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

Christophe Djemiel[1], Damien Plassard[2], Sébastien Terrat[1], Olivier Crouzet[3], Joana Sauze[4], Samuel Mondy[1], Virginie Nowak[1], Lisa Wingate[4], Jérôme Ogée[4], Pierre-Alain Maron[1]*

[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

Phone: +33 (0)380 69 34 46

Fax: +33 (0)380 693 224

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

Abstract : 200 words max ➔ Delete 30/40 words

Abstract:

~~The study of~~Studying the ecology of photosynthetic eukaryotic microalgae and prokaryotic cyanobacteria communities requires molecular tools to complement the historical technique of morphological observations. These tools being developed ~~lay~~ rely on specific genetic markers and~~, hence~~ require~~ing~~ the development of specialised databases to achieve taxonomic assignment. Here, we set up a reference database, called µgreen-db, for the plastidial 23S rRNA gene. 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~~S~~ searches to obtain the best sequence recovery. At present, 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 based on the NCBI and AlgaeBase taxonomy. In addition the µgreen-db is also available using ~~based on~~ the PR^2/SILVA taxonomy ~~is also available with~~containing 2,217 sequences (399 unique *genera* and 696 unique *species*). ~~By u~~Using the µgreen-db, we were able to assign 98.5% of the sequences at the phyla level for~~of~~ the V5 domain? of the 23S rRNA plastid gene obtained by metabarcoding after amplification from soil extracted DNA~~. This~~, thus highlighting~~s~~ the good coverage of database. The µgreen-db database is accessible at http://microgreen-23sdatabase.ea.inra.fr.

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, freshwater and brackish, airborne, plants and animals, including extreme environments such as polar regions or deserts)[1-5]. 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[6]) and may represent a potentially rich reservoir for diverse, natural biosynthetic products[7].

Soil microalgae primarily belong to three main groups: the prokaryotic cyanobacteria and two eukaryotic algae including the green algae and the diatoms[8]. Cyanobacteria, formerly called 'blue-green algae', are prokaryotes, monophyletic and belong to the bacterial domain[9,10]. 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 (Rhodophyta) 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

known as Myzozoa)], also belong to the Viridiplantae group but have a secondary endosymbiotic origin[11-16].

The diversity and composition of the microbial photosynthetic community can be used as a bioindicator of soil quality[17] 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[5,16,18]. In addition a better understanding of microbial photosynthetic community diversity can help understand their function and contribution to C cycling, notably in marine[19] and dryland[20] 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[21,22]. Recently, several studies have estimated the diversity of indigenous photosynthetic microbial communities in various environments using metabarcoding coupled to High-Throughput Sequencing (HTS)[3,23-28]. 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)[28-33]. Various hypervariable regions (*e.g.* V4, V8-V9) of the 18S rRNA gene are commonly used[34]. However, the 23S rRNA gene presents several advantages over the other markers. In particular its length and higher sequence variability provide a better phylogenetic resolution compared to small rRNA subunits[35,36]. 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[37,38]. 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[31]. Moreover, the UPA has a length (~ 410 bp) suitable for HTS Technologies[37], such as Illumina[39]. The use of UPA can also be used in addition to other markers thereby obtaining a comprehensive overview of microbial diversity[27,31,33,40].

Major collaborative projects and studies at the international level (*e.g.* UniEuk, EukRef) are currently underway to propose a classification of microbial eukaryotes that will serve as a reference for a universal taxonomy[41-43]. The proposed tools are mainly deployed on the 18S gene that allows the targeting of eukaryotic photosynthetic organisms but not the cyanobacteria group.

Metabarcoding still remains the fastest and cheapest method to study ~~the~~ microbial diversity and community structures. However, it requires reference databases, updated with a good coverage of organisms, a ~~good~~ high level of sequence quality and curated taxonomy to achieve taxonomic assignment of obtained sequences[44]. 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[45]; PR^2, a protist small subunit ribosomal reference database[46]; PhytoREF, a reference database of the plastidial 16S rRNA gene of photosynthetic eukaryotes[47]; R-Syst::diatom, that gathers the 18S rRNA gene and rbcL diatom

sequences[48]; and DINOREF, a reference database of 18S rRNA for the dinoflagellates[49]. Recently, Sherwood et al.[27] made ~~available~~ a database available (http://scholarspace.manoa.hawaii.edu/handle/10125/42782) that groups 97,194 UPA and 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. A 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[50]. Thus to our knowledge, no 23S rRNA database exists to date that fulfil~~l~~s all of 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~~coherent with the PR^2/SILVA, NCBI and AlgaeBase databases. In the µgreen-db and ~~only for~~those of the NCBI and AlgaeBase databases, 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~~help avoid 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).

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 (Fig. 2A). The mean sequence~~s~~ length is between 800 bp and 4,000 bp with 2,271 sequences longer than 800 bp (Fig. 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]AccessionNumber.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, ~~the~~ AlgaeBase or PR^2/SILVA taxonomy.

Taxonomic validation – Taxonomic composition of μgreen-db

Following the initial 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) obsolete accession number (8 sequences) or (iii) removal or loss of the accession number (2 sequences). 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 (Fig. 3). Coverage at the class and genus level was slightly better with ~~The~~ PR^2/SILVA compared to AlgaeBase. For sequences assigned from the NCBI database, we obtained 86% and 42% coverage at the Phylum and Class rank respectively (Fig. 3). We obtained 9 phyla through the 4 supergroups (Terrabacteria, Excavata, Archaeplastida, and SAR) but 14% of the sequences were without taxonomic assignment at this rank (Fig. 4A). For our sequences assigned from the AlgaeBase database, we obtained 100% coverage at the phylum level (Fig. 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 (Fig. 4B). The most represented phylum was Cyanobacteria with 939 sequences, followed by Euglenozoa (349 sequences), and Chlorophyta (314 sequences) while Bacillariophyta was less represented (54 sequences) (Fig. 4B).

2,283 sequences (*i.e.* 98% of the total sequences) could be assigned up to genus rank with 442 unique *genera* (the top 3 of the most represented *genera* were: 207 Prochlorococcus, 120 Chroococcidiopsis, 90 Synechococcus, all types of cyanobacteria). A total of 1,590 sequences have species level affiliation including 736 unique *species* (not including uncultured and * .sp) with the NCBI and AlgaeBase taxonomy.

The µgreen-db based on the PR^2/SILVA taxonomy contains 2,217 of the 2,326 sequences found, distributed across seven groups (Bacteria, Stramenopiles, Hacrobia, Alveolata, Rhizaria, Excavata, Archaeplastida) (Fig. 4C) with 399 unique *genera* and 696 unique *species* available (with the same top 3 as previously) ~~representing~~ as well as 93.3 % of Cyanobacteria, 97% of photosynthetic eukaryotic algae and 92.8% of non-vascular land plants.

Databases finalisation

This database is now fully operational and can now be used to perform ~~a~~ taxonomic assignments in metabarcoding projects. Using the universal primer pair to amplify the 23S rRNA V region (UPA)[38,51] on our database, we obtain 1,500 out of the 2,366 sequences with a PCR *in silico* (Supplementary Fig. S1). Several formatted files have been generated for metabarcoding data analysis (https://zenodo.org/record/3385760#.XW-NptPVLUI).

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 assign sequence datasets generated from a set of indigenous soil phototrophic microbial communities obtained from a soil that was exposed to two contrasted light~~ening~~ conditions (dark *vs.* light). The Shannon diversity indices calculated from the OTU dataset highlighted a higher diversity in the dark treatment~~ness~~ compared to that of the light ~~he 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 appropriate 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

Code de champ modifié

Commenté [LW1]: You probably want to update the info on the webpage to remove the submitted to Molecular Ecology Resources

the assigned sequences after exposing ~~of~~ the soil to ~~the~~ light (Fig. 5A, Supplementary 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 in the dark treatment, decreased to 5.9 ± 0.7; 4.4 ± 0.9; and 4.3 ± 0.8%, respectively after light exposure (Fig. 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 (Fig. 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 achieved 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[31,52,53] and to continuously improve barcoding databases. This combining of techniques is particularly powerful for identifying key relationships that underpin the construction of trait-based knowledge that can be used to improve our functional understanding of photosynthetic microbial communities.

μgreen-db is a new resource gathering 23S rRNA sequences associated with their taxonomy. We have paid special attention to taxonomy by providing three different sources ~~from~~ spanning the PR^2/SILVA, NCBI and AlgaeBase databases and with a full lineage from the kingdom/phylum levels to the *species* level, allowing an efficient taxonomic assignment. Nevertheless, μgreen-db is not a phylogenetic or taxonomic authority and provides only taxonomy data from various sources (PR^2/SILVA, NCBI and AlgaeBase) with a full lineage from the supergroup/kingdom/phylum levels to the species level, allowing a complete taxonomic assignment for bioinformatic analysis. 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)[51,54]. 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 there was an assignment for almost all sequences at the genus level, only 86% of the sequences were assigned at the phylum level and 42% at the *class* level. We also noticed that there could be diverging rankings between the PR^2/SILVA, NCBI ~~database~~ and AlgaeBase databases. For example, ~~for~~ the cryptomonads group was~~,~~ 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, until very recently, the consensus classification for the eukaryotes[55] did not use any ranks at all and the cryptomonads are thus listed just as 'Cryptophyceae' (not phylum/division Cryptophyceae or *class* Cryptophyceae) but should be classified now to order rank[43]. 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[56].

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 assign 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

297 provided evidence for a strong impact of photoperiod illumination on the composition
298 and diversity of the phototrophic microbial community.
299 Under long-term dark incubation, the dominant eukaryotic microalgae can be
300 related to *species* having a mixotrophic strategy to remain active in the dark, and/or
301 to *species* better able to overcome unfavorable lighting conditions through the switch
302 to dormant forms and/or the production of resistant forms. A number of the algae taxa
303 detected in the dark conditioned soil across the dominant phyla (Chlorophyta,
304 Bacillariophyta and Ochrophyta) are able to modulate their metabolism from
305 phototrophic to heterotrophic with ~~assimilating~~ assimilation of dissolved organic
306 carbon depending on prevalent environmental conditions[57,58]. Such trophic and
307 flexible metabolic strategies are an important competitive advantage in soils, where
308 light can rapidly become a limiting factor for obligate autotrophs[59] during
309 photosynthetic growth, as reported in lakes[60]. In our study, the dominance of some
310 eukaryotic *classes* of microalgae under continuous dark conditions stressed that they
311 may be equally adapted to survive using obligate chemoheterotrophic metabolism. In
312 contrast, Cyanobacteria with strict mixotrophic capacities may not be as able to grow
313 efficiently using chemoheterotrophy under long periods of time[61]. Moreover, the
314 relatively strong occurrence of certain *species* (*e.g. Vaucheriaceae*), currently not
315 considered as mixotrophs[62] may result from an ability of these organisms to switch to
316 a dormant stage during unfavorable conditions and produce resistant~~ce~~ forms (*i.e.*
317 zygospore, akinetes, zoospores). Such forms of resistance or dispersal stages have
318 been reported for a wide range of Cyanobacteria and eukaryotic algae[63].

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 promote cyanobacteria over eukaryotic green algae[64,65]. 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. The µgreen-db now paves the way for future studies investigating the community and functional ecology of photosynthetic organisms in soils.

In conclusion, our results demonstrate that µgreen-db represents a powerful tool to assign 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 to enrich this open access database by adding new sequences ~~but also~~and assimilating any changes in accession, by updating NCBI accession numbers and taxonomy from various sources. 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.

Methods

Retrieval of plastidial 23S rDNA sequences from public databases

342 We developed several strategies to recover the maximum number and diversity of
343 sequences (Fig. 1). Plastidial 23S rRNA sequences in cyanobacteria, algae and
344 bryophytes were retrieved from SILVA r123 (June 2016)[45]. We also retrieved 23S
345 chloroplast sequences from various organisms (*e.g.* algae, bryophytes, angiosperms)
346 from a Comparative RNA Web Site and Project led by the Gutell Lab at the University
347 of Texas at Austin (www.rna.ccbb.utexas.edu/DAT/3C/Alignment/)[66]. Another set of
348 sequences was also recovered from NCBI with the Gene (the list of different queries
349 is available in the Supp data 1 file). We also used various BLAST (with a megablast
350 approach and set the maximum target parameter of 1,000) to improve the sequence
351 recovery. We first performed a BLAST with a 23S rRNA sequence from a close
352 organism on plastid genomes. We then performed a second BLAST by taking a
353 sequence query in the nr/nt database and retrieved all the returned sequences. From
354 these sequences, we performed recursively a phylogenetic tree in order to know what
355 sequence was furthest away every time. We thereafter aligned the sequences using
356 Muscle (Mega7)[67] and reconstructed the phylogenetic tree using a maximum-
357 likelihood method[68]. To improve the exhaustivity of µgreen-db, we performed another
358 BLAST against the WGS database of NCBI
359 (http://www.ncbi.nlm.nih.gov/genbank/wgs/), selecting sequences with a score bit
360 greater than 1000 and belonging to the targeted organisms, and performed a final
361 BLAST from the sequences obtained against the 23S rRNA sequence file.
362 Sequences corresponding to taxa not present in the 23S rRNA sequence file were
363 then selected and added to the sequence dataset (based on less than 97% identity).
364

Code de champ modifié

Sequence verification

According to the origins of the sequences, we conducted a series of different filters in order to retain only plastid sequences (Fig. 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[69]. 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 assigned to Angiosperms from the CRW database. High length sequences (more than 5,000 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 (Fig. 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 Unnamed_rank (*e.g.* p__Unnamed_rank). As non-vascular land plants are not represented in AlgaeBase,

Code de champ modifié

Code de champ modifié

388 we assigned the Plantae Kingdom from the NCBI taxonomy to these sequences and
389 made modifications at the phylum level. All these sequences were assigned to the
390 phylum Streptophyta from the NCBI taxonomy. However as Streptophyta is an
391 Infrakingdom subdivised into three phyla in AlgaeBase, we have assigned Bryopsida,
392 Polytrichopsida, Sphagnopsida, Tetraphidopsida, Takakiopsida, Andreaeobryopsida,
393 Andreaeopsida, Oedipodiopsida classes to the Bryophyta phylum;
394 Jungermanniopsida, Marchantiopsida, Haplomitriopsida *classes* to the
395 Marchantiophyta phylum, and Anthocerotopsida and Leiosporocerotopsida classes to
396 the Anthocerotophyta phylum. To format at the PR^2/SILVA taxonomy, the full lineage
397 are constructed with the genus name from NCBI by searching into the PR^2 database
398 (https://github.com/pr2database/pr2database) and SILVA taxonomy archive
399 (https://www.arb-silva.de/no_cache/download/archive/current/Exports/taxonomy/).
400

Code de champ modifié

401 **Databases finalisation – Construction of the µgreen-db database**
402 The database is available in two forms: from tabular flat files, and from a website
403 (http://microgreen-23sdatabase.ea.inra.fr) (Fig. 1). The tabular flat files were
404 formatted with a custom homemade script. The web interface was built using Bulma
405 (https://bulma.io), a modern and open source CSS framework based on Flexbox with
406 a custom template. The website uses PHP (v7.2.7) to communicate with the MySQL
407 database, providing back-end storage of sequences and taxonomy by using queries
408 and Javascript to make it more dynamic and user-friendly. We have estimated the
409 hypothetical coverage of primers conventionally used to study the diversity of
410 algae[38,51] by performing an *in silico* PCR amplification.

Code de champ modifié

Code de champ modifié

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, C_{org} = 11.5 g kg^{-1}, N_{tot} = 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[70] 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 dm^3 glass jars. Three microcosms were coated with aluminium foil to prevent the development of phototrophic organisms (dark condition), and three microcosms were conditioned under a day/night cycle (light 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[71]. 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 characterise algae diversity was amplified using the primers p23SrV_f1 (5′GGACAGAAAGACCCTATGAA3′) and p23SrV_r1 (5′TCAGCCTGTTATCCCTAGAG3′)[38]. 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 $MgSO_4$ (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

455 (100 µl) (Roche) was used for the elution. Sequencing was then carried out on an
456 Illumina MiSeq (GenoScreen, France).
457
458 *Bioinformatics sequence analysis*
459 To perform the raw data analysis of the 23S plastid rDNA amplicons generated from
460 the soil samples, we used the GnS-PIPE pipeline (availability:
461 https://zenodo.org/record/1123425#.W82vmDVR2OE)[72]. The different steps have
462 already been described previously[73]. After preprocessing filtering and chimera
463 checking, all samples were normalized at 31.650 sequences. The taxonomic
464 affiliation was performed using the µgreen-db and the USEARCH program (v6.0.307;
465 www.drive5.com/usearch) with specific parameters (-maxhits 15, -maxaccepts 0, and
466 maxrejects 0). The microbial DNA sequence data sets supporting the results in this
467 article are available at the EBI ENA with accession PRJEB30252.
468 To access the putative number of amplifications and the coverage of the
469 different taxons, we achieved *in silico* PCR from µgreen-db we used the mothur
470 software (v.1.40.5) with the *pcr.seqs* command and allowed zero mismatches
471 between each of the primer pairs. Graphic representations were produced using
472 custom scripts based on Highcharts facilities (http://www.highcharts.com/).
473

474 **Acknowledgements**
475 This project has received funding from the Agence National de la Recherche (ANR,
476 ORCA project award no. ANR-13-BS06- 0005-01). JS was jointly funded by a PhD
477 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 contributions

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 and Zenodo repository (https://zenodo.org/record/3385760#.XW-NptPVLUI).

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

Code de champ modifié

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

1. Elbert, W. *et al.* Contribution of cryptogamic covers to the global cycles of carbon and nitrogen. *Nature Geoscience* **5**, 459-462, doi:10.1038/ngeo1486 (2012).
2. Ramanan, R., Kim, B. H., Cho, D. H., Oh, H. M., & Kim, H. S. Algae-bacteria interactions: Evolution, ecology and emerging applications. *Biotechnology Advances* **34**, 14-29, doi:10.1016/j.biotechadv.2015.12.003 (2016).
3. Rippin, M., Lange, S., Sausen, N., & Becker, B. Biodiversity of biological soil crusts from the Polar Regions revealed by metabarcoding. *FEMS Microbiology Ecology* **94**, 1-15, doi:10.1093/femsec/fiy036 (2018).
4. Tesson, S. V. M., Skjøth, C. A., Šantl-Temkiv, T., & Löndahl, J. Airborne Microalgae: Insights, Opportunities, and Challenges. *Applied and Environmental Microbiology* **82**, 1978-1991, doi:10.1128/AEM.03333-15 (2016).
5. Zancan, S., Trevisan, R., & Paoletti, M. G. Soil algae composition under different agro-ecosystems in North-Eastern Italy. *Agriculture, Ecosystems and Environment* **112**, 1-12, doi:10.1016/j.agee.2005.06.018 (2006).
6. Azam, F., & Malfatti, F. Microbial structuring of marine ecosystems. *Nature Reviews Microbiology* **5**, 782-791, https://doi.org/10.1038/nrmicro1747 (2007).
7. Schenk, P. M. *et al.* Second Generation Biofuels: High-Efficiency Microalgae for Biodiesel Production. *BioEnergy Research* **1**, 20-43, doi:10.1007/s12155-008-9008-8 (2008).
8. Hoffmann, L. Algae of terrestrial habitats. *The Botanical Review* **55**, 77-105, doi:10.1007/BF02858529 (1989).
9. Palinska, K. A., & Surosz, W. Taxonomy of cyanobacteria: A contribution to consensus approach. *Hydrobiologia* **740**, 1-11, doi:10.1007/s10750-014-1971-9 (2014).
10. Soo R.M. et al. An expanded genomic representation of the phylum cyanobacteria. *Genome Biology and Evolution* **6**,1031-1045, doi:10.1093/gbe/evu073 (2014).
11. Andersen, R. A. Diversity of eukaryotic algae. *Biodiversity and Conservation* **1**, 267-292, doi:10.1007/BF00693765 (1992).
12. Bhattacharya, D., & Medlin, L. Algal Phylogeny and the Origin of Land Plants. *Plant Physiology* **116**, 9-15, doi:10.1104/pp.116.1.9 (1998).
13. Clerck, O., Bogaert, K. A., & Leliaert, F. Diversity and Evolution of Algae. *Genomic Insights Into the Biology of Algae* **64**, doi:10.1016/B978-0-12-391499-6.00002-5 (2012).
14. Keeling, P. J. Diversity and evolutionary history of plastids and their hosts. *American Journal of Botany* **91**, 1481-1493, doi:10.3732/ajb.91.10.1481(2004).
15. Leliaert, F. *et al.* Phylogeny and Molecular Evolution of the Green Algae. *Critical Reviews in Plant Sciences*, **31**, 1-46. doi:10.1080/07352689.2011.615705 (2012).
16. Lowe, R. L., & LaLiberte, G. D. *Benthic Stream Algae: Distribution and Structure*. *Methods in Stream Ecology: Third Edition* (Vol. 1). Elsevier Inc. doi:10.1016/B978-0-12-416558-8.00011-1 (2017).
17. Pipe, A. E., & Shubert, L. E. The use of algae as indicators of soil fertility. *Algae as Ecological Indicators. Academic Press, London*, 213-233. (1984).

Mis en forme : Anglais (Royaume-Uni)

Mis en forme : Français (France)

18. Sauvage, T., Schmidt, W. E., Suda, S., & Fredericq, S. A metabarcoding framework for facilitated survey of endolithic phototrophs with tufA. *BMC Ecology* **16**, 1-21, doi:10.1186/s12898-016-0068-x (2016).
19. Hügler, M., & Sievert, S. M. Beyond the Calvin Cycle: Autotrophic Carbon Fixation in the Ocean. *Annual Review of Marine Science* **3**, 261-289, doi:10.1306/06210404037(2011).
20. Muñoz-Rojas, M. *et al.* Cyanobacteria inoculation enhances carbon sequestration in soil substrates used in dryland restoration. *Science of the Total Environment* ***636***, 1149-1154, doi:10.1016/j.scitotenv.2018.04.265 (2018).
21. Luo, W., Pflugmacher, S., Pröschold, T., Walz, N., & Krienitz, L. Genotype versus Phenotype Variability in Chlorella and Micractinium (Chlorophyta, Trebouxiophyceae). *Protist* **157**, 315-333, doi:10.1016/j.protis.2006.05.006 (2006).
22. Proschold, T., & Leliaert, F. Systematics of the green algae: conflict of classic and modern approaches BT - Unravelling the algae: the past, present, and future of algal systematics. *Unravelling the Algae: The Past, Present, and Future of Algal Systematics* **75**, 124-153, Retrieved from papers2://publication/uuid/7B8D0095-F34D-4006-A354-E35FA472816E (2007).
23. Cho, D. H. *et al.* Microalgal diversity fosters stable biomass productivity in open ponds treating wastewater. *Scientific Reports* ***7***, 1-11, doi:10.1038/s41598-017-02139-8 (2017).
24. Kim, E., Harrison, J. W. et al. Newly identified and diverse plastid-bearing branch on the eukaryotic tree of life. *Proceedings of the National Academy of Sciences* **108**, 1496-1500, doi:10.1073/pnas.1013337108 (2011).
25. Oliveira, M. C. *et al.* High-throughput sequencing for algal systematics. *European Journal of Phycology* **53**, 256-272, doi:10.1080/09670262.2018.1441446 (2018).
26. Seppey, C. V. W. *et al.* 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, doi:10.1016/j.soilbio.2017.05.002 (2017).
27. Sherwood, A. R., Dittbern, M. N., Johnston, E. T., & Conklin, K. Y. A metabarcoding comparison of windward and leeward airborne algal diversity across the Koʻolau mountain range on the island of Oʻahu, Hawaiʻi 1. *Journal of Phycology* **53**, 437-445, doi:10.1111/jpy.12502 (2017).
28. Vasselon, V., Domaizon, I., Rimet, F., Kahlert, M., & Bouchez, A. Application of high-throughput sequencing (HTS) metabarcoding to diatom biomonitoring: Do DNA extraction methods matter? *Freshwater Science* **36**, 162-177, doi:10.1086/690649 (2017).
29. Eriksson, K. M. *et al.* Community-level analysis of psbA gene sequences and irgarol tolerance in marine periphyton. *Applied and Environmental Microbiology* **75**, 897-906, doi:10.1128/AEM.01830-08 (2009).
30. Hall, J. D., Fucikova, K., Lo, C., Lewis, L. A., & Karol, K. G. An assessment of proposed DNA barcodes in freshwater green algae. *Cryptogamie Algologie* **31**, 529-555, doi:10.1111/gcbb.12105 (2010).

31. Marcelino, V. R., & Verbruggen, H. Multi-marker metabarcoding of coral skeletons reveals a rich microbiome and diverse evolutionary origins of endolithic algae. *Scientific Reports* **6**, 1-9, doi:10.1038/srep31508 (2016).
32. Saunders, G. W., & Kucera, H. An evaluation of rbcL , tufA , UPA , LSU and ITS as DNA barcode markers for the marine green macroalgae INTRODUCTION. *Algologie* **31**, 487-528. (2010).
33. Sherwood, A. R., Conklin, K. Y., & Liddy, Z. J. What's in the air? Preliminary analyses of Hawaiian airborne algae and land plant spores reveal a diverse and abundant flora. *Phycologia* **53**, 579-582, doi:10.2216/14-059.1 (2014).
34. Bradley, I.M., Pinto, A.J., & Guest, J.S. Design and Evaluation of Illumina MiSeq-Compatible, 18S rRNA Gene-Specific Primers for Improved Characterization of Mixed Phototrophic Communities. *Applied and Environmental Microbiology* **82,** 5878-5891; DOI: 10.1128/AEM.01630-16 (2016).
35. Gutell, R. R., Larsen, N., & Woese, C. R. Lessons from an evolving rRNA: 16S and 23S rRNA structures from a comparative perspective. *Microbiological Reviews* **58**, 10-26, doi:10.1038/468755a (1994).
36. Pei, A. *et al.* Diversity of 23S rRNA Genes within Individual Prokaryotic Genomes. *PLoS ONE* ***4***, e5437, doi:10.1371/journal.pone.0005437 (2009).
37. Presting, G. G. Identification of conserved regions in the plastid genome: implications for DNA barcoding and biological function. *Canadian Journal of Botany* **84**, 1434-1443, doi:10.1139/b06-117 (2006).
38. Sherwood, A. R., & Presting, G. G. Universal primers amplify a 23S rDNA plastid marker in eukaryotic algae and cyanobacteria. *Journal of Phycology* **43**, 605-608, doi:10.1111/j.1529-8817.2007.00341.x (2007).
39. Lentendu, G. *et al.* Effects of long-term differential fertilization on eukaryotic microbial communities in an arable soil: A multiple barcoding approach. *Molecular Ecology* **23**, 3341-3355, doi:10.1111/mec.12819 (2014).
40. Sherwood, A. R., Kurihara, A., Conklin, K. Y., Sauvage, T., & Presting, G. G. The Hawaiian Rhodophyta Biodiversity Survey (2006-2010): a summary of principal findings. *BMC Plant Biology* **10**, 258, doi:10.1186/1471-2229-10-258 (2010).
41. Berney, C. *et al.* UniEuk: Time to Speak a Common Language in Protistology!. J. *Eukaryot. Microbiol.* **64**, 407-411, doi:10.1111/jeu.12414 (2017).
42. del Campo, J. *et al.* EukRef: Phylogenetic curation of ribosomal RNA to enhance understanding of eukaryotic diversity and distribution. *PLOS Biology* **16**, e2005849. https://doi.org/10.1371/journal.pbio.2005849 (2018)
43. Adl, S. M. *et al.* Revisions to the classification, nomenclature, and diversity of eukaryotes. *J. Eukaryot. Microbiol.* **66**, 4-119, doi:10.1111/jeu.12691 (2019).
44. Balvočiūtė, M., & Huson, D. H. SILVA, RDP, Greengenes, NCBI and OTT — how do these taxonomies compare? *BMC Genomics* **18**, 114. doi:10.1186/s12864-017-3501-4 (2017).
45. Quast, C. *et al.* The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Research*, *41*(Database issue), D590-6, doi:10.1093/nar/gks1219 (2013).
46. Guillou, L. *et al.* The Protist Ribosomal Reference database (PR2): a catalog of unicellular eukaryote Small Sub-Unit rRNA sequences with curated taxonomy. *Nucleic Acids Research* **41**, D597-D604, doi:10.1093/nar/gks1160 (2012).

Mis en forme : Anglais (Royaume-Uni)

47. Decelle, J. *et al.* PhytoREF: A reference database of the plastidial 16S rRNA gene of photosynthetic eukaryotes with curated taxonomy. *Molecular Ecology Resources* **15**, 1435-1445, doi:10.1111/1755-0998.12401(2015).
48. Rimet, F. *et al.* R-Syst::diatom: an open-access and curated barcode database for diatoms and freshwater monitoring. *Database*, *2016* (August 2018), baw016. doi:10.1093/database/baw016 (2016).
49. Mordret, S. et al. dinoref: A curated dinoflagellate (Dinophyceae) reference database for the 18S rRNA gene. *Molecular Ecology Resources* **18**, 974-987, doi:10.1111/1755-0998.12781 (2018).
50. Rossetto Marcelino, V., & Verbruggen, H. Reference datasets of tufA and UPA markers to identify algae in metabarcoding surveys. *Data in Brief* **11**, 273-276, doi:10.1016/j.dib.2017.02.013 (2017).
51. Yoon, T. H. *et al.* Development of a cost-effective metabarcoding strategy for analysis of the marine phytoplankton community. *PeerJ*, *4*, e2115, doi:10.7717/peerj.2115 (2016).
52. Groendahl, S., Kahlert, M., & Fink, P. The best of both worlds: A combined approach for analyzing microalgal diversity via metabarcoding and morphology-based methods. *PLoS ONE* **12**, 1-15, doi:10.1371/journal.pone.0172808 (2017).
53. Zou, S. *et al.* How DNA barcoding can be more effective in microalgae identification: a case of cryptic diversity revelation in *Scenedesmus* (Chlorophyceae). *Scientific Reports* **6**, 36822, doi:10.1038/srep36822 (2016).
54. Yilmaz, P., Kottmann, R., Pruesse, E., Quast, C., & Glöckner, F. O. Analysis of 23S rRNA genes in metagenomes - A case study from the Global Ocean Sampling Expedition. *Systematic and Applied Microbiology* **34**, 462-469, doi:10.1016/j.syapm.2011.04.005 (2011).
55. Adl, S. M. *et al*. The revised classification of eukaryotes. *Journal of Eukaryotic Microbiology* **59**, 429-493, doi:10.1111/j.1550-7408.2012.00644.x (2012).
56. Guiry, M.D. & Guiry, G.M. AlgaeBase. World-wide electronic publication, National University of Ireland, Galway, http://www.algaebase.org; searched on 201 (2018).
57. Jones, R. I. Mixotrophy in planktonic protists: an overview. *Freshwater Biology* **45**, 219-226, doi:10.1046/j.1365-2427.2000.00672.x (2000).
58. Parker, B. C. Facultative Heterotrophy in Certain Soil Algae from the Ecological Viewpoint. *Ecology* **42**, 381-386, doi:10.2307/1932089 (1961).
59. Starks, T., Shubert, L., & Trainor, F. Ecology of soil algae: a review. *Phycologia* **20**, 65-80, doi:10.2216/i0031-8884-20-1-65.1 (1981).
60. Porter, K. G. Phagotrophic phytoflagellates in microbial food webs. *Hydrobiologia* **159**, 89-97, doi:10.1007/BF00007370 (1988).
61. Rippka, R. Photoheterotrophy and chemoheterotrophy among unicellular blue-green algae. *Archiv Für Mikrobiologie* **87**, 93-98, doi:10.1007/BF00424781 (1972).
62. Kvíderová, J., Souquieres, C. E., & Elster, J. Ecophysiology of photosynthesis of Vaucheria sp. mats in a Svalbard tidal flat. *Polar Science*, doi:10.1016/j.polar.2018.11.006 (2018).
63. Agrawal, S. C. Factors affecting spore germination in algae - review. *Folia Microbiologica* **54**, 273-302, doi:10.1007/s12223-009-0047-0 (2009).

64. Shields, L. M., & Durrell, L. W. Algae in relation to soil fertility. *The Botanical Review* **30**, 92-128, doi:10.1007/BF02858614 (1964).
65. Starks, T. L., & Shubert, L. E. Colonization and Succession of Algae and Soil-Algal Interactions Associated With Disturbed Areas. *Journal of Phycology* **18**, 99-107, doi:10.1111/j.1529-8817.1982.tb03162.x (1982).
66. Cannone, J. J. *et al*. The Comparative RNA Web (CRW) Site: An online database of comparative sequence and structure information for ribosomal, intron, and other RNAs. *BMC Bioinformatics* **3**, 15, doi:10.1186/1471-2105-3-2 (2002).
67. Kumar, S., Stecher, G., & Tamura, K. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Molecular Biology and Evolution* **33**, 1870-1874, doi:10.1093/molbev/msw054 (2016).
68. Felsenstein, J. Evolutionary trees from DNA sequences: A maximum likelihood approach. *Journal of Molecular Evolution* **17**, 368-376. doi:10.1007/BF01734359 (1981).
69. Nawrocki, E. P., & Eddy, S. R. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* **29**, 2933-2935, doi:10.1093/bioinformatics/btt509 (2013).
70. Sauze, J. *et al.* The interaction of soil phototrophs and fungi with pH and their impact on soil CO_2, $CO^{18}O$ and OCS exchange. *Soil Biology and Biochemistry* **115**, 371-382, doi:10.1016/j.soilbio.2017.09.009 (2017).
71. Terrat, S. *et al.* Mapping and predictive variations of soil bacterial richness across France. *PLoS ONE* **12**, 5-8, doi:10.1371/journal.pone.0186766 (2017).
72. Terrat, S. *et al.* Molecular biomass and MetaTaxogenomic assessment of soil microbial communities as influenced by soil DNA extraction procedure. *Microbial Biotechnology* **5**, 135-41, doi:10.1111/j.1751-7915.2011.00307.x (2012).
73. Terrat, S. *et al.* Meta-barcoded evaluation of the ISO standard 11063 DNA extraction procedure to characterize soil bacterial and fungal community diversity and composition. *Microbial Biotechnology* **8**, 131-142, doi:10.1111/1751-7915.12162 (2015).

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 PR^2/SILVA, NCBI and AlgaeBase taxonomy.

Figure 4: Sequence distribution of the µgreen-db database at the Phylum level and grouped by Kingdom or supergroups. (A) Based on NCBI taxonomy according to Adl *et al.* (2012)[55] for the group classification, (B) Based on AlgaeBase taxonomy, (C) Based on PR^2 and SILVA taxonomy.

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

Additional Information

Supplementary information

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.

Supplementary Table 1 : Indexes of diversity of the soil photosynthetic microbial communities according to the lightning treatment.

Supplementary Table 2 : Occurence (A) and relative abundance (B) of the main phyla of soil photosynthetic microorganisms according to lightning treatment.

Supplementary Table 3 : Occurence (A) and relative abundance (B) of the main genera of soil photosynthetic microorganisms according to lightning treatment.

Supplementary Table 4: Identifcation of phyla shared between different soil samples.

Supplementary Table 5: Identifcation of genera shared between different soil samples.

Competing interests: The Authors declare no competing interests