceanica* IMET1

TF_ID	TF family	Methods of identification
s140.g4548	AP2	Ortholog of WR11
s271.g8196	AP2	Ortholog of WR11
s295.g8604	AP2	Ortholog of WR11
s082.g3001	AP2	co-expression analysis
s247.g6616	HB-other	co-expression analysis
s295.g8571	C3H	co-expression analysis
s354.g10201	MYB_related	co-expression analysis
s247.g6812	E2F/DP	co-expression analysis
s140.g4738	MYB	co-expression analysis
s009.g534	C3H	co-expression analysis
s009.g795	MYB_related	co-expression analysis
s043.g2022	E2F/DP	co-expression analysis
s354.g10025	MYB	co-expression analysis
s043.g1664	MYB_related	co-expression analysis
s124.g6222	MYB_related	co-expression analysis
s127.g4328	NF-YC	co-expression analysis
s185.g5880	MYB_related	co-expression analysis
s330.g9297	C3H	co-expression analysis
s267.g7604	CPP	co-expression analysis
s259.g7292	NF-YC	co-expression analysis
s267.g7530	HSF	co-expression analysis
s026.g1365	NF-YC	co-expression analysis
s145.g5102	bZIP	co-expression analysis
s247.g6694	bZIP	co-expression analysis
s043.g1754	HSF	co-expression analysis
s145.g4903	NF-YC	co-expression analysis
s104.g3807	C3H	co-expression analysis
s267.g7598	LFY	co-expression analysis
s007.g251	NF-YC	co-expression analysis
s245.g6460	E2F/DP	co-expression analysis

Functional enrichment analysis of predicted TFBSs. We next probed the functions of the predicted TFBS-motifs in IMET1 via two different approaches (Methods). In the “TFBS-enrichment analysis”, those TFBSs enriched in promoters of specific functional gene clusters in IMET1, such as genes involved in the TAG (triacylglycerols) biosynthetic pathways and the photosynthesis pathways, were identified, which thus allowed pinpointing the TFBS-motifs specifically associated with these functions. Glycerolipids and specifically triacylglycerols (TAG) are the main target products in algae-based biofuels¹⁶. Totally 36 genes and 44 genes were found to be involved in “TAG assembly pathway” and “Fatty acid biosynthesis pathway” respectively in *N. oceanica* IMET1 (Supplemental Dataset S3). Nanno_M49 and Nanno_M51 were enriched in the genes cluster of “TAG assembly pathway”. Nanno_M55 was enriched in the genes cluster of “Fatty acid biosynthesis”. These TFBSs motifs were thus proposed to be implicated in lipid accumulation pathways in *Nannochloropsis*. Photosynthesis is also an essential pathway for biomass accumulation and biofuel production in *Nannochloropsis*. In total, 41 nuclear genes were found related to photosynthesis in IMET1, which encode components of photosynthetic linear electron transport chain, including photosystem (PS) I reaction center and extrinsic proteins, PS II reaction center and extrinsic proteins, chlorophyll binding proteins and photosynthetic electron transfer proteins (Supplemental Dataset S6). Four TFBS motifs were enriched, including Nanno_M4, Nanno_M5, Nanno_M19 and Nanno_M35, in these 41 genes, which thus may potentially be the binding sites of TFs that are related to photosynthesis.

In the Gene Ontology (GO)-enrichment analysis, the enriched GO terms and GO-slms (Generic GO slms)⁵² in the target genes of each predicted TFBS motif were identified to reveal the main function of the TFBS motif. For the 68 “most conserved” TFBSs motifs, totally 40 GO-slms on “biological process” level and 36 GO-slms on “molecu-

lar function” level were found enriched (Supplemental Dataset S7). Several GO-slms on the “biological process” level were enriched in multiple TFBSs motifs, including cellular amino acid metabolic process (GO:0006520), chromosome organization (GO:0051276) and sulfur compound metabolic process (GO:0006790). Two TFBS motifs (Nanno_M5, Nanno_M35) were likely involved in photosynthesis pathways, as the GO-slim of “photosynthesis” (GO:0015979) is significantly enriched in the target genes of these motifs. Six TFBS motifs were found involved in lipid synthesis pathways, as the GO-slim of “lipid metabolic process” (GO:0006629) is significantly enriched in the target genes of these motifs. In addition, “carbohydrate metabolic process” (GO:0005975) was enriched in target genes of three TFBSs motifs, while “cellular nitrogen compound metabolic process” (GO:0034641) in target genes of five TFBSs motifs (Supplemental Dataset S7). Thus these TFBSs motifs might also contribute to lipid accumulation process in *Nannochloropsis* spp.

A preliminary TF-TFBS interacting network of *Nannochloropsis oceanica* IMET1. Construction of a preliminary TF-TFBS interacting network for *N. oceanica* IMET1. To probe the relationship between these predicted TFs and TFBSs, the 125 TFs and the 68 “most-conserved” TFBSs in IMET1 were compared to the reference plant TF-TFBS pairs in TRANSFAC (Methods). The analysis yield 78 TF-TFBS interaction pairs that involved 35 TFs and 14 TFBSs motifs (Supplemental Dataset S8). Then the genes whose promoter sequences contained one or more of these 14 TFBSs motifs were identified, which were considered as the targets of the corresponding TFs in the interaction pairs. As a result, 18992 regulatory connections for 35 TFs that target 2801 genes were predicted. To reduce the false positive rate, for each regulatory connection, we computed the Pearson product-moment correlation coefficient and its statistical significance (p -value) based on the time-course transcriptomic data of IMET1 under N-depletion conditions (Methods). Only those with a significant correlation (p -value ≤ 0.05) are preserved. In the end, 2,386 regulatory connections between a TF and a gene in IMET were identified, among which 1,315 are positively correlated and 1,071 negatively correlated (Figure 4, Dataset S9).

We next compared our network with those connections between the TFs and their target genes in the *Arabidopsis thaliana* regulatory network database AtRegNet⁵³ (Methods). As the result, 76 connections in the IMET1 network were supported by AtRegNet, with 11 of them supported by the “confirmed” connections in AtRegNet (Table 2).

There are on average 68 gene targets per TF in the IMET1 network. However the number of gene targets for each TF varies widely (ranging from 1 to 250). Several TFs regulate only a small number of genes such as s043.g1656 (one target) and s355.g10347 (five targets), while others control a large number of targets in the network such as s259.g7362 (250 targets), s043.g2022 (230 targets) and s247.g6812 (221 targets). For example, the 250 targets of s259.g7362 encode proteins involved in biosynthetic process, cellular nitrogen compound metabolic process, transport, small molecule metabolic process, cellular amino acid metabolic process, etc., suggesting that this TF might regulate a wide range of functions in *Nannochloropsis*. Despite the lack of experimental evidence, it is possible that such TFs might be the “master regulators” in the IMET1 regulatory network.

The IMET1 regulatory network revealed 11 TFs that are potentially involved in the transcriptional regulation of TAG biosynthesis pathways (Figure 5). Among them, TFs of the bZIP family were dominant (five such genes). The binding site of s259.g7362[bZIP] was present in the promoter regions of four genes including the Acyl-CoA-binding proteins (ACBP) and 3-Ketoacyl-ACP synthase (KAS) in fatty acid biosynthesis pathway, the Long chain acyl-CoA synthetases (LCFACS) converting free fatty acid to acyl-CoA, and the Lysophospholipid acyltransferase (LPAT) transferring acyl-CoA to

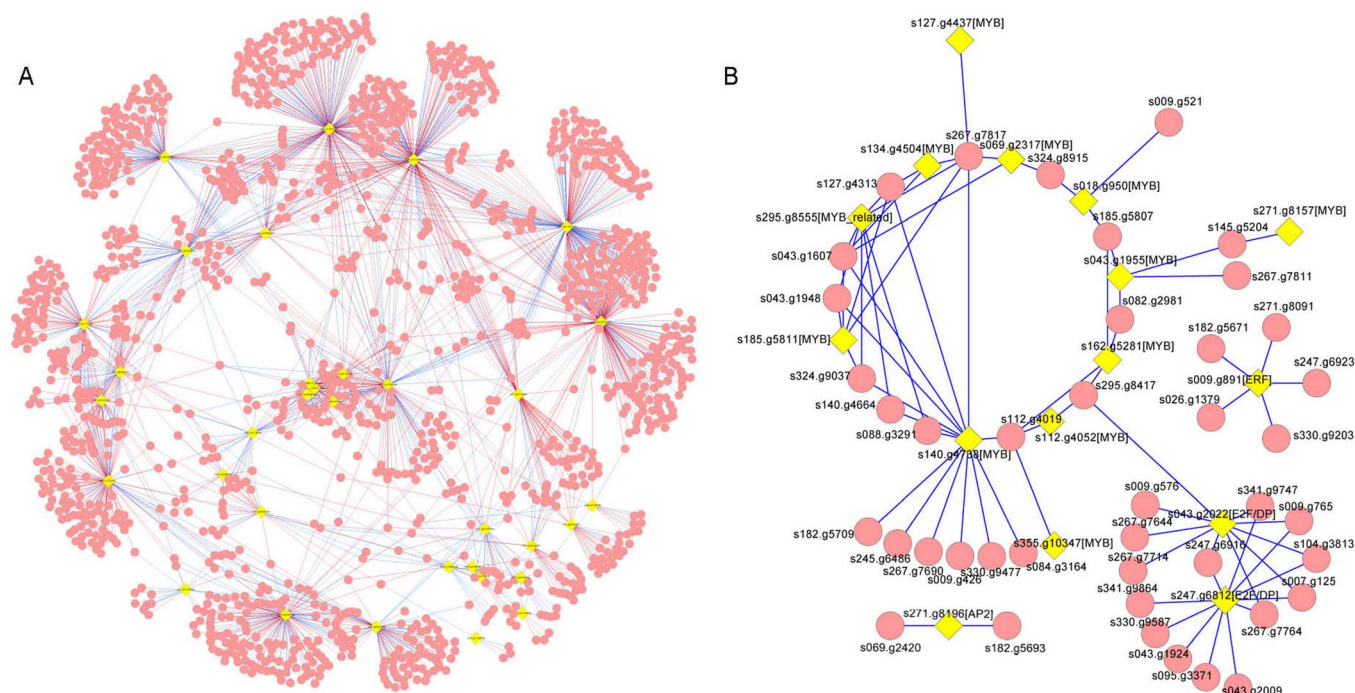


Figure 4 | An initial regulatory network of *N. oceanica* IMET1 based on the predicated TFs and TFBSs. (A) In this regulatory network, each “yellow quadrangle” represents a TF gene and each “red dot” represents a target gene. The length of the line between a TF and a target is the absolute value of correlation coefficient minus by one. (B) Regulatory connections in this sub-network are all verified by AtRegNet. Each “yellow quadrangle” represents a TF gene and each “red dot” represents a target gene. Network visualization was via the software package Cytoscape.

Lysophosphatidic acid to form Phosphatidic acid. Moreover, all these four genes were potentially up-regulated by s259.g7362[bZIP], in that the Pearson product-moment correlation coefficient between each of the genes and this TF is positive based on the time-course transcriptomic data. These observations suggested a potential co-regulation mechanism where this TF simultaneously controls the transcript levels of multiple enzymes to produce TAG under nitrogen-depletion conditions (Figure 5).

Such co-regulation by a single TF was also found in additional sets of genes that are directly involved in TAG-synthesis. Examples included the co-regulation of Lysophospholipid acyltransferase (*LPAT*) and Enoyl-ACP reductase (*ENR*) by s009.g891[ERF], and that of Phosphatidic acid phosphatase (*PAP*), Long chain acyl-CoA synthetases (*LCFACS*) and Type I 3-Ketoacyl-ACP synthase (*FAS-1*) by s295.g8604[AP2] (Figure 5). On the other hand, a number of genes might be each regulated by multiple TFs, such as *LPAT* (connected to three TFs) and the *DGAT-2A* (controlled by four TFs).

To allow readily access by the research community, the predicted TFs and TFBSs motifs in *Nannochloropsis* spp. and the preliminary regulatory network of *N. oceanica* IMET1 are displayed in a public website (<http://www.singlecellcenter.org/en/NannoRegulationDatabase/home.htm>).

Discussion

Genome-wide identification of TFs and TFBSs is one first step for dissecting gene regulation networks in oleaginous microalgae, and serves as the foundation of directed genetic engineering to enhance the lipid-synthesis process. Here employing *Nannochloropsis oceanica* IMET1 as a model, we present one of the first genome-wide TFs and TFBSs maps in oleaginous microalgae. Furthermore, a preliminary global regulation network that links TFs to their target genes was constructed.

Nannochloropsis spp. are promising feedstock for biofuel production. Genome sequencing and gene annotation studies in *Nannochloropsis* have revealed genes involved in lipid production in this

microalgae^{16,18,19}. However, it's crucial to identify TFs involved in lipid related pathways in this species, which serve as the master controls of gene regulation. From the 125 TFs identified in *N. oceanica* IMET1, we predicted 30 TFs which might be related to gene regulation of lipid synthesis processes, three of which are WR11 orthologs, while others are detected based on gene expression correlation analysis of mRNA-Seq data. These TFs, albeit the presence of false positives, can serve as the primary focuses for experimental tests.

The TF-family profiles of *Nannochloropsis* spp. revealed significant divergence of TF-family profiles among *Nannochloropsis*, *Chlorophyta*, *Rhodophyta* and higher plants. Within microalgae, the TF-family profiles of *Nannochloropsis* strains are most similar to those of red algae, and relatively distinct from those of green algae, which is consistent with their organismal phylogeny. TF-family profiles of green algae are more similar to those of higher plants, which is also consistent with organismal evolution in plants¹. Specifically, several TF families such as SBP and WRKY, which are usually large in size and vital for the gene regulation in higher plants, exist in green algae while are absent in *Nannochloropsis* spp. and red algae. SBPs form a major family of plant-specific TFs related to flower development⁵⁴, thus their specific emergence in green algae might underlie the development of the distinct reproduction modes in terrestrial plants. The WRKY family of TFs is one of the largest TF-families in plants and they are integral parts of signalling webs that modulate many plant processes⁵⁵, thus its emergence in green algae might contribute to the formation of cell signaling in modern land plants.

Moreover, a significant expansion of TF families from microalgae to terrestrial plants was observed, as evidenced by the 16 higher-plant specific TF families that were absent in the microalgae lineages. These higher-plant specific families are mainly involved in advanced regulation processes for multicellular organisms, such as mediation of auxin signaling by NAC to promote lateral root development in *A. thaliana*⁵⁶ and modulation of phyA-signaling homeostasis by FAR1 in higher plants⁵⁷. Confirmation of the expansion and elucidation of

Table 2 | Regulatory connections predicted in *N. oceanica* IMET1 that were verified by the confirmed connections in AtRegNet

Regulatory connections in the predicted <i>N. oceanica</i> IMET1 network (TF-target)	Regulatory connections in AtRegNet (TF-target)
s009.g891[ERF]-s026.g1379	AT4G36920-AT1G51460
s009.g891[ERF]-s026.g1379	AT4G36920-AT5G06530
s009.g891[ERF]-s182.g5671	AT4G36920-AT1G51460
s009.g891[ERF]-s182.g5671	AT4G36920-AT5G06530
s009.g891[ERF]-s247.g6923	AT4G36920-AT1G51460
s009.g891[ERF]-s247.g6923	AT4G36920-AT5G06530
s009.g891[ERF]-s271.g8091	AT4G36920-AT1G51460
s009.g891[ERF]-s271.g8091	AT4G36920-AT5G06530
s009.g891[ERF]-s330.g9203	AT4G36920-AT1G51460
s009.g891[ERF]-s330.g9203	AT4G36920-AT5G06530
s043.g1955[MYB]-s082.g2981	AT1G14350-AT3G54180
s043.g1955[MYB]-s082.g2981	AT1G14350-AT3G48750
s043.g1955[MYB]-s082.g2981	AT3G27920-AT5G64960
s043.g1955[MYB]-s082.g2981	AT2G02820-AT3G54180
s043.g1955[MYB]-s082.g2981	AT2G02820-AT3G48750
s140.g4738[MYB]-s267.g7690	AT3G27920-AT1G77670
s162.g5281[MYB]-s082.g2981	AT1G14350-AT3G54180
s162.g5281[MYB]-s082.g2981	AT1G14350-AT3G48750
s162.g5281[MYB]-s082.g2981	AT3G27920-AT5G64960
s162.g5281[MYB]-s082.g2981	AT2G02820-AT3G54180
s162.g5281[MYB]-s082.g2981	AT2G02820-AT3G48750
s271.g8196[AP2]-s069.g2420	AT4G36920-AT1G51460
s271.g8196[AP2]-s069.g2420	AT4G36920-AT5G06530
s271.g8196[AP2]-s182.g5693	AT4G36920-AT1G11050
s271.g8196[AP2]-s182.g5693	AT4G36920-AT3G45780

its mechanism and functional implication will be facilitated by analysis of genomes of additional plants and algae and experimental characterization of TFs in model plant and algal species. Previous studies reveal that alterations in the expression of TF-encoding genes serve as a major source of the diversity and changes that underlie evolution⁶. Here, the correlation between TF-family profiles and phylogenetic relationship in plants, as well as the distinct TF-family profiles between microalgae and land plants, suggested that the TF-family profiles also play a potential role in organismal evolution.

In this study, identification of TFBS motifs was carried out via comparative genomics of the six *Nannochloropsis* strains. Here we built connections of predicted TFs and TFBS motifs by comparison with reference plant TF-“TFBS motif” pairs in TRANSFAC, using correlation in gene expression to filter potential false positives. This network likely only represents a certain portion of the regulatory connections in this species, due to the limited number of experimental-confirmed TF-“TFBS motif” pairs in TRANSFAC. Regarding lipid accumulation processes in IMET1, 11 TFs were predicted to be involved in the transcriptional regulation of TAG biosynthesis pathway, and several genes in the pathway appeared to be regulated by multiple TFs. Efforts are currently ongoing to experimentally verify these predicted regulatory links.

In summary, this preliminary global regulatory network in an oleaginous microalga should help to prioritize the interactions between TF and their target genes for in-depth interrogation of regulatory links of interest. Moreover, these *in silico* efforts can guide experimental approaches such as Chip-Seq¹³ that promise to unveil the intricate regulatory interactions that underpin the robust TAG biosynthesis in this and related microalgae.

Methods

Gene models of six sequenced strains from five *Nannochloropsis* species.

Genomes, genes and proteins of *N. oceanica* IMET1, *N. oculata* CCMP525, *N. granulata* CCMP529, *N. oceanica* CCMP531 and *N. salina* CCMP537 were available at NCBI (BioProject accessions for genomes: PRJNA202418 for IMET1, PRJNA65107 for CCMP525, PRJNA65111 for CCMP529, PRJNA65113 for CCMP531 and PRJNA62503 for CCMP537)¹⁸. Genomes, genes and proteins of *N. gaditana* CCMP526 and *N. oceanica* CCMP1779 were downloaded from <http://www.nannochloropsis.genomeprojectsolutions-databases.com/> and <http://www.bmb.msu.edu/Nannochloropsis.html>, respectively. In total, 9052, 11973 and 9915 genes were obtained from CCMP526, CCMP1779 and IMET1 genomes, respectively.

nannochloropsis.genomeprojectsolutions-databases.com/ and <http://www.bmb.msu.edu/Nannochloropsis.html>, respectively. In total, 9052, 11973 and 9915 genes were obtained from CCMP526, CCMP1779 and IMET1 genomes, respectively.

TFs prediction was carried out in IMET1, CCMP526 and CCMP1779, mainly because the genome assemblies and gene annotations for these three strains are of higher quality than the other strains. To improve the accuracy of TFs prediction, genomes of all available strains except CCMP1779 were employed for TFBSs identification. The “promoter sequence” of a gene was defined as the upstream 1 kb sequence relative to the translation start site of the gene. “Translation start sites” instead of “transcription start sites” was adopted, because “translation start sites” could be obtained more reliably than “transcription start sites” especially in the draft genomes. Promoter sequences of all these *Nannochloropsis* strains can be accessed via our website (<http://www.singlecellcenter.org/en/NannoRegulationDatabase/home.htm>).

Prediction of TFs based on the method in PlantTFDB. To identify the TFs in *N. oceanica* IMET1, *N. oceanica* CCMP1779 and *N. gaditana* CCMP526, HMMER 3.0⁵⁸ was employed to search the characteristic domains of plant TFs in the proteins of each strain following the methods of PlantTFDB¹⁴. Details are as follows: (i) a total of 64 HMM models were used to identify the TF-related domains, of which 53 models were collected from Pfam 24.0⁵⁹ and 11 models were built by the PlantTFDB authors; (ii) HMMsearch in HMMER 3.0 package was employed to search the TF-related domains in the proteins of each strain, with *e*-value of 0.01 and the “domain-specific bit-score” from PlantTFDB as the threshold for domain identification; (iii) Each TF candidate was assigned into a specific TF-family based on the “family assignment rules” described in PlantTFDB. Details about the “64 HMM models”, “domain-specific bit-score” and “family assignment rules” used in TF identification can be found in¹⁴.

Phylogenetic tree for 36 plant species. The 18S sequences of 36 plant species were downloaded from Silva database (<http://www.arb-silva.de/>; released on August 23, 2013)⁶⁰. The phylogenetic tree was constructed as follows: (i) multiple sequence alignment was carried out by MUSCLE (-maxiters 100)²⁹ based on all 18S sequences; (ii) based on the alignment results, the NJ tree was constructed by MEGA5.2³¹ with bootstrap test (100 replicates), setting *Homo sapiens* as the outlier.

Identification of lipid-related TFs via co-expression analysis based on time-series transcriptome dataset. Our IMET1 genome annotation revealed 118 genes that were related to lipid-synthesis pathways (such as TAG assembly pathway, fatty acid desaturase pathway and fatty acid biosynthesis pathway). The correlation coefficient between each of the 125 TF genes and each of the 118 lipid-synthesis-related genes was calculated based on the temporal dynamics of transcripts in the triplicate cultures over the six time-points (3, 4, 6, 12, 24, and 48 h) under the nitrogen-depletion condition²³. A correlation was considered significant if the absolute value of the coefficient was over 0.8 and *p*-value not higher than 0.05 (Method below). A TF gene was considered as lipid-synthesis related if its transcript level during the time-course was significantly correlated with those of at least 30% of the 118 lipid-synthesis related genes in IMET1.

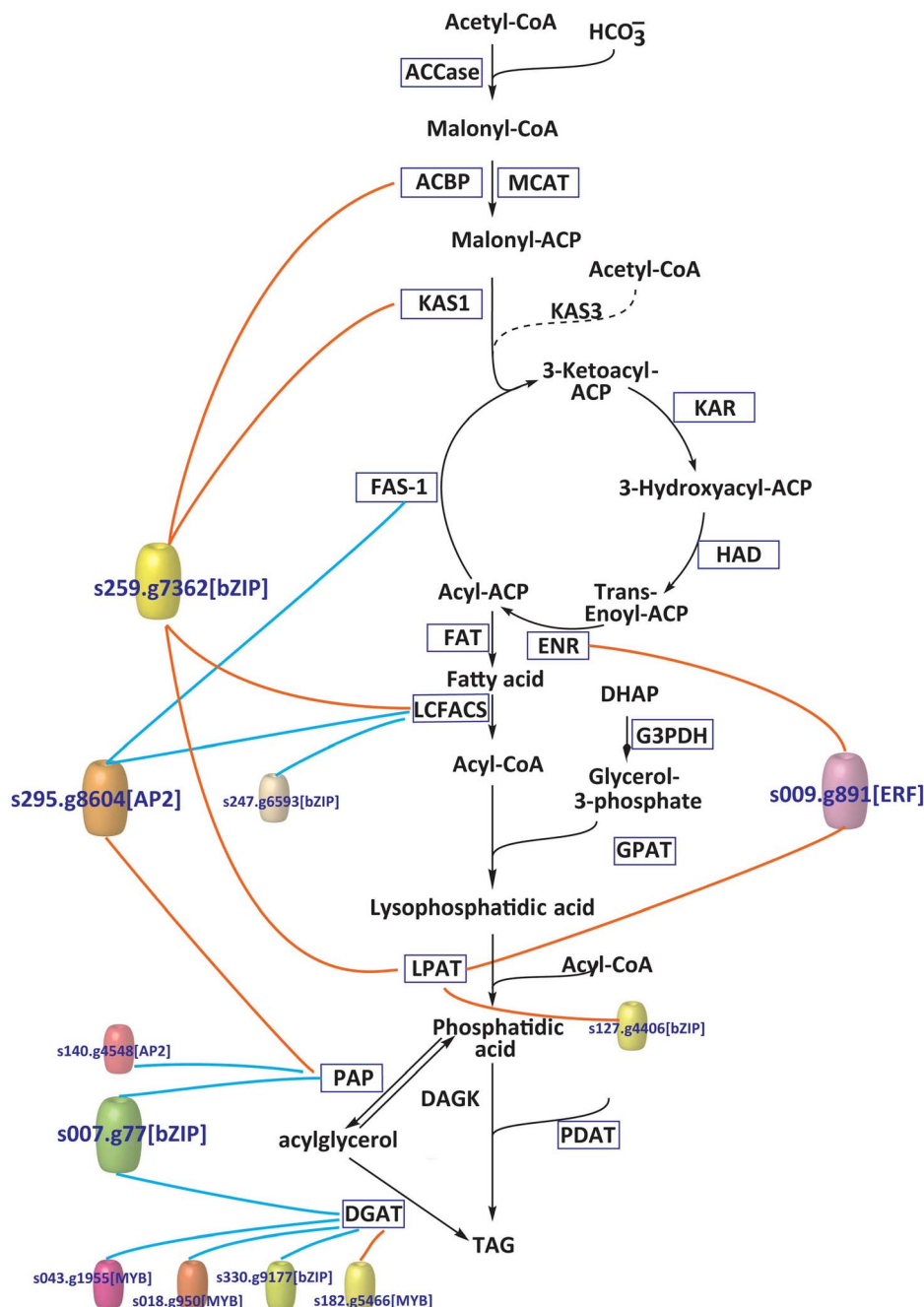


Figure 5 | The proposed transcriptional regulatory network of TAG biosynthesis in *N. oceanica* IMET1. This network reveals the genes involved in TAG biosynthesis pathway in *N. oceanica* IMET1, as well as those TFs predicted to regulate these genes. Each colored cylinder represents a TF. Red lines: “up-regulated”. Blue lines: “down-regulated”. Abbreviations: ACBP, Acyl-CoA-binding proteins; KAS, 3-Ketoacyl-ACP synthase; FAS-1, Type I 3-Ketoacyl-ACP synthase; ENR, Enoyl-ACP reductase; LCFACS, Long chain acyl-CoA synthetases; LPAT, Lyso-phospholipid acyltransferase; PAP, Phosphatidic acid phosphatase; DGAT, Diacylglycerol acyltransferase.

Pearson correlation coefficient (r) and p -value were used in assessing statistical significance in the prediction of lipid-synthesis-related TFs and the filtration of predicted regulatory connections based on the time-series transcriptome dataset of *N. oceanica* IMET1. The Pearson correlation coefficient (r) of two genes (their transcript level designated as x and y at the time point of i over the totally six time points sampled) was calculated as below:

$$r = \frac{\sum_{i=1}^6 (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^6 (x_i - \bar{x})^2 (y_i - \bar{y})^2}} \quad (1)$$

The Pearson correlation p -value of two genes were also calculated based on the Pearson’s correlation coefficient and a Student t distribution with four ($4 = 6 - 2$) degrees of freedom. The $\text{cor}()$ and $\text{cor.test}()$ functions in the R “stats” package were employed for the calculations above⁶¹.

TFBSs prediction based on comparative genomics of six *Nannochloropsis* strains. To identify the TFBSs in *Nannochloropsis*, we devised an improved pipeline based on phylogenetic footprinting. This pipeline utilized the MERCED¹⁷ algorithm as the computational core, and improved it by extending the comparative genomics method of two species into multi-species, and by applying parallelization programming to improve computing performance. The pipeline has four main steps: (i) IMET1 was chosen as “reference genome”, while genomes of other five strains were considered as “query genomes”. PSI-BLAST was performed (E-value cutoff of $1E-5$) between proteins of IMET1 and those of each query genome to get the “reciprocal best hit pairs”. Orthologous gene groups of all strains were defined as the intersection of all these five “strain-pair” orthologous gene (protein) sets. (ii) Substitution matrix was constructed between IMET1 genome and each query genome to describe the neutral evolution rate of nucleotides, based on the four-fold degenerate sites in orthologous proteins for strain pairs. MUSCLE (version 3.8)²⁹ was employed to align each of the orthologous proteins and obtain all the four-fold degenerate sites



with the same amino acid in the alignment. Additional details on the substitution matrix calculation method were as previously published⁴⁶. (iii) To identify the conserved k -mers (with the degree of conservation evaluated by statistical significance calculated from the corresponding nucleotide substitution matrices) in *Nannochloropsis* spp., we first identified conserved k -mers between IMET1 and one “query genome”, by adopting the method used in MERCED¹⁷. Then the “conserved k -mers” shared by more *Nannochloropsis* strains than our pre-set cutoff were obtained. For example, the “most-conserved TFBS” were defined as the k -mers shared by all six strains. Details about “conserved k -mers” identification in the promoters of the strain pairs were described in MERCED¹⁷. The k -mers’ length was chosen as 8 bp in our study, because the most dominant length of motifs in the TRANSFAC database is eight and several previous studies have successfully identified meaningful motifs in plants and other species using 8-mers^{17,62,63}. (iv) Hierarchical clustering algorithm⁴⁷ was applied to cluster the “conserved k -mers” to obtain TFBS motifs, which can be degenerated in some sites, employing a position weight matrix to represent the TFBS motif for each cluster⁶⁴.

TFBS enrichment analysis for genes clusters of specific functions. TFBS enrichment analysis was performed on gene clusters related to “photosynthesis”, “TAG assembly pathway” and “Fatty acid biosynthesis pathway” in *N. oceanica* IMET1. Genes related to photosynthesis were extracted from the genome based on blast-NR annotation results (E-value cutoff of 1E-05). Genes related to “TAG assembly pathway” and “Fatty acid biosynthesis pathway” were obtained from lipid pathway reconstruction of *N. oceanica* IMET1 based on lipid pathway in *Chlamydomonas reinhardtii* and *Saccharomyces cerevisiae*²³. The statistical significance of each TFBS motif in the functional gene cluster was calculated as follow. For example, for “TFBS Nanno_M0” (Dataset S4) in “photosynthesis genes cluster”, let “N” be the total number of genes in *N. oceanica* IMET1, “n” be the number of genes in “photosynthesis genes cluster”, “M” be the total number of target genes for Nanno_M0, and “m” be the number of target genes for Nanno_M0 in “photosynthesis genes cluster”. Then the p -value of the Nanno_M0 in “photosynthesis genes cluster” can be estimated based on the hypergeometric test:

$$P = \sum_{i=m}^{\min(n,M)} \frac{C(M,i)C(N-M,n-m)}{C(N,n)} \quad (2)$$

in which $C(x,y)$ is the combinational number of choosing y items out of x items. All the TFBSs with sufficient statistical significance (p -value ≤ 0.05) and the ratio (i.e., relative abundance of the targets in the total number of genes of the cluster) over 0.1 were selected as the enriched TFBSs.

Gene ontology (GO) enrichment analysis for each predicted TFBS motif. Gene ontology (GO) enrichment analysis was carried out for each TFBS motif to investigate its possible functions. GO annotation for all the genes of IMET1 was carried out by InterProScan⁶⁵. GO terms were mapped to the GO slim (Generic GO slim) hierarchy proposed by the GO consortium based on the goslim_generic.obo (version 1.2) file (<http://www.geneontology.org/GO.slims.shtml>). The statistical significance of each GO-slim term in the target genes of one given TFBS motif was calculated as follows. For example, for GO-slim term “lipid metabolic process (GO:0006629)” and motif “Nanno_M0”, let “N” be the total number of genes with GO annotation, “n” be the number of genes in targets of “Nanno_M0” with GO annotation, “M” be the total number of genes belonging to GO-slim term “lipid metabolic process (GO:0006629)”, and “m” be the number of genes in targets of “Nanno_M0” belonging to GO-slim term “lipid metabolic process (GO:0006629)”. Then, the p -value of the “lipid metabolic process (GO:0006629)” in target genes of “Nanno_M0” can be estimated based on the hypergeometric test:

$$P = \sum_{i=m}^{\min(n,M)} \frac{C(M,i)C(N-M,n-m)}{C(N,n)} \quad (3)$$

in which $C(x,y)$ is the combinational number of choosing y items out of x items. All GO-slim terms with sufficient statistical significance (p -value ≤ 0.05) and the ratio (i.e., relative abundance of the targets in the total number of genes of the cluster) over 0.1 were selected as the enriched terms.

Identification of TF-TFBS interaction pairs for constructing an initial regulation network in *N. oceanica* IMET1. To establish the genome-wide regulatory connections between our predicted TFs and target genes, we developed a method via comparison with the TRANSFAC database²⁴. First, our computationally determined TFBS motifs were mapped to all those 199 experimentally verified plant TFBS motifs in TRANSFAC using STAMP (E-value cutoff 1E-5)⁴⁹. For each of the TFBS motifs, the best-hit plant motif in TRANSFAC was defined as its “similar motif”. Next, the TF proteins binding to this “similar motif” were then extracted from the TRANSFAC database; and thus the orthologous proteins of these TF proteins were obtained from the collection of our predicted TF proteins (using PSI-BLAST; E-value cutoff at 1E-5)⁴⁵. Based on this procedure, all possible interaction pairs between our predicted TFs and TFBS motifs could be obtained. The target genes of a predicted TF were thus defined as all those genes whose promoters contained one or more target TFBSs.

Method to compare our network to the *Arabidopsis thaliana* regulatory network database AtRegNet. AtRegNet contains information on physical direct regulatory interactions between 8131 target genes, 64 TFs and three TF complexes, connected by 11355 edges, among which 769 interactions were classified as “confirmed

connections” (<http://arabidopsis.med.ohio-state.edu/moreNetwork.html>). Bidirectional Blast search was performed for the TFs and the target genes between the two networks respectively (E-value cutoff at 1E-05). A connection in the IMET1 network was considered “supported” by the corresponding one in AtRegNet if both the two criteria were met: (i) the TF in an IMET1 connection was the ortholog of the TF in the corresponding connection in AtRegNet; (ii) the target gene in the IMET1 connection was also the ortholog of the target gene in the corresponding connection in AtRegNet.

- Falkowski, P. G. *et al.* The evolution of modern eukaryotic phytoplankton. *Science* **305**, 354–360 (2004).
- Wijffels, R. H. & Barbosa, M. J. An outlook on microalgal biofuels. *Science* **329**, 796–799 (2010).
- Georgianna, D. R. & Mayfield, S. P. Exploiting diversity and synthetic biology for the production of algal biofuels. *Nature* **488**, 329–335 (2012).
- Li, Y. *et al.* *Chlamydomonas* starchless mutant defective in ADP-glucose pyrophosphorylase hyper-accumulates triacylglycerol. *Metab. Eng.* **12**, 387–391 (2010).
- Li, Y., Han, D., Hu, G., Sommerfeld, M. & Hu, Q. Inhibition of starch synthesis results in overproduction of lipids in *Chlamydomonas reinhardtii*. *Biotechnol Bioeng.* **107**, 258–268 (2010).
- Carroll, S. B. Endless forms: the evolution of gene regulation and morphological diversity. *Cell.* **101**, 577–80 (2000).
- Riechmann, J. *et al.* Arabidopsis transcription factors: genome-wide comparative analysis among eukaryotes. *Science* **290**, 2105–2110 (2000).
- Babu, M. M. & Teichmann, S. A. Evolution of transcription factors and the gene regulatory network in *Escherichia coli*. *Nucleic Acids Res.* **31**, 1234–1244 (2003).
- Kellis, M., Patterson, N., Endrizzi, M., Birren, B. & Lander, E. S. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* **423**, 241–254 (2003).
- Stark, A. *et al.* Discovery of functional elements in 12 *Drosophila* genomes using evolutionary signatures. *Nature* **450**, 219–232 (2007).
- Galas, D. J. & Schmitz, A. DNAase footprinting a simple method for the detection of protein-DNA binding specificity. *Nucleic Acids Res.* **5**, 3157–3170 (1978).
- Garner, M. M. & Revzin, A. A gel electrophoresis method for quantifying the binding of proteins to specific DNA regions: application to components of the *Escherichia coli* lactose operon regulatory system. *Nucleic Acids Res.* **9**, 3047–3060 (1981).
- Jothi, R., Cuddapah, S., Barski, A., Cui, K. & Zhao, K. Genome-wide identification of *in vivo* protein-DNA binding sites from ChIP-Seq data. *Nucleic Acids Res.* **36**, 5221–5231 (2008).
- Zhang, H. *et al.* plantTFDB 2.0: update and improvement of the comprehensive plant transcription factor database. *Nucleic Acids Res.* **39**, D1114–D1117 (2011).
- Pérez-Rodríguez, P. *et al.* PlnTFDB: updated content and new features of the plant transcription factor database. *Nucleic Acids Res.* **38**, D822–D827 (2010).
- Vieler, A. *et al.* Genome, functional gene annotation, and nuclear transformation of the heterokont oleaginous alga *Nannochloropsis oceanica* CCMP1779. *PLoS Genet* **8**, e1003064 (2012).
- Ding, J., Li, X. & Hu, H. Systematic prediction of *cis*-regulatory elements in the *Chlamydomonas reinhardtii* genome using comparative genomics. *Plant Physiol.* **160**, 613–623 (2012).
- Wang, D. *et al.* *Nannochloropsis* genomes reveal evolution of microalgal oleaginous traits. *PLoS Genet* **10**, e1004094 (2014).
- Radakovits, R. *et al.* Draft genome sequence and genetic transformation of the oleaginous alga *Nannochloropsis gaditana*. *Nat Commun.* **3**, 686 (2012).
- Wei, L. *et al.* *Nannochloropsis* plastid and mitochondrial phylogenomes reveal organelle diversification mechanism and intragenus phylogeny strategy in microalgae. *BMC Genomics* **14**, 534 (2013).
- Sukenik, A. Ecophysiological considerations in the optimization of eicosapentaenoic acid production by *Nannochloropsis* sp. (Eustigmatophyceae). *Bioresour Technol.* **35**, 263–269 (1991).
- Wang, D. M., Lu, Y. D., Huang, H. & Xu, J. Establishing oleaginous microalgae research models for consolidated bioprocessing of solar energy. *Adv Biochem Eng Biotechnol.* **128**, 69–84 (2012).
- Li, J. *et al.* Choreography of transcriptomes and lipidomes of *Nannochloropsis* reveals the mechanisms of oil synthesis in microalgae. *Plant Cell tpc.* **113**.121418 (2014).
- Matys, V. *et al.* TRANSFAC® and its module TRANSCOMP®: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res.* **34**, D108–D110 (2006).
- Harrison, S. C. A structural taxonomy of DNA-binding domains. *Nature* **353**, 715–719 (1991).
- Dubos, C. *et al.* MYB transcription factors in *Arabidopsis*. *Trends Plant Sci.* **15**, 573–581 (2010).
- Rosinski, J. A. & Atchley, W. R. Molecular evolution of the Myb family of transcription factors: evidence for polyphyletic origin. *J Mol Evol.* **46**, 74–83 (1998).
- Jiang, C., Gu, J., Chopra, S., Gu, X. & Peterson, T. Ordered origin of the typical two- and three-repeat Myb genes. *Gene* **326**, 13–22 (2004).
- Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).



30. Saitou, N. & Nei, M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol.* **4**, 406–425 (1987).
31. Tamura, K. *et al.* MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol Biol Evol.* **28**, 2731–2739 (2011).
32. Pruesse, E. *et al.* SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res.* **35**, 7188–7196 (2007).
33. Wang, H. W. *et al.* The soybean Dof-type transcription factor genes, GmDof4 and GmDof1, enhance lipid content in the seeds of transgenic *Arabidopsis* plants. *Plant J.* **52**, 716–729 (2007).
34. Focks, N. & Benning, C. Wrinkled1: a novel, low-seed-oil mutant of *Arabidopsis* with a deficiency in the seed-specific regulation of carbohydrate metabolism. *Plant Physiol.* **118**, 91–101 (1998).
35. Baud, S., Wuilleme, S., To, A., Rochat, C. & Lepiniec, L. Role of WRINKLED1 in the transcriptional regulation of glycolytic and fatty acid biosynthetic genes in *Arabidopsis*. *Plant J.* **60**, 933–947 (2009).
36. Cernac, A. & Benning, C. WRINKLED1 encodes an AP2/EREB domain protein involved in the control of storage compound biosynthesis in *Arabidopsis*. *Plant J.* **40**, 575–585 (2004).
37. Bailey, T. L. *et al.* MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.* **37**, W202–W208 (2009).
38. Sinha, S. & Tompa, M. YMF: a program for discovery of novel transcription factor binding sites by statistical overrepresentation. *Nucleic Acids Res.* **31**, 3586–3588 (2003).
39. Fang, F. & Blanchette, M. FootPrinter3: phylogenetic footprinting in partially alignable sequences. *Nucleic Acids Res.* **34**, W617–W620 (2006).
40. Blanchette, M. & Tompa, M. FootPrinter: a program designed for phylogenetic footprinting. *Nucleic Acids Res.* **31**, 3840–3842 (2003).
41. Cai, X. *et al.* Systematic identification of conserved motif modules in the human genome. *BMC Genomics* **11**, 567 (2010).
42. Lenhard, B. *et al.* Identification of conserved regulatory elements by comparative genome analysis. *J Biol.* **2**, 13 (2003).
43. Xie, X. *et al.* Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. *Nature* **434**, 338–345 (2005).
44. Das, M. K. & Dai, H.-K. A survey of DNA motif finding algorithms. *BMC Bioinformatics* **8**, S21 (2007).
45. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).
46. Li, W.-H., Wu, C.-I. & Luo, C.-C. A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. *Mol Biol Evol.* **2**, 150–174 (1985).
47. Sokal, R. R. A statistical method for evaluating systematic relationships. *Univ Kans Sci Bull* **38**, 1409–1438 (1958).
48. Higo, K., Ugawa, Y., Iwamoto, M. & Higo, H. PLACE: a database of plant cis-acting regulatory DNA elements. *Nucleic Acids Res.* **26**, 358–359 (1998).
49. Mahony, S. & Benos, P. V. STAMP: a web tool for exploring DNA-binding motif similarities. *Nucleic Acids Res.* **35**, W253–W258 (2007).
50. Bruex, A. *et al.* A gene regulatory network for root epidermis cell differentiation in *Arabidopsis*. *PLoS Genet* **8**, e1002446 (2012).
51. Shen, Q. & Ho, T. Functional dissection of an abscisic acid (ABA)-inducible gene reveals two independent ABA-responsive complexes each containing a G-box and a novel cis-acting element. *Plant Cell* **7**, 295–307 (1995).
52. Ashburner, M. *et al.* Gene Ontology: tool for the unification of biology. *Nat Genet.* **25**, 25–29 (2000).
53. Yilmaz, A. *et al.* AGRIS: the *Arabidopsis* gene regulatory information server, an update. *Nucleic Acids Res.* **39**, D1118–D1122 (2011).
54. Yamasaki, K. *et al.* A novel zinc-binding motif revealed by solution structures of DNA-binding domains of *Arabidopsis* SBP-family transcription factors. *J Mol Biol.* **337**, 49–63 (2004).
55. Rushton, P. J., Somssich, I. E., Ringler, P. & Shen, Q. J. WRKY transcription factors. *Trends Plant Sci.* **15**, 247–258 (2010).
56. Xie, Q., Frugis, G., Colgan, D. & Chua, N.-H. *Arabidopsis* NAC1 transduces auxin signal downstream of TIR1 to promote lateral root development. *Genes Dev.* **14**, 3024–3036 (2000).
57. Lin, R. *et al.* Transposase-derived transcription factors regulate light signaling in *Arabidopsis*. *Science* **318**, 1302–1305 (2007).
58. Finn, R. D., Clements, J. & Eddy, S. R. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res.* **39**, W29–W37 (2011).
59. Finn, R. D. *et al.* The Pfam protein families database. *Nucleic Acids Res.* **38**, D211–D222 (2010).
60. Quast, C. *et al.* The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* **41**, D590–D596 (2013).
61. R Development Core Team. *R: A Language and Environment for Statistical Computing*. (R Foundation for Statistical Computing, Vienna, Austria, 2012).
62. Mariño-Ramírez, L., Spouge, J. L., Kanga, G. C. & Landsman, D. Statistical analysis of over-represented words in human promoter sequences. *Nucleic Acids Res.* **32**, 949–958 (2004).
63. Yamamoto, Y. Y. *et al.* Prediction of transcriptional regulatory elements for plant hormone responses based on microarray data. *BMC Plant Biol.* **11**, 39 (2011).
64. Stormo, G. D. & Hartzell, G. W. Identifying protein-binding sites from unaligned DNA fragments. *Proc Natl Acad Sci U S A.* **86**, 1183–1187 (1989).
65. Hunter, S. *et al.* InterPro in 2011: new developments in the family and domain prediction database. *Nucleic Acids Res.* **40**, D306–D312 (2012).

Acknowledgments

We are grateful to Gao Ge from Peking University for discussion and help for the methodology on TFs in *Nannochloropsis*. This work was supported by National Basic Research Program from Ministry of Science and Technology of China (2011CB200902; 2012CB721101), International Research Collaboration Program (31010103907) and Young Investigator Program (61103167) from National Natural Science Foundation of China and International Innovation Partnership Program and Solar Energy Research Initiative from Chinese Academy of Sciences.

Author contributions

J.X. and K.N. designed the research; J.H. performed data analysis; J.L. contributed to collection of the transcriptome data; D.W. contributed to Figure 5; G.J. designed the website; J.H., J.X. and K.N. wrote the paper.

Additional information

Supplementary information accompanies this paper at <http://www.nature.com/scientificreports>

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Hu, J.Q. *et al.* Genome-wide identification of transcription factors and transcription-factor binding sites in oleaginous microalgae *Nannochloropsis*. *Sci. Rep.* **4**, 5454; DOI:10.1038/srep05454 (2014).



This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder in order to reproduce the material. To view a copy of this license, visit <http://creativecommons.org/licenses/by-nc-sa/4.0/>