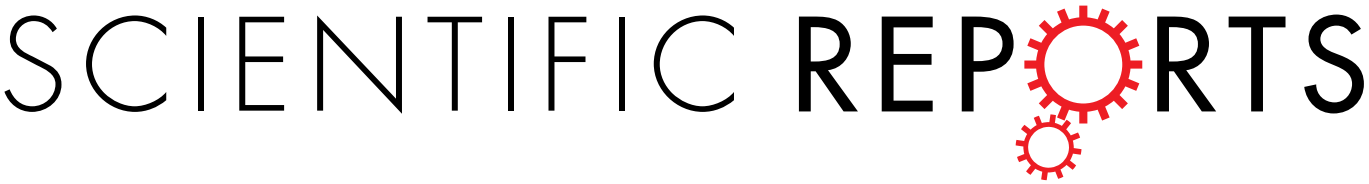

OPEN

How DNA barcoding can be more effective in microalgae identification: a case of cryptic diversity revelation in *Scenedesmus* (Chlorophyceae)

Received: 15 June 2016
Accepted: 21 October 2016
Published: 09 November 2016

Shanmei Zou, Cong Fei, Chun Wang, Zhan Gao, Yachao Bao, Meilin He & Changhai Wang

Microalgae identification is extremely difficult. The efficiency of DNA barcoding in microalgae identification involves ideal gene markers and approaches employed, which however, is still under the way. Although *Scenedesmus* has obtained much research in producing lipids its identification is difficult. Here we present a comprehensive coalescent, distance and character-based DNA barcoding for 118 *Scenedesmus* strains based on *rbcL*, *tufA*, ITS and 16S. The four genes, and their combined data *rbcL* + *tufA* + ITS + 16S, *rbcL* + *tufA* and ITS + 16S were analyzed by all of GMYC, P ID, PTP, ABGD, and character-based barcoding respectively. It was apparent that the three combined gene data showed a higher proportion of resolution success than the single gene. In comparison, the GMYC and PTP analysis produced more taxonomic lineages. The ABGD generated various resolution in discrimination among the single and combined data. The character-based barcoding was proved to be the most effective approach for species discrimination in both single and combined data which produced consistent species identification. All the integrated results recovered 11 species, five out of which were revealed as potential cryptic species. We suggest that the character-based DNA barcoding together with other approaches based on multiple genes and their combined data could be more effective in microalgae diversity revelation.

Microalgae are diverse and ubiquitous in aquatic and some terrestrial habitats. They play a crucial role in the global ecosystem for hundreds of millions of years[1,2]. The revelation of biodiversity for microalgae is significant to nature conversation, food safety and better understanding the patterns of ecosystem functioning. The Chlorophyta form a large and morphologically diverse clade of marine, freshwater and terrestrial green algae. Although Chlorophyta have a long history of study, they are still poorly understood taxonomically and phylogenetically due to their much diversity, especially for microalgae. Microscopic green algae are mainly identified based on the general shape of their vegetative cells, the position of chloroplasts and pyrenoids, and the ultrastructural characteristics[3,4]. However, identification of microalgae can be very difficult since most species lack obvious structural features and some of the observable characteristics are variable within species. Since the identification of microalgae typically requires the use of a microscope, sometimes at very high magnification, taxonomy of it is somewhat inaccessible to non-specialists and sometimes rapid identification of some species even by microscopy is impossible. Even worth, the number of taxonomists is declining seriously.

The genus *Scenedesmus* (Chlorophyceae) is one of the most common freshwater genera, which can be as ideal microalgae for producing biofuel owing to the substantial amounts of lipids, proteins and carbohydrates. Most species of *Scenedesmus* are found across the world. *Scenedesmus* includes all autosporic coccal green algae with flat or curved coenobia[5], species of which are poor in characteristics and are differentiated mainly by cell shape or coenobial habitus[6]. The extremely diverse morphologies make identification of *Scenedesmus* very difficult[7]. Currently, there are 74 taxonomically accepted species of *Scenedesmus*[8], but not determinate. It is hard to distinguish them just based on the limited obscure characters. At present, there are few studies about taxonomic

College of Resources and Environmental Science, Nanjing Agricultural University, Nanjing 210095, PR China. Correspondence and requests for materials should be addressed to C.W. (email: chwang@njau.edu.cn)

assignments of *Scenedesmus* using molecular tools[5,9], and these studies just used single gene or limited analysis. It is urgent to give a revision to the classification of *Scenedesmus*.

DNA barcoding is currently a widely used and effective tool for fast and accurate species identification[10–15]. The efficient "universal barcode gene" across all forms of life is a key factor for success application of DNA barcoding. However, the weakest spot of DNA barcoding is the obvious fact that no gene can serve as an ideal barcode for all forms of life, i.e., be always invariant within species but different among species. While the application of cytochromeoxidase- I (COI) has been highly successful in a wide range of animal taxa[16], the attempts to employ a single barcode for plants identification remains a vain hope for a longtime. The cpDNA two-locus combination *rbcL* + *matK* has been recommended as the universal DNA barcode for land plants[17]. However, the discriminatory power of the *rbcL* + *matK* sequence combination is still very far from the usually higher (though variable) rate of over 90% success of COI in animals[16,18]. Moreover, the *matk* is absent in algae. Thus, the efficient "DNA barcodes" for algae are more ambiguous, and it seems more effective to employ multiple genes for barcoding algae. Several gene loci, e.g. *rbcL*, ITS and *tufA*, have been recommend as the promising DNA barcodes for some green algae[19,20].

Members of the barcoding community have put forward several different methods of distinguishing species, including the coalescent, distance and character-based methods. Traditional DNA barcoding[10] constructs Neighbor-Joining (NJ), Bayesian or Maximum likelihood trees for species identification, and calculates a genetic distance between species and assigns a cutoff value (the 'barcode gap') to divide OTUs into species. The distinct clades in a phylogenetic tree are often interpreted as species. However, monophyly of a set of taxa can occur by chance within a larger panmictic group as a result of the coalescent process. Recently, the P ID(Liberal) method of species delimitation is advanced for the exploration of species boundaries which are identified by deep divergences in phylogenetic trees[21]. Another popular tree-based approach, the General Mixed Yule-coalescent[22,23], is a species delimitation method that estimates species boundary directly from branching rates in a phylogenic tree rather than actual sequence data and attempts to statistically model the point on a time calibrated (ultrametric) phylogeny. The single-threshold approach is generally preferred for GMYC analysis[24]. The poisson tree process (PTP) model is another tree-based method that distinguishes specimens in both populations and species level using coalescence theory[25]. It has been proposed that P ID, GMYC and PTP approaches could be as complementary analysis to the phylogenetic tree identification of traditional DNA barcoding[26–28]. For distance-based barcoding approach, due to the absence of a "barcoding gap", the specimen identification based on intraspecific variation vs. interspecific divergence has already been shown to be impossible for some taxonomic groups, especially for the plants[29–31]. Recently, a new distance method, called Automatic Barcode Gap Discovery (ABGD), is proposed as an automatic procedure that sorts the sequences into hypothetical species based on the barcode gap which can be observed whenever the divergence among organisms belonging to the same species is smaller than divergence among organisms from different species[32]. Another new barcoding method, the character-based barcoding which is different from phylogenetic tree and distance analysis, is based on the fundamental concept that members of a given taxonomic group share diagnostic characters that are absent from comparable groups[33,34]. It can provide better resolution in species identification and cryptic species revelation of some organisms (including some plants) in several cases where distance-based methods fail to distinguish species[12,31].

Here we present the comprehensive DNA barcoding taxonomic assignment of *Scenedesmus* based on four gene loci and their combined data, the *rbcL* gene (encodes the large subunit of Rubisco), the *tufA* gene (encoding elongation factor), the ITS (internal transcribed spacer region) and 16S. This study represents one of the efforts to use DNA barcode data as a taxonomic tool for exploring biodiversity of microalgae. We employ a novel combination of methods to reach this goal, examining the congruence of OTUs (operational taxonomic units) resulting from coalescent (P ID, GMYC and PTP), distance (ABGD) and character-based DNA barcoding. The objectives of this study are: (1) to identify the confused *Scenedesmus* strains; (2) to uncover the cryptic species in *Scenedesmus*; (3) to test the efficiency of multiple gene markers and barcoding approaches; (4) to evaluate how DNA barcoding can be more effective in microalgae diversity revelation.

Results

A total of 68 *rbcL*, 80 ITS, 64 16S and 54 *tufA* sequences of *Scenedesmus* samples and outgroups were analyzed (Supplementary Table 1). The samples from this study were selected from many regions of China (Fig. 1). The newly obtained sequences from this study were submitted to the GenBank Barcode database with accession numbers KT777944- KT778122 and KT818697- KT818720. The *rbcL* sequence had a length of 1323 bp with 740 variable nucleotide sites (55.9%), 605 of which were non-synonymous substitutions. The ITS sequence had a length of 1354 bp with 582 variable nucleotide sites (43.0%), 514 of which were non-synonymous substitutions. The 16S sequence had a length of 436 bp with 209 variable nucleotide sites (48.0%), 179 of which were non-synonymous substitutions. The *tufA* sequence had a length of 789 bp with 459 variable nucleotide sites (58.2%), 393 of which were non-synonymous substitutions.

Single marker barcoding. Generally, the NJ, Bayesian and Maximum Likelihood trees of *rbcL* recovered consistent groups (Fig. 2, Supplementary Fig. 1 and Supplementary Fig. 2), including the potential cryptic lineages (e.g *Scenedesmus deserticola* I,II,III and *Scenedesmus obliquus* I,II,III). As can be seen in Supplementary Fig. 3, Supplementary Fig. 4, Supplementary Fig. 5, Supplementary Fig. 6, Supplementary Fig. 7, Supplementary Fig. 8, Supplementary Fig. 9, Supplementary Fig. 10 and Supplementary Fig. 11, the NJ, Bayesian and Maximum Likelihood trees of *tufA*, ITS and 16S also retrieved generally consistent groups. However, it could also be seen that some clades could not be separated clearly in the *rbcL*, ITS, 16S and *tufA* phylogenetic trees, e.g. the *S. deserticola* I and *S. deserticola* II clades.

The distance variation of *rbcL*, ITS, *tufA* and 16S among taxa assignments by P ID, ABGD, GMYC, PTP and CAOS analysis were conducted. The results showed that the mean intraspecific distance of *rbcL* was from 0% to

Figure 1. Map showing the locations from which *Scenedesmus* strains were obtained from China (shown as dark dots). The map was created using Quantum GIS 1.8.0 (http://www.qgis.org/) based on a map from Natural Earth (version 2.0.0).

1.64% while the mean interspecific distances was from 0.4% to 23.0% (Supplementary Table 2). All the pairwise distance of *rbcL* ranged from 0% to 32.5%. The mean intraspecific distance of ITS ranged from 0% to 4.08% while the mean interspecific distances was from 1.1% to 21.70% (Supplementary Table 3). All the pairwise distance of ITS ranged from 0% to 24.3%. For 16S and *tufA*, the mean intraspecific distance ranged from 0% to 2.84% and 0% to 4.54% respectively while the mean interspecific distances was from 0% to 35.1% and 1.9% to 38.0% respectively (Supplementary Table 4 and Supplementary Table 5). The pairwise distance of 16S and *tufA* ranged from 0% to 36.7% and 0% to 38.0% respectively. All these distance variation showed that a DNA "barcoding gap" did not existed in *rbcL*, 16S, ITS and *tufA* sequences.

Based on the distance-based approach ('Barcode-gap analysis') as implemented in the software ABGD, different groups as candidate species were produced for *rbcL*, ITS, 16S and *tufA* gene sequences. The ABGD analysis of *rbcL*, ITS, 16S and *tufA* did not always produce consistent genetic groups for all species (Fig. 2, Supplementary Fig. 3, Supplementary Fig. 6 and Supplementary Fig. 9). For *rbcL*, the ABGD analysis revealed 18 genetic groups when using restrictive values with priori genetic distance thresholds between 0.46 and 0.77% (Fig. 2 and Supplementary Fig. 12). This result was generally consistent with the ABGD analysis of ITS in which 16 groups were produced with prior maximal distance 1.29% (Supplementary Fig. 6 and Supplementary Fig. 13). In both analyses, *S. deserticola* was split into several groups. All the specimens studied were split into more groups in ABGD analysis of 16S where 21 groups were revealed with priori genetic distance thresholds between 0.28 and 0.46% (Supplementary Fig. 9 and Supplementary Fig. 14). However, the ABGD analysis of *tufA* could not separate most specimens in which only 10 groups were produced with priori genetic distance thresholds between 0.28 and 2.15% (Supplementary Fig. 3 and Supplementary Fig. 15).

Based on the Bayesian analysis, the tree-based hypotheses were reevaluated for species hypothesis testing by P ID species boundary delimitation. Most candidate species were recovered as monophyletic clades in all of *rbcL*, ITS, *tufA* and 16S genes except *S. deserticola* I which was not monophyletic in P ID analysis of *rbcL*, ITS and 16S (Fig. 2, Supplementary Fig. 3, Supplementary Fig. 6 and Supplementary Fig. 9). Several cryptic clades of *S. obliquus* were also not monophyletic in 16S P ID analysis. All delimited species of *rbcL*, ITS, *tufA* and 16S sequences possessed a P ID (Liberal) value $P > 0.5$ except two clades in 16S analysis (Supplementary Table 6, Supplementary Table 7, Supplementary Table 8 and Supplementary Table 9).

On the whole, the specimens analyzed were oversplitted by the GMYC model in comparison with the ABGD and P ID analysis in *rbcL*, ITS and *tufA* genes (Fig. 2, Supplementary Fig. 3, Supplementary Fig. 6, Supplementary Fig. 9, Supplementary Fig. 16, Supplementary Fig. 17 and Supplementary Fig. 18). The results of single threshold analysis for the *rbcL*, ITS and *tufA* gene suggested 25, 25 and 19 groups respectively. In 16S gene dataset ten GMYC entities recovered was generally congruent with the P ID species boundary delimitation Supplementary Fig. 19.

For bPTP approach, the maximum-likelihood identification was showed since it produced better resolution than bayesian identification. The taxonomic clades produced by bPTP approach was variable among *rbcL*, ITS, *tufA* and 16S genes (Fig. 2, Supplementary Fig. 6, Supplementary Fig. 20, Supplementary Fig. 21, Supplementary Fig. 22 and Supplementary Fig. 23). It was apparent that the *rbcL* and ITS genes generated better identification than that of *tufA* and 16S genes. The *tufA* and 16S genes did not distinguish most species (Supplementary Fig. 22 and Supplementary Fig. 23). For *rbcL*, it recognized 19 independent entities, some of which were consistent with

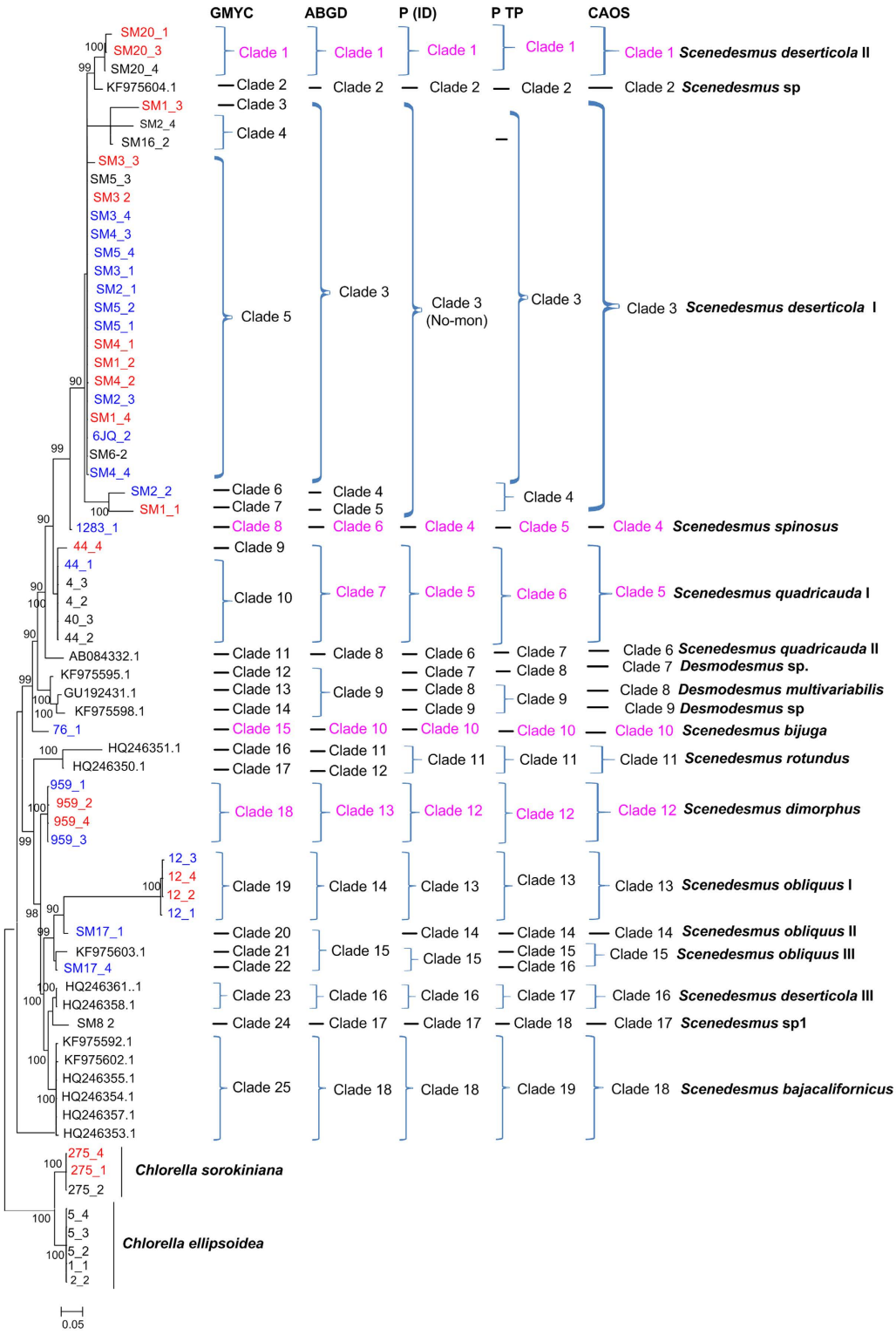

Figure 2. Bayesian phylogenetic tree for the *rbcL* gene. Vertical brace on the right indicate the clades detected by the tree-based GMYC, PID, PTP and the distance-based ABGD approach, the character-based CAOS and the taxa assignment. The clades highlighted in pink were also detected by 16S, ITS and *tufA* gene loci. For samples colored in red, 16S, ITS and *tufA* sequences were also available. For specimens colored in blue, two of 16S, ITS and *tufA* sequences were available.

the groups revealed by GMYC, ABGD or P ID analyses (Fig. 2). For ITS, 11 clades were identified by PTP analysis, which were not completely consistent with the identification of other methods (Supplementary Fig. 6 and Supplementary Fig. 21).

Species (Cryptic lineage number, Genbank number)	Positions																										
	172	183	211	214	220	398	456	458	503	545	557	563	622	644	750	762	790	873	938	947	976	989	1007	1016	1025	1031	1046
Scenedesmus deserticola I	A	G	G	G	A	T	T	A	T	T	T	C	G	T	T	T	G	C	T	T	T	C	T	T	T	G	T
Scenedesmus deserticola II	A	G	G	G	G	T	C	T	T	T	C	C	G	T	T	T	G	C	T	T	T	C	T	-	T	C	C
Scenedesmus deserticola III	T	G	A	G	T	A	T	A	C	T	C	T	T	A	T	A	G	T	A	T	T	C	C	T	C	C	C
Scenedesmus sp. (KF975604)	A	G	G	G	A	T	T	A	T	T	C	C	C	A	T	T	G	T	T	T	T	C	T	T	T	T	T
Scenedesmus spinosus	T	G	A	G	T	G	C	T	T	C	C	C	G	T	T	T	G	C	T	T	T	C	T	T	T	T	T
Scenedesmus bijuga	T	G	G	A	G	G	T	A	C	C	C	C	G	T	C	A	G	T	T	A	T	C	C	C	C	C	T
Scenedesmus quadricauda I	T	G	A	A	T	G	C	T	T	C	C	C	G	A	T	T	T	T	T	A	T	C	T	C	C	C	C
Scenedesmus quadricauda II	T	G	G	A	G	G	C	T	C	T	C	C	G	T	C	A	T	T	T	A	T	C	T	C	C	T	C
Desmodesmus multivariabilis	T	G	A	A	T	A	T	A	C	C	T	C	G	A	C	T	G	C	T	A	T	C	C	C	C	C	T
Desmodesmus sp. (KF975598)	A	A	G	G	T	A	T	A	C	C	T	C	G	A	C	T	G	C	C	T	G	T	T	T	T	T	C
Desmodesmus sp. (KF975595)	T	G	G	A	A	A	T	A	C	T	T	C	G	A	C	T	G	C	T	A	T	C	C	C	C	C	C
Scenedesmus rotundus	A	G	A	G	T	A	T	A	T	C	C	C	T	T	C	T	G	C	A	A	T	C	T	C	C	C	C
Scenedesmus bajacalifornicus	T	G	A	G	T	A	T	A	C	T	C	C	C	A	T	A	G	T	A	A	T	C	C	C	C	C	C
Scenedesmus dimorphus	T	G	A	G	G	A	C	T	C	C	C	C	T	T	C	A	G	C	A	A	T	C	C	C	C	-	-
Scenedesmus sp1	T	G	A	G	T	A	C	G	C	T	C	C	T	A	T	A	G	C	A	T	T	C	C	C	C	C	C
Scenedesmus obliquus I	T	G	G	G	A	A	T	A	T	T	C	T	C	T	G	C	A	A	C	T	A	A	-	-	-	-	-
Scenedesmus obliquus II	T	G	G	G	A	A	T	A	T	T	C	C	C	T	T	A	T	C	T	T	T	C	T	T	T	C	-
Scenedesmus obliquus III	T	G	G	G	A	A	T	A	T	T	C	C	C	T	T	A	T	C	A	A	T	C	T	C	C	C	C

Table 1. Combinations of diagnostic nucleotides for each of the 18 *Scenedesmus* taxa recovered in Fig. 2 by CAOS. Nucleotide numbers refer to 27 selected positions on the *rbcL* sequences.

The relatively congruent defined *Scenedesmus* groups based on ABGD, P ID, PTP and GMYC analysis, as well as the morphological characters, were analyzed for searching for diagnostic characters. A total of 18, 14, 14 and 15 clades recovered by *rbcL*, ITS, *tufA* and 16S genes were analyzed respectively by character-based DNA barcoding (Fig. 2, Supplementary Fig. 3, Supplementary Fig. 6 and Supplementary Fig. 9). It was found that all the *Scenedesmus* groups including the possible cryptic lineages and unknowns were clearly distinguished in the character-based DNA barcoding. In the *rbcL* gene region of 18 *Scenedesmus* taxa recovered in Fig. 2, character states at 27 nucleotide positions were detected (Table 1). All the 18 clades revealed a unique combination of character states at 27 nucleotide positions with more than three CAs, including the cryptic lineages, unknowns and species represent by single specimens. As can be seen in Supplementary Table 10, 14 clades recovered in Supplementary Fig. 6, also revealed a unique combination of character states at 28 nucleotide positions with more than three CAs, for ITS sequences. The 16S character states for the 15 *Scenedesmus* clades (Supplementary Fig. 9) were shown in Supplementary Table 11. At 25 nucleotide positions of the 16S gene region more than four CAs were revealed for each clade. The *tufA* character-based DNA barcode were shown in Supplementary Table 12, in which 14 defined *Scenedesmus* clades recovered in Supplementary Fig. 3 revealed a unique combination of character states at 39 positions.

Comparison of species delimitation by traditional DNA barcoding, ABGD, P ID, GMYC, PTP and character-based methods in four gene loci. As seen in Fig. 2, Supplementary Fig. 3, Supplementary Fig. 6 and Supplementary Fig. 9, relatively congruent clades were recovered by *rbcL*, ITS, 16S and *tufA* genes. For example, the species *Scenedesmus deserticola* II, *Scenedesmus deserticola* I, *Scenedesmus bijuga, Scenedesmus dimorphus*, *Scenedesmus quadricauda* and *Scenedesmus bajacalifornicus* were recovered as monophyletic clades in all of *rbcL*, ITS, 16S and *tufA* genes by more than three methods of Bayesian trees, ABGD, P ID, GMYC, PTP and character-based. However, on the other hand, the Bayesian trees, ABGD, P ID, GMYC, PTP and character-based methods did not always generate consistent clades in each of *rbcL*, ITS, 16S and *tufA* genes. As a whole, the GMYC method produced more OTUs than other methods. The PTP approach just could identify most species in *tufA* and ITS genes. It was apparent that the character-based method recovered consistent clades among *rbcL*, ITS, 16S and *tufA* gene loci. In comparison with *rbcL*, 16S and ITS, the *tufA* showed a higher intraspecific and interspecific divergence, more consistent groups among GMYC, ABGD and P ID methods, and more diagnostic characters.

Combined markers barcoding. The combined data of *rbcL* + ITS + 16S + *tufA*, *rbcL* + *tufA* and ITS + 16S were analyzed respectively, based on the ABGD, P ID, GMYC, PTP and character-based methods. The *rbcL*, ITS, 16S and *tufA* sequences were all available for *S. deserticola* I, *S. deserticola* II, *S. quadricauda*, *S. obliquus*, and *S. dimorphus* (Fig. 3). The NJ, Bayesian and Maximum Likelihood trees of *rbcL* + ITS + 16S + *tufA* separated the five clades clearly, with high support. The ABGD analysis revealed 4 genetic groups when using restrictive values with priori genetic distance thresholds between 2.15% (Fig. 3 and Supplementary Fig. 24), which did not distinguish *Scenedesmus deserticola* I and *Scenedesmus deserticola* II, and *S. obliquus*, and *S. dimorphus*. The five species were divided into 8 genetic groups by the GMYC model (Supplementary Fig. 25), which was consistent with the single gene result that GMYC model could separate one species as more clades. P ID species boundary delimitation revealed all the several species as monophyletic clades. The delimited species of *rbcL* + ITS + 16S + *tufA* sequences possessed a P ID (Liberal) value $P>0.9$ (Supplementary Table 13). As the GMYC analysis, the PTP approach also divided the *Scenedesmus deserticola* I into several clades (Fig. 3 and Supplementary Fig. 26). The

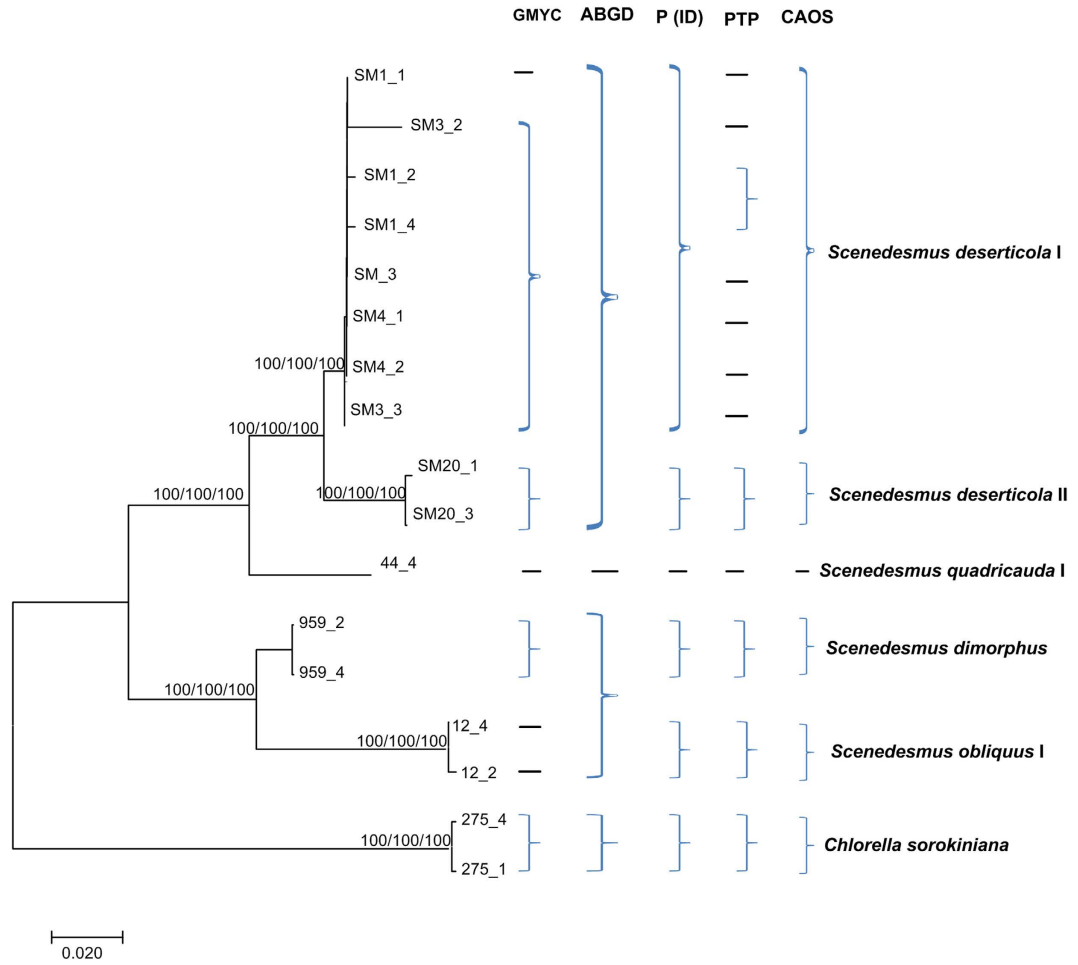

Figure 3. Bayesian phylogenetic tree for the *rbcL* + ITS + 16S + *tufA* data. The NJ and Maximum Likelihood bootstrap were also indicated. Vertical brace on the right indicate the clades detected by the tree-based GMYC, PID, PTP, the distance-based ABGD approach and the character-based CAOS assignment.

Species (cryptic lineage number)	Position																		
	35	137	149	195	278	383	1578	1602	1622	1737	2367	2535	2556	2906	2923	3227	3628	3646	3647
Scenedesmus deserticola II	C	T	T	G	T	T	A	C	A	A	C	A	A	T	T	A	T	C	T
Scenedesmus deserticola I	T	C	T	G	T	T	A	T	A	G	T	A	A	C	A	A	T	A	T
Scenedesmus quadricauda I	G	A	A	T	T	T	A	C	T	A	C	A	C	T	C	A	T	T	T
Scenedesmus dimorphus	T	A	A	A	T	A	A	C	A	T	A	T	T	C	T	A	C	G	A
Scenedesmus obliquus I	T	A	A	A	C	A	C	A	G	T	A	C	T	C	T	T	C	G	G

Table 2. Combinations of diagnostic nucleotides for each of the 5 *Scenedesmus* taxa recovered in Fig. 3 by CAOS analysis. Nucleotide numbers refer to 19 selected positions on the *rbcL* + ITS + 16S + *tufA* sequences.

character-based barcoding separated the five species more clearly (Table 2). It was indicated that every species revealed in Table 2 possessed more than 7 character attributes in only 19 positions.

The *rbcL* + *tufA* sequences revealed similar resolution to *rbcL* + ITS + 16S + *tufA*. Both *rbcL* and *tufA* sequences were all available for 7 species (Fig. 4). The 7 clades were also clearly separated by NJ, Bayesian and Maximum Likelihood analysis of *rbcL* + *tufA* with strong support. The ABGD analysis revealed 8 genetic groups when using restrictive values with priori genetic distance thresholds between 0.1–0.45% (Fig. 4 and Supplementary Fig. 27). The 7 species were also oversplitted by the GMYC model (Supplementary Fig. 28). All the several species were also revealed as monophyletic clades by P ID species boundary delimitation, and the delimited species of *rbcL* + *tufA* sequences possessed a P ID (Liberal) value $P > 0.8$ (Supplementary Table 14). It was apparent that the PTP method divided *S. deserticola* I into more separate clades (Supplementary Fig. 29). As *rbcL* + ITS + 16S + *tufA* recovered, the character-based barcoding also separated the seven species clearly (Table 3). Seven defined *Scenedesmus* clades recovered in Fig. 4 revealed a unique combination of character states at 29 positions with more than 8 character attributes.

The identification of ITS + 16S sequences by GMYC, P ID, PTP, ABGD and character-based methods were showed in Fig. 5. It was indicated that generally the ABGD, P ID and character-based barcoding recovered same

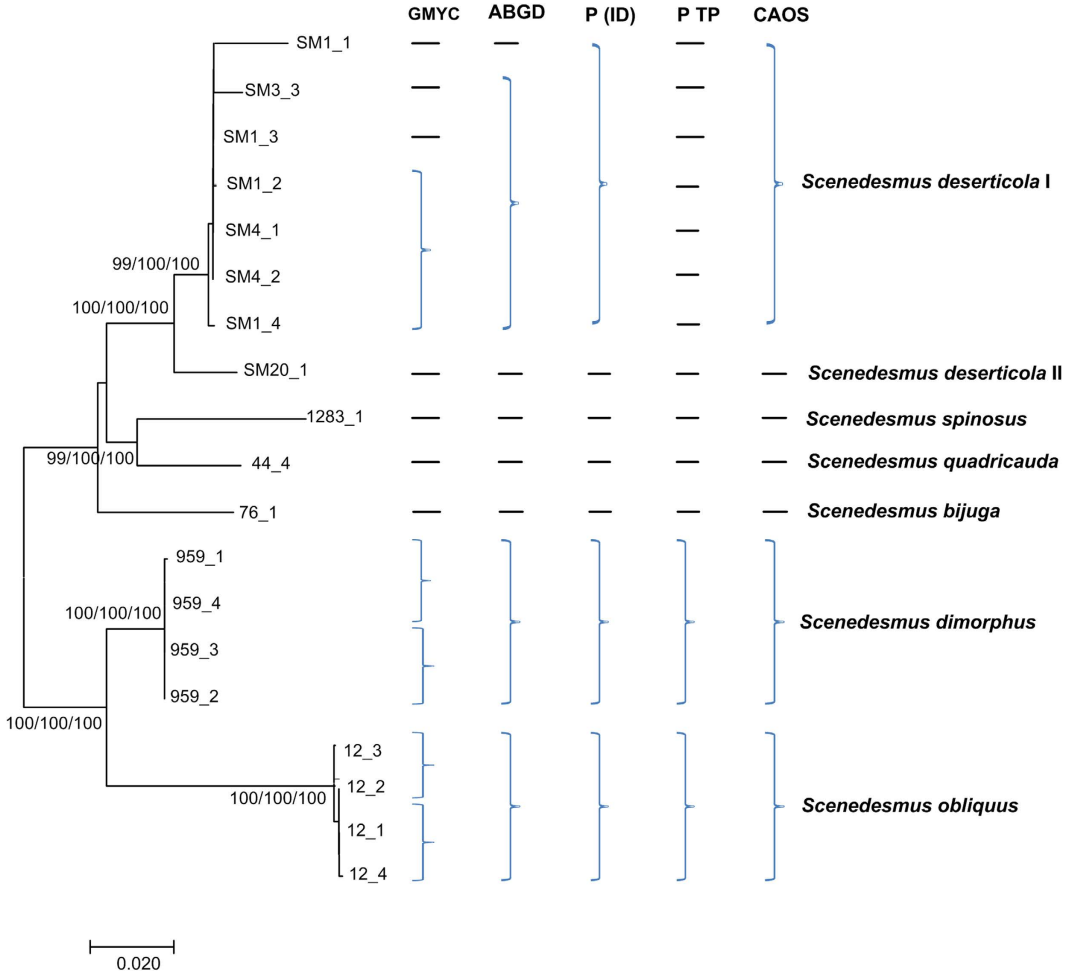

Figure 4. Bayesian phylogenetic tree for the *rbcL* + *tufA* data. The NJ and Maximum Likelihood bootstrap were also indicated. Vertical brace on the right indicate the clades detected by the tree-based GMYC, PID, PTP, the distance-based ABGD approach and the character-based CAOS assignment.

Species (cryptic lineage number)	Position																													
	118	222	753	789	793	811	853	874	881	899	904	906	938	945	948	951	978	990	1004	1015	1033	1075	1365	1383	1485	1608	1620	1746	1767
Scenedesmus deserticola I	T	A	T	A	G	C	A	C	C	G	T	T	G	A	T	T	T	C	C	T	G	C	C	T	A	T	C	T	T
Scenedesmus deserticola II	T	G	T	A	G	C	A	C	C	G	T	T	G	G	T	A	T	C	C	C	C	A	T	C	G	T	C	T	T
Scenedesmus spinosus	T	T	T	A	G	C	A	C	C	G	T	T	G	A	T	T	T	C	C	T	T	A	A	C	A	A	A	A	T
Scenedesmus quadricauda I	A	T	T	A	T	C	A	T	C	A	C	T	G	A	A	A	A	G	A	C	C	A	G	T	A	T	T	T	A
Scenedesmus bijuga	T	G	C	T	G	A	A	T	T	T	C	T	C	A	A	A	A	C	C	C	C	T	A	C	A	T	C	T	T
Scenedesmus dimorphus	A	G	C	A	G	C	T	C	T	T	C	C	C	T	A	T	T	C	C	C	-	-	T	T	T	T	T	T	A
Scenedesmus obliquus I	T	A	G	G	A	T	C	A	A	A	A	G	T	C	C	C	C	T	T	A	A	-	T	T	G	C	T	C	A

Table 3. Combinations of diagnostic nucleotides for each of the 7 *Scenedesmus* taxa recovered in Fig. 4 by CAOS analysis. Nucleotide numbers refer to 29 selected positions on the *rbcL* + *tufA* sequences.

resolution where seven genetic lineages were clearly separated, including the potential cryptic lineages *S. deserticola* and *S. obliquus*. The ABGD analysis revealed 8 genetic groups. The P ID species boundary delimitation of ITS + 16S sequences revealed all species as monophyletic clades, and the delimited species possessed a P ID (Liberal) value $P > 0.6$ (Supplementary Table 15). Both the GMYC model and PTP analysis generated more genetic lineages compared with the ABGD, P ID and the character-based methods (Fig. 5, Supplementary Fig. 30 and Supplementary Fig. 31). The character-based barcoding separated the seven taxonomic lineages clearly (Fig. 5, Table 4). It was indicated that every clade revealed in Table 4 possessed more than 5 character attributes in only 15 positions.

Discussion

Identification of microalgae species is often difficult based on the morphological characters due to their tiny body, unobvious structural features and variable characters within species. DNA barcoding has developed to be a useful tool for species discrimination. However, how DNA barcoding can be more effective in microalgae

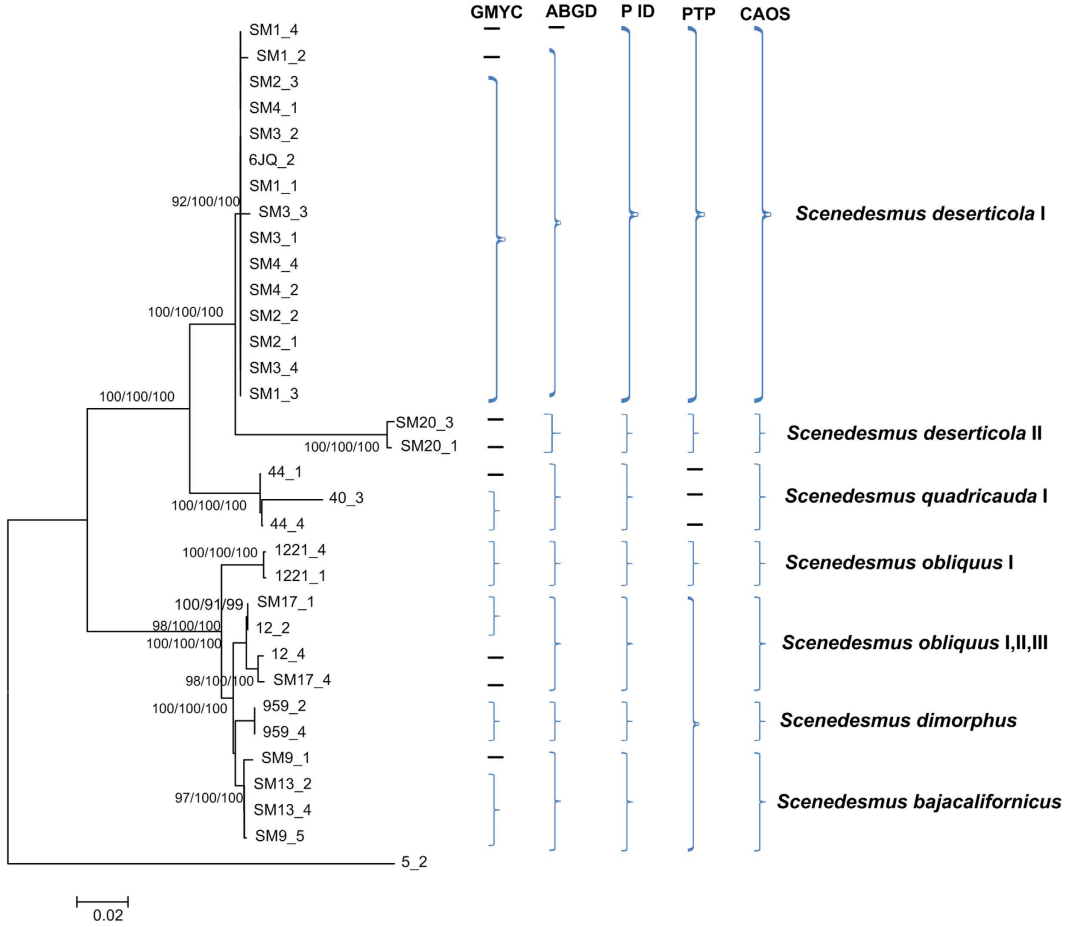

Figure 5. Bayesian phylogenetic tree for the ITS + 16S data. The NJ and Maximum Likelihood bootstrap were also indicated. Vertical brace on the right indicate the clades detected by the tree-based GMYC, PID, PTP, the distance-based ABGD approach and the character-based CAOS assignment.

Species (Cryptic lineage number)	Position														
	507	789	793	820	821	822	865	866	867	1289	1416	1440	1522	1534	1535
Scenedesmus deserticola I	C	T	C	T	C	A	A	C	C	G	A	C	C	C	T
Scenedesmus deserticola II	C	C	C	T	A	A	A	C	A	G	A	C	A	A	T
Scenedesmus quadricauda I	C	T	T	T	C	A	T	A	A	G	T	T	A	T	T
Scenedesmus acuminatus	T	C	A	G	T	G	A	A	C	A	A	A	A	G	A
Scenedesmus dimorphus	C	C	T	T	C	T	T	A	T	A	A	A	G	G	A
Scenedesmus obliquus	T	C	T	T	C	T	C	A	C	G	A	A	A	G	G
Scenedesmus bajacalifornicus	C	C	T	C	T	C	T	A	C	A	A	G	A	G	A

Table 4. Combinations of diagnostic nucleotides for each of the 7 *Scenedesmus* taxa recovered in Fig. 5 by CAOS analysis. Nucleotide numbers refer to 15 selected positions on the ITS +16S sequences.

diversity revelation by selecting suitable markers and barcoding approaches is ambiguous. The identification of *Scenedesmus* by morphological characters is often confused, which hinders us from selecting optimal *Scenedesmus* strains for producing biofuel. In this study, the morphological characters were initially used to identify the strains. Some strains could easily be identified to species level. However, although some strains could be identified to species level they seemed to be potential cryptic species due to the various morphological characters. Here we employ multiple genes to assign *Scenedesmus* species based on various barcoding methods, and evaluate their congruent results.

Generally, different gene markers produced consistent resolution for species discrimination. Most species, including the unknowns, were separated clearly in barcoding analysis of *rbcL*, ITS, 16S, *tufA* and the three combined data. For example, the species *S. bajacalifornicus*, *S. dimorphus* and *S. quadricauda* were all clearly distinguished from other species by P ID, ABGD, GMYC and CAOS analysis of *rbcL*, ITS and 16S sequences, or the three combined data. More importantly, some species which were identified as potential cryptic species complexes were also divided into several separate lineages by the gene sequences. As can be seen in Figs 2, 3, 4 and 5,

and Supplementary Fig. 3, Supplementary Fig. 6 and Supplementary Fig. 9, *S. deserticola* were clearly retrieved as separate clades in all barcoding analysis of *rbcL*, ITS, 16S, *tufA* and the three combined data, including the sequences from Genbank. Additionally, *S. obliquus*, *S. quadricauda*, *S. spinosus* and *S. acuminatus* were also recovered as separate clades. These species might be as overlooked cryptic species. Since *S. deserticola*, *S. obliquus*, *S. quadricauda* and *S. acuminatus* are all considered to be promising candidates for biodiesel production[35,36] their correct identification is significant to the application as biodiesel feedstock. In sum, the molecular study of DNA barcoding here gave new insights into the taxonomic assignment of *Scenedesmus*.

Since mitochondrial genes are not suitable for barcoding plantae[37] the chloroplast and nuclear genomes with high substitution rates could be employed to search for plant barcodes. Although *rbcL* + *matK* are proposed as candidates of DNA barcode loci for plants it has been proved that the *matK* or *rbcL* alone can not be as a suitable universal barcode[37,38]. In the present study, the four gene loci and the combined data generally produced congruent clades among Bayesian, ABGD, GMYC, P ID, PTP and character-based analysis, respectively. By comparison, all of the four-marker combination of *rbcL* + ITS + 16S + *tufA* and two-marker combination of *rbcL* + *tufA* and ITS + 16S showed a much higher proportion of resolution success than the single genes, including the more consistent groups among GMYC, ABGD, PTP, and P ID analysis, and many more diagnostic characters. Among the four genes, the *tufA* generally produced better resolution than other genes, also including the higher intraspecific and interspecific divergence, more consistent groups among GMYC, ABGD and P ID analysis, and many more diagnostic characters.

This study represents one of the first efforts to examine the congruence of barcoding results from multiple delimitation methods. The P ID, PTP and Character methods were particularly included in this study in comparison with previous barcoding evaluation[26,27]. The traditional barcoding analysis, including the phylogenetic (NJ, Bayesian and Maximum Likelihood analysis) and intra and interspecific distance analysis, was first conducted to discriminate species. Due to the drawback of monophyly-based species identification[39,40] it is more likely the phylogenetic trees could be used as the initial step to identify putative independently-evolving lineages. It has been proposed that an optimal path to understand species boundaries is starting with a tree-based framework to develop the initial species hypotheses where distinct clades can be identified as divergent monophyletic population clusters[27]. In this study, as a whole, the NJ, Bayesian and Maximum Likelihood phylogenetic trees produced consistent topology for each marker of *rbcL*, ITS, 16S, *tufA*, *rbcL* + *ITS* + *16S* + *tufA*, *rbcL* + *tufA* and ITS + 16S. For all of *rbcL*, ITS, 16S and *tufA* sequences, although the interspecific distance was generally higher than the intraspecific distance, there was no barccoding gap between them. That is, the minimum interspecific distance is not higher than the maximum intraspecific distance, which contradicts the criterion of species identification with sequences distance[17]. In this context, multiple species identification methods should be incorporated to barcoding species.

Recently, it was proposed that incorporation of multiple lines of methodologies were more useful for barcoding species[26–28,41]. In these previous studies, one or two gene loci were conducted through incorporation of several barcoding approaches. In this study, four gene loci and their combined data were employed to give more evidence on the species identification. It was indicated that: (1) the GMYC model and PTP analysis generated more genetic groups; (2) the ABGD approach always recovered various genetic groups among the single marker and combined marker data; (3) as expected, all the four single data and three combined data produced consistent groups in the character-based analysis. Similar to previous studies[22,26,42,43], GMYC typically generates more OTUs (operational taxonomic units) than other methods for *rbcL*, ITS, *tufA* and the combined data, and errors in the ultrametric gene tree that underpins the analysis will influence final results. The GMYC results, produced by 16S gene which may not be suitable for barcoding plantae due to the very low rates of substitution[37], however, were relatively congruent with the resolution produced by phylogenetic analysis and other barcoding methods. It could be inferred that the GYMC method which has a strong theoretical basis may be more suitable for analyzing gene sequences evolving slowly. The PTP analysis generated various resolution among the single genes and the combined data. As the GMYC resolved, the PTP method also generated more taxonomic groups and oversplitted some species in *rbcL* and the combined data. On the other hand, the PTP analysis of *tufA*, ITS and 16S could not discriminate most species. The ABGD generated diverse outcomes among the four gene loci and the combined data, which not only over-split some certain lineages but also clustered together some lineages. As a whole, some species could be separated by ABGD analysis in all markers, but some groups generated by ABGD were not consistent with analysis of other barcoding methods. For all of *rbcL*, ITS, 16S and *tufA* genes, P ID (Liberal) could recognize most taxonomic species inferred from the phylogeetic trees, including the potentially cryptic lineages. However, only for the *tufA* and combined data sequences, all the diverged lineages were recovered as monophyletic clades. In *rbcL*, ITS and 16S genes, some diverged lineages were not recovered as monophyletic clades, e.g. the potentially cryptic lineage *S. deserticola* I. P ID (Liberal) species designation probabilities were found to be moderate significant ($P > 50\%$) for all redefined species except some species in 16S. In sum, none of P ID, ABGD, GMYC and PTP approaches produced completely congruent clades among the single and the combined genes, but to some extent, they still could provide useful information for identification of some species.

Based on the integrated results of initial morphological characters, traditional barcoding (NJ, Bayesian and Maximum Likelihood analysis), GMYC, ABGD, P ID and PTP analysis, the putative species were confirmed by character-based method. As expected, the character-based analysis generated relatively congruent results of species discrimination in single marker of *rbcL*, ITS, 16S, *tufA* and the three combined data. All species revealed by character-based analysis possessed more than three unique character attributes. Most importantly, most taxonomic groups recovered by the character analysis were consistent with the morphological identification. Some species that were difficultly identified by morphological characters could be confirmed in character-based analysis, especially for the potential cryptic lineages. All taxonomic groups analyzed by *rbcL*, ITS, 16S, *tufA*, *rbcL* + ITS + 16S + *tufa*, *rbcL* + *tufA* and ITS + 16S, including the potential cryptic species, possessed unique simple identifying character states in character-based barcoding. Some species that could not be

identified consistently with traditional barcoding, GMYC, ABGD, P ID or PTP methods could be detected by character-based method, e.g. *S. deserticola* I in *rbcL* and ITS analysis, *S. dimorphus* and *S. bajacalifornicus* in *tufA* analysis, *S. obliquus* IV in ITS analysis and *S. obliquus* I,II,III in 16S analysis. The three combined data particularly distinguished the species clearly with more attribute characters by the character-based barcoding. Generally, the groups recovered by P ID were congruent with the character analysis. Therefore, this study proved that the character-based method showed more advantages and was the most effective barcoding approach for identifying microalgae. It may be an optimal option to first combine multiple barcoding approaches to test primary species and then confirm the taxonomic assignments by the character-based method.

Conclusions

Here we report the comprehensive molecular taxonomic identification of *Scenedesmus* to give a test that how DNA barcoding can be more effective in microalgae diversity revelation based on *rbcL*, ITS, 16S, *tufA*, *rbcL* + ITS + 16S + *tufA*, *rbcL* + *tufA* and ITS + 16S, with GMYC, P ID, PTP, ABGD and character-based barcoding approaches. First of all, the comprehensive results gave new insights into the taxonomic assignment of *Scenedesmus*, including the discrimination of most *Scenedesmus* species and the revelation of potential cryptic species. Five species, *S. deserticola*, *S. obliquus*, *S. quadricauda*, *S. spinosus* and *S. acuminatus* which were divided into several separate clades in multiple barcoding analysis of the single and combined data, could be as potential cryptic species. The three combined data showed a much higher proportion of resolution success than the single data. The traditional barcoding, GMYC, P ID, PTP and ABGD analysis of single genes generated various resolution. The character-based barcoding was proved to be the most effective approach for distinguishing species, which produced consistent species discrimination in all single and combined data and could distinguish the species clearly. After the initial morphological identification, it may be an optimal option to first combine multiple barcoding approaches to test primary microalgae species and then confirm the taxonomic assignments by the character-based method based on the single and combined data of multiple genes.

Methods

Algal sampling, culturing and morphological identification. The *Scenedesmus* green microalgae strains studied were collected from different environmental regions of China, e.g. the freshwaters and terrestrial areas (Fig. 1). The strains were isolated following Andersen (2005). The nonaxenic strains were grown in 250 mL flask containing 200 mL Bourelly medium at an irradiance of 40 umol m^{-2} s^{-1} with 14:10 h light: dark cycle at 20 °C. A detailed list of taxa studied, including the species name and distribution, was shown in Supplementary Table 1.

Firstly the *Scenedesmus* samples collected were identified by available morphological characters using microscope. Strains that had similar morphological characters and were difficult identified were just labeled as potential cryptic spices which would be further analyzed by the barcoding. Finally, 11 species were identified as known and 2 species were identified as unknown.

Molecular protocols and alignment. DNA extractions were performed using the Qiagen DNEasy Plant Extraction kit (Qiagen Inc., Valencia, CA, USA). The *rbcL*, *tufA*, ITS and 16S barcode regions were amplified using either universal primers from previous studies[44–47] or primers designed in the course of this study (Supplementary Table 16). PCR reactions were carried out in a total volume of 25 μL, using 2 × Taqman PCR MasterMix. PCR conditions for all primer sets were as follows: 95 °C for 3 min, primer-specific annealing temperatures for 45 s, 72 °C for 1 min; 35 cycles of 95 °C for 30 s, primer-specific annealing temperatures for 45s, 72 °C for 1 m, with a final extension of 72 °C for 1 min. The PCR products that provided a single band of sufficient intensity after running a 1.5% agarose gel were sent to the Beijing Genomics Institute (BGI) for bidirectional sequencing.

All sequences were manually edited using the program Sequencher 4.5 (Genecodes Corporation, Ann Arbor, MI). Sequences were aligned with MAFFT 6.717[48], followed by minor adjustment if needed. Kimura 2-Parameter corrected distances[10] between specimens were calculated with MEGA 5[49]. After edition, the *rbcL*, *tufA*, ITS and 16S sequences were combined as *rbcL* + *tufA* + ITS + 16S, the *rbcL* and *tufA* sequences were combined as *rbcL* + *tufa*, and the ITS and 16S sequences were combined as ITS + 16S.

Data analysis. The *rbcL*, *tufA*, ITS and 16S sequences were analyzed respectively. Then the combination of *rbcL* + *tufA* + ITS + 16S, *rbcL* + *tufA* and ITS + 16S was analyzed.

The NJ analyses were conducted using Kimura 2-parameter (K2P) distance model as recommended by Hebert *et al.*[10] in MEGA 5.0[49] with bootstrap values (1000 replications. Bayesian trees of *rbcL*, *tufA*, ITS and 16S were generated in MrBayes v.3.1.2[50]. Nucleotide substitution models of each gene for Bayesian analyses were selected separately using the Akaike Information Criterion (AIC) as implemented in the jModeltest v.0.1.1[51]. The most appropriate models for Bayesian analyses were GTR + G for *rbcL*, GTR + G for ITS, TVMef + I + G for 16S, GTR + G for *tufa*, GTR + G for *rbcL* + *tufA* + ITS + 16S, HKY for *rbcL* + *tufa* and GTR for ITS + 16S. Four chains were run twice in parallel for 10^5 generations with a sample frequency of 1/1,000. Maximum Likelihood trees were inferred from *rbcL*, *tufA*, ITS and 16S datasets by employing PhyML 3.0[52]. To assess the distance variation, the analyses of intra- and interspecific divergences were conducted among the final taxa assignments based on all P ID, ABGD, GMYC, PTP and CAOS analyses.

To assess species boundary hypotheses across the Bayesian gene tree, the Species Delimitation plugin[21] within Geneious Pro v5.5.4 (Biomatters; http://www.geneious.com) was investigated. Geneious is a bioinformatics desktop software package produced by Biomatters Ltd (http://www.biomatters.com). P ID(Liberal) in Geneious, represents the probability of making a correct identification of an unknown specimen by measuring the genetic variation found within its putative species group[27]. Maximum Likelihood trees were inferred from *rbcL*, *tufA*, ITS, 16S, *rbcL* + *tufA* + ITS + 16S, *rbcL* + *tufA* and ITS + 16S datasets by employing PhyML 3.0[52].

A linearised Bayesian phylogenetic tree was firstly calculated in BEAST[53] employing a Yule pure birth model tree prior. Settings in BEAUTi v. 1.7.1 were: substitution models for each gene, empirical base frequencies, four gamma categories, all codon positions partitioned with unlinked base frequencies and substitution rates. An uncorrelated relaxed lognormal clock model was used with rate estimated from the data and ucldmean parameter with uniform prior to value 0 as a lower and 10 as an upper boundary. All other settings were left as defaults. The length of MCMC chain was 40 000 000 sampling every 4000. All BEAST runs were executed in Bioportal[54], and the ESS values and trace files of runs were evaluated in Tracer v1.5.0. Two independent runs were merged using Log-Combiner v1.7.1 with 20% burn-in. Maximum clade credibility trees with a 0.5 posterior probability limit, and node heights of target tree were constructed in TreeAnnotator v1.7.1. Single-threshold GMYC analyses was conducted in R[55] using the APE[56] and SPLITS[57] packages.

The Automated Barcode Gap Discovery (ABGD) method (available at http://wwwabi.snv.jussieu.fr/public/abgd/) was used to statistically detect barcode gaps and identify distinct clusters of DNA sequences. The prior for the maximum value of intraspecific divergence was set between 0.001 and 0.1.

For Poisson tree process model (PTP), since the ultrametric trees are not required as input this coalescent-based method is very fast. This method is implemented in a web server (http://species.h-its.org/).

The character-based identification was conducted in characteristic attribute organization system (CAOS) and CAOS-Analyzer (http://bol.uvm.edu/caos-workbench/)[34,58]. The CAOS algorithm extracts characteristic attributes (CAs) for each clade at branching node within a guide tree that is first produced from a given dataset[33]. The incorporated NEXUS datasets of *rbcL*, *tufA*, ITS,16S, *rbcL* + *tufA* + ITS + 16S, *rbcL* + *tufA* and ITS + 16S NJ trees and their DNA data matrix were produced in MacClade v4.06[59], and were carried out in CAOS system. The characteristic attributes at the nucleotide positions where the most variable sites can distinguish all the taxa were listed.

References

1. O'Kelly, C. J. The Origin and Early Evolution of Green Plants. *Evolution of Primary Producers in the Sea* **73,** 287–309 (2007).
2. Leliaert, F. *et al.* Phylogeny and Molecular Evolution of the Green Algae. *Crit Rev Plant Sci.* **31,** 1–46 (2012).
3. Watanabe, S. & Floyd, G. L. Considerations on the systematics of coccoid green algae and related organisms based on the ultrastructure of swarmers. (1996).
4. Skaloud, P., Neustupa, J., Radochova, B. & Kubinova, L. Confocal microscopy of chloroplast morphology and ontogeny in three strains of Dictyochloropsis (Trebouxiophyceae, Chlorophyta). *Phycologia* **44,** 261–269, doi: 10.2216/0031-8884(2005)44[261:cmocma]2.0.co;2 (2005).
5. Hegewald, E. & Wolf, M. Phylogenetic relationships of Scenedesmus and Acutodesmus (Chlorophyta, Chlorophyceae) as inferred from 18S rDNA and ITS-2 sequence comparisons. *Plant Syst Evol.* **241,** 185–191, doi: 10.1007/s00606-003-0061-7 (2003).
6. Hindák, F. *Studies on the chlorococcal algae, Chlorophyceae*. (VEDA, Pub. House of the Slovak Academy of Sciences, 1990).
7. Lürling, M. The smell of water: grazer-induced colony formation in Scenedesmus. *Universiteit Wageningen* **77,** 246–248 (1999).
8. Guiry, M. D. *et al.* AlgaeBase: an on-line resource for Algae. *Cryptogamie Algol.* **35,** 105–115, doi: 10.7872/crya.v35.iss2.2014.105 (2014).
9. An, S. S., Friedl, T. & Hegewald, E. Phylogenetic relationships of Scenedesmus and Scenedesmus-like coccoid green algae as inferred from ITS-2 rDNA sequence comparisons. *Plant Biol.* **1,** 418–428, doi: 10.1055/s-2007-978535 (1999).
10. Hebert, P. D. N., Cywinska, A., Ball, S. L. & DeWaard, J. R. Biological identifications through DNA barcodes. *P Roy Soc B-Biol Sci.* **270,** 313–321, doi: 10.1098/rspb.2002.2218 (2003).
11. Ratnasingham, S. & Hebert, P. D. N. BOLD: The Barcode of Life Data System (www.barcodinglife.org). *Molecular Ecology Notes* **7,** 355–364, doi: 10.1111/j.1471-8286.2006.01678.x (2007).
12. Reid, B. N. *et al.* Comparing and combining distance-based and character-based approaches for barcoding turtles. *Mol Ecol Notes* **11,** 956–967, doi: 10.1111/j.1755-0998.2011.03032.x (2011).
13. Zou, S., Li, Q. & Kong, L. Monophyly, Distance and Character-Based Multigene Barcoding Reveal Extraordinary Cryptic Diversity in Nassarius: A Complex and Dangerous Community. *Plos One* **7,** doi: 10.1371/journal.pone.0047276 (2012).
14. Krawczyk, K., Szczecinska, M. & Sawicki, J. Evaluation of 11 single-locus and seven multilocus DNA barcodes in Lamium L. (Lamiaceae). *Mol Ecol Resour* **14,** 272–285, doi: 10.1111/1755-0998.12175 (2014).
15. Chakraborty, C., Doss, C. G. P., Patra, B. C. & Bandyopadhyay, S. DNA barcoding to map the microbial communities: current advances and future directions. *Appl Microbio Biot* **98,** 3425–3436, doi: 10.1007/s00253-014-5550-9 (2014).
16. Hebert, P. D. N., deWaard, J. R. & Landry, J.-F. DNA barcodes for 1/1000 of the animal kingdom. *Biol Letters* **6,** 359–362, doi: 10.1098/rsbl.2009.0848 (2010).
17. Hollingsworth, P. M. *et al.* A DNA barcode for land plants. *P Natl Acad Sci USA* **106,** 12794–12797, doi: 10.1073/pnas.0905845106 (2009).
18. Ran, J.-H., Wang, P.-P., Zhao, H.-J. & Wang, X.-Q. A Test of Seven Candidate Barcode Regions from the Plastome in Picea (Pinaceae). *J Integr Plant Biol.* **52,** 1109–1126, doi: 10.1111/j.1744-7909.2010.00995.x (2010).
19. Saunders, G. W. & Kucera, H. An evaluation of rbcL, tufA, UPA, LSU and ITS as DNA barcode markers for the marine green macroalgae. *Cryptogamie Algol* **31,** 487–528 (2010).
20. Hall, J. D., Fucikova, K., Lo, C., Lewis, L. A. & Karol, K. G. An assessment of proposed DNA barcodes in freshwater green algae. *Cryptogamie Algol* **31,** 529–555 (2010).
21. Masters, B. C., Fan, V. & Ross, H. A. Species delimitation - a geneious plugin for the exploration of species boundaries. *Mol Ecol Resour* **11,** 154–157, doi: 10.1111/j.1755-0998.2010.02896.x (2011).
22. Talavera, G., Dinca, V. & Vila, R. Factors affecting species delimitations with the GMYC model: insights from a butterfly survey. *Methods Ecol Evol.* **4,** 1101–1110, doi: 10.1111/2041-210x.12107 (2013).
23. Fujisawa, T. & Barraclough, T. G. Delimiting Species Using Single-Locus Data and the Generalized Mixed Yule Coalescent Approach: A Revised Method and Evaluation on Simulated Data Sets. *Syst Biol.* **62,** 707–724, doi: 10.1093/sysbio/syt033 (2013).
24. Brewer, M. S., Spruill, C. L., Rao, N. S. & Bond, J. E. Phylogenetics of the millipede genus Brachycybe Wood, 1864 (Diplopoda: Platydesmida: Andrognathidae): Patterns of deep evolutionary history and recent speciation. *Mol Phylogenet Evol.* **64,** 232–242, doi: 10.1016/j.ympev.2012.04.003 (2012).
25. Zhang, J. J., Kapli, P., Pavlidis, P. & Stamatakis, A. A general species delimitation method with applications to phylogenetic placements. *Bioinformatics* **29,** 2869–2876, doi: 10.1093/bioinformatics/btt499 (2013).
26. Kekkonen, M. & Hebert, P. D. N. DNA barcode-based delineation of putative species: efficient start for taxonomic workflows. *Mol Ecol Resour* **14,** 706–715, doi: 10.1111/1755-0998.12233 (2014).
27. Hamilton, C. A., Hendrixson, B. E., Brewer, M. S. & Bond, J. E. An evaluation of sampling effects on multiple DNA barcoding methods leads to an integrative approach for delimiting species: A case study of the North American tarantula genus Aphonopelma (Araneae, Mygalomorphae, Theraphosidae). *Mol Phyl Evol.* **71,** 79–93, doi: 10.1016/j.ympev.2013.11.007 (2014).

28. Weiss, M., Macher, J. N., Seefeldt, M. A. & Leese, F. Molecular evidence for further overlooked species within the Gammarus fossarum complex (Crustacea: Amphipoda). *Hydrobiologia* **721,** 165–184, doi: 10.1007/s10750-013-1658-7 (2014).
29. Yang, J.-B., Wang, Y.-P., Moeller, M., Gao, L.-M. & Wu, D. Applying plant DNA barcodes to identify species of Parnassia (Parnassiaceae). *Mol Ecol Resour* **12,** 267–275, doi: 10.1111/j.1755-0998.2011.03095.x (2012).
30. Ashfaq, M., Asif, M., Anjum, Z. I. & Zafar, Y. Evaluating the capacity of plant DNA barcodes to discriminate species of cotton (Gossypium: Malvaceae). *Mol Ecol Resour* **13,** 573–582, doi: 10.1111/1755-0998.12089 (2013).
31. Jaen-Molina, R. *et al.* Molecular taxonomic identification in the absence of a 'barcoding gap': a test with the endemic flora of the Canarian oceanic hotspot. *Mol Ecol Resour* **15,** 42–56, doi: 10.1111/1755-0998.12292 (2015).
32. Puillandre, N., Lambert, A., Brouillet, S. & Achaz, G. ABGD, Automatic Barcode Gap Discovery for primary species delimitation. *Mol Ecol* **21,** 1864–1877, doi: 10.1111/j.1365-294X.2011.05239.x (2012).
33. Rach, J., DeSalle, R., Sarkar, I. N., Schierwater, B. & Hadrys, H. Character-based DNA barcoding allows discrimination of genera, species and populations in Odonata. *P Roy Soc B-Biol Sci.* **275,** 237–247, doi: 10.1098/rspb.2007.1290 (2008).
34. Sarkar, I. N., Planet, P. J. & Desalle, R. CAOS software for use in character-based DNA barcoding. *Mol Ecol Resour* **8,** 1256–1259, doi: 10.1111/j.1755-0998.2008.02235.x (2008).
35. Li, T., Wan, L., Li, A. & Zhang, C. Responses in growth, lipid accumulation, and fatty acid composition of four oleaginous microalgae to different nitrogen sources and concentrations. *Chin J Oceanol Limn* **31,** 1306–1314, doi: 10.1007/s00343-013-2316-7 (2013).
36. Anand, J. & Arumugam, M. Enhanced lipid accumulation and biomass yield of Scenedesmus quadricauda under nitrogen starved condition. *Bioresource Technol* **188,** 190–194, doi: 10.1016/j.biortech.2014.12.097 (2015).
37. Meier, R., Shiyang, K., Vaidya, G. & Ng, P. K. L. DNA barcoding and taxonomy in diptera: A tale of high intraspecific variability and low identification success. *Syst Biol* **55,** 715–728, doi: 10.1080/10635150600969864 (2006).
38. Li, X. *et al.* Plant DNA barcoding: from gene to genome. *Biol Rev* **90,** 157–166, doi: 10.1111/brv.12104 (2015).
39. Goldstein, P. Z. & DeSalle, R. Phylogenetic species, nested hierarchies, and character fixation. *Cladistics* **16,** 364–384 (2000).
40. Knowles, L. L. & Carstens, B. C. Delimiting species without monophyletic gene trees. *Syst Biol* **56,** 887–895, doi: 10.1080/10635150701701091 (2007).
41. Yu, Z., Li, Q., Kong, L. & Yu, H. Utility of DNA Barcoding for Tellinoidea: A Comparison of Distance, Coalescent and Character-based Methods on Multiple Genes. *Mar Biotechnol.* **17,** 55–65, doi: 10.1007/s10126-014-9596-6 (2015).
42. Esselstyn, J. A., Evans, B. J., Sedlock, J. L., Khan, F. A. A. & Heaney, L. R. Single-locus species delimitation: a test of the mixed Yule-coalescent model, with an empirical application to Philippine round-leaf bats. *P Roy Soc B-Biol Sci.* **279,** 3678–3686, doi: 10.1098/rspb.2012.0705 (2012).
43. Miralles, A. & Vences, M. New Metrics for Comparison of Taxonomies Reveal Striking Discrepancies among Species Delimitation Methods in Madascincus Lizards. *Plos One* **8,** doi: 10.1371/journal.pone.0068242 (2013).
44. Burja, A. M., Tamagnini, P., Bustard, M. T. & Wright, P. C. Identification of the green alga, Chlorella vulgaris (SDC1) using cyanobacteria derived 16S rDNA primers: targeting the chloroplast. *Fems Microbiol Lett* **202,** 195–203, doi: 10.1016/s0378-1097(01)00306-8 (2001).
45. Fama, P., Wysor, B., Kooistra, W. & Zuccarello, G. C. Molecular phylogeny of the genus Caulerpa (Caulerpales, Chlorophyta) inferred from chloroplast tufA gene. *J Phycol* **38,** 1040–1050, doi: 10.1046/j.1529-8817.2002.t01-1-01237.x (2002).
46. Sun, X., Xiao-Wei, W. U., Xing-Wen, L. I. & Pei, L. Q. Molecular identification of Chlorella strains based on sequence analysis of nuclear rDNA ITS and chloroplast rbcL gene. *JFSC* **33,** 565–571 (2009).
47. Bock, C., Proeschold, T. & Krienitz, L. Two new Dictyosphaerium-morphotype lineages of the Chlorellaceae (Trebouxiophyceae): Heynigia gen. nov. and Hindakia gen. nov. *Eur J Phycol.* **45,** 267–277, doi: 10.1080/09670262.2010.487920 (2010).
48. Katoh, K., Asimenos, G. & Toh, H. In *Bioinformatics for DNA Sequence Analysis* Vol. 537 *Methods in Molecular Biology* (ed Posada, D.) 39–64 (2009).
49. Tamura, K. *et al.* MEGA5: Molecular Evolutionary Genetics Analysis Using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods. *Mol Biol Evol.* **28,** 2731–2739, doi: 10.1093/molbev/msr121 (2011).
50. Ronquist, F. & Huelsenbeck, J. P. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19,** 1572–1574, doi: 10.1093/bioinformatics/btg180 (2003).
51. Posada, D. jModelTest: Phylogenetic model averaging. *Mol Biol Evol.* **25,** 1253–1256, doi: 10.1093/molbev/msn083 (2008).
52. Guindon, S. *et al.* New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. *Syst Biol.* **59,** 307–321, doi: 10.1093/sysbio/syq010 (2010).
53. Drummond, A. J. & Rambaut, A. BEAST: Bayesian evolutionary analysis by sampling trees. *Bmc Evol Biol.* **7,** doi: 10.1186/1471-2148-7-214 (2007).
54. Kumar, S. *et al.* AIR: A batch-oriented web program package for construction of supermatrices ready for phylogenomic analyses. *Bmc Bioinformatics* **10,** doi: 10.1186/1471-2105-10-357 (2009).
55. Team, C. R. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, 2012 (2012).
56. Paradis, E., Claude, J. & Strimmer, K. APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics* **20,** 289–290, doi: 10.1093/bioinformatics/btg412 (2004).
57. Ezard, T. SPecies' LImits by Threshold Statistics.
58. Bergmann, T., Hadrys, H., Breves, G. & Schierwater, B. Character-based DNA barcoding: a superior tool for species classification. *Berl Munch Tierarztl* **122,** 446–450, doi: 10.2376/0005-9366-122-446 (2009).
59. Mindell, D. P. MacClade: Analysis of Phylogeny and Character Evolution. Version 3.0 Wayne P. Maddison David R. Maddison. *Auk* **111,** 1035–1036 (1994).

Acknowledgements

The financial support from the China Postdoctoral Science Foundation (2014M561661, 2015T80558), Natural Science Foundation of Jiangsu Province (BK20150680) and National Natural Science Foundation of China (NSFC) (31600294) was gratefully acknowledged.

Author Contributions

S.Z. designed the research. S.Z., C.F., Chun. W. and Changhai. W. performed all the molecular work and analysis. Y.B., Z.G., and M.H. collected samples from different locations. S.Z. wrote the paper in discussion with all authors.

Additional Information

Supplementary information accompanies this paper at http://www.nature.com/srep

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

How to cite this article: Zou, S. *et al.* How DNA barcoding can be more effective in microalgae identification: a case of cryptic diversity revelation in *Scenedesmus* (Chlorophyceae). *Sci. Rep.* **6**, 36822; doi: 10.1038/srep36822 (2016).