非モデル緑藻の光化学系解析のための RNA seq 解析

亀尾辰砂 [1,2], 高林厚史 [1,2]

[1] 北海道大学低温科学研究所
[2] 北海道大学大学院環境科学院
〒060-0819 札幌市北区北 19 条西 8 丁目

RNA seq analysis for photosystem analysis of non-model green algae

Shinsa Kameo[1,2], Atsushi Takabayashi[1,2]

[1]Institute of Low Temperature Science
[2]Graduate School of Environmental Science, Hokkaido University,
Kita-19, Nishi-8, Kita-ku, Sapporo, Hokkaido, 060-0819 Japan

Keywords: green algae, Iso-seq, non-model organism, photosystem, RNA seq

DOI: 10.24480/bsj-review.15b3.00261

１．はじめに

約 20 年にわたる次世代シークエンサーの絶え間ない技術革新により，ゲノム解析や RNA seq 解析のコストは私たちのような wet な研究室の手の届くところまで降りてきた。一方で，データ解析の手法やノウハウについてはまだ十分に行き渡ったとは言えないのが現状と思われる。

私たちの研究室では長年にわたり，モデル光合成生物を用いた逆遺伝学的な解析を中心にクロロフィル代謝や光合成の研究を行ってきた。しかし近年では光合成生物の環境適応能の多様性に興味が移りつつあり，非モデル光合成生物を用いた研究が研究室の軸になってきた。そこで問題になったのが配列情報の不足であった。端的に言えば，私たちの解析に配列情報が必須であるものの，私たちにはその配列情報を得るためのノウハウがなかったのである。そこで，私たちは，この５年間，手探りで RNA seq を進めてきた。試行錯誤を繰り返し，ロングリードやショートリードを組み合わせた RNA seq 解析を行ってきた結果，私たちは，非モデル光合成生物の遺伝子/タンパク質の配列や機能を予測するとともに，トランスクリプトームの変動を見ることも可能になった。これらの解析結果は，私たちの研究プロジェクトの基盤となっている。

本稿では，私たちが研究材料としている非モデル光合成生物の魅力や研究意義，そして RNA seq 解析を利用した研究成果を紹介するとともに，現行の解析スキームについても紹介する。

２．モデル生物を用いた光合成研究が主流の中，非モデル緑藻であるプラシノ藻類やストレプト藻類に着目した理由

ゲノム解読が行われる前のモデル真核光合成生物は，クロレラ，エンドウ，ホウレンソウなど培養や栽培が容易であり，同時に，葉緑体やチラコイド膜の単離が容易な生物種であっ

た。そのような研究状況の中で，最初に真核光合成生物のゲノム解読が光合成研究に大きな影響を与えたのは，タバコおよびゼニゴケの葉緑体ゲノムの解読（Shinozaki et al. 1986; Ohyama et al. 1986）だろう。葉緑体ゲノムの解読により光化学系（PS）I や II の反応中心タンパク質群などの配列情報が明らかになったことに加え，その後確立された葉緑体ゲノムの形質転換系を組みわせることで，葉緑体ゲノムコードの機能未知遺伝子群の解析も大きく進展することになった。

次の飛躍は言うまでもなく 2000 年のシロイヌナズナの全ゲノム解読である（Arabidopsis Genome Initiative 2000）。配列情報が得られること，形質転換系が容易であること，またノックアウトラインが容易に入手できること，など今日に至るまでモデル植物の代表であるシロイヌナズナは瞬く間に光合成研究のモデル植物として利用されるようになり，数多くの研究がこの植物を材料として行われた。また，光合成研究のモデル緑藻としてはクラミドモナスが広く用いられている。

また近年特筆すべきこととして，CryoEM 技術の発展により，より幅広い生物種で光化学系の構造解析が行われるようになった。特に陸上植物やコア緑藻類の光化学系については精力的に構造解析が行われており，その結果としてコケ植物の PSI の構造が維管束植物とは異なること，コア緑藻類の光化学系には予想以上の多様性が見られること，などが明らかになってきた（Suga and Shen 2020; Bai et al. 2021）。しかし，依然として，ストレプト藻類やプラシノ藻類の光合成の研究は進んでおらず，光化学系の構造が明らかになっているのは，プラシノ藻 *Ostreococcus tauri* の PSI（Ishii et al. 2023）のみである。そこで，私たちは，ストレプト藻類やプラシノ藻類の光化学系の解析を進めている。

では，光合成研究材料としてのストレプト藻類やプラシノ藻類の魅力はどこにあるのだろうか？まずストレプト藻類に関しては，陸上植物の祖先であるということは大きな魅力である（図１）。陸上植物の誕生は地球の生命史の中でも最大級のイベントであり，陸上環境や大気環境を大きく変えることになった。陸上植物へと進化したのは淡水性緑藻のストレプト藻類の一系統であると考えられている (Bowles et al. 2020）。しかし，ストレプト藻類から顕花植物への進化に伴う光合成システムの変遷については未解明の点が多い。緑藻と顕花植物の中間的な形質を持つコケ植物の光化学系の解析はまさにその観点から興味深いものであり，実際にコケ植物の光化学系の構造には維管束植物やコア緑藻類との違いも見出されてきた(Iwai et al. 2015; Pinnola et al. 2018）。ではコケ植物とストレプト藻類の光化学系に違いはあるのだろうか？私たちは，後述のように，最も早く分岐した「古い」ストレプト藻類の *Mesostigma viride* の光化学系はコア緑藻類と陸上植物の中間的な形質を持っていると考えている。そして，その後の陸上植物に至るまでの進化の道筋において光化学系がどのような変遷を遂げているのかは不明である。そのため，筆者は，今後ストレプト藻類からコケ植物への進化に伴う光合成システムの変遷についても，研究が進展する中で，興味や関心を集めるのではないかと考えている。

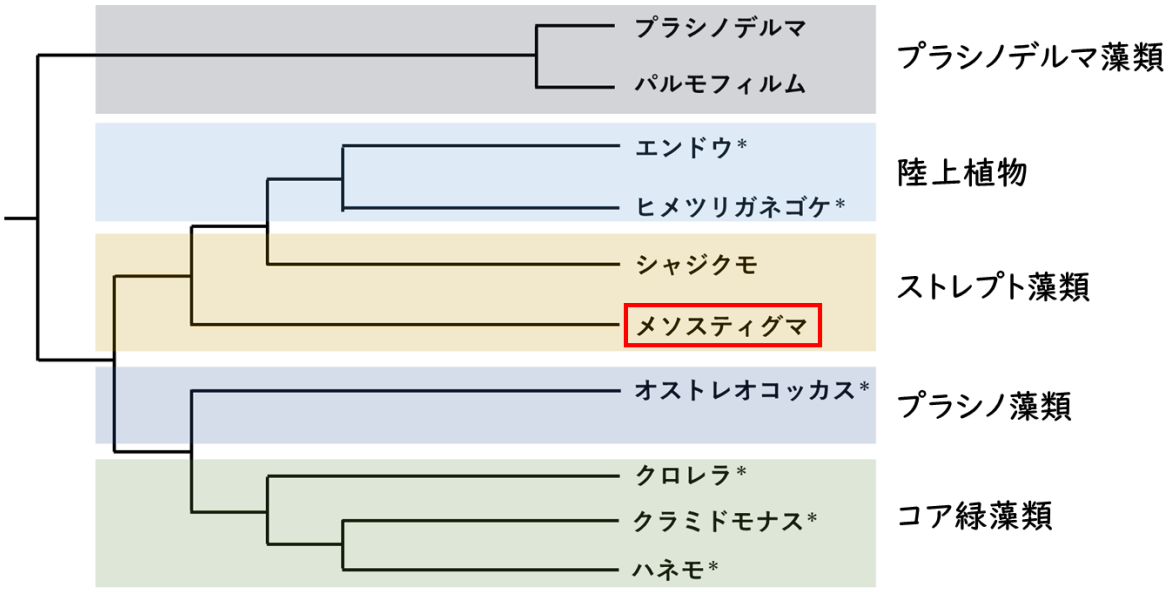

図 1　緑色植物（緑藻と陸上植物）の系統関係

アスタリスクは PSI もしくは PSII の構造が既に公開されている種を示している。プラシノ藻類ではオストレオコッカス（*Ostreococcus tauri*）の PSI のみが公開されており，ストレプト藻類に関しては現時点では光化学系の構造が公開された種はない。

一方のプラシノ藻類は，コア緑藻類の祖先にあたる系統群であり，おもに海洋性の緑藻である（図 1）。光合成という観点から特に興味深いのは，光合成色素の多様性である。まず，カロテノイド色素の多様性は光合成生物の中でも特筆すべきレベルである（Latasa et al. 2004）。また，クロロフィルの代謝中間体でありクロロフィル *c* とよく似た吸収スペクトルを持つジビニルプロトクロロフィリド（MgDVP）を例外的に集光に用いることも知られている（Ishii et al. 2023）。このような光合成色素の多様性はおそらくは海洋の多様な光環境への適応機構を反映したものである。また，集光アンテナタンパク質にもユニークな特徴があり，LHCP（prasinophyte-specific LHC）はプラシノ藻類に特有の LHC（light-harvesting complex）タンパク質である（Ishii et al. 2023）。この LHCP は通常の LHC よりも多くのカロテノイド色素を結合すると報告されており，弱光環境の海洋での集光に有利な LHC であると考えられる。一方で LHCP を持たないプラシノ藻類も存在すること，カロテノイドの多様性，などから判断すると，プラシノ藻類の集光系には高度な多様性があるのではないかと予想される。

このように，ストレプト藻類やプラシノ藻類には他の緑藻や陸上植物と異なる興味深い特長があるものの，現時点では知見が不足している。そして，そのことが緑藻の光化学系の進化や環境適応機構についての理解を妨げている。そこで，著者らはストレプト藻類やプラシノ藻類の中で，比較的初期に分岐した種にフォーカスして，光化学系の構造解析を進めている。

３．ショートリードを用いた RNA seq 解析とロングリードを用いた RNA seq 解析

非モデル緑藻の光化学系の解析を行う上で問題になるのはゲノム情報がないことである。緑藻のゲノムプロジェクトは急速に広がり着実に進んでいるものの，まだ陸上植物と比べるとゲノムが明らかになった種は限られている。そのため，著者らは興味のある緑藻の光合成遺伝子の配列を入手するために，ロングリードとショートリードを組み合わせた解析を行っている。

まず，ショートリードを用いた RNA seq 解析のメリットは，リード数が多く，配列正確性が高く，コストが安いことである。一方デメリットとしてはリード長が短いため，非モデル生物で RNA の全長を読むためには *de novo* assembly が必要になることが挙げられる。*De novo* assembly はマシンパワーを要求する解析であり，一般的なノート PC では解析が難しい。また，遺伝子の全長を得ることは容易ではなく，しかもキメラ配列を比較的多く含むことにも注意が必要である。

一方，ロングリードを用いた RNA seq 解析（Isoform sequencing: Iso-seq）の特徴は，全長 cDNA が得られることである。配列の正確性はショートリードに比べれば低いものの，エラーはランダムに生じるため，クラスタリングによって得られたコンセンサス配列については高い正確性を得ることが可能である。ショートリードに比べてリード数は劣るため，発現量の低い mRNA の情報を得ることは難しいものの，光合成遺伝子のように比較的発現量の多い遺伝子についてはより長い配列情報を得ることが期待できる。

筆者は非モデル生物の光合成の解析のため，両方の解析を組み合わせることが多い。具体的には， Iso-seq 解析で構築した全長 cDNA をテンプレートとし，ショートリードを疑似マッピングすることで遺伝子発現解析を行うのが一例である。また， *de novo* assembly 解析は Iso-seq 解析では検出できなかった（比較的発現量が低い）遺伝子の検出や配列の正確性の検証にも有用である。

４．ロングリードを用いた RNA seq 解析の実際

筆者は PacBio sequel を利用した Iso-seq 解析をよく行っているので，ここではそれについて紹介する。まず，私たちが材料としている樹木や藻類の中には市販のキットで total RNA を抽出すると，多糖類の混入が多く，解析に十分な品質が得られない場合もめずらしくない。そこで私たちがよく利用しているのは，ニッポンジーン社の Assist buffer と ISOSPIN Plant RNA を組み合わせた抽出キットである。このキットは多糖類などが多く含まれているサンプルに対して非常に有効であり，実際にカエデなどの樹木類や緑藻類で成果を挙げている。また，さらに精製が必要な場合には Qiagen 社などのクリーンアップキットを使っている。

Iso-seq を受託解析する際には，複数の（～４種類の）生物種のサンプルを同時に依頼することが多い。これは１種あたりの解析コストを下げるためであり，著者らの研究対象が発現量の比較的多い光合成遺伝子であるため，仮にリード数が 1/4 になっても目的遺伝子が得られることを期待してのことである。ただ，リード数が多いとは言えない Iso-seq において，さらにリード数を下げることには葛藤もあり，それが理由で目的の遺伝子が検出できないこと

もあれば，断片的な遺伝子配列しか入手できないこともある。研究目的に応じて個別に判断する必要があると思われる。

データ解析も受託することが可能であるが，解析コストを下げるためデータ解析を頼まず，成果物の bam ファイルを自分で解析することが多い。その際，筆者は，Windows11 の WSL2（Windows Subsystem for Linux）に Ubuntu 環境を導入し，PacBio 社の Iso-seq パイプライン version 3（https://github.com/PacificBiosciences/pbbioconda）を利用して解析している。このパイプラインは Anaconda/Miniconda の Bioconda チャンネルに登録されているが，python3 ではなく python 2 を必要とするため，仮想環境に python 2.7 を導入する必要がある。Iso-seq 解析で得られた full-length cDNA 配列は GeneMarkS-T（Tang et al. 2015）や TransDecoder（https://github.com/TransDecoder/TransDecoder）などのソフトウエアでタンパク質アミノ酸配列に転換し，BLASTP（Camacho et al. 2009）や Diamond（Buchfink et al. 2021）などの相同検索ソフトを用いて自分の研究対象である光合成タンパク質を抽出している。もしくは，eggNOG-mapper（Cantalapiedra et al. 2021）や KofamKOALA（Aramaki et al. 2020）などのソフトウエアを利用して Gene Ontology（GO）や KEGG Orthology（KO）などの機能アノテーションを行い， full-length cDNA 配列をテンプレートとしショートリードを用いた発現解析に利用している。

５．ショートリードを用いた RNA seq 解析の実際

著者らはロングリードに加えて Illumina 社の NovaSeq や MGI 社の DNBSEQ などを利用したショートリードを用いた解析も行っている。目的の 1 つは *de novo* assembly を用いた遺伝子探索であり，著者らは Iso-seq 解析に加えてこれも併用している。例えば，筆者の主なターゲットの *LHC* 遺伝子群では *LHCB7* などの発現量の低い遺伝子が Iso-seq で単離できず *de novo* assembly で検出できることはめずらしくない。また，複数の手法で遺伝子の存在や配列を検証できること自体が有益である。

De novo assembly の際には，ソフトウエアによって結果が異なるため，複数のソフトウエアを用いて結果を比較するように心がけている。また TransRate（Smith-Unna et al. 2016）で *de novo* assembly で組み立てた contig の質を評価し，質の高い contig のみを下流の解析に用いるようにしている。得られた contig の機能推定については，ロングリードの際に用いたソフトウエアを用いて同様の解析を行っている。

６．*Mesostigma viride* の光化学系の解析結果（Iso-seq 解析の実際と応用）

ここではロングリードを用いた Iso-seq 解析の具体例として *Mesostigma viride*（*M. viride*）の解析結果を紹介したい。先述の通り *M. viride* は陸上植物の祖先でもあるストレプト藻類の 1 種であり，その中でも最も早く分岐した「古い」ストレプト藻類である。筆者らはストレプト藻類から陸上植物への進化に伴う光化学系の特に集光系の分子進化に興味があったため，解析を始めた。現在では *M. viride* のゲノムが公開されているが，筆者らが解析を始めた際にはゲノムが公開されていなかった。そのため，*M. viride* の光化学系がどのような LHC を集光アンテナとして利用しているのかを調べるために RNA seq 解析による *LHC* 遺伝子の探索を

試みた。具体的には，*M. viride* の total RNA を用いて，PacBio sequel による Iso-seq 解析を行った後，その full-length cDNA 配列の中から既存の *LHC* 遺伝子と有意な相同性を持つ遺伝子を探索し，LHC タンパク質群の系統樹を作成することで，*M. viride* が持つ *LHC* 遺伝子を推定した（Aso et al. 2021）。

緑藻類の持つ *LHC* 遺伝子セットを比較すると，*M. viride* は同じストレプト藻類とは大きく異なることが明らかになった。例えば，*M. viride* はクラミドモナスの *CrLHCA2* 遺伝子のホモログ（*algae LHCA2*）や *CrLHCA9* 遺伝子のホモログ（*algae LHCA9*）を持つが，これらの *LHC* 遺伝子群はプラシノ藻類やコア緑藻類に保存されているものの他のストレプト藻類には見られない。また，一部のプラシノ藻類のみに見られる *LHCP* 遺伝子も持つが，これも他のストレプト藻類には見られない。逆に，ストレプト藻類や陸上植物によく保存されているシロイヌナズナ *LHCB6* 遺伝子のホモログを持たない。この結果は，*M. viride* の LHC セットがストレプト藻類ではなく，むしろプラシノ藻類の LHC セットに近いことを示しており，*M. viride* の系統的な位置を踏まえると，*M. viride* が他のストレプト藻類から分岐した後で，ストレプト藻類が植物型の LHC セットを持つようになったことを示唆している。なお，シャジクモ類よりも早く分岐した *Klebsormidium nitens* は植物型の LHC セットを持つ。つまり，*K. nitens* が他のストレプト藻類から分岐する前にすでに植物型の LHC セットを獲得していたと考えられる。これらのことから，*M. viride* はストレプト藻類としてはユニークな光化学系の集光系を持つことが明らかになった（図 2）。

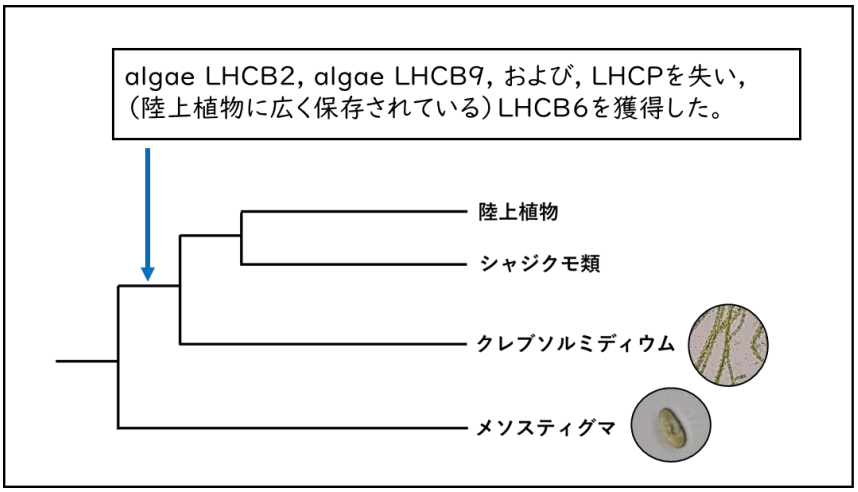

図 2　メソスティグマとクレブソルミディウムの LHC セットの違い

メソスティグマの LHC セットに関してはプラシノ藻類との共通点が多く見られたが，クレブソルミディウムの LHC セットは陸上植物とよく似ている。このことから，メソスティグマが他のストレプト藻類から分岐した後，クレブソルミディウムが分岐する前に，陸上植物型の LHC セットを獲得したのだろうと考えている。

次に，*M. viride* の LHC タンパク質群が PSI と PSII のどちらに結合しているのかを調べるために，Clear-Native（CN）-PAGE で分離した PSI と PSII のバンドのタンパク質組成を質量分析で解析した（Aso et al. 2021）。その際，ペプチドの同定のためのデータベースには，Iso-

seq 解析を利用して構築したタンパク質配列とすでに公開されていた葉緑体ゲノムデータを用いた。その結果から，PSI に結合する LHC タンパク質群と PSII に結合する LHCII タンパク質群を推定し，既知の PSI や PSII の構造と比較することで *M.viride* の光化学系の集光系の構造を予測した（図 3）。LHCP は PSI ではなく PSII に結合していることを見出した。オストレオコッカスと異なり LHCP は 3 量体を形成しているのではなく，マイナーLHCII として機能しているのだろうと予想している。また algae LHCA2 と algae LHCA9 は陸上植物には見られないことから，クレブソルミディウムへの進化の過程で失われたのだろうと考えている。

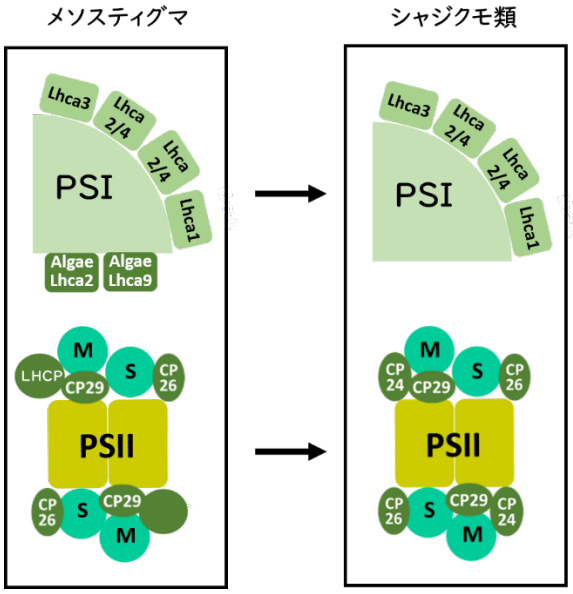

図 3　メソスティグマと他のシャジクモ類の光化学系の構造の違い（予測）

7．まとめ

光化学系の色素やアンテナの多様性は光合成生物の光環境適応に重要な役割を担うと考えられている。近年の CryoEM 技術の進展により，光化学系の構造解析は飛躍的に進んでいるものの，ゲノムプロジェクトが進んでいる現在においても，ゲノム情報が公開された緑藻種の数は決して十分ではなく，光化学系の構造解析の一つの制約になっている。本研究の RNA seq，系統解析，Native-PAGE，MS 解析を組み合わせた解析は光化学系に限らず，タンパク質複合体のタンパク質組成を予測するうえで強力な手法と考えており，今後は，*de novo* assembly もしくは Iso-seq で構築したタンパク質アミノ酸配列の長さや精度が CryoEM 解析による光化学系の構造解析のためにどの程度有効であるのかどうかを試みていきたい。同時に，さらなるゲノムプロジェクトの発展にも期待したい。

引用文献

Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature 408:796–815. doi: 10.1038/35048692

Aramaki T, Blanc-Mathieu R, Endo H, Ohkubo K, Kanehisa M, Goto S, Ogata H. (2020) KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score threshold. Bioinformatics 36:2251–2252. doi: 10.1093/bioinformatics/btz859

Aso M, Matsumae R, Tanaka A, Tanaka R, Takabayashi A. (2021) Unique Peripheral Antennas in the Photosystems of the Streptophyte Alga *Mesostigma viride*. Plant Cell Physiol 62:436–446. doi: 10.1093/pcp/pcaa172

Bai T, Guo L, Xu M, Tian L (2021) Structural Diversity of Photosystem I and Its Light-Harvesting System in Eukaryotic Algae and Plants. Front Plant Sci 12:781035. doi: 10.3389/fpls.2021.781035

Bowles AMC, Bechtold U, Paps J (2020) The Origin of Land Plants Is Rooted in Two Bursts of Genomic Novelty. Curr Biol 30:530-536.e2. doi: 10.1016/j.cub.2019.11.090

Buchfink B, Reuter K, Drost HG (2021) Sensitive protein alignments at tree-of-life scale using DIAMOND. Nat Methods 18:366–368. doi: 10.1038/s41592-021-01101-x

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. (2009) BLAST+: architecture and applications. BMC Bioinformatics 10:421. doi: 10.1186/1471-2105-10-421

Cantalapiedra CP, Hernández-Plaza A, Letunic I, Bork P, Huerta-Cepas J. (2021) eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. Mol Biol Evol 38:5825–5829. doi: 10.1093/molbev/msab293

Ishii A, Shan J, Sheng X, Kim E, Watanabe A, Yokono M, Noda C, Song C, Murata K, Liu, et al. (2023) The photosystem I supercomplex from a primordial green alga *Ostreococcus tauri* harbors three light-harvesting complex trimers. Elife 12:e84488. doi: 10.7554/eLife.84488

Iwai M, Yokono M, Kono M, Noguchi K, Akimoto S, Nakano A. (2015) Light-harvesting complex Lhcb9 confers a green alga-type photosystem I supercomplex to the moss *Physcomitrella patens*. Nat Plants 1:14008. doi: 10.1038/nplants.2014.8

Latasa M, Scharek R, Gall FL, Guillou L (2004) Pigment suites and taxonomic groups in Prasinophyceae. J Phycol 40:1149–1155. doi: 10.1111/j.1529-8817.2004.03136.x

Ohyama K, Fukuzawa H, Kohchi T, Shirai H, Sano T, Sano S, Umesono K, Shiki Y, Takeuchi M, Chang Z, et al (1986) Chloroplast gene organization deduced from complete sequence of liverwort *Marchantia polymorpha* chloroplast DNA. Nature 322:572–574. doi: 10.1038/322572a0

Pinnola A, Alboresi A, Nosek L, Barozzi F, Kouřil R, Dall'Osto L, Aro EM, et al. (2018) A LHCB9-dependent photosystem I megacomplex induced under low light in *Physcomitrella patens*. Nat Plants 4:910–919. doi: 10.1038/s41477-018-0270-2

Shinozaki K, Ohme M, Tanaka M, Wakasugi T, Hayashida N, Matsubayashi T, Zaita N, Chunwongse J, Obokata J, Yamaguchi-Shinozaki K, et al. (1986) The complete nucleotide sequence of the tobacco chloroplast genome: its gene organization and expression. EMBO J 5:2043–2049. doi: 10.1002/j.1460-2075.1986.tb04464.x

Smith-Unna R, Boursnell C, Patro R, Hibberd JM, Kelly S. (2016) TransRate: reference-free quality assessment of de novo transcriptome assemblies. Genome Res 26:1134–1144. doi: 10.1101/gr.196469.115

Suga M, Shen JR (2020) Structural variations of photosystem I-antenna supercomplex in response to adaptations to different light environments. Curr Opin Struct Biol 63:10–17. doi: 10.1016/j.sbi.2020.02.005

Tang S, Lomsadze A, Borodovsky M (2015) Identification of protein coding regions in RNA transcripts. Nucleic Acids Res 43:e78. doi: 10.1093/nar/gkv227