**µgreen-db: a reference database of the plastidial 23S rRNA gene of photosynthetic eukaryotic algae and cyanobacteria**

Christophe Djemiel1, Damien Plassard2, Sébastien Terrat1, Olivier Crouzet3, Joana Sauze4, Samuel Mondy1, Virginie Nowak1, Lisa Wingate4, Jérôme Ogée4, Pierre-Alain Maron1🟏

*1 Agroécologie, AgroSup Dijon, INRA, Univ. Bourgogne Franche-Comté, Dijon, France*

*2 Plateforme GenomEast, IGBMC, CNRS UMR7104, Illkirch, France*

*3 Univ Paris Saclay, AgroParisTech, UMR ECOSYS, INRA, F-78206 Versailles, France*

*4 INRA, Bordeaux Science Agro, UMR 1391 ISPA, 33140 Villenave d’Ornon, France*

🟏 Corresponding author: pierre-alain.maron@inra.fr

Téléphone: 00 33 380 69 34 46

Fax: 00 33 380 693 224

Address: UMR Agroecologie, 17 rue de Sully, 21065 Dijon, France

**Abstract:**

Photosynthetic eukaryotic microalgae and prokaryotic cyanobacteria are key ecological players in biogeochemical cycles both in aquatic and terrestrial habitats. To study the evolution and ecology of these organisms *in situ*, molecular tools are necessary to complement the historical and widely applied technique f morphological observation. The development of metabarcoding has enabled the construction of large generalist databases of prokaryotes or eukaryotes with historical taxonomic molecular markers to perform sequence affiliation. However, the use of more specific markers that target taxonomic groups more closely has now been developed, hence requiring the development of specialised databases to achieve taxonomic assignment of detected organisms. Here, we set up a reference database for the plastidial 23S rRNA gene used to characterise the group of photosynthetic eukaryotic algae and prokaryotic cyanobacteria. The sequences were retrieved from either generalist (NCBI, SILVA) or Comparative RNA Web (CRW) databases, in addition to using a more original approach involving recursive BLAST searches to obtain the best sequence recovery. At present, this database, called µgreen-db, includes 2,326 plastidial 23S rRNA sequences spanning four Kingdoms (Eubacteria, Chromista, Protozoa and Plantae) encompassing 442 unique *genera* and 736 *species* of eukaryotic algae, cyanobacteria and non-vascular land plants. In order to test the effectiveness of this database, we sequenced domain V of the 23S rRNA plastid gene (commonly known as the Universal Plastid Amplicon) from two contrasting soil samples, and performed a taxonomic assignment of the OTUs. The µgreen-db database is accessible through the URL <http://microgreen-23sdatabase.ea.inra.fr>.

**Keywords:**

23S rRNA gene, cyanobacteria, eukaryotic algae, sequence database, metabarcoding, universal plastid amplicon

**Introduction**

Photosynthetic microalgae and cyanobacteria can be found inhabiting diverse aquatic and terrestrial habitats thanks to their advanced abilities to adapt to a range of challenging environmental conditions (*e.g.*, soils, marine, airborne, plants and animals, including extreme environments such as polar regions or deserts) (Andersen 1992; Zancan et al., 2006; Tomitani et al., 2006; Seckbach, 2007; Elbert et al., 2012; Ramanan et al., 2015; Tesson et al., 2016; Seppey et al., 2017; Rippin et al., 2018). These ubiquitous microorganisms play essential ecological roles in the global carbon and nitrogen cycles and also contribute to the production of atmospheric oxygen. As primary producers, they form the base of trophic networks (*e.g.* microbial loop in aquatic ecosystems, Azam et al, 1983) and may represent a potentially rich reservoir for diverse, natural biosynthetic products (Schenk et al., 2008).

Soil microalgae primarily belong to three main groups: the prokaryotic cyanobacteria and two eukaryotic algae including the green algae and the diatoms (Hoffmann, 1989). Cyanobacteria, commonly called ‘blue-green algae’, are prokaryotes with a monophyletic distribution inside the eubacteria (Palinska et al., 2014). Eukaryotic algae represent a polyphyletic assemblage including several lineages that evolved from a primary common endosymbiosis: the main group of green algae (Viridiplantae) belongs to a well-supported monophyletic group subdivided in two major groups, the Chlorophyta and the Streptophyta [this second group includes Charophyta and the land plants; the red algae (Rodophyta) and the glaucophytes (Glaucophyta)]. Other lineages such as euglenids (Euglenozoa), chlorarachniophytes (Cercozoa), cryptomonads (Cryptophyta), haptophytes (Haptophyta or brown algae), stramenopiles [Bacillariophyta (or diatoms) and Ochrophyta)], and Dinoflagellates (Miozoa, also know as Myzozoa)], also belong to the Viridiplantae group but have a secondary endosymbiotic origin (Andersen 1992; Bhattacharya & Medlin 1998; De Clerck 2012; Keeling 2004; Leliaert et al., 2012; Lowe et al., 2017).

The diversity and composition of the microbial photosynthetic community can be used as a bioindicator of soil quality (Pipe and Schubert, 1984) and the presence of invasive species. Microbial photosynthetic communities can also help identify and monitor the involvement of specific groups in the biodegradation of environmental pollutants (Zancan et al., 2006; Sauvage et al., 2016; Lowe et al., 2017). In addition a better understanding of microbial photosynthetic community diversity can help understand their function and contribution to C cycling, notably in marine (Hüegler and Sievert, 2011) and dryland (Muñoz-Rojas et al., 2018) ecosystems.

During the past century, a large body of knowledge on microalgae taxonomy has been gathered from microscopic observations, providing valuable information for a complementary trait approach. However, in the past twenty years, phylogenetic analyses have demonstrated that an approach based on morphological determination alone is somewhat artificial for most of the microalgal genera and should be revised (Pröschold and Leliaert, 2007; Luo et al., 2006). Recently several studies have estimated the diversity of indigenous photosynthetic microbial communities in various environments using metabarcoding coupled to High-Throughput Sequencing (HTS) (Sherwood et al., 2010; Seppey et al., 2017, Cho et al., 2017; Rippin et al., 2018, Kim et al., 2011; Vasselon et al., 2017; Oliveira et al., 2018). So far a range of molecular markers have been used to describe cyanobacteria and eukaryotic algae diversity with varying degrees of resolution (*e.g.* 16S/18S/23S rRNA, *tuf*A, *psbA*, rbcL, ITS) (Eriksson et al., 2009; Saunders & Kucera 2010; Hall 2010; Sherwood et al., 2014; Rossetto Marcelino 2016, Vasselon et al., 2017). Amongst these markers the 23S rRNA gene presents several advantages over the other markers. In particular its length and higher sequence variability provides a better phylogenetic resolution compared to small rRNA subunits (Gutell et al., 1994; Pei et al., 2009). More precisely, domain V of the 23S rRNA gene, known as the Universal Plastid Amplicon (UPA), allows the targeting of organisms containing plastids with a remarkable universality, covering most photosynthetic microbial groups (Presting 2006; Sherwood & Presting 2007). For cyanobacteria, this marker also seems to be promising as it provides better coverage of community diversity than either 16S/18S rDNA or *tuf*A, (Rossetto Marcelino & Verbruggen 2016). Moreover, the UPA has a length (~ 410 bp) suitable for HTS Technologies (Presting 2006), such as Illumina (Lentendu et al., 2014). The use of the 23S rRNA gene can also be used in addition to other markers thereby obtaining a comprehensive overview of microbial diversity (Sherwood et al., 2014; Rossetto Marcelino & Verbruggen 2016, Sherwood et al., 2010; Sherwood et al., 2017).

Metabarcoding still remains the fastest and cheapest method to study microbial diversity and community structures. However, it requires reference databases, updated with a good coverage of organisms, a good sequence quality and curated taxonomy to achieve taxonomic assignment of obtained sequences (Balvočiūtė & Huson 2017). There are already several generalist or specialist databases that include some groups of algae with curated taxonomy. The most popular databases are: SILVA, that groups SSU and LSU rRNA genes from eukaryotic and prokaryotic organisms (Quast et al., 2012); PR2, a protist small subunit ribosomal reference database (Guillou et al., 2012); PhytoREF, a reference database of the plastidial 16S rRNA gene of photosynthetic eukaryotes (Decelle et al., 2015); R-Syst::diatom, that gathers the 18S rRNA gene and rbcL diatom sequences (Rimet et al., 2016); and DINOREF, a reference database of 18S rRNA for dinoflagellates (Mordret et al. 2018). Recently, Sherwood et al. (2017) developed a new database (<http://scholarspace.manoa.hawaii.edu/handle/10125/42782>) that groups 97,194 UPA or LSU amplicon sequences from their own project, including sequences that are not present in SILVA. However, these sequences are mainly assigned to the Bacteria domain (75% of the total sequences with only < 1% assigned to Cyanobacteria) whilst within the 10% of eukaryotic sequences, 80% are associated to the Metazoa group. Moreover, the taxonomy is not completely standardized and therefore difficult to use for HTS analyses. Another reference database of the UPA marker exists containing only algae-related taxa as well as standardized taxonomy however, it includes much fewer sequences (573 sequences) than the other UPA database described above (Rossetto Marcelino & Verbruggen 2017). To our knowledge, no 23S rRNA database exists to date that fulfills all the important criteria (*i.e.* good coverage of organisms, good sequence quality and curated taxonomy) for the metabarcoding study of indigenous algae communities.

Here, we propose a new reference database of plastid sequences in eukaryotes and cyanobacteria. This database, called µgreen-db, was constructed from various sources (SILVA, CRW, BLAST or extracted from genomes) to be the most representative. When possible, the complete sequence of the 23S rRNA gene is provided, allowing users greater flexibility to create, for example, their own primers for environmental metabarcoding studies. The taxonomy associated with the sequences is based on the NCBI and AlgaeBase websites. In green-db, sequences of non-vascular land plants are also provided with the aim of improving the study of algae communities in soil environments where mosses and liverworts (Bryophytes) can be abundant and where bryophyte sequences can be consequently co-amplified with algal sequences because of the similarity of the plastidial 23S rRNA gene between algae and mosses. Thus the inclusion of sequences related to bryophyte taxa will allow avoiding orphaned sequences and improve the recovery of taxonomic information from sequence datasets. This database is open-source and can be downloaded from the website (http://microgreen-23sdatabase.ea.inra.fr).

**Materials and Methods**

*Retrieval of plastidial 23S rDNA sequences from public databases*

We developed several strategies to recover the maximum number and diversity of sequences (Figure 1). Plastidial 23S rRNA sequences in cyanobacteria, algae and bryophytes were retrieved from SILVA r123 (June 2016) (Quast et al., 2012). We also retrieved 23S chloroplast sequences from various organisms (*e.g.* algae, bryophytes, angiosperms) from a Comparative RNA Web Site and Project led by the Gutell Lab at the University of Texas at Austin ([www.rna.ccbb.utexas.edu/DAT/3C/Alignment/](http://www.rna.ccbb.utexas.edu/DAT/3C/Alignment/)) (Cannone et al., 2002). Another set of sequences was also recovered from NCBI with the Gene (the list of different queries is available in the Supp data 1 file). We also used various BLAST (with a megablast approach and set the maximum target parameter of 1000) to improve the sequence recovery. We first performed a BLAST with a 23S rRNA sequence from a close organism on plastid genomes. We then performed a second BLAST by taking a sequence query in the nr/nt database and retrieved all the returned sequences. From these sequences, we performed recursively a phylogenetic tree in order to know what sequence was furthest away every time. We thereafter aligned the sequences using Muscle (Mega7) (Kumar et al., 2016) and reconstructed the phylogenetic tree using a maximum-likelihood method (Felsenstein 1981). To improve the exhaustivity of µgreen-db, we performed another BLAST against the WGS database of NCBI (<http://www.ncbi.nlm.nih.gov/genbank/wgs/>), selecting sequences with a score bit greater than 1000 and belonging to the targeted organisms, and performed a final BLAST from the sequences obtained against the 23S rRNA sequence file. Sequences corresponding to taxa not present in the 23S rRNA sequence file were then selected and added to the sequence dataset (based on less than 97% identity).

*Sequence verification*

According to the origins of the sequences, we conducted a series of different filters in order to retain only plastid sequences (Figure 1). Regarding the sequences originating from SILVA, we only kept sequences with a length higher than 700 bp and a quality ≥75 %. For the sequences recovered from other databases, we also carried out a verification of the secondary structure, with the INFERNAL tool (Nawrocki & Eddy, 2013). Finally we checked the non-redundancy of the sequences to retain only unique sequences. For each sequence found in both SILVA and BLAST databases, we checked whether the sequence was included in the "BLAST" sequence *(i.e.* at the identity level). If it was not the case, we aligned them and kept the least fragmented sequence. We also removed the sequences affiliated to Angiosperms from the CRW database. High length sequences (more than 5000 bp) were also deleted.

*Taxonomic validation – The taxonomic framework of* µgreen-db

The NCBI and AlgaeBase taxonomy were both retrieved to provide users the choice for further analyses (Figure 1). To obtain a standardized taxonomy in the form of phylum, class, order, family, genus and species, we recovered the *taxonID* from the accession number of NCBI (<ftp://ftp.ncbi.nih.gov/pub/taxonomy/accession2taxid/>) and used the *taxonkit* tool (<http://github.com/shenwei356/taxonkit>) to retrieve the full lineage. The AlgaeBase taxonomy was also used to obtain more information at the kingdom level. Where no rank information was available, we ascribed the abbreviation rank followed by two underscores plus no\_rank (*e.g.* p\_\_no\_rank). As non-vascular land plants are not represented in AlgaeBase, we assigned the Plantae Kingdom from the NCBI taxonomy to these sequences and made modifications at the *phylum* level. All these sequences were assigned to the *phylum* Streptophyta from the NCBI taxonomy. However as Streptophyta is an Infrakingdom subdivised into three *phyla* in AlgaeBase, we have assigned Bryopsida, Polytrichopsida, Sphagnopsida, Tetraphidopsida, Takakiopsida, Andreaeobryopsida, Andreaeopsida, Oedipodiopsida *classes* to the Bryophyta *phylum*; Jungermanniopsida, Marchantiopsida, Haplomitriopsida *classes* to the Marchantiophyta *phylum*, and Anthocerotopsida and Leiosporocerotopsida *classes* to the Anthocerotophyta *phylum*.

*Database finalisation – Construction of the* µgreen-db *database*

The database is available in two forms: from tabular flat files, and from a website (http://microgreen-23sdatabase.ea.inra.fr) (Figure 1). The tabular flat files were formatted with a custom homemade script. The web interface was built using Bulma ([https://bulma.io](https://bulma.io/)), a modern and open source CSS framework based on Flexbox with a custom template. The website uses PHP (v7.2.7) to communicate with the MySQL database, providing back-end storage of sequences and taxonomy by using queries and Javascript to make it more dynamic and user-friendly. We have estimated the hypothetical coverage of primers conventionally used to study the diversity of algae (Sherwood 2007; Yoon et al., 2016) by performing an *in silico* PCR amplification.

*Metabarcoding validation*

**Soil sampling, experimental design**

Soil samples were taken from the top 10 cm of a luvisol with a decarbonated sandy A horizon (pH = 8.2, Corg = 11.5 g kg‑1, Ntot = 0.83 g kg‑1) located in the north of Paris used for conventional cropping, with a wheat/maize rotation. Soil was sampled and incubated either under a 16 h light/24h photoperiod or continuous dark conditions, as described previously (Sauze et al., 2017) to obtain contrasted phototrophic microbial communities. Briefly, after sieving at 5 mm and homogenizing the soil, 6 microcosms were set up by placing 400 g of fresh soil weighed at 80% of its water holding capacity in 0.825 dm3 glass jars. Three microcosms were coated with aluminum foil to prevent the development of phototrophic organisms (darkness condition), and three microcosms were conditioned under a day/night cycle (lightness condition) consisting of a 16 h light/24h photoperiod using LED lighting with an intensity of about 200 μmol photons m−2 s−1 in the visible range to promote the growth of the native phototrophic organisms. After 40 days of incubation at 20°C with regular monitoring of soil moisture, a soil aliquot was sampled from each of the six microcosms and stored at ‑40°C before DNA extraction.

**Soil microbial DNA extraction, 23S rRNA gene amplification and Illumina sequencing** Microbial DNA was extracted and purified from 1 g of each soil sampled, using the GnSGII procedure described previously (Terrat et al., 2017). Crude DNA extracts were quantified by agarose gel electrophoresis before being purified using a GENECLEAN turbo kit (MpBiomedical) and quantified using a QuantiFluor staining kit (Promega) prior to further investigation.

A 23S rRNA gene fragment targeting the V5 domain to characterse algae diversity was amplified using the primers p23SrV\_f1 (5′GGACAGAAAGACCCTATGAA3′) and p23SrV\_r1 (5′TCAGCCTGTTATCCCTAGAG3′) (Sherwood and Presting, 2007). Amplifications were carried out in a total volume of 25 μl using 5 μl of DNA (10 ng), 10 μl of buffer solution 10x with 20 mM MgSO4 (Promega), 0.4 μl of dNTPs (25 mM, DNTPack 250U Roche), 2 μl (10 μM, Eurogentec) of each primer, 0.5 μl of Taq polymerase (5U/μl Taq PFU, Promega), 1.25 μl of T4 gene 32 (500 μg/mL, MP Biomedical) and 11.35 μl of water. PCR1 conditions were: 2 min at 94°C, followed by 35 cycles of 45 s at 94°C, 45 s at 63°C, and 1 min at 72°C, and final elongation for 10 min at 72°C. The PCR products were then purified using a MinElute PCR purification kit (Qiagen) and quantified using a QuantiFluor staining kit (Promega). A second PCR of seven cycles was then duplicated for each sample under similar PCR conditions, with purified PCR products as matrix (10 ng of DNA was used for a 25 µl mix of PCR) and dedicated fusion primers (‘p23SrV\_f1/MID,’ ‘p23SrV\_r1/MID) integrating the required keys, and multiplex identifiers at the 5′ extremities. All duplicated PCR products were then pooled, purified using a MinElute PCR purification kit (Qiagen), and quantified using a QuantiFluor staining kit (Promega). For all libraries, equal amounts from 29 samples were pooled and then cleaned to remove excess nucleotides, salts, and enzymes using the Agencourt AMPure XP system (Beckman Coulter Genomics). TE buffer (100 µl) (Roche) was used for the elution. Sequencing was then carried out on an Illumina MiSeq (GenoScreen, France).

**Bioinformatics sequence analysis**

To perform the raw data analysis of the 23S plastid rDNA amplicons generated from the soil samples, we used the GnS-PIPE pipeline (availability: https://zenodo.org/record/1123425#.W82vmDVR2OE) (Terrat et al., 2012). The different steps have already been described previously (Terrat et al., 2015). After preprocessing filtering and chimera checking, all samples were normalized at 31.650 sequences. The taxonomic affiliation was performed using the μgreen-db and the USEARCH program (v6.0.307; www.drive5.com/usearch) with specific parameters (-maxhits 15, -maxaccepts 0, and maxrejects 0). The microbial DNA sequence data sets supporting the results in this article are available at the EBI ENA with accession PRJEB30252.

To access the putative number of amplifications and the coverage of the different taxons, we achieved *in silico* PCR from green-db we used the mothur software (v.1.40.5) with the pcr.seqs command and allowed zero mismatches between each of the primer pairs. Graphic representations were produced using custom scripts based on Highcharts facilities (http://www.highcharts.com/).

**Results**

*Overview of* µgreen-db

The µgreen-db currently contains 2,326 non-redundant sequences including 440 complete, 1,658 incomplete, and 228 environmental plastidial 23S rDNA sequences (Figure 2 A). The mean sequence length is between 800 bp and 4,000 bp with 2,271 sequences longer than 800 bp (Figure 2B).

The µgreen-db provides a reference file containing all the sequences in fasta format. For each sequence, the associated identifier is in the following form: [C or I or E]AccessNumber.Letter(if duplicate).start.end;AllLineage where ‘C’ signifies complete, ‘I’ incomplete and ‘E’ environmental. We also provide a set of two files, a reference sequence file with a unique identifier in fasta format and another with the complete taxonomy that can be used easily in the most popular metabarcoding pipelines (e.g. Mothur, QIIIME, GnS-PIPE) with either the NCBI or the AlgaeBase taxonomy.

*Taxonomic validation – Taxonomic composition of* green-db

Following the intial retrieval of the database sequences in June 2016, a further update of the entire taxonomy was completed using NCBI in August 2018. During this update we encountered three scenarios for each sequence these included (i) no change in taxonomy, (ii) obsoletion of the accession number or (iii) removal or loss of the access number. For the specific cases of (ii) we updated the accession number whilst in the case of (iii) we removed these particular sequences from our database.

Taxonomic coverage (corresponding to the percentage of sequences for a given rank) was higher with AlgaeBase than with NCBI (Figure 3). For sequences assigned from the NCBI database, we obtained 86% and 42% coverage at the *Phylum* and *Class* rank respectively (Figure 3). We obtained 9 phyla through the 4 supergroups (Terrabacteria, Excavata, Archaeplastida, and SAR) but 14% were without taxonomic assignment at this rank (Figure 4A). For our sequences assigned from the AlgaeBase database, we obtained 100% coverage at the *phylum* level (Figure 3), with 1 *phylum* for the Eubacteria kingdom, 6 phyla for the Chromista kingdom, 1 for the Protozoa kingdom and 7 for the Plantae kingdom, of which 4 were algae, and 1 *phylum* Chromerida that has no kingdom affiliation as yet (Figure 4B). The most represented *phylum* was Cyanobacteria with 939 sequences, followed by Euglenozoa (349 sequences), and Chlorophyta (314 sequences) while Bacillaryophyta was less represented (54 sequences) (Figure 4B).

More than 2,283 sequences (*i.e.* 98% of the total sequences) could be assigned up to genus rank with 442 unique *genera* (the ranked top 3 were: Prochlorococcus': 207, 'g\_\_Chroococcidiopsis': 120, 'g\_\_Synechococcus': 90 all types of cyanobacteria). A total of 1,590 sequences have species level affiliation including 736 unique *species* (not including uncultured and \* .sp).

*Database finalisation*

This database is now fully operational and can now be used to perform taxonomic assignments in metabarcoding projects. Using the universal primer pair to amplify the 23S rRNA V region (UPA) (Sherwood 2007; Yoon et al., 2016) on our database, we obtain 1,500 out of the 2,366 sequences with a PCR *in silico* (Figure S1). Several formatted files have been generated for metabarcoding data analysis.

*Description of the µgreen-db web interface*

The µgreen-db is also available *via* a web interface (http://microgreen-23sdatabase.ea.inra.fr). Access to all data is provided *via* this interface and simply allows searches for taxa of interest. This website also permits downloading of the latest sequence and/or taxonomy files. Finally, various information on the construction of this database, statistics and news are also accessible through this website.

*Metabarcoding validation*

We tested the ability of µgreen-db to affiliate sequence datasets generated from a set of indigenous soil phototrophic microbial communities obtained from a soil that was exposed to two contrasted light conditions (darkness *vs.* lightness). The Shannon diversity indices calculated from the OTU dataset highlighted a higher diversity in the darkness compared to the enlightened treatment (*H’*=3.1 ± 0.1 vs. *H’*=2.6 ± 0.1, respectively) (Table S1). This decrease of diversity was associated to a lower richness (441.3 ± 41.2 *vs.* 378.3 ± 31.4 OTUs) and a lower evenness (0.51 ± 0.01 *vs.* 0.43 ± 0.02) of the community after exposure to light. Interestingly, µgreen-db allowed suitable affiliation of 96% and 98.5% of the sequence datasets at the *phylum* and genus level, respectively. Examination of the taxonomic affiliation of the sequences also revealed a broad diversity of the phototrophic soil microbial community, with 11 phyla and 149 unique *genera* detected. As observed for the diversity metrics, light conditions significantly shaped the composition of the phototrophic community. Most markedly, at the *phylum* level, Cyanobacteria became highly dominant, increasing from 4 ± 2.4% to 72.0 ± 1.8% of the affiliated sequences after exposing the soil to the light (Figure 5A, Table S1). In the same way, sequences related to Charophyta increased from 1.4 ± 0.5 to 8.4 ± 1.8% following light exposure. In contrast, Chlorophyta, Bacillariophyta and Ochrophyta, which represented 39.75 ± 2.37; 29.2 ± 3.1; and 17.4 ± 0.2% of the sequences affiliated from the dark treatment, decreased to 5.9 ± 0.7; 4.4 ± 0.9; and 4.3 ± 0.8%, respectively after light exposure (Figure 5A). Also, the Miozoa *phylum* disappeared in the light treatment. Typically all phyla were consistently found in all three sample replicates, with the exception of Anthocerotophyta that was detected in only one of the three replicates belonging to the light treatment (Fig. 5A). The clear taxonomic separation of dark and light treatments was also observed at the genus level (Figure 5B). The increase of Cyanobacteria in the light was mainly caused by the stimulation of three *genera*: Microcoleus, Nodosilinea and Synechococcus. Klebsormidium was the only genus explaining the increase of the Charophyta *phylum* in response to light. In contrast, there was a higher contribution of Chlorophyta, Bacillariophyta and Ochrophyta in the dark treatment caused by the higher occurrence of genera such as Chlorella and Ettlia (for Chlorophyta); Eunotia (for Bacillariophyta) and Ectocarpus, Nannochloropsis and Vaucheria (for Ochrophyta).

**Discussion**

The study of algae and cyanobacteria diversity can now be acheived using either morphological identification through microscopy observations or molecular tools that analyse genetic markers and provide taxonomic affiliation. It is even recommended to combine these two methods or to use multiple molecular markers to obtain improved coverage of the *species* present (Zou et al., 2016; Rossetto Marcelino & Verbruggen 2016; Groendahl et al., 2017) and to continuously improve barcoding databases. This combining of techniques is particularly powerful for identifying key relationships that underpin the construction trait-based knowledge that can be used to improve our functional understanding of photosynthetic microbial communities.

µgreen-db is a new resource gathering sequences associated with their taxonomy. We have paid special attention to taxonomy by providing two different sources from NCBI and AlgaeBase with a full lineage from the kingdom/*phylum* levels to the *species* level, allowing an efficient taxonomic assignment. One limitation of using the plastidial 23S rRNA gene in metabarcoding studies is that few sequences are available from public databases (*e.g.* GenBank, SILVA) (Yilmaz et al., 2011; Yoon et al., 2016). This explains why it was necessary to retrieve our sequences using several strategies to obtain the most diverse database possible. In addition, to retrieve the sequences from various databases we implemented a strategy of recursive BLAST with phylogenetic tree construction to improve our spectrum of organisms. Consequently, we were able to recover more than 1,500 sequences and to significantly increase the total number of sequences in our reference database.

Regarding the taxonomic assignment of sequences from NCBI, we obtained contrasting results. Although we obtained a good assignment at the genus level, we only affiliated 86% of our sequences at the *phylum* level and 42% at the *class* level. We also noticed that there could be diverging rankings between the NCBI database and AlgaeBase. For example for the cryptomonads group, placed at the *class* level for NCBI and at the *phylum* level for AlgaeBase. The classification of this particular group remains widely debated and explains why we opted to propose both affiliations, allowing the user to decide. Indeed, the most recent consensus classification for the eukaryotes (Adl et al., 2012) does not use any ranks at all and the cryptomonads are thus listed just as 'Cryptophyceae' (not *phylum* Cryptophyceae or *division* Cryptophyceae or *class* Cryptophyceae). Another example is the Phaeophyceae group that is associated at the *phylum* level for NCBI and at the *class* level for AlgaeBase. As stated on their website, the NCBI taxonomy database is not an authoritative source for nomenclature or classification. For this reason, we recommend using the taxonomy from the AlgaeBase, because it provides manual curation, and offers a very complete bibliography for each taxon (Guiry 2018).

Analysis of the environmental soil samples allowed validation of the power of µgreen-db to characterise the taxonomic composition of indigenous phototrophic microbial communities. We were able to affiliate 98.5 of the sequences at the *phylum* level and 96% at the genus level, highlighting the good coverage of the phototrophic diversity in the database. From a biological point of view, our results provided evidence for a strong impact of photoperiod illumination on the composition and diversity of the phototrophic microbial community. Concomitantly to these strong changes, a very high increase of the soil chlorophyll*-a* concentrations was found in the light treatment compared to soils incubated in the dark (data not shown), underlying the significant photosynthetic growth of micro-phototroph biomass stimulated by the light.

Under long-term dark incubation, the dominant eukaryotic microalgae can be related to *species* having a mixotrophic strategy to remain active in the dark, and/or to *species* better able to overcome unfavorable lighting conditions through the switch to dormant forms and/or the production of resistant forms. A number of the algae taxa detected in the dark conditioned soil across the dominant phyla (Chlorophyta, Bacillaryphyta and Ochrophyta) are able to modulate their metabolism from phototrophic to heterotrophic assimilating dissolved organic carbon depending on prevalent environmental conditions (Parker, 1961; Jones, 2000). Such trophic and flexible metabolic strategies are an important competitive advantage in soils, where light can rapidly become a limiting factor for obligate autotrophs (Stark et al., 1981) during photosynthetic growth as reported in lakes (Porter, 1988). In our study, the dominance of some eukaryotic *classes* of microalgae under continuous dark conditions stressed that they may be equally adapted to survive using obligate chemoheterotrophic metabolism. In contrast, Cyanobacteria with strict mixotrophic capacities may not be as able to grow efficiently using chemoheterotrophy under long periods of time (Rippka, 1972). Moreover, the relatively strong occurrence of certain *species* (*e.g.* *Vaucheriaceae*), currently not considered as mixotrophs (Kviderová et al., 2018) may result an ability of these organisms to switch to a dormant stage during unfavourable conditions and produce resistance forms (*i.e.* zygospore, akinetes, zoospores). Such forms of resistance or dispersal stages have been reported for a wide range of Cyanobacteria and eukaryotic algae (Agrawal, 2009).

During the photoperiod treatment, the strong development of numerous cyanobacterial taxa over-competing eukaryotic algae might also be explained partially by the high alkalinity of the studied soil (pH = 8.2). Alkaline soils are known to favor cyanobacteria over eukaryotic green algae (Shields and Durrell, 1964; Starks et Shubert, 1982). Under our experimental conditions (optimum water content, temperature and light) cyanobacteria that have relatively faster growing strategies with shorter generation times than eukaryotic algae, may have been favoured over microalgae. This could potentially explain why the soil surface became overrun by cyanobacteria and contributed to the lower diversity indices observed under light conditions.

In conclusion, our results demonstrate that µgreen-db represents a powerful tool to affiliate the plastidial 23S rRNA genes of photosynthetic eukaryotic algae and cyanobacteria in soil environments. Future improvements to the database will consist of setting up regular routines (how regular every yr?) to enrich this open access database by adding new sequences but also assimilating any changes in accession, by updating NCBI access numbers and taxonomy. We also encourage the future community of users to engage with the curators of the database to report any errors found either in the database or on the website or via the website portal or directly by email to the corresponding author.

**Acknowledgements**

This project has received funding from the Agence National de la Recherche (ANR, ORCA project award no. ANR-13-BS06- 0005-01). JS was jointly funded by a PhD scholarship from the INRA departments EFPA and EA and the ANR project ORCA. This project has also received funding from the European Research Council (ERC) under the European Union’s Seventh Framework Programme (FP7/2007-2013) (grant agreement No. 338264).

**Author contribution**

PAM and ST and OC designed the study; JS, LW, OC and JO performed the soil study and provided the environmental 23S rRNA datasets; CD, DP, ST and SM performed bioinformatics; CD, ST, OC and PAM wrote the first draft of the manuscript. All authors contributed to the final editing.

**Data accessibility**

µgreen-db is available in flat files at url: http://microgreen-23sdatabase.ea.inra.fr.

The microbial DNA sequencing data sets supporting the results in this article are available at the EBI ENA with accession number PRJEB30252.

**ORCID**

Christophe DJEMIEL <https://orcid.org/0000-0002-5659-7876>

Samuel Mondy, <https://orcid.org/0000-0002-9203-6398>

Jérôme Ogée, <https://orcid.org/0000-0002-3365-8584>

Lisa Wingate, https://orcid.org/0000-0003-1921-1556

**References**

Andersen, R.A. (1992). Diversity of eukaryotic algae. Biodiversity & Conservation, 1(4), 267-292.

Sauze, J., Ogée, J., Maron, P.A., Crouzet, O., Nowak, V., Wohl, S., Kaisermann, A., Jones, S., & Wingate, L. (2017). The interaction of soil phototrophs and fungi with pH and their impact on soil CO2, CO18O and OCS exchange. Soil Biology and Biochemistry, 115, 371-382.

Zancan, S., Trevisan, R., & Paoletti, M.G. (2006). Soil algae composition under different agro-ecosystems in North-Eastern Italy. Agriculture, ecosystems & environment, 112(1), 1-12.

Tomitani, A., Knoll, A.H., Cavanaugh, C.M., & Ohno, T. (2006). The evolutionary diversification of cyanobacteria: molecular–phylogenetic and paleontological perspectives. Proceedings of the National Academy of Sciences, 103(14), 5442-5447.

Seckbach, J. (Ed.). (2007). Algae and cyanobacteria in extreme environments (Vol. 11). Springer Science & Business Media.

Elbert, W., Weber, B., Burrows, S., Steinkamp, J., Büdel, B., Andreae, M.O., & Pöschl, U. (2012). Contribution of cryptogamic covers to the global cycles of carbon and nitrogen. Nature Geoscience, 5(7), 459.

Ramanan, R., Kim, B.H., Cho, D.H., Oh, H.M., & Kim, H.S. (2016). Algae–bacteria interactions: evolution, ecology and emerging applications. Biotechnology advances, 34(1), 14-29.

Tesson, S.V., Skjøth, C.A., Šantl-Temkiv, T., & Löndahl, J. (2016). Airborne Microalgae: Insights, Opportunities and Challenges. Applied and environmental microbiology, AEM-03333.

Seppey, C.V., Singer, D., Dumack, K., Fournier, B., Belbahri, L., Mitchell, E.A., & Lara, E. (2017). Distribution patterns of soil microbial eukaryotes suggests widespread algivory by phagotrophic protists as an alternative pathway for nutrient cycling. Soil Biology and Biochemistry, 112, 68-76.

Rippin, M., Lange, S., Sausen, N., & Becker, B. (2018). Biodiversity of biological soil crusts from the Polar Regions revealed by metabarcoding. FEMS microbiology ecology, 94(4), fiy036.

Palinska, K.A., & Surosz, W. (2014). Taxonomy of cyanobacteria: a contribution to consensus approach. Hydrobiologia, 740(1), 1-11.

Bhattacharya, D., & Medlin, L. (1998). Algal phylogeny and the origin of land plants. Plant Physiology, 116(1), 9-15.

De Clerck, O., Bogaert, K.A., & Leliaert, F. (2012). Diversity and evolution of algae: primary endosymbiosis. In Advances in botanical research (Vol. 64, pp. 55-86). Academic Press.

Keeling, P.J. (2004). Diversity and evolutionary history of plastids and their hosts. American journal of botany, 91(10), 1481-1493.

Leliaert, F., Smith, D.R., Moreau, H., Herron, M.D., Verbruggen, H., Delwiche, C.F., & De Clerck, O. (2012). Phylogeny and molecular evolution of the green algae. Critical Reviews in Plant Sciences, 31(1), 1-46.

Lowe, R.L., & LaLiberte, G.D. (2017). Benthic stream algae: distribution and structure. In Methods in Stream Ecology, Volume 1 (Third Edition) (pp. 193-221).

Sauvage, T., Schmidt, W. E., Suda, S., & Fredericq, S. (2016). A metabarcoding framework for facilitated survey of endolithic phototrophs with tuf A. BMC ecology, 16(1), 8.

Kim, E., Harrison, J.W., Sudek, S., Jones, M.D., Wilcox, H.M., Richards, T.A., ... & Archibald, J. M. (2011). Newly identified and diverse plastid-bearing branch on the eukaryotic tree of life. Proceedings of the National Academy of Sciences, 201013337.

Vasselon, V., Domaizon, I., Rimet, F., Kahlert, M., & Bouchez, A. (2017). Application of high-throughput sequencing (HTS) metabarcoding to diatom biomonitoring: Do DNA extraction methods matter?. Freshwater Science, 36(1), 162-177.

Oliveira, M.C., Repetti, S.I., Iha, C., Jackson, C. J., Díaz-Tapia, P., Lubiana, K.M.F., ... & Verbruggen, H. (2018). High-throughput sequencing for algal systematics. European Journal of Phycology, 1-17.

Sherwood, A.R., Kurihara, A., Conklin, K.Y., Sauvage, T., & Presting, G.G. (2010). The Hawaiian Rhodophyta Biodiversity Survey (2006-2010): a summary of principal findings. BMC plant biology, 10(1), 258.

Cho, D.H., Choi, J.W., Kang, Z., Kim, B.H., Oh, H.M., Kim, H.S., & Ramanan, R. (2017). Microalgal diversity fosters stable biomass productivity in open ponds treating wastewater. Scientific Reports, 7(1), 1979.

Saunders, G.W., & Kucera, H. (2010). An evaluation of rbcL, tufA, UPA, LSU and ITS as DNA barcode markers for the marine green macroalgae. Cryptogamie Algologie, 31(4), 487.

Hall, J.D., Fučíková, K., Lo, C., Lewis, L.A., & Karol, G. (2010). An assessment of proposed DNA barcodes in freshwater green algae. Cryptogamie, Algologie, 31(4), 529-555.

Sherwood, A.R., Conklin, K.Y., & Liddy, Z.J. (2014). What's in the air? Preliminary analyses of Hawaiian airborne algae and land plant spores reveal a diverse and abundant flora. Phycologia, 53(6), 579-582.

Marcelino, V.R., & Verbruggen, H. (2016). Multi-marker metabarcoding of coral skeletons reveals a rich microbiome and diverse evolutionary origins of endolithic algae. Scientific reports, 6, 31508.

Gutell, R.R., Larsen, N., & Woese, C.R. (1994). Lessons from an evolving rRNA: 16S and 23S rRNA structures from a comparative perspective. Microbiological reviews, 58(1), 10-26.

Pei, A., Nossa, C.W., Chokshi, P., Blaser, M.J., Yang, L., Rosmarin, D.M., & Pei, Z. (2009). Diversity of 23S rRNA genes within individual prokaryotic genomes. PloS one, 4(5), e5437.

Presting, G.G. (2006). Identification of conserved regions in the plastid genome: implications for DNA barcoding and biological function. Botany, 84(9), 1434-1443.

Sherwood, A.R., & Presting, G.G. (2007). Universal primers amplify a 23S rDNA plastid marker in eukaryotic algae and cyanobacteria. Journal of phycology, 43(3), 605-608.

Yilmaz, P., Kottmann, R., Pruesse, E., Quast, C., & Glöckner, F.O. (2011). Analysis of 23S rRNA genes in metagenomes–a case study from the Global Ocean Sampling Expedition. Systematic and applied microbiology, 34(6), 462-469.

Balvočiūtė, M., & Huson, D.H. (2017). SILVA, RDP, Greengenes, NCBI and OTT—how do these taxonomies compare?. BMC genomics, 18(2), 114.

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., ... & Glöckner, F.O. (2012). The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic acids research, 41(D1), D590-D596.

Guillou, L., Bachar, D., Audic, S., Bass, D., Berney, C., Bittner, L., ... & Del Campo, J. (2012). The Protist Ribosomal Reference database (PR2): a catalog of unicellular eukaryote small sub-unit rRNA sequences with curated taxonomy. Nucleic acids research, 41(D1), D597-D604.

Decelle, J., Romac, S., Stern, R.F., Bendif, E.M., Zingone, A., Audic, S., ... & Gourvil, P. (2015). Phyto REF: a reference database of the plastidial 16S rRNA gene of photosynthetic eukaryotes with curated taxonomy. Molecular ecology resources, 15(6), 1435-1445.

Rimet, F., Chaumeil, P., Keck, F., Kermarrec, L., Vasselon, V., Kahlert, M., ... & Bouchez, A. (2016). R-Syst:: diatom: an open-access and curated barcode database for diatoms and freshwater monitoring. Database, 2016.

Mordret, S., Piredda, R., Vaulot, D., Montresor, M., Kooistra, W.H., & Sarno, D. (2018). DinoREF: a curated dinoflagellate (Dinophyceae) reference database for the 18S rRNA gene. Molecular ecology resources. doi: 10.1111/1755-0998.12781

Sherwood, A.R., Dittbern, M.N., Johnston, E.T., & Conklin, K.Y. (2017). A metabarcoding comparison of windward and leeward airborne algal diversity across the Ko ‘olau mountain range on the island of O'ahu, Hawai ‘i1. Journal of phycology, 53(2), 437-445.

Marcelino, V.R., & Verbruggen, H. (2017). Reference datasets of tufA and UPA markers to identify algae in metabarcoding surveys. Data in brief, 11, 273-276.

Cannone, J.J., Subramanian, S., Schnare, M.N., Collett, J.R., D'Souza, L.M., Du, Y., ... & Pande, N. (2002). The comparative RNA web (CRW) site: an online database of comparative sequence and structure information for ribosomal, intron, and other RNAs. BMC bioinformatics, 3(1), 2.

Kumar, S., Stecher, G., & Tamura, K. (2016). MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. Molecular biology and evolution, 33(7), 1870-1874.

Yoon, T.H., Kang, H.E., Kang, C.K., Lee, S.H., Ahn, D.H., Park, H., & Kim, H.W. (2016). Development of a cost-effective metabarcoding strategy for analysis of the marine phytoplankton community. PeerJ, 4, e2115.

Zou, S., Fei, C., Wang, C., Gao, Z., Bao, Y., He, M., & Wang, C. (2016). How DNA barcoding can be more effective in microalgae identification: a case of cryptic diversity revelation in Scenedesmus (Chlorophyceae). Scientific reports, 6, 36822.

Groendahl, S., Kahlert, M., & Fink, P. (2017). The best of both worlds: A combined approach for analyzing microalgal diversity via metabarcoding and morphology-based methods. PloS one, 12(2), e0172808.

Nawrocki, E.P., & Eddy, S.R. (2013). Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics, 29(22), 2933-2935.

Terrat, S., Horrigue, W., Dequietd, S., Saby, N.P., Lelièvre, M., Nowak, V., ... & Ranjard, L. (2017). Mapping and predictive variations of soil bacterial richness across France. PloS one, 12(10), e0186766.

Adl, S.M., Simpson, A.G., Lane, C.E., Lukeš, J., Bass, D., Bowser, S.S., ... & Heiss, A. (2012). The revised classification of eukaryotes. Journal of Eukaryotic Microbiology, 59(5), 429-514.

Guiry, M.D. & Guiry, G.M. (2018). AlgaeBase. World-wide electronic publication. http://www. algaebase. Org.

Terrat, S., Plassart, P., Bourgeois, E., Ferreira, S., Dequiedt, S., Adele-Dit-De-Renseville, N., ... & Ranjard, L. (2015). Metabarcoded evaluation of the ISO standard 11063 DNA extraction procedure to characterize soil bacterial and fungal community diversity and composition. Microbial biotechnology, 8(1), 131-142.

Terrat, S., Christen, R., Dequiedt, S., Lelièvre, M., Nowak, V., Regnier, T., ... & Bispo, A. (2012). Molecular biomass and MetaTaxogenomic assessment of soil microbial communities as influenced by soil DNA extraction procedure. Microbial biotechnology, 5(1), 135-141.

Azam, F., Fenchel, T., Field, J.G., Gray, J.S., Meyer-Reil, L.A., & Thingstad, F. (1983). The Ecological Role of Water-Column Microbes in the Sea. Marine Ecology Progress Series, 10, 257–263. doi:10.3354/meps010257.

Schenk, P.M., Thomas-Hall, S.R., Stephen, E., Marx U.C., Mussgnug, J.H., Posten, C., Kruse, O., & Hankamer, B. (2008). Second generation biofuels: high-efficiency microalgae for biodiesel production. BioEnergy Research, 1, 20–43.

Hoffmann, L. (1989). Algae of terrestrial habitats. Bot Rev,55, 77-105.

Pipe, A.E., & Shubert, L.E. (1984). The use of algae as indicators of soil fertility. *In* Shubert LE (ed.), Algae as Ecological Indicators. Academic Press, London, pp, 213-233.

Hüegler, M., & Sievert, S.M. (2011). Beyond the Calvin cycle: autotrophic carbon fixation in the ocean. Annu.Rev. Mar. Sci. 3, 261-289. doi:10.1146/annurev-marine-120709-142712

Muñoz-Rojas, M., Román, J.R., Roncero-Ramos, B., Erickson, T.E., Merritt, D.J., Aguila Carricondo, P., Cantón Y. (2018). Cyanobacteria inoculation enhances carbon sequestration in soil substrates used in dryland restoration, Science of The Total Environment, 636, 1149-1154.

Pröschold, T., & Leliaert, F.( 2007). Systematics of the green algae: conflict of classic and modern approaches. In: Brodie J, Lewis J, editors. Unravelling the algae: the past, present, and future of the algae systematics. London, UK: Taylor and Francis; 2007. pp. 123-153.

Luo, W., Pflugmacher, S., Pröschold, T., Walz, N., & Krienitz, L. (2006). Genotype versus phenotype variability in Chlorella and Micractinium (Chlorophyta, Trebouxiophyceae). Protist, 157, 315–333. pmid:16843061

Eriksson, K.M., Clarke, A.K., Franzen, L.G., Kuylenstierna, M., Martinez, K., & Blanck H. (2009). Community-Level Analysis of *psbA* Gene Sequences and Irgarol Tolerance in Marine Periphyton. Appl. Environ. Microbiol. 75, 897-906; DOI: 10.1128/AEM.01830-08

Lentendu, G., Wubet, T., Chatzinotas, A., Wilhelm, C., Buscot, F., & Schlegel, M. (2014). Effects of long-term differential fertilization on eukaryotic microbial communities in an arable soil: a multiple barcoding approach, Molecular ecology, 23, 3341 -3355.

Parker, B.C. (1961). Facultative Heterotrophy in Certain Soil Algae from the Ecological Viewpoint. Ecology, 42, 381-386. doi:[10.2307/1932089](https://doi.org/10.2307/1932089)

Jones R.I. (2000). Mixotrophy in planktonic protists: an overview. Freshwater Biology, 45, 219-226.

Porter, K.G. (1988). Phagotrophic phytoflagellates in microbial food webs. Hydrobiologia, 159, 89–97.

Rippka R. (1972). Photoheterotrophy and chemoheterotrophy among unicellular blue-green algae. Arch. Mikrobiol. 87,93-98.

Kvíderová, J., Souquieres, C. E., & Elster J. (2018). Ecophysiology of photosynthesis of Vaucheria sp. mats in a Svalbard tidal flat, in Press, Accepted Manuscript, doi.org/10.1016/j.polar.2018.11.006

Agrawal, S.C. (2009). Factors Affecting Spore Germination in Algae – Review. Folia Microbiol. 54, 273–302.

Shields, L.M., & Durrell, L.W. (1964). Algae in Relation to Soil Fertility. Botanical Review, 30, 92-128.

Starks, T.L., & Shubert, L.E. (1982). Colonization and succession of algae and soil-algal interactions associated with disturbed areas. Journal of Phycology 18, 99-107. DOI: 10.1111/j.0022-3646.1982.00099.x

Starks, T.L., Shubert, L.E., & Trainor, F.R. (1981). Ecology of soil algae – a review. Phycologia. 20, 65-80. DOI: 10.2216/i0031-8884-20-1-65.1

Felsenstein, J. (1981). A likehood approach to character weighting and what it tells us about parsimony and compatibility. Biological Journal of the Linnean Society. 16, 183-196. DOI: 10.1111/j.1095-8312.1981.tb01847.x

**Figures legends**

Figure 1: Workflow describing the different steps performed to generate the curated and annotated 23S rDNA reference database constructed from various databases and methods.

Figure 2: Pie chart and histograms showing (A) the origin and number, and (B) the length of the plastidial 23S rDNA sequences available in the database.

Figure 3: Taxonomic coverage at different ranks from the NCBI and AlgaeBase taxonomy.

Figure 4: Sequence distribution of the database at the Phylum level and grouped by Kingdom. (A) Taxonomic assignment of the sequences of µgreen-db with the NCBI database according to Adl et al. (2012) for the group classification, (B) Taxonomic assignment of the sequences of µgreen-db with the Algae database.

Figure 5: Relative sequence abundance of algae and cyanobacteria at phylum (A) and genus (B) level.

**Supp data**

Supp data File 1: List of commands used to retrieve, filter and construct the µgreen-db.

Supp data Figure 1 ; Number of sequences amplified by in silico PCR from µgreen-db based on AlgaeBase taxonomy and using different primer pairs from literature with 0 mismatch.