Journal of *Marine Science and Engineering*

MDPI

Article

A CHEMTAX Study Based on Picoeukaryotic Phytoplankton Pigments and Next-Generation Sequencing Data from the Ulleungdo–Dokdo Marine System of the East Sea (Japan Sea): Improvement of Long-Unresolved Underdetermined Bias

Myung Jin Hyun [1,2], Jongseok Won [1,3], Dong Han Choi [1,3], Howon Lee [1], Yeonjung Lee [1,3], Charity Mijin Lee [4], Chan Hong Park [5] and Jae Hoon Noh [1,2,3,*]

[1] Marine Ecosystem Research Center, Korea Institute of Ocean Science and Technology, Busan 49111, Republic of Korea
[2] Department of Ocean Science, University of Science and Technology, Daejeon 34113, Republic of Korea
[3] Department of Convergence Study on the Ocean Science and Technology, Ocean Science and Technology School, Korea Maritime and Ocean University, Busan 49112, Republic of Korea
[4] Ocean Law Research Center, Korea Institute of Ocean Science and Technology, Busan 49111, Republic of Korea
[5] Dokdo Research Center, East Sea Research Institute, Korea Institute of Ocean Science and Technology, Uljin 36315, Republic of Korea
* Correspondence: jhnoh@kiost.ac.kr; Tel.: +82-(0)51-664-3260

Abstract: The CHEMTAX program has been widely used to estimate community composition based on major pigment concentrations in seawater. However, because CHEMTAX is an underdetermined optimization algorithm, underdetermined bias has remained an unsolved problem since its development in 1996. The risk of producing biased results increases when analyzing the picophytoplankton community; therefore, this study tested a new method for avoiding biased CHEMTAX results using the picophytoplankton community around the East Sea (Japan Sea). This method involves building a linear model between pigment concentration data and community composition data based on DNA sequencing to predict the pigment range for each operational taxonomic unit, based on the 95% prediction interval. Finally, the range data are transformed into an initial ratio and ratio limits for CHEMTAX analysis. Three combinations of initial ratios and ratio limits were tested to determine whether the modeled initial ratio and ratio limit could prevent underdetermined bias in the CHEMTAX estimates; these combinations were the modeled initial ratio and ratio limit, the modeled initial ratio with a default ratio limit of 500 s, and an initial ratio from previous research with the default ratio limit. The final ratio and composition data for each combination were compared with Bayesian compositional estimator-based final ratio and composition data, which are robust against underdetermined bias. Only CHEMTAX analysis using the modeled initial ratio and ratio limit was unbiased; all other combinations showed significant signs of bias. Therefore, the findings in this study indicate that ratio limits and the initial ratio are equally important in the CHEMTAX analysis of biased datasets. Moreover, we obtained statistically supported initial ratios and ratio limits through linear modeling of pigment concentrations and 16s rDNA composition data.

Keywords: CHEMTAX; next-generation sequencing (NGS); underdetermined bias; linear modeling; initial ratio; ratio limit; Bayesian compositional estimator (BCE); East Sea (Japan Sea); Ulleung Basin

check for updates

Citation: Hyun, M.J.; Won, J.; Choi, D.H.; Lee, H.; Lee, Y.; Lee, C.M.; Park, C.H.; Noh, J.H. A CHEMTAX Study Based on Picoeukaryotic Phytoplankton Pigments and Next-Generation Sequencing Data from the Ulleungdo–Dokdo Marine System of the East Sea (Japan Sea): Improvement of Long-Unresolved Underdetermined Bias. *J. Mar. Sci. Eng.* **2022**, *10*, 1967. https://doi.org/10.3390/jmse10121967

Academic Editor: Feng Zhou

Received: 1 November 2022
Accepted: 8 December 2022
Published: 10 December 2022

1. Introduction

CHEMTAX (chemical taxonomy) is a program that allocates chlorophyll *a* (Chl-*a*) into taxa of interest, usually at the class level [1]. It has been widely used in marine ecosystem research (i.e., [2–4]) because the Chl-*a* is an indicator for phytoplankton biomass, and CHEMTAX directly fractionates it into taxa. However, because the algorithm of CHEMTAX is inherently underdetermined [1], the algorithm is always at risk of yielding

biased results. Moreover, this risk increases when CHEMTAX is used to estimate Chl-*a* content of picophytoplankton (PPP) because the community composition of PPP tends to be complex; its pigment ratios are also complex. Therefore, CHEMTAX must first identify the dominant taxa of PPP based on pigment and next-generation sequencing (NGS) data. Moreover, high complexity in terms of pigment ratio would increase the severity of underdetermined bias; Latasa [5] noted that the presence of a shared pigment among multiple taxa increases the risk of bias. This problem is a major barrier to the use of CHEMTAX; resolution or improvement of this inherent limitation is a challenge for biological oceanographers.

PPP mainly comprise two groups of small phytoplankton (prokaryotic picophytoplankton [P-PPP] and eukaryotic picophytoplankton [E-PPP]) that are characterized by small size (cell diameter $\leq$ 3 μm); these are the most abundant phytoplankton component in marine environments [6,7]. PPP may comprise >50% of total phytoplankton in terms of chlorophyll or biomass in some environments. P-PPP consists of two genera, *Synechococcus* and *Prochlorococcus*, which exhibit differing spatial distributions worldwide. *Synechococcus* is ubiquitously distributed across all marine environments. In contrast, *Prochlorococcus* is mainly observed between 40° S and 40° N [8,9]. E-PPP are responsible for 20% of the ocean's primary production and carbon biomass [10,11]. CHEMTAX-based research concerning E-PPP has the following limiting factors; small size ($\leq$3 μm), morphological simplicity, high diversity, and marker pigment overlap among taxa. Particularly, little is known regarding the composition and distribution of E-PPP communities at large spatial and temporal scales [12].

The East Sea (also known as the Japan Sea; hereafter, the East Sea) is a typical mid-latitude, semi-enclosed marginal sea in the northwestern Pacific Ocean surrounded by Korea, Japan, and Russia. The East Sea is connected to the Sea of Okhotsk in the north by the Soya and Tatar Straits, to the Pacific Ocean in the east by the Tsugaru Strait, and to the South Sea in the south by the Korea Strait. A well-defined sub-polar front at approximately 37–40° N is observed in the East Sea mainly because of seawater circulation through these four straits [13]. The front separates a warm water mass from the East Korea Warm Current and a cold water mass from the North Korea Cold Current, which branches off of the Liman Current. The Ulleung Basin is a major feature located in the southwestern region of the East Sea. This basin covers a large area of approximately 100 km in the north–south direction and 150 km in the east–west direction. The Ulleung Basin is a marine system containing Ulleungdo and Dokdo, and seasonal sub-polar fronts are sometimes present. A recent study revealed that the Ulleung Basin has environmental characteristics of a biological hotspot, with high primary phytoplankton productivity [14,15]. Additionally, some studies have shown that the primary production and chlorophyll contributions of PPP are essential to ecosystem changes [16,17], and the Ulleung Basin frequently undergoes ecosystem changes in response to climate change [18–21].

The advantageous features of PPP in aquatic environments [22] suggest that PPP would increase in aquatic communities with global warming and related climate changes [23]. CHEMTAX-based studies have been conducted in the East Sea using phytoplankton pigment data along a north–south longitudinal observation line [2] and from the Ulleung Basin [24,25]. These CHEMTAX-based studies did not use size-fractionated phytoplankton pigment data, and the results provided only partial information regarding the distributional characteristics of E-PPP. Thus, although CHEMTAX is useful, CHEMTAX-based studies of E-PPP in the East Sea have been inadequate and limited. Advanced research based on modern technological approaches is urgently needed to improve the applicability of CHEMTAX to this region.

The underdetermined bias of CHEMTAX remains an unresolved problem. CHEMTAX analysis may not always be limited by underdetermined bias, as demonstrated by Latasa [5]; using an artificially produced dataset and eight intentionally distorted initial ratios, that study tested whether the CHEMTAX final ratios converged at the true ratio. After 10 successive runs, using the final ratio from the previous run as the initial ratio for the next

run, all eight pigment ratios tended to converge around the true value. Although the overall pigment ratios tended to converge, minor pigments shared among multiple taxa did not converge at a specific point; instead, a biased pattern was observed with decreased accuracy for taxa with shared pigments. In the same context, it is challenging to constrain CHEMTAX analysis to the pico-size class because the pico-sized phytoplankton community often contains a non-negligible portion of Chl-*a* from small Ochrophyta, such as Pelagophyceae and Chryophyceae. Because these taxa share some pigments (e.g., fucoxanthin and 19′-butanoyloxyfucoxanthin) with Bacillariophyceae and Prymnesiophyceae, the probability of bias increases, as noted by Latasa [5]; thus, the CHEMTAX results become unreliable.

To address the above challenges, Van den Meersche et al. [26] developed the Bayesian compositional estimator (BCE) as a new optimization algorithm to overcome the underdetermined bias of CHEMTAX. BCE is a statistically advanced algorithm and the first algorithm designed to overcome the underdetermined bias of CHEMTAX. Nonetheless, the limitations of BCE have hindered its adoption in research fields other than CHEMTAX-related research; one of these limitations is the acquisition of different results for each run of the algorithm, even when using the same data and settings, because part of the algorithm depends on a random walk [27].

Community composition analysis using next-generation sequencing (NGS) techniques can represent the PPP community at the operational taxonomic unit (OTU) level [28]. This representation is challenging for CHEMTAX because the NGS approach quantitatively targets specific DNA sequences for each OTU (usually a gene representing the small subunit of ribosomal RNA) to analyze such communities (e.g., [29]). However, the number of target sequence copies present in a cell may vary among OTUs or ecotypes, and biases arising from the polymerase chain reaction process may distort community structures [30]. Therefore, quantitative analysis based on the NGS technique presumably involves some bias. Furthermore, NGS quantifies the number or ratio of targeted sequences, which is another disadvantage compared with CHEMTAX, which directly quantifies the concentration or ratio of Chl-*a*.

This study aims to improve CHEMTAX analysis by removing the bias associated with the underdetermined CHEMTAX algorithm through the adoption of NGS data. In this study, we assess whether the NGS results and pigment concentration data are linearly correlated, then predict the possible range of pigment ratios using a linear model. Statistically reasonable initial ratio and ratio limit matrices can be produced based on the range. The CHEMTAX results are validated by comparison with BCE results to confirm avoidance of the underdetermined bias. Finally, we suggest an appropriate method for analyzing PPP communities using CHEMTAX in the East Sea area, and we demonstrate the importance of using appropriate ratio limits for CHEMTAX analysis.

2. Methods

2.1. Survey Area

Sampling for this study was conducted using RV Eardo from February 2018 to April 2020. The research cruises occurred in February, April, June, August, and October to reflect the area's seasonal changes. Samples were collected from 6 regular stations (47, 50, 45, 30, 16, and 20) representing the Ulleung Basin, and 0 to 3 (A, B, and C) additional stations for comparison during each cruise (Figure 1). Detailed information concerning the surveyed stations is provided in Table 1.

2.2. Sample Collection

At each station, seawater was collected from the surface and sub-surface chlorophyll maximum (SCM) layer using a rosette sampler equipped on the conductivity–temperature–depth recorder (SBE 911, Sea-bird Scientific, Bellevue, WA, USA). The SCM layer was identified from fluorescence data acquired using the conductivity–temperature–depth recorder. The collected seawater was pre-filtered through a 3-μm polycarbonate membrane filter (GVS Filter Technology, Bologna, Italy); only particles smaller than 3 μm were retained.

Next, samples for determination of pigment concentration were collected by filtering 2 L of the pre-filtered seawater through GF/F filters (Whatman plc, Buckinghamshire, UK), and samples for DNA sequence analysis were collected by filtering 1 L of pre-filtered seawater into a 0.2-µm Supor polyethersulfone membrane filter (Pall Corporation, Port Washington, NY, USA). Then, 1 mL of sodium chloride–Tris–ethylenediaminetetraacetic acid buffer was injected into each DNA sample prior to storage. Samples were stored in liquid nitrogen during the cruises and during transport to the laboratory, then stored in a freezer (−80 °C).

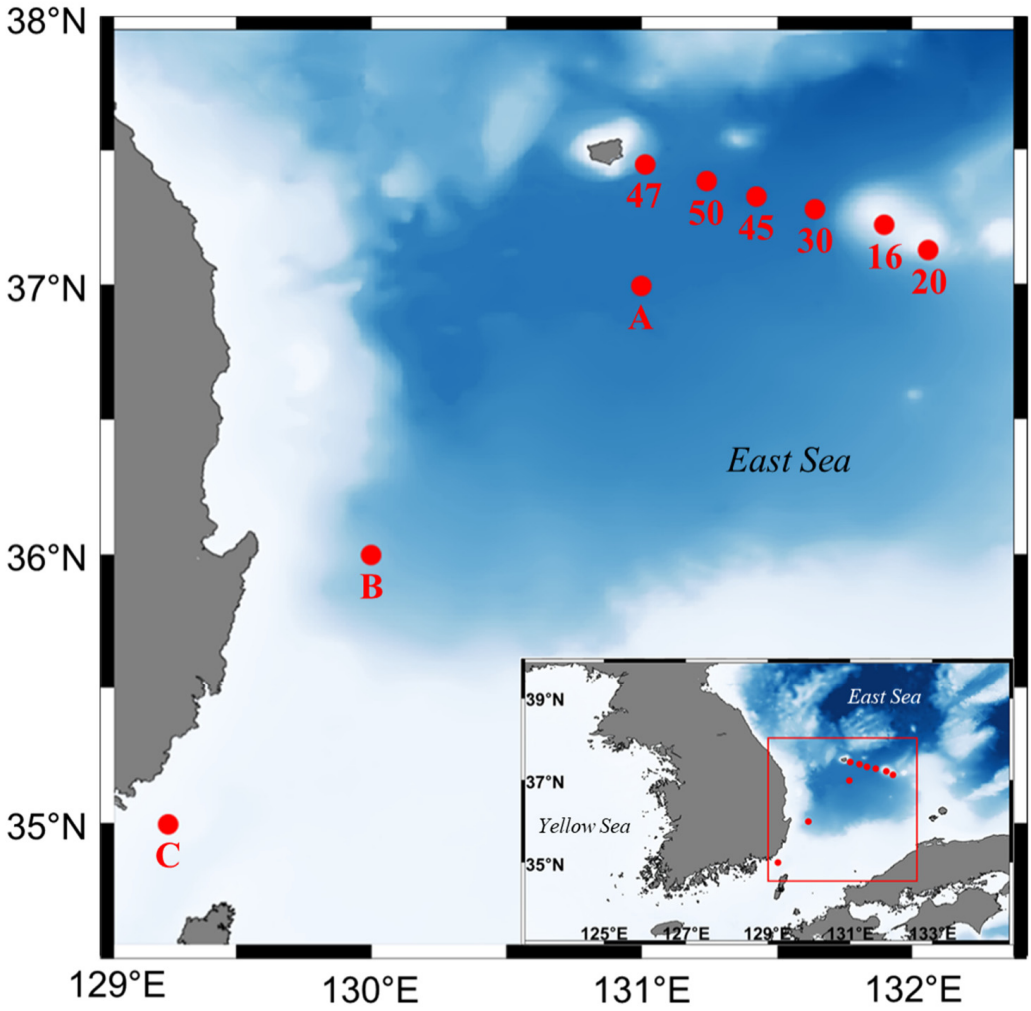

Figure 1. Station Map. The red circles on the map represent sampling stations. Stations A, B, C are reference stations.

Table 1. Research stations and sampling depths (m) surveyed during each cruise. Surface water and SCM-layer water were sampled.

Station	Latitude	Longitude	February 2018	August 2018	June 2019	October 2019	April 2020
St. 47	37.451	131.013	0, 35	0, 40	0, 30	0, 20	0, 20
St. 45	37.391	131.240	0, 20	0, 30	0, 35	0, 24	0, 30
St. 50	37.331	131.422	0, 20	0, 35	0, 50	0, 50	0, 20
St. 30	37.284	131.640	0, 20	0, 35	0, 40	0, 40	0, 20
St. 16	37.226	131.897	0, 25	0, 16	0, 30	0, 30	0, 20
St. 20	37.134	132.060	0, 15	0, 46	0, 45	0, 40	0, 20
St. A	37.000	131.000	0, 20	0, 36		0, 20	0, 15
St. B	36.020	130.015	0, 18	0, 36		0, 30	0, 20
St. C	35.000	129.250	0, 52	0, 40			

2.3. Determination of Pigment Concentrations Using HPLC

Pigment samples were freeze-dried before extraction to maximize the extraction efficiency. Then, they were extracted by soaking in 4 mL of aqueous acetone solution (5:95 *v:v*), wrapped with aluminum foil to prevent exposure to light, and stored in a refrigerator (4 °C) for 24 h. The extracts were filtered through 0.2-μm polytetrafluoroethylene syringe filters (Hyundai Micro, Seoul, Korea) to ensure that no contaminants were injected into the HPLC system. Then, 1-mL aliquots of extract were pipetted into brown amber vials, and 400 μL of HPLC-grade water were added for water packing.

The HPLC system (LC-2030c 3D, Shimadzu Corporation, Kyoto, Japan) was used to separate and quantify pigment concentrations as described by Zapata et al. [31]. For separation, reverse-phase chromatography was conducted using a C8 column (150 × 4.6 mm, 3.5 μm particle size, 100 Å pore size, Waters Corporation, Milford, MA, USA), whereas quantification was performed with the 440-nm chromatogram measured by a photodiode-array detector. Wavelengths from 370 to 800 nm were also measured to confirm the purity of each peak.

The factors for converting peak area to pigment concentrations were obtained prior to analysis (at least once per year) using a calibration curve determined from standard pigments (DHI LAB, Hørsholm, Denmark). Furthermore, to facilitate peak identification, a mixture of standard pigments was run as the first and last sample daily during HPLC operation.

2.4. Community Composition Analysis of Eukaryotes Using NGS

Cell lysis was executed in accordance with the protocol established by Somerville et al. [32]. Lysozyme, sodium dodecyl sulfate, and proteinase K were added to dissolve microbial cells. DNA extraction was conducted using phenol–chloroform–isoamyl alcohol (25:24:1, *v:v:v*) and chloroform–isoamyl alcohol (24:1) extraction procedures [33]. DNA purification was then conducted using spin columns (Biofact, Daejeon, Korea) with AW1 and AW2 washing buffer solutions (Qiagen, Hilden, Germany) [34].

To amplify the target sequence (the V3–V4 region of 16S rDNA), polymerase chain reaction was conducted as described by Choi et al. [29] using the primers PLA491F and PLA907R. The products were purified using AMPure XP beads (Beckman Coulter, Brea, CA, USA), subjected to polymerase chain reaction [35], and purified again. Then, the products were sequenced on the Illumina MiSeq 2 × 300 bp paired-end platform at ChunLab (Seoul, Korea). The resulting nucleotide sequences were analyzed using Mothur software v.1.39.5 [29,36,37].

2.5. Linear Modeling between Pigment Concentrations and Community Composition Data

Pigment concentration data and NGS data have essential differences that must be resolved before regression is performed. First, the pigment concentration data include pigments from both eukaryotic and prokaryotic phytoplankton, whereas the NGS data only contain information regarding eukaryotic phytoplankton. Second, pigment concentrations are absolute values, whereas NGS data are percentages. Therefore, pretreatment of pigment data is needed prior to regression.

Pretreatment was conducted as follows. First, zeaxanthin and divinyl Chl-*a* originate only from the cyanobacteria; therefore, the removal of their concentrations resolves some mismatches (Figure 2a). Second, Chl-*a* is present in both eukaryotic and prokaryotic cells. To eliminate Chl-a from prokaryotic cells, BCE analysis was conducted prior to regression analysis. Cyanobacterial Chl-*a* was calculated with the constrained least squares method using the best-fitting pigment ratios obtained in the BCE run. In this procedure, the bce package [38] and lsei package [39] of R [40] were used for calculation. The BCE output was not directly used to represent cyanobacterial Chl-*a* because of the low reproducibility of BCE (discussed in greater depth in the Discussion).

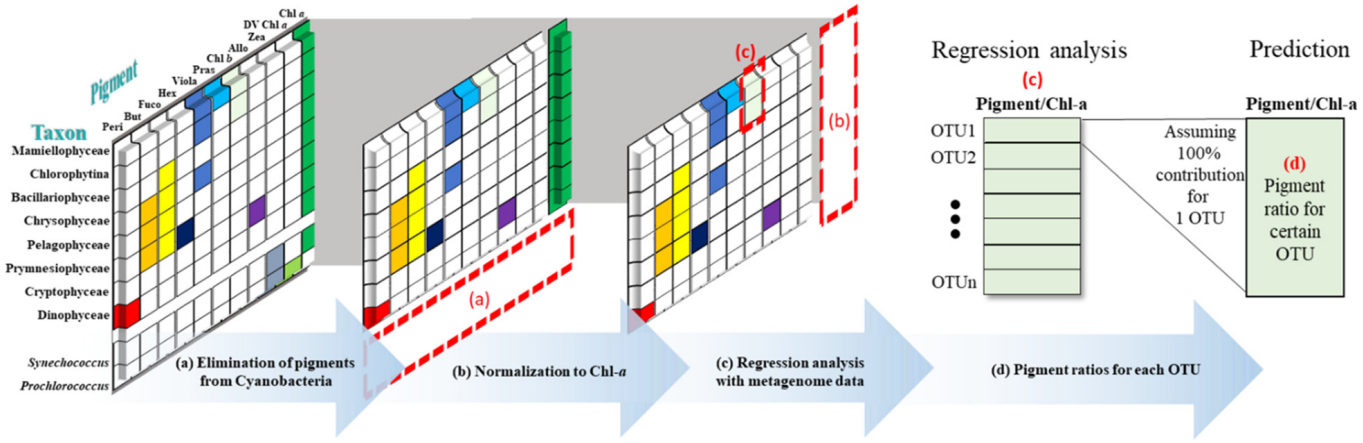

Figure 2. Diagram of pigment ratio modeling. To construct a linear model that related pigment concentrations to DNA sequence composition data, preprocessing was conducted to eliminate pigments from Cyanobacteria (**a**) and normalize to Chl-*a* (**b**). Then, the linear model was constructed (**c**) using the non-negative least squares method [41], and the range for each pigment ratio was determined from 95% prediction intervals (**d**).

The pigment concentrations of eukaryotic phytoplankton were then normalized to Chl-a to obtain ratio data (Figure 2b). Finally, multiple linear regression analysis with no intercept was conducted using the pigment ratio data as the dependent variable and NGS OTU data as independent variables (Figure 2c,d). The stat package of R version 4.1.4 [40] was used to perform this analysis. Statistical indices were then assessed to determine whether the linear model was valid. Because an excessive number of variables in a linear model can reduce regression validity, only OTUs with a mean ratio >1% or maximum ratio >5% were included.

Because few negative coefficients were obtained from multiple linear regression due to multicollinearity, the non-negative least squares (NNLS) method, which constrains negative coefficients obtained from multiple linear regression, was used to eliminate negative coefficients [41]. The following formula was used for NNLS:

$$\frac{Pigment}{Chlorophyll\ a} = a \times OTU_a + b \times OTU_b + \cdots + n \times OTU_n \quad (1)$$

Based on the NNLS results, the situation in which one OTU comprises 100% of the species composition was used to calculate the pigment ratio for the CHEMTAX ratio matrix. For example, the modeled pigment ratio for OTU_a would be:

$$\frac{Pigment}{Chlorophyll\ a} = a \times 100 + b \times 0 + \cdots + n \times 0 = 100a \quad (2)$$

Because it is not possible for one OTU to actually contribute 100% of the community, this approach generates extrapolation error, which causes the best-fitting coefficients to become unreliable. As a reliable alternative, 95% prediction intervals were used [42]. Negative values produced for an interval's lower boundary were replaced with zeroes because the actual pigment ratio could never be negative.

The pigment ratio ranges obtained from the model were converted into ranges of pigment ratios for CHEMTAX, using a weighted average for each boundary. When the ranges of all elements of the pigment ratio for CHEMTAX were determined, the initial ratio and ratio limit were calculated as follows:

$$Initial\ Ratio = \sqrt{(lower\ boundary) * (higher\ boundary)} \quad (3)$$

$$RLM = \frac{higher\ boundary}{initial\ ratio} \times 100 - 100 \quad (4)$$

2.6. CHEMTAX Analysis

The data used for CHEMTAX analysis were divided into clusters according to cruise to reflect seasonal differences. Data from the surface and SCM layer were not separated because no significant differences were detected in the linear modeling, except in August 2018; therefore, only data from August 2018 were divided into two clusters according to depth. Consequently, the data were divided into six clusters and analyzed separately to generate different final ratios.

The clusters used three different combinations of initial ratios and ratio limits: (1) linear modeling-based initial ratios and ratio limits, (2) a linear modeling-based initial ratio with a default ratio limit of 500, and (3) an initial ratio determined from previous research [27] (Table 2) with the default ratio limit. In total, six clusters based on these three combinations were tested, producing 18 final ratios. Note that linear modeling between pigment data and eukaryotic DNA sequence composition data could not be used to predict the pigment ratio associated with Cyanobacteria. Finally, the initial ratios of combinations (1) and (2) and ratio limits of combination (1) for zeaxanthin and divinyl Chl-*a* were determined from the 95% prediction intervals of BCE.

Table 2. Initial ratio matrix which was built based on previous work. The matrix was made by reconstructing the pigment ratio statistics data from Roy et al. [27] to best fit the present study. The Chrysophyceae * was included in June 2019 and August 2018 (Surface) clusters, Pelagophyceae ** was included in October 2019 cluster, and Prochlorococcus *** was Included in August 2018 (Surface), August 2018 (SCM), and October 2019. Other taxa are included in all clusters).

Class/Pigment	Peridinin	ButFuco	Fuco	HexFuco	Prasino	Viola	Allo	Zea	Chl-*b*	DV Chl-*a*	Chl-*a*
Mamiellophyceae	0	0	0	0	0.248	0.054	0	0.059	0.764	0	1
Chlorophytina	0	0	0	0	0	0.081	0	0.011	0.686	0	1
Bacillariophyceae	0	0	0.776	0	0	0	0	0	0	0	1
Chrysophyceae *	0	0	0.15	0	0	0.07	0	0	0	0	1
Pelagophyceae **	0	0.847	0.365	0	0	0	0	0	0	0	1
Cryptophyceae	0	0	0	0	0	0	0.277	0	0	0	1
Prymnesiophyceae	0	0.1335	0.309	0.675	0	0	0	0	0	0	1
Dinophyceae	0.838	0	0	0	0	0	0	0	0	0	1
Synechococcus	0	0	0	0	0	0	0	0.868	0	0	1
Prochlorococcus ***	0	0	0	0	0	0	0	0.389	0	1	0

Abbreviations: ButFuco = 19′-Butanoyloxyfucoxanthin; Fuco = Fucoxanthin; HexFuco = 19′-Hexanolyoxyfucoxanthin; Prasino = Prasinoxanthin; Viola = Violaxanthin; Allo = Alloxanthin; Zea = Zeaxanthin; Chl-*b* = Chlorophyll *b*; DV Chl-*a* = Divinyl Chlorophyll *a*; Chl-*a* = Chlorophyll *a*. Same abbreviation used in Tables S1–S5.

The CHEMTAX settings were established as reported by Latasa [5]: iteration limit of 5000, epsilon limit of 0.0001, initial step size of 25, step ratio of 2, cutoff step of 30,000, and bounded relative weighting. Identical settings were used for all 18 CHEMTAX runs.

2.7. Bayesian Compositional Estimator (BCE)

The BCE package [38] in R was used to manipulate the final ratio estimates for the 6 clusters. In this paper, the BCE input ratio was used as the initial ratio and the best-fit ratio was used as the final ratio to avoid terminology-related confusion; however, these terms do not exactly align with the terms used by Van den Meersche et al. [26]. The initial ratios presented in Table 2 were used for BCE, and 100,000 iterations were conducted for each estimation run. The values of jmpA and jmpX, basic settings of BCE analysis, were finely adjusted to obtain acceptance rates of 30–80% and minimize the impact of autocorrelation. The final settings applied and resulting acceptance rates are listed in

Table 3. The outputlengths parameter was set to 100, and other setting values used the default values of the BCE1 function in the BCE package.

Table 3. Final settings used for BCE analysis and the resulting acceptance rates.

Cluster	jmpA	jmpX	Iteration	Acceptance
February 2018	0.035	0.035	100,000	76.04%
August 2018 (Surface)	0.025	0.025	100,000	69.35%
August 2018 (SCM)	0.025	0.025	100,000	72.50%
June 2019	0.030	0.030	100,000	80.00%
October 2019	0.039	0.039	100,000	66.65%
April 2020	0.031	0.031	100,000	74.21%

3. Results

3.1. Statistical Indices and the Validity of Linear Modeling

The *p*-values for regression analysis were <0.001 for all pigments analyzed, supporting the validity of linear modeling. Adjusted R^2 values were >0.85 for most pigments. However, the adjusted R^2 values for peridinin and alloxanthin were low: 0.1233 and 0.1830, respectively. These low R^2 values suggest that peridinin and alloxanthin are more strongly influenced by environmental factors, whereas other pigments are mainly influenced by species composition. The *p*-values and adjusted R^2 values are presented in Table 4.

Table 4. *p*-values and adjusted R^2 values of multiple linear regression models.

Pigment	Adjusted R^2	*p*-Value
Peridinin	0.1233	0.0009
ButFuco	0.8925	<0.0001
Fuco	0.9031	<0.0001
Prasino	0.8538	<0.0001
Viola	0.8685	<0.0001
HexFuco	0.8547	<0.0001
Allo	0.1830	<0.0001
Chl-*b*	0.6765	<0.0001

3.2. The 95% Prediction Intervals for Pigment Ratios and the Initial Ratio and Ratio Limit Matrices (RLM) Constructed to Cover the Weighted Average of Lower and Upper Boundaries

The pigment ratio ranges determined by the 95% prediction intervals of NNLS are presented in Table 5. The lower and upper boundaries of the CHEMTAX pigment ratio were determined from the weighted averages of OTU pigment ratio boundaries. Then, CHEMTAX initial ratios and RLMs were determined to cover the whole ranges. The initial ratio and RLM for the February 2018 cluster are shown in Table 6, and the initial ratio and RLM combinations for other clusters are presented in Tables S1–S5 in the supplementary material. CHEMTAX analysis was conducted based on the values presented in those tables.

3.3. Final Ratios Obtained Using Three Different Combinations of Initial Ratios and Ratio Limits

As described in the Methods, three combinations of initial ratio and ratio limit values were used; the NGS model-based initial ratio and ratio limit, the NGS model-based initial ratio with the default ratio limit, and an initial ratio from previous research [27] (Table 2) with the default ratio limit. Linear regression analyses between CHEMTAX and BCE, final ratios were used to determine which combinations exhibited bias.

Among the CHEMTAX final ratios tested, the use of a model-based initial ratio and ratio limit produced a result similar to the BCE final ratio (R^2 = 0.982), indicating the greatest resistance to bias (Figure 3). In contrast, the use of a model-based initial ratio with the default ratio limit produced a low R^2 of 0.326, indicating some degree of bias. Finally, the use of the initial ratio from previous research [27] with the default ratio limit led to severe bias.

Table 5. NNLS-based 95% prediction intervals for pigment ratios.

CHEMTAX Taxon	Nearest Species for Each OTU	Peridinin	ButFuco	Fuco	Prasino	Viola	HexFuco	Allo	Chl-*b*
Mamiellophyceae	*Ostreococcus* sp.	-	-	-	0.044–0.069	0.016–0.025	-	-	0.163–0.382
	Micromonas sp.	-	-	-	0.023–0.185	0–0.042	-	-	0–1.367
	Micromonas pusilla	-	-	-	0.043–0.176	0–0.029	-	-	0–0.824
	Bathycoccus sp.	-	-	-	0.025–0.225	0–0.023	-	-	0.433–2.302
	Mamiellaceae sp.	-	-	-	0–0.140	0–0.061	-	-	0–1.004
Chlorophytina	Chlorellaceae sp.	-	-	-	-	0.008–0.045	-	-	0–0.658
Cryptophyceae	Pyrenomonadales sp.	-	-	-	-	-	-	0.183–0.498	-
Dinophyceae	*Amphidinium testudo*	0.240–0.892	-	-	-	-	-	-	-
Prymneisophyceae	Chrysochromulinaceae sp.	-	0.068–0.498	0.154–0.730	-	-	1.072–1.832	-	-
	Phaeocystis globosa	-	0–0.206	0.233–0.831	-	-	0–0.382	-	-
	Prymnesiophyceae sp.	-	0–0.221	0–0.446	-	-	0–1.005	-	-
	Phaeocystaceae sp.	-	0.076–0.456	0.156–0.757	-	-	0.530–1.523	-	-
	Prymnesiales sp.	-	0–0.248	0–0.304	-	-	0–0.720	-	-
	Braarudosphaeraceae sp.	-	0.535–1.229	0–0.844	-	-	0–1.401	-	-
	Emiliania huxleyi	-	0–0.428	0–0.739	-	-	0–1.084	-	-
	Chrysochromulinaceae sp.	-	0–0.481	0–0.508	-	-	0–2.312	-	-
Bacillariophyceae	Cymatosiraceae sp.	-	-	0.151–0.582	-	-	-	-	-
Chrysophyceae	Synurophyceae sp.	-	0–0.375	0–1.016	-	0–0.102	-	-	-
	Chrysophyceae sp.	-	0.455–2.724	0–2.718	-	0.378–0.739	-	-	-
	Chrysophyceae sp.	-	0–0.524	0–0.674	-	0–0.146	-	-	-
Pelagophyceae	Pelagophyceae sp.	-	0.388–1.264	0.162–1.341	-	-	-	-	-

Abbreviations: ButFuco = 19′-Butanoyloxyfucoxanthin; Fuco = Fucoxanthin; HexFuco = 19′-Hexanolyoxyfucoxanthin; Prasino = Prasinoxanthin; Viola = Violaxanthin; Allo = Alloxanthin; Zea = Zeaxanthin; Chl-*b* = Chlorophyll *b*; DV Chl-*a* = Divinyl Chlorophyll *a*; Chl-*a* = Chlorophyll *a*.

Table 6. Initial ratio and ratio limit for the February 2018 cluster.

February 2018	Taxa	Peridinin	ButFuco	Fuco	HexFuco	Prasino	Viola	Allo	Zea	Chl-*b*	DV Chl-*a*	Chl-*a*
Initial ratio	Mamiellophyceae	0	0	0	0	0.062	0.012	0	0.051	0.315	0	1
	Chlorophytina	0	0	0	0	0	0.019	0	0.072	0.226	0	1
	Bacillariophyceae	0	0	0.296	0	0	0	0	0	0	0	1
	Cryptophyceae	0	0	0	0	0	0	0.302	0	0	0	1
	Prymnesiophyceae	0	0.192	0.177	0.396	0	0	0	0	0	0	1
	Dinophyceae	0.463	0	0	0	0	0	0	0	0	0	1
	Synechococcus	0	0	0	0	0	0	0	0.519	0	0	1
RLM	Mamiellophyceae	0	0	0	0	139	226	0	89	229	0	0.1
	Chlorophytina	0	0	0	0	0	137	0	48	191	0	0.1
	Bacillariophyceae	0	0	96	0	0	0	0	0	0	0	0.1
	Cryptophyceae	0	0	0	0	0	0	65	0	0	0	0.1
	Prymnesiophyceae	0	282	299	259	0	0	0	0	0	0	0.1
	Dinophyceae	93	0	0	0	0	0	0	0	0	0	0.1
	Synechococcus	0	0	0	0	0	0	0	6	0	0	0.1

Abbreviations: ButFuco = 19′-Butanoyloxyfucoxanthin; Fuco = Fucoxanthin; HexFuco = 19′-Hexanolyoxyfucoxanthin; Prasino = Prasinoxanthin; Viola = Violaxanthin; Allo = Alloxanthin; Zea = Zeaxanthin; Chl-*b* = Chlorophyll *b*; DV Chl-*a* = Divinyl Chlorophyll *a*; Chl-*a* = Chlorophyll *a*.

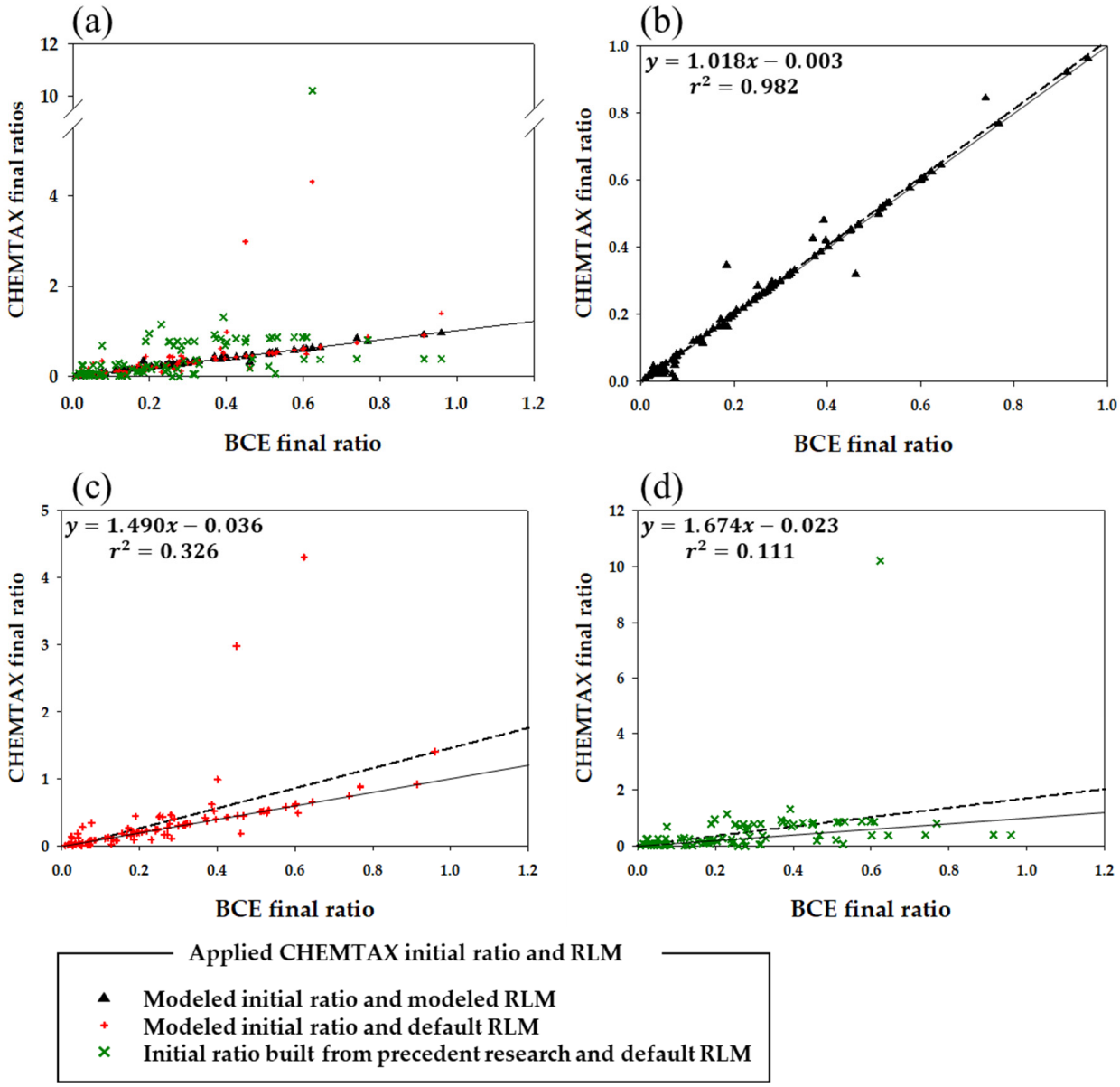

Figure 3. Regression analysis of (**a**) all BCE and CHEMTAX final ratios and those obtained using (**b**) modeled initial ratios and ratio limits, (**c**) modeled initial ratios with default ratio limits, and (**d**) initial ratios built from precedent research [27] (Table 2) with default ratio limits. Dashed and solid lines are regression and 1:1 lines, respectively.

3.4. Taxonomic Composition Based on Final Ratios

In this study, we investigated how the bias of CHEMTAX distorts community composition data. The BCE final ratio was converted into composition data using the constrained least squares method [39] (i.e., the tool CHEMTAX uses to convert the final ratio into composition data), and the results were compared with the output from CHEMTAX. The BCE best-fit results were not used to obtain composition data for reasons that are explained in the Discussion.

The CHEMTAX composition data obtained using the model-based initial ratio and ratio limit produced the highest R^2 of 0.989, indicating that these data were most similar to BCE final ratio-based composition data (Figure 4). When the default ratio limit was used, the R^2 value was reduced to 0.755, revealing a bias of decreased values for more abundant taxa and increased values for less abundant taxa. This finding indicates that the selection of

appropriate initial ratio and ratio limit values is essential for CHEMTAX analysis. Finally, CHEMTAX analysis based on previous research [27] produced highly biased data.

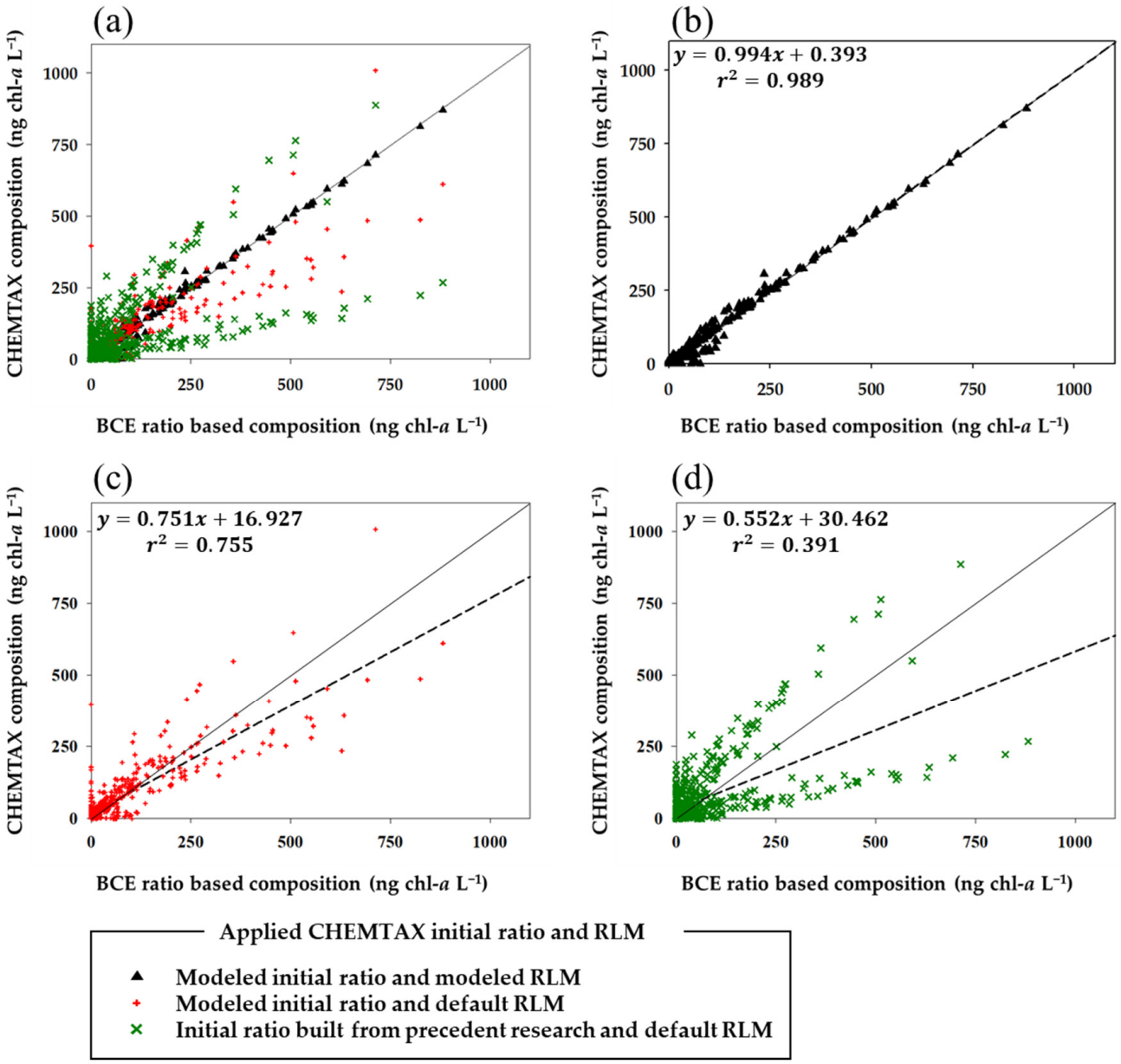

Figure 4. Regression analysis of composition data derived from BCE final ratios and (**a**) all CHEMTAX final ratios and those based on (**b**) modeled initial ratios and ratio limits, (**c**) modeled initial ratios and the default ratio limit, (**d**) and initial ratios built from precedent research [27] (Table 2) with default ratio limits. Dashed and solid lines are regression and 1:1 lines, respectively.

4. Discussion

4.1. Preventing Biased Output from CHEMTAX Estimation

In summary, CHEMTAX uses an underdetermined optimization algorithm and is therefore easily biased, resulting in unreliable composition data. Because restricting the coefficient variation range to have only one global minimum in root mean square (RMS) can prevent the result from having underdetermined bias, Mackey et al. included settings to restrict the variation range, which are the initial ratio and RLM [1]. They recommended that the initial ratio be close to the actual pigment ratio, but RLM does not seem to be as

important as the initial ratio, and they suggested the default RLM, which is a matrix with 500s for all elements.

However, our investigation found that the elements in the default RLM are too big to prevent the CHEMTAX analysis from giving biased results when analyzing the picophytoplankton community in the East Sea. If the dataset contains apparent global minimum in RMS even when using the default RLM, simply running CHEMTAX several times successively would improve the accuracy of CHEMTAX analysis, as suggested by Latasa [5]; however, no suggestions have been made for analysis in the absence of prominent global minima.

Accordingly, this study aimed to establish linear models between pigment ratios and 16S rDNA sequence composition, determine the statistically supported initial ratio and multiple linear regression value using 95% prediction intervals, and then conduct a CHEMTAX run. The resulting CHEMTAX final ratios were compared with the BCE final ratio, which is free of bias, to determine whether the CHEMTAX runs could successfully avoid underdetermined bias.

As noted in the Results, CHEMTAX with finely adjusted RLM values showed strong agreement with BCE, whereas the CHEMTAX with default RLM values showed poor agreement with BCE. These findings confirm that linear modeling between pigment ratios and 16S rDNA sequence composition data using the NNLS method, and 95% prediction intervals can effectively determine initial ratios and RLM values that prevent bias in CHEMTAX analysis. Furthermore, these results indicate that BCE is an effective optimization algorithm for the acquisition of bias-free final ratios.

4.2. The Application of BCE Alone for Chemotaxonomic Analysis Is Not Yet Recommended

In this study, the BCE algorithm was used as an indicator to determine whether range-limited CHEMTAX analysis successfully mitigated bias. If BCE is free of underdetermined bias, complete replacement of CHEMTAX with BCE would be recommended. BCE is a good optimization algorithm that allows researchers to omit the challenging process of selecting rational values for the initial ratio and RLM. However, the BCE algorithm has some critical issues that are difficult to address for now.

The most important issue with the BCE method is reproducibility. At a particular step of BCE optimization, known as the sampling step, the algorithm depends on random probability. Thus, different results are obtained from each run when BCE analysis is repeated multiple times with the same data and settings. Figure 5 show the results of correlation analysis between two BCE runs with the data and settings used in this study. In terms of both final ratio and composition, the two BCE analysis runs produced differing results.

In terms of the final ratio, the two BCE runs clearly differed but showed some degree of consistency, with an R^2 value of 0.911 and a formula near the 1:1 line (Figure 5a). Even tiny errors in the final ratio affected the composition results; however, considering the error in biased CHEMTAX results, BCE appears to be a reliable analysis method for determining the final ratio.

In contrast, the BCE best-fit results for composition data showed no correlation between the two runs, with an R^2 of 0.016, indicating that the BCE composition results are unreliable. The reproducibility problem is more extensive for composition results than for the final ratio because BCE estimates composition data independently of the final ratio [26].

From the perspective of the basic framework of chemotaxonomy, in which community composition is estimated from chemical marker concentrations, a rational approach would be to first determine the final ratio, and then calculate community composition from the final ratio. CHEMTAX and most other methods determine community composition in this manner, including methods based on the inverse simultaneous equation [43], multiple linear regression [44,45], and Excel Solver [46]. All of these methods determine the pigment ratio first, and then calculate community composition.

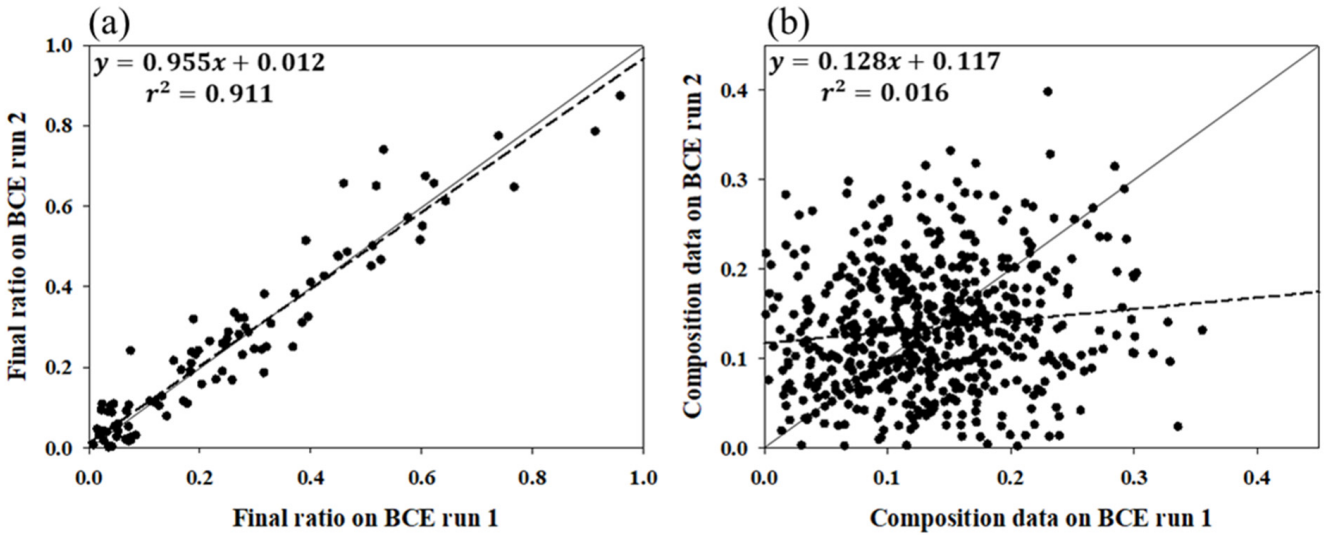

Figure 5. Comparison of the results of two BCE runs. (**a**) Comparison of final ratios between BCE runs 1 and 2, and (**b**) composition data for BCE runs 1 and 2. Dashed and solid lines are regression and 1:1 lines, respectively.

However, BCE avoids this general approach. The pigmentation of phytoplankton varies according to light intensity and wavelength, nutritive conditions, detailed species composition, and many other variables (e.g., [27,47–49]). Thus, each sample has a different pigment ratio, but conventional approaches (including CHEMTAX) use one pigment ratio for all samples. This simplification inevitably produces errors [1,26,27]. In contrast, BCE allows the pigment ratio and composition data to be independently estimated using Bayesian statistics.

This strategy of independent estimation causes the reproducibility problem inherent to BCE, resulting in composition data with high uncertainty (Figure 5b). Because in the optimization procedures in pigment ratio, the elements that BCE optimizes are the number of taxon-pigment combinations in the pigment ratio table: 14 in the February 2018 cluster in this research (see Table 6), for example. This number does not increase when the sample number increase. On the other hand, the number of elements for composition data BCE optimize are equivalent to the sample numbers times the targeted taxa number. For example, say we have seven taxonomic groups in the pigment ratio table and ten samples; the BCE need to optimize 70 elements for composition data. When we increase the sample number to 20, it becomes 140 elements.

Since the BCE algorithm is designed to test which pigment ratio matrices and composition data matrices are best fitting among the randomly produced matrices (the matrices' numbers are equivalent to iteration numbers), increasing the number of elements would result in decreasing probability that the BCE algorithm can find the best-fitting results from the limited iteration numbers. This is why the pigment ratio appears to be stable, while the composition data fluctuate every run: the increasing number of elements would harm the accuracy of composition data. Therefore, even when BCE optimization produces a bias-free final ratio, the composition data may be unreliable. Accordingly, we obtained composition data based on the BCE final ratio when investigating how the biased final ratio affects composition data, as shown in Figure 5 and in the Results.

Fortunately, a few methods exist to reduce the reproducibility problem that affects BCE composition data. First, confidence intervals are a good alternative to the best-fit values for composition data. Van den Meersche et al. [26] optimized BCE to produce a range of results using confidence intervals rather than exact values for each parameter.

Second, using valid prior information and fine-tuning the covariance matrix could significantly improve the accuracy of BCE estimation. Although the BCE sampling step relies on a random walk, the function is not entirely random; the starting point and standard

deviation can be fixed to restrict the variance of the random function. By manipulating specific settings, the users can improve BCE estimation. However, the BCE algorithm is more computationally intense than CHEMTAX and thus requires greater effort. Further intensification of the analytical method could increase the difficulty of analysis and cause the results to become more subjective, thereby making CHEMTAX users reluctant to apply the BCE algorithm. Thus, use of the BCE algorithm to make ecological inferences is challenging, although BCE is a more statistically advanced algorithm than CHEMTAX.

Despite these weaknesses, BCE could be a powerful tool for improving chemotaxonomic analysis, if it is used properly. The BCE algorithm has potential for future improvement; for example, the current Metropolis–Hastings sampling method could be replaced with a more advanced method, such as Gibbs sampling. Furthermore, CHEMTAX and BCE are complementary methods. Therefore, it is possible to determine whether a CHEMTAX final ratio is biased via comparison with the BCE final ratio; it is also possible to constrain the CHEMTAX pigment range using the BCE confidence intervals applied in this study. In conclusion, complementary use of BCE and CHEMTAX will improve chemotaxonomic estimation.

4.3. Advantages and Disadvantages of Chemotaxonomic Quantification Methods

CHEMTAX and BCE have advantages and disadvantages. We performed linear modeling of pigment ratio and 16S rDNA composition data to analyze the advantages of both methods and confirmed a linear relationship between 16S rDNA composition data and pigment concentrations (Table 7). These results provide new insight into the statistical support for initial ratios and ratio limit matrices (RLM).

Table 7. Advantages and disadvantages of chemotaxonomic quantification methods.

Advantages	Normal CHEMTAX	Bayesian Compositional Estimator	NGS Data Supported CHEMTAX
Robust to bias	No	Yes	Yes
Reproducibility on final ratio	Yes	No, but the deviation is acceptable	Yes
Reproducibility on composition	Yes	No	Yes
Evidence for taxa selection	No	No	Yes
Evidence for clustering	No	No	Yes

However, the least-squares method applied in this study is known to be vulnerable to multicollinearity. Accordingly, there may be some degree of bias in the coefficients determined using this method. We used prediction intervals rather than coefficients to mitigate bias caused by multicollinearity; however, this method does not perfectly eliminate bias. Nonetheless, our comparison of BCE's best-fit ratio and the linear model-supported CHEMTAX final ratio confirmed that it is an accurate analysis.

Another weakness could come from DNA copy number variability: It is well known that DNA copy numbers vary among individuals' genomes [50], possibly producing bias for using it quantitatively. However, the present study targeted plastidic 16s rDNA, which has less variability than nuclear 18S rDNA [51]. In addition, this research focused on small cell-sized phytoplankton, which reduced the interspecific DNA copy number variation [52,53]. Most importantly, the statistical indices presented in Table 4 show that the copy number variation rarely impacts the results in our investigation.

Alternatively, DNA composition data could provide critical clues for determining which taxonomic groups should be included in CHEMTAX analysis. Beta diversity data based on DNA composition data could represent practical evidence of prior clustering, which diminishes the error derived from using the same final ratios.

In conclusion, we propose that adopting 16S rDNA composition data into CHEMTAX analysis could significantly improve the performance (Table 7) and avoid the underdetermined bias of CHEMTAX, and eliminate the reproducibility problem of BCE. Furthermore,

the proposed method provides critical evidence for taxon selection and prior clustering, which could also improve the accuracy of CHEMTAX analysis.

5. Conclusions

We verified that the combination with the modeled initial ratio and RLM successfully avoids underdetermined bias while the CHEMTAX results with the default RLM suffer from bias (Figures 3 and 4). Although this research tested only the picophytoplankton community in the East Sea, CHEMTAX analysis for other phytoplankton communities may have a high probability of being biased; therefore, we strongly recommend that it is necessary to confirm if the CHEMTAX analysis successfully avoids bias before applying to general phytoplankton group, at least, apply a narrow RLM other than the default RLM. The linear model explained in this study is a great way to narrow the RLM, but BCE is a good alternative. However, caution is needed when using the direct composition data from the BCE analysis; i.e., the reproducibility needs to be checked.

To produce reproducible composition data from the BCE analysis, our suggestions are as follows; minimizing the sample size for each run by; sub-clustering the samples that have similar pigment ratios; using the information from the prior distribution data; fine-tuning the covariance matrix; using bigger iteration steps; and, using advanced sampling techniques other than the Metropolis-Hastings. Unfortunately, these suggestions are often difficult for ecologists to adopt because they require much greater statistical understanding than the CHEMTAX analysis. Albeit, from the present study, our work showed that to advance the CHEMTAX with enhanced BCE algorithm is essential to marine ecosystem study and would provide an improved methodological option for biological oceanographers.

Supplementary Materials: The following supporting information can be downloaded at https://www.mdpi.com/article/10.3390/jmse10121967/s1. Table S1: Initial ratio and ratio limit for the surface of the August 2018 cluster. Table S2: Initial ratio and ratio limit for SCM for the August 2018 cluster. Table S3: Initial ratio and ratio limit for the June 2019 cluster. Table S4: Initial ratio and ratio limit for the October 2019 cluster. Table S5: Initial ratio and ratio limit for the April 2020 cluster.

Author Contributions: Conceptualization, J.H.N. and M.J.H.; survey and sampling, M.J.H., J.W. and H.L.; data curation, M.J.H., J.W. and H.L.; funding acquisition, J.H.N. and C.H.P.; investigation, M.J.H., J.H.N. and Y.L.; methodology, J.H.N., M.J.H., H.L. and J.W.; project administration, J.H.N., Y.L. and C.H.P.; resources, J.H.N., Y.L. and D.H.C.; supervision, J.H.N.; visualization, M.J.H., J.W., H.L. and J.H.N.; writing—original draft, M.J.H. and J.H.N.; writing—review and editing, J.H.N., C.M.L. and D.H.C. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the project A Sustainable Research and Development of Dokdo (PG52911) funded by the Ministry of Oceans and Fisheries (MOF) of Korea. This study is also part of the project (PM63280) awarded by the Korea Institute of Marine Science & Technology Promotion (KIMST) also funded by MOF.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Acknowledgments: We wish to thank the editor and three other reviewers for their critical comments and encouragement. We also would like to acknowledge the captain and crew onboard the R/V Eardo for their assistance in this study.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Mackey, M.; Mackey, D.; Higgins, H.; Wright, S. CHEMTAX-a program for estimating class abundances from chemical markers: Application to HPLC measurements of phytoplankton. *Mar. Ecol. Prog. Ser.* **1996**, *144*, 265–283. [CrossRef]
2. Kim, T.-H.; Lee, Y.-W.; Kim, G. Hydrographically mediated patterns of photosynthetic pigments in the East/Japan Sea: Low N: P ratios and cyanobacterial dominance. *J. Mar. Syst.* **2010**, *82*, 72–79. [CrossRef]

3. Swan, C.M.; Vogt, M.; Gruber, N.; Laufkoetter, C. A global seasonal surface ocean climatology of phytoplankton types based on CHEMTAX analysis of HPLC pigments. *Deep. Sea Res. Part I: Oceanogr. Res. Pap.* **2016**, *109*, 137–156. [CrossRef]
4. Wright, S.W.; van den Enden, R.L.; Pearce, I.; Davidson, A.T.; Scott, F.J.; Westwood, K.J. Phytoplankton community structure and stocks in the Southern Ocean (30-80 degrees E) determined by CHEMTAX analysis of HPLC pigment signatures. *Deep. -Sea Res. Part II-Top. Stud. Oceanogr* **2010**, *57*, 758–778. [CrossRef]
5. Latasa, M. Improving estimations of phytoplankton class abundances using CHEMTAX. *Mar. Ecol. Prog. Ser.* **2007**, *329*, 13–21. [CrossRef]
6. Waterbury, J.B.; Watson, S.W.; Guillard, R.R.L.; Brand, L.E. Widespread occurrence of a unicellular, marine, planktonic, cyanobacterium. *Nature* **1979**, *277*, 293–294. [CrossRef]
7. Chisholm, S.W.; Olson, R.J.; Zettler, E.R.; Goericke, R.; Waterbury, J.B.; Welschmeyer, N.A. A novel free-living prochlorophyte abundant in the oceanic euphotic zone. *Nature* **1988**, *334*, 340–343. [CrossRef]
8. Partensky, F.; Blanchot, J.; Vaulot, D. Differential distribution and ecology of Prochlorococcus and Synechococcus in oceanic waters: A review. *Bull. -Inst. Oceanogr. Monaco-Numero Spec.* **1999**, *19*, 457–476.
9. Partensky, F.; Hess, W.R.; Vaulot, D. Prochlorococcus, a marine photosynthetic prokaryote of global significance. *Microbiol. Mol. Biol. Rev.* **1999**, *63*, 106–127. [CrossRef]
10. Richardson, T.L.; Jackson, G.A. Small phytoplankton and carbon export from the surface ocean. *Science* **2007**, *315*, 838–840. [CrossRef]
11. Buitenhuis, E.T.; Li, W.K.; Vaulot, D.; Lomas, M.; Landry, M.; Partensky, F.; Karl, D.; Ulloa, O.; Campbell, L.; Jacquet, S. Picophytoplankton biomass distribution in the global ocean. *Earth Syst. Sci. Data* **2012**, *4*, 37–46. [CrossRef]
12. Lepère, C.; Vaulot, D.; Scanlan, D.J. Photosynthetic picoeukaryote community structure in the South East Pacific Ocean encompassing the most oligotrophic waters on Earth. *Environ. Microbio.* **2009**, *11*, 3105–3117. [CrossRef] [PubMed]
13. Kim, D.; Ji, R.; Park, H.J.; Feng, Z.; Jang, J.; Lee, C.L.; Kang, Y.-H.; Kang, C.-K. Impact of Shifting Subpolar Front on Phytoplankton Dynamics in the Western Margin of East/Japan Sea. *Front. Mar. Sci.* **2021**, *8*, 790703. [CrossRef]
14. Joo, H.; Son, S.; Park, J.-W.; Kang, J.J.; Jeong, J.-Y.; Lee, C.I.; Kang, C.-K.; Lee, S.H. Long-term pattern of primary productivity in the East/Japan Sea based on ocean color data derived from MODIS-aqua. *Remote Sens.* **2015**, *8*, 25. [CrossRef]
15. Yoo, S.; Park, J. Why is the southwest the most productive region of the East Sea/Sea of Japan? *J. Mar. Syst.* **2009**, *78*, 301–315. [CrossRef]
16. Jang, H.-K.; Youn, S.-H.; Joo, H.; Kim, Y.; Kang, J.-J.; Lee, D.; Jo, N.; Kim, K.; Kim, M.-J.; Kim, S. First Concurrent Measurement of Primary Production in the Yellow Sea, the South Sea of Korea, and the East/Japan Sea, 2018. *J. Mar. Sci. Eng.* **2021**, *9*, 1237. [CrossRef]
17. Joo, H.; Son, S.; Park, J.-W.; Kang, J.J.; Jeong, J.-Y.; Kwon, J.-I.; Kang, C.-K.; Lee, S.H. Small phytoplankton contribution to the total primary production in the highly productive Ulleung Basin in the East/Japan Sea. *Deep. Sea Res. Part II: Top. Stud. Oceanogr.* **2017**, *143*, 54–61. [CrossRef]
18. Jung, S. Asynchronous responses of fish assemblages to climate-driven ocean regime shifts between the upper and deep layer in the Ulleung basin of the East Sea from 1986 to 2010. *Ocean. Sci. J.* **2014**, *49*, 1–10. [CrossRef]
19. Kim, S.-L.; Yu, O.-H. Understanding the Spatial and Temporal Distribution and Environmental Characteristics of Polychaete Assemblages in the Coastal Waters of Ulleungdo, East Sea of Korea. *J. Mar. Sci. Eng.* **2021**, *9*, 1310. [CrossRef]
20. Choi, H.; Hwang, J.; Kim, G.; Shin, K.-H. Seasonal Trophic Dynamics of Sinking Particles in the Ulleung Basin of the East Sea (Japan Sea): An Approach Employing Nitrogen Isotopes of Amino Acids. *Front. Mar. Sci.* **2022**, *9*, 520. [CrossRef]
21. Belkin, I.M. Rapid warming of large marine ecosystems. *Prog. Oceanogr.* **2009**, *81*, 207–213. [CrossRef]
22. Maranon, E. Cell size as a key determinant of phytoplankton metabolism and community structure. *Ann. Rev. Mar. Sci.* **2015**, *7*, 241–264. [CrossRef] [PubMed]
23. Daufresne, M.; Lengfellner, K.; Sommer, U. Global warming benefits the small in aquatic ecosystems. *Proc. Natl. Acad. Sci. USA* **2009**, *106*, 12788–12793. [CrossRef] [PubMed]
24. Lee, Y.-W.; Park, M.-O.; Kim, Y.-S.; Kim, S.-S.; Kang, C.-K. Application of photosynthetic pigment analysis using a HPLC and CHEMTAX program to studies of phytoplankton community composition. *Sea J. Korean Soc. Oceanogr.* **2011**, *16*, 117–124.
25. Lee, M.; Kim, Y.-B.; Park, C.-H.; Baek, S.-H. Characterization of Seasonal Phytoplankton Pigments and Functional Types around Offshore Island in the East/Japan Sea, Based on HPLC Pigment Analysis. *Sustainability* **2022**, *14*, 5306. [CrossRef]
26. Van den Meersche, K.; Soetaert, K.; Middelburg, J.J. A Bayesian compositional estimator for microbial taxonomy based on biomarkers. *Limnol. Oceanogr. Meth.* **2008**, *6*, 190–199. [CrossRef]
27. Roy, S.; Llewellyn, C.A.; Egeland, E.S.; Johnsen, G. *Phytoplankton Pigments: Characterization, Chemotaxonomy and Applications in Oceanography*; Cambridge University Press: Cambridge, UK, 2011.
28. Mäki, A.; Salmi, P.; Mikkonen, A.; Kremp, A.; Tiirola, M. Sample preservation, DNA or RNA extraction and data analysis for high-throughput phytoplankton community sequencing. *Front. Microbiol.* **2017**, *8*, 1848. [CrossRef]
29. Choi, D.H.; An, S.M.; Chun, S.; Yang, E.C.; Selph, K.E.; Lee, C.M.; Noh, J.H. Dynamic changes in the composition of photosynthetic picoeukaryotes in the northwestern Pacific Ocean revealed by high-throughput tag sequencing of plastid 16S rRNA genes. *FEMS Microbiol. Ecol.* **2016**, *92*, fiv170. [CrossRef]
30. Grossart, H.P.; Massana, R.; McMahon, K.D.; Walsh, D.A. Linking metagenomics to aquatic microbial ecology and biogeochemical cycles. *Limnol. Oceanogr.* **2020**, *65*, S2–S20. [CrossRef]

31. Zapata, M.; Rodríguez, F.; Garrido, J.L. Separation of chlorophylls and carotenoids from marine phytoplankton: A new HPLC method using a reversed phase C8 column and pyridine-containing mobile phases. *Mar. Ecol. Prog. Ser.* **2000**, *195*, 29–45. [CrossRef]
32. Somerville, C.C.; Knight, I.T.; Straube, W.L.; Colwell, R.R. Simple, rapid method for direct isolation of nucleic acids from aquatic environments. *Appl. Environ. Microbiol.* **1989**, *55*, 548–554. [CrossRef]
33. Sambrook, J.; Russell, D.W. Purification of nucleic acids by extraction with phenol: Chloroform. *Cold Spring Harb. Protoc.* **2006**, *2006*, pdb.prot4455.
34. Yang, W.; Noh, J.H.; Lee, H.; Lee, Y.; Choi, D.H. Weekly Variation of Prokaryotic Growth and Diversity in the Inner Bay of Yeong-do, Busan. *Ocean. Polar Res.* **2021**, *43*, 31–43.
35. Illumina, 16s Metagenomic Sequencing Library Preparation. 2013. Available online: https://sapac.support.illumina.com/downloads/16s_metagenomic_sequencing_library_preparation.html (accessed on 1 December 2022).
36. Kozich, J.J.; Westcott, S.L.; Baxter, N.T.; Highlander, S.K.; Schloss, P.D. Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq Illumina sequencing platform. *Appl. Environ. Microbiol.* **2013**, *79*, 5112–5120. [CrossRef]
37. Schloss, P.D.; Westcott, S.L.; Ryabin, T.; Hall, J.R.; Hartmann, M.; Hollister, E.B.; Lesniewski, R.A.; Oakley, B.B.; Parks, D.H.; Robinson, C.J. Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl. Environ. Microbiol.* **2009**, *75*, 7537–7541. [CrossRef]
38. Van den Meersche, K.; Soetaert, K. BCE: Bayesian Composition Estimator: Estimating Sample (Taxonomic) Composition from Biomarker Data. R Package Version 2.1. 2014. Available online: https://search.r-project.org/CRAN/refmans/BCE/html/BCE-package.html (accessed on 1 December 2022).
39. Wang, Y.; Lawson, C.; Hanson, R. Lsei: Solving Least Squares or Quadratic Programming Problems under Equality/Inequality Constraints. R Package Version 1.3-0. 2020. Available online: https://cran.r-project.org/web/packages/lsei/DESCRIPTION (accessed on 1 December 2022).
40. R Core Team. *R: A Language and Environment for Statistical Computing*; Foundation for Statistical Computing: Vienna, Austria, 2021; Available online: https://www.R-project.org/ (accessed on 1 December 2022).
41. Lawson, C.L.; Hanson, R.J. *Solving Least Squares Problems*; SIAM: Philadelphia, PA, USA, 1995.
42. Fox, J.; Weisberg, S. *An R Companion to Applied Regression*; Sage Publications: Thousand Oaks, CA, USA, 2018.
43. Everitt, D.; Wright, S.; Volkman, J.; Thomas, D.; Lindstrom, E. Phytoplankton community compositions in the western equatorial Pacific determined from chlorophyll and carotenoid pigment distributions. *Deep. Sea Res. Part A Oceanogr. Res. Pap.* **1990**, *37*, 975–997. [CrossRef]
44. Gieskes, W.; Kraay, G. Dominance of Cryptophyceae during the phytoplankton spring bloom in the central North Sea detected by HPLC analysis of pigments. *Mar. Biol.* **1983**, *75*, 179–185. [CrossRef]
45. Uitz, J.; Claustre, H.; Morel, A.; Hooker, S.B. Vertical distribution of phytoplankton communities in open ocean: An assessment based on surface chlorophyll. *J. Geophys. Res. Ocean.* **2006**, *111*. [CrossRef]
46. Not, F.; Latasa, M.; Scharek, R.; Viprey, M.; Karleskind, P.; Balagué, V.; Ontoria-Oviedo, I.; Cumino, A.; Goetze, E.; Vaulot, D. Protistan assemblages across the Indian Ocean, with a specific emphasis on the picoeukaryotes. *Deep. Sea Res. Part I Oceanogr. Res. Pap.* **2008**, *55*, 1456–1473. [CrossRef]
47. Falkowski, P.G.; Raven, J.A. *Aquatic Photosynthesis*; Princeton University Press: Priceton, NJ, USA, 2013.
48. Wu, J.; Sunda, W.; Boyle, E.A.; Karl, D.M. Phosphate depletion in the western North Atlantic Ocean. *Science* **2000**, *289*, 759–762. [CrossRef] [PubMed]
49. Van Lenning, K.; Latasa, M.; Estrada, M.; Sáez, A.G.; Medlin, L.; Probert, I.; Véron, B.; Young, J. Pigment signatures and phylogenetic relationships of the pavlovophyceae (haptophyta) 1. *J. Phycol.* **2003**, *39*, 379–389. [CrossRef]
50. Lavrinienko, A.; Jernfors, T.; Koskimäki, J.J.; Pirttilä, A.M.; Watts, P.C. Does intraspecific variation in rDNA copy number affect analysis of microbial communities? *Trends Microbiol.* **2021**, *29*, 19–27. [CrossRef]
51. Decelle, J.; Romac, S.; Stern, R.F.; Bendif, E.M.; Zingone, A.; Audic, S.; Guiry, M.D.; Guillou, L.; Tessier, D.; Le Gall, F. Phyto REF: A reference database of the plastidial 16S rRNA gene of photosynthetic eukaryotes with curated taxonomy. *Mol. Ecol. Resour.* **2015**, *15*, 1435–1445. [CrossRef] [PubMed]
52. Godhe, A.; Asplund, M.E.; Härnström, K.; Saravanan, V.; Tyagi, A.; Karunasagar, I. Quantification of diatom and dinoflagellate biomasses in coastal marine seawater samples by real-time PCR. *Appl. Environ. Microbiol.* **2008**, *74*, 7174–7182. [CrossRef] [PubMed]
53. Vaulot, D.; Eikrem, W.; Viprey, M.; Moreau, H. The diversity of small eukaryotic phytoplankton ($\leq$3 µm) in marine ecosystems. *FEMS Microbiol. Rev.* **2008**, *32*, 795–820. [CrossRef] [PubMed]