Avoiding quantification bias in metabarcoding: application of a cell biovolume correction factor in diatom molecular biomonitoring

Valentin Vasselon[*], Agnès Bouchez[*], Frédéric Rimet [*], Stéphan Jacquet[*], Rosa Trobajo[†], Méline Corniquel[*], Kálmán Tapolczai [*], Isabelle Domaizon[*]

[*]*CARRTEL, French National Institute for Agricultural Research (INRA), University of Savoie Mont Blanc, 75 bis Avenue de Corzent, 74200, Thonon-les-Bains, France,* [†] *Aquatic Ecosystems, Institute for Food and Agricultural Research and Technology (IRTA), Crta de Poble Nou Km 5.5, Sant Carles de la Ràpita, Catalunya, Spain.*

Corresponding author: Valentin Vasselon; 75 bis Avenue de Corzent, 74200, Thonon-les-Bains, France; +33 (0)4 50 26 78 29; valentin.vasselon@inra.fr

Running headline: Improvement of diatom HTS quantification

Abstract

1. In recent years, remarkable progress has been made in developing environmental DNA metabarcoding. However, its ability to quantify species relative abundance remains uncertain, limiting its application for biomonitoring. In diatoms, although the *rbc*L gene appears to be a suitable barcode for diatoms, providing relevant qualitative data to describe taxonomic composition, improvement of species quantification is still required.

2. Here, we hypothesized that *rbc*L copy number is correlated with diatom cell biovolume (as previously described for the 18S gene) and that a correction factor (CF) based on cell biovolume should be applied to improve taxa quantification. We carried out a laboratory experiment using pure cultures of 8 diatom species with contrasted cell biovolumes in order to (i) verify the relationship between *rbc*L copy numbers (estimated by qPCR) and diatom cell biovolumes, and (ii) define a potential CF. In order to evaluate CF efficiency, five mock communities were created by mixing different amounts of DNA from the 8 species, and were sequenced using HTS and targeting the same *rbc*L barcode.

3. As expected, the correction of DNA reads proportions by the CF improved the congruence between morphological and molecular inventories. Final validation of the CF was obtained on environmental

samples (metabarcoding data from 80 benthic biofilms) for which the application of CF allowed differences between molecular and morphological water quality indices to be reduced by 47 %.

4. Overall, our results highlight the usefulness of applying a CF factor, which is effective in reducing over-estimation of high biovolume species, correcting quantitative biases in diatom metabarcoding studies and improving final water quality assessment.

Keywords: Benthic diatom, Biovolume correction factor, Freshwater ecosystems, Gene copy number variation, Quantitative metabarcoding

Introduction

DNA metabarcoding allows species present in an environmental sample to be detected using a short DNA marker specific for a particular taxonomic group (Taberlet *et al.* 2012). Combined with High-Throughput Sequencing (HTS), hundreds of samples can be analyzed at the same time, offering an alternative to microscopy with higher resolution and accuracy, while being faster and cheaper (Stein *et al.* 2014). This is particularly interesting for freshwater biomonitoring, in which thousands of river samples have to be analyzed annually and management actions applied quickly (Keck *et al.* 2017). The European Water Framework Directive (WFD, European Council 2000) has implemented the use of benthic diatoms, among other biological indicators (fishes, macroinvertebrates, macrophytes and phytoplankton), for the assessment of aquatic ecosystem integrity. The different biotic diatom indices that have been developed are based on the relative abundances and the ecological values (sensitivity and tolerance to pollutants) of the species observed in rivers and lakes systems (*e.g.* Rimet 2012). Different studies have already revealed the potential application of diatom metabarcoding in freshwater quality assessment (Kermarrec *et al.* 2014; Visco *et al.* 2015; Vasselon *et al.* 2017a,b; Apothéloz-Perret-Gentil *et al.* 2017). However, discrepancies between DNA metabarcoding and microscopy have been observed in species composition and relative abundance (Zimmermann *et al.* 2015). This drawback is likely to affect the congruence

between morphological and DNA metabarcoding quality index values and, *in fine*, the ecological assessment.

With respect to qualitative aspects, the incompleteness of the reference databases, the choice of the DNA marker and the efficiency of the PCR primers have been identified as important biases affecting species detection using DNA metabarcoding (Pawlowski *et al.* 2016). For benthic diatoms, the *rbc*L gene has proved to be an appropriate taxonomic marker for biomonitoring (Kermarrec *et al.* 2013, 2014, Vasselon *et al.* 2017a,b) and a well-curated barcode reference library is already available in open-access to assign species names to *rbc*L sequences (R-Syst::diatom, Rimet *et al.* 2016). However, no clear relationship has yet been demonstrated between the relative species abundances obtained by DNA metabarcoding with the *rbc*L barcode and those obtained by morphological observations (Rimet *et al.* 2014). As quantification of diatom species is required by the WFD for quality index calculation, more investigation is needed to understand and correct biases affecting diatom quantification based on HTS data.

Species quantification based on HTS data can be estimated from the number of DNA sequences (*i.e.* reads) assigned to each species, from which relative abundances can be calculated. Previous studies have documented a variety of problems that may affect the proportions of DNA reads obtained with HTS (Amend, Seifert & Bruns 2010; Deagle *et al.* 2013; Tan *et al.* 2015; Thomas *et al.* 2016; Pawlowski *et al.* 2016), including biological biases (*e.g.* gene copy number variation, tissue cell density, cell biovolume), technical biases (*e.g.* DNA extraction, PCR amplification), and biases linked to HTS itself (*e.g.* library construction, HTS technology used, bioinformatics treatments). Variation of gene copy number per cell constitutes a major bias known to affect the proportion of DNA-read found for each species present in complex assemblages; this has been demonstrated for macroinvertebrates (Elbrecht, Peinert & Leese 2017), fish, amphibians (Evans *et al.* 2016), oligochaetes (Vivien, Lejzerowicz & Pawlowski 2016), foraminifera (Weber & Pawlowski 2013), and microbial assemblages (Angly *et al.* 2014). However, to the best of our knowledge, no study has yet evaluated gene copy number variation bias on diatom metabarcoding quantification. While tissue cell density and species biomass are major biases likely to affect DNA metabarcoding quantification of multicellular organisms like macroinvertebrates (Elbrecht &

Leese 2015) or fish (Evans *et al.* 2016), diatoms are unicellular organisms for which gene copy number is mainly affected by the number of genomes and the number of gene copies per genome. This may be particularly true for non-nuclear markers like the chloroplast-encoded *rbc*L gene. Godhe *et al.* (2008) reported a clear correlation between the 18S gene copy number per cell with diatom cell length and biovolume, suggesting that the cell biovolume could be a proxy for the gene copy number. Keeping in mind that diatom biovolume varies from 10^1 to 10^9 µm^3 (Snoeijs, Busse & Potapova 2002), gene copy number may vary greatly between the smallest and the biggest diatom species, affecting metabarcoding quantification.

For all the reasons mentioned above, we hypothesized that a quantification correction factor (CF) based on diatom cell biovolume should be necessary to correct DNA read proportions to provide species quantification more comparable to microscopical counts. In order to confirm this hypothesis, we firstly conducted experiments on 8 pure diatom cultures to examine whether variation in *rbc*L gene copy number per cell correlates with morphological characteristics (*e.g.* biovolume, cell length), from which a CF might be calculated. Secondly, the efficiency of the proposed CF was tested on (i) mock communities made by mixing known proportions of the 8 diatom species cultures, and (ii) environmental diatom assemblages from rivers previously sequenced (Vasselon *et al.* 2017b) and for which data are available online (Vasselon *et al.* 2017b dataset, http://doi.org/10.5281/zenodo.400160). Last, the capacity of the CF to improve the ecological assessment of rivers was tested by comparing water quality index values calculated from molecular data with corrected abundances to those calculated from classical morphological abundances.

Methods

Evaluation of the quantification bias and development of a quantification correction factor (CF)

To evaluate whether the *rbc*L copy number per cell varies between diatom species, strains from 8 freshwater diatom species were selected from the Thonon Culture Collection (TCC; http://www6.inra.fr/carrtel-collection_eng/) **(Table 1)**. The 8 species were chosen for their contrasted morphological (size and cell biovolume), cytological (*e.g.* chloroplast number) and phylogenetic

characteristics **(Table 1)**. Cell dimensions (width, length, thickness) of the 8 diatom species were measured under light microscopy (1000× magnification) using a minimum of 10 specimens per species. Then, appropriate geometrical models were applied to calculate their cell biovolume (Sun & Liu 2003) **(Table 1)**. The 8 diatom cultures were cultivated in triplicate in 40 mL sterile DV medium (Rimet *et al.* 2014) using 50 mL Nunc™ EasYFlasks™ (Thermo Fisher Scientific, Waltham, Massachusetts). Flasks were placed on a rotating platter (4 rpm) in a controlled thermostatic room (21 ± 2°C, 14h light/10h dark cycle, light intensity of ca. 100 µmol quanta m^{-2} s^{-1}). Flasks were inoculated in order to reach a concentration of ≈ 100 cells/mL at the beginning of the experiment for each species, except for *Ulnaria ulna* for which a concentration of ≈ 1000 cells/mL was used (due to its low growth rate). The growth of the 8 diatom cultures was followed during 40 days, except for *Pinnularia viridiformis* for which the survey lasted 73 days, due to its low growth rate. Cell concentrations, proportions of live/dead cells and *rbc*L gene copy concentrations per mL of media were measured for each culture at 7 sampling times (referred to as T0 to T6) **(Fig. 1)**.

Diatom cell concentrations and proportions of live/dead cells were obtained by counting at least 400 specimens using inverted microscopy (×1000 magnification) and the standard Utermöhl technique (European Committee for Standardization (CEN) 2006) **(Fig. 1)**. The proportion of live/dead cells was estimated by considering cells without visible intracellular contents as dead. Only living cells were taken into account to calculate the diatom cell concentration per mL of media. Flow cytometry using Sytox-Green was also used to confirm the microscopical data (not shown).

*Rbc*L copy number per mL was estimated by qPCR. From each cultivation replicate, 10 mL of culture was centrifuged at 17,000 x *g* for 30 min **(Fig.1)**. Total DNA was extracted from the resulting pellet using a protocol based on GenEluteTM-LPA DNA precipitation (Sigma-Aldrich, St Louis, Missouri) as previously described (Vasselon *et al.* 2017a). Then, qPCR assays were performed for each of the 8 diatom species on DNA extracted at all 7 sampling times and with each of the 3 replicates, using the QuantiTect SYBR Green PCR Kit (Life Technologies, Carlsbad, USA) and the Rotor-Gene Q (Qiagen, Hilden, Germany). A short 312 bp region of the *rbc*L gene (the same as was used for HTS sequencing) was targeted using primers used by (Vasselon *et al.* 2017b) and described in **Table S1**. qPCR reactions were performed following the method

used by Vasselon *et al.* (2017a), using a final volume of 25 µL using mix preparation and reaction conditions as described in **Table S1**. A fluorescence threshold of 0.01 was used to allow comparison of qPCR assays, denoising and determination of the cycles' threshold (Ct). Data analysis was performed using the Rotor-Gene Q Series software (version 2.3.1) and the *rbc*L copy per mL of media was determined.

Finally, the number of *rbc*L gene copies per diatom cell was calculated for the 8 diatom species by dividing the *rbc*L concentration (qPCR data) by the living cell concentration (microscopy data). A Kruskal-Wallis test was performed using R (R Development core team 2013) to determine if the *rbc*L gene copy number per diatom cell varied significantly between the 8 diatom species. Then, we tested the level of correlation between the number of *rbc*L gene copies per diatom cell and several morphological characteristics of the diatom cells **(Table 1)**. Variables that did not approximate normal distributions were log transformed. Pearson correlation coefficients were calculated between the gene copy number per cell and the diatom cell morphological characteristics. This correlation was represented by a linear model.

Validation of the quantification CF to mock and environmental HTS data

Mock communities The calculated CF was applied to metabarcoding data obtained from controlled diatom mock communities. 5 mock communities (M1 to M5) were created by mixing DNA extracted from each of the 8 diatom species sampled during their exponential growth phase, and for which the correspondence between cell abundances (microscopy) and qPCR counts was known. For each of the 5 mock communities, the volume of DNA used for 7 species was kept unchanged (1 µL) and only the volume of DNA of *P. viridiformis* varied as followed: M1 = 0.2 µL, M2 = 0.4 µL, M3 = 0.8 µL, M4 = 1.6 µL, M5 = 3.2 µL. This resulted in contrasted *rbc*L proportions of the 8 species among the 5 mock communities. Then, HTS sequencing of the *rbc*L 312 bp fragment was performed on 3 replicates of the 5 mock communities. The 15 corresponding libraries were prepared following the method described by Vasselon *et al.* (2017a) with the same primers and PCR reaction conditions as those used for *rbc*L qPCR **(Table S1),** changing only the cycle number to 30. Each library was diluted to 100 pm and all 15 were pooled together for one HTS run

159 performed on the PGM Ion Torrent machine by the “Plateforme Génome Transcriptome” (PGTB, Bordeaux,
160 France).

161 The sequencing platform provided a unique fastq file for each of the 15 libraries containing
162 demultiplexed DNA reads without the sequencing adapters. Quality filtering of DNA reads was performed
163 using the Mothur software (Schloss *et al.* 2009) and bioinformatics process described previously (Vasselon
164 *et al.* 2017a,b). Finally, a taxonomy was assigned to each DNA read with the “classify.seqs” command
165 (Mothur) using default parameters with a confidence threshold of 85% and the R-Syst::diatom library
166 (Rimet *et al.* 2016, version updated in January 2015 and available upon request) as a *rbc*L reference library.
167 A molecular taxonomic list with the associated read numbers assigned to each of the 8 diatom species was
168 obtained for each of the 5 mock communities and used for subsequent analysis.

169 The quantification CF defined for the *rbc*L gene was then applied to the molecular taxonomic lists
170 for the 5 mock communities by dividing the read number for each species by its corresponding CF. Both the
171 uncorrected and corrected HTS relative abundances of species from the 5 mock communities were then
172 compared to the relative abundances obtained using microscopy.

173 ***Environmental diatom assemblages*** To evaluate the efficiency of the CF to improve metabarcoding
174 quantification from environmental samples, we used *rbc*L HTS data obtained from (Vasselon *et al.* 2017b),
175 corresponding to 80 benthic diatom samples collected from rivers in tropical island of Mayotte, Indian
176 Ocean (Vasselon *et al.* 2017b dataset, http://doi.org/10.5281/zenodo.400160). A CF was calculated for
177 each species (or genus when the species level was not reached) detected in molecular inventories of the
178 rivers of Mayotte island using a generalised average of the morphological information (*e.g.* biovolume,
179 length) available in the R-Syst::diatom library and applied to HTS data. Corrected molecular inventories
180 were produced for all the 80 river samples using the CF. The impact of the CF on diatom taxa abundance
181 rank in the molecular inventories was assessed by comparing original and corrected molecular diatom
182 inventories. Then, the Specific Pollution-sensitivity Index (SPI) used for ecological assessment was
183 calculated for each sample based on the corrected diatom molecular inventories using the Omnidia 5
184 software (Lecointe, Coste & Prygiel 1993, library 5.3 2015) and compared to the morphological SPI values

for all river samples (Vasselon *et al.* 2017b). Pearson correlation was used to evaluate the strength of correlations between original or corrected molecular SPI values and the morphological SPI values. Wilcoxon Signed Rank tests were conducted to determine whether the difference between the molecular and the morphological SPI (ΔSPI) varied significantly when using the original or the corrected molecular data for the molecular SPI calculation.

Results

*Variation of rbc*L *gene copy number between diatom species*

Cell and *rbc*L gene concentrations were measured, by inverted microscopy and qPCR respectively, for the 8 diatom species at different cultivation stages corresponding to 7 sampling points (T0 to T6). Information has been summarized in **Tables S2** and **S3**. As the 8 diatom species reached the beginning of the stationary phase at the sampling time T2 (*i.e.* between 13 and 31 days of cultivation), only the [cell] and the [gene copy] values obtained for the T0, T1 and T2 sampling times were used for further analysis. The calculated mean values of the *rbc*L gene copy number per cell for each diatom species varied between 0.5 and 130 copies per cell **(Fig. 2)**. The Kruskal-Wallis test revealed that the *rbc*L copy number per cell was significantly different ($p < 0.001$) between the 8 diatom species.

Development of quantification CFs

The *rbc*L copy number per cell was highly correlated with cell biovolume ($r = 0.97$, $p < 0.001$), length ($r = 0.82$, $p < 0.001$), width ($r = 0.94$, $p < 0.001$) and thickness ($r = 0.96$, $p < 0.001$). The correlation between the *rbc*L copy number per cell and the cell biovolume followed a linear model **(Fig. 3)**. Assuming that this linear relation based on 8 diatom species is applicable to all diatom species, the equation of this model allows calculation of an estimate of the relative *rbc*L copy number per cell as soon as the biovolume of the cell is known, and thus to define a CF specific to each species. Such quantification CFs were calculated for each of the 8 diatom species of the mock communities **(Table 2)** and varied from 0.6 for *Achnanthidium minutissimum* to 78.5 for *P. viridiformis*. For each of the diatom taxa found in the environmental samples,

CFs were also calculated using the biovolume information available for each taxa (from Rsyst::diatom library) **(Table S4)** and varied over a wider range, from 0.03 for *Fistulifera saprophila* to 649.8 for *Rhopalodia gibba*.

Application of CFs to mock and environmental HTS data

953,082 DNA reads were produced from the 15 libraries corresponding to the 5 DNA mock communities (3 replicates per mock). Following the bioinformatics quality filtering steps, 385,367 DNA reads were retained. A molecular taxonomic list was then created by removing DNA reads which remained unclassified (0.43 % of the reads) or assigned to different taxa than the 8 diatom species present in the mock communities (0.004 % of the reads) **(Table S5)**. The proportions of *P. viridiformis* reads in the 5 mock communities varied from 9 % in M1 to 57 % in M5 **(Fig. 4A)** while observed cell proportions were lower; ≈ 0.03 % in M1 and 0.55 % in M5 **(Fig. 4B)**. The application of the CF on DNA reads counts of the 8 species changed their relative abundances in the 5 mock communities **(Fig. 4A)**. The rank of the 8 species was also affected; for example, in M5 the application of the CF changed the proportion of *P. viridiformis* from 57 % to 4 % and the proportion of *A. minutissimum* from 4 % to 42 %. The correspondence between morphological and molecular relative abundances was highly improved by applying the CF on the HTS data **(Fig. 4A, 4B)**.

From the 80 environmental samples previously sequenced (Vasselon *et al.* 2017b), a molecular taxonomic list based on assigned DNA reads was produced including 23 families (75.1 % of total reads assigned), 39 genera (72 % of total reads assigned) and 66 diatom species (40.7 % of total reads assigned). From this list, 84 diatom taxa, including taxa assigned at the genus and the species level, were used to calculate the SPI freshwater quality index. CFs calculated from cell biovolumes for those 84 taxa were then applied to correct the quantification of the environmental molecular inventories **(Table S4)**. The proportions and ranks of the dominant taxa were affected by the application of the CFs **(Fig. 5)**. For example, the application of CFs reduced the relative abundances of *Eunotia* and *Ulnaria* from 31.9 % to 3.3 % and 11.7 % to 2.3 %, respectively, making them more congruent with cell proportions observed with

microscopy (3.1% for *Eunotia* and 0.4 % for *Ulnaria*). The correlation between the morphological and the molecular SPI values for all river samples previously described ($r = 0.72$, $p < 0.001$) was slightly improved using SPI values based on inventories with corrected abundances ($r = 0.77$, $p < 0.001$). The application of the CF to correct the HTS quantification reduced significantly ($p < 0.001$) the differences between the molecular and morphological SPI values by 47 % (ΔSPI reduced to 1.9 on average compared to 3.6 before correction, corresponding to 37.3 % and 21.2 % of error respectively) **(Fig. 6)**.

Discussion

Species quantification based on DNA metabarcoding is challenging for most of taxonomic groups as technical and biological biases affect DNA reads proportions. In order to limit those biases, several attempts were done to apply a CF on metabarcoding data, as shown for fishes (Thomas *et al.* 2016), bacteria and archea (Angly *et al.* 2014) or oligochaetes (Vivien, Lejzerowicz & Pawlowski 2016). For those studies, application of the CF, whether for correcting single (Angly *et al.* 2014) or multiple sources of quantification biases (Thomas *et al.* 2016), improved taxa quantification from metabarcoding data compare to morphological one. The result is generally a change in the ranks of the dominant taxa which affect directly the community structure and can lead to different ecological interpretations. For example, the application of a CF on metabarcoding data obtained from aquatic oligochaetes samples improved the freshwater quality assessment based on molecular index calculation (Vivien, Lejzerowicz & Pawlowski 2016). However, the development of CF can be challenging depending on the organism studied, as it requires finding a clear relationship between DNA reads and specimen proportions. This may be impossible due to accumulation of quantification biases (*e.g.* cell density, cell biomass, gene copy number). Nevertheless, the use of CF can be advantageous for organisms with a high variation of the DNA reads proportions between taxa (*e.g.* several log) and where a limited number of biases are involved like diatoms.

Correlation between *rbc*L gene copy number and diatom cell biovolume: impacts on HTS quantification

The copy number of the *rbc*L gene present in one diatom cell is affected by 3 parameters: (i) the number of chloroplasts per cell, (ii) the number of genomes per chloroplast and (iii) the number of copies of the *rbc*L gene per chloroplast genome (Ersland, Aldrich & Cattolico 1981; Treusch *et al.* 2012). (i) For benthic diatoms, the chloroplast number per cell is quite stable inside a single genus with variations ranging from 1 to ≈ 8 chloroplast(s) per cell from a genus to another (Round, Crawford & Mann 1990), even if some centric genera may have tens of chloroplasts (*e.g. Melosira*, *Cyclotella*). (ii) Regarding the chloroplast genome number per cell, higher plants can contain up to thousands of copies of chloroplast genome per cell (Bendich 1987; Rauwolf *et al.* 2010) while unicellular algae generally exhibit a lower number of copies. For example, *Olisthodiscus luteus* (Raphidophyceae), *Chlamydomonas reinhardtii* (Chlorophyceae), *Phaeodactylum tricornutum* (pennate diatom) and *Thalassiosira pseudonana* (centric diatom) contain respectively around 650, 80, 137 and 55 genome copies per cell (Ersland, Aldrich & Cattolico 1981; Koop *et al.* 2007; Gruber 2008; von Dassow *et al.* 2008). (iii) Finally, there is only 1 copy of the *rbc*L gene per chloroplast genome (*e.g.* Sabir *et al.* 2014), as in higher plants (Gutteridge & Gatenby 1995).

Thus, the *rbc*L copy number may vary from tens to hundreds of copies per diatom cell. Our estimations are within this range with a maximum of 130 copies estimated for *P. viridiformis*. However, our method underestimates the *rbc*L gene copy number since 0.5 copy per cell was estimated for *A. minutissimum* (so implying that some cells have no *rbc*L copy). This may result from certain variability inherent to the estimation of gene copy number by qPCR and the quantification of cells by microscopical counts. Our results demonstrate, however, that the *rbc*L copy number varies significantly between the 8 diatom species used in this study, according to the different diatom cell characteristics tested. In particular, we found a significant linear relationship between the *rbc*L copy number and the cell biovolume. Although the size of the chloroplasts could not be estimated in this study, we assume that the increase of the cell biovolume is accompanied by an increase of the chloroplast biovolume (as shown by Okie, Smith & Martin-Cereceda 2016), inducing an increase of DNA quantity and chloroplast genome copies per chloroplast as shown by Rauwolf *et al.* (2010).

The correlation we found between the *rbcL* copy number and the diatom cell biovolume suggests that the relative abundance of diatom species with high cell biovolume is likely to be over-represented in metabarcoding data compared to microscopical counts. This is confirmed by the HTS data obtained for the mock communities, where diatom species with high cell biovolume are over-represented (*e.g. P. viridiformis*) and diatom species with low cell biovolume are under-represented (*e.g. A. minutissimum*). The relative abundance of *P. viridiformis* in the mock communities was negligible compared to other species, and doubling its proportion did not change its rank: the species remained the least abundant taxon within the morphological inventory. However, due to its high cell biovolume (10^4 µm^3) and relatively high *rbc*L copy number per cell, a marked over-representation of this species within the molecular inventory was observed. A CF was thus defined to correct these quantitative biases and was verified on mock communities and environmental samples.

Current potential and limits of the quantification CF

The use of the same *rbc*L primers for the qPCR assays and the HTS enabled us to generate a specific CF well suited to correct *rbc*L metabarcoding quantifications. Its application to the HTS data of the mock communities allowed us to obtain comparable species proportions in morphological and molecular based approaches of mock communities. This was also confirmed with the Mayotte river samples, for which the quantification CF resulted in a better congruence between DNA reads and cells proportions, reducing the over-representation of high biovolume *Eunotia* and *Ulnaria* species. Furthermore, SPI calculation based on corrected metabarcoding data gives SPI values more comparable to SPI values obtained from morphological data, suggesting that it may be possible to replace morphological by molecular monitoring for the ecological assessment of Mayotte rivers. In the same way, (Vivien, Lejzerowicz & Pawlowski 2016) have shown that application of a CF to correct DNA reads proportions allows a more accurate estimation of oligochaete proportions, improving quality index calculation and quality assessment of watercourse sediments. Our results confirm that water quality index based on diatom metabarcoding and DNA read proportions are directly affected by gene copy number variation, and show the potential value of

integrating CFs into molecular SPI calculation. However, as the biovolume–copy number relationship was based on only 8 diatom species and the efficiency of the resulting CFs validated on only one HTS dataset, further experiments including more species and larger datasets will be required to develop and fully validate CFs for use in molecular biomonitoring.

The CF developed in the present study assumes that gene copy number is constant in each taxon. However, gene copy number may vary with the physiological status of the cell and stage of the life cycle, since in most diatoms cell volume decreases during the vegetative phase. The physiological status varies with cell cycle progression; additionally several factors may affect the physiological status of diatoms like changes in environmental conditions (*e.g.* nutrient availability, pollutants, temperature …) (Pandey *et al.* 2017). Altered physiological status of a given population is generally characterized by a higher proportion of damaged cells. The compromised/damaged cells are characterized by alteration of membrane integrity, degradation of the photosynthetic pigments or fragmentation of genomic DNA (Zetsche & Meysman 2012; Znachor *et al.* 2015). Variations of DNA integrity and chloroplast physiology between cells of a given population can impact directly the *rbc*L gene copy number per cell and thus DNA metabarcoding quantification. (Eberhard, Drapier & Wollman 2002) showed that chloroplast genome copy number is reduced when the green alga *Chlamydomonas reinhardtii* is cultivated under phototrophic conditions compared to cultivation in mixotrophic conditions. Limitation by mineral nutrients may also have an impact; for instance iron limitation can reduce the number of the chloroplast per cell (from 4 to 2) and their size in the marine diatom *Thalassiosira oceanica* (Hustedt) Hasle et Heimdal (Lommer *et al.* 2012). Variation of the cell physiological state was not taken into account in developing CFs for diatom metabarcoding. However, during our experiments we discriminated live and dead cells; we observed that their respective proportions did not affect significantly the correlation between the gene copy number per cell and the cell biovolume (Fig. S1). Further experiments should be performed to evaluate the impact on the final CFs of *rbc*L gene copy number variation linked to physiological status.

The biovolume of each diatom species is required to apply the CF and hence correct the quantification in metabarcoding datasets. Several reference databases provide biovolume information for

a lot diatom species (*e.g.* Rimet *et al.* 2016), but they do not generally account for biovolume variability, which is a complicating factor in diatoms because of the peculiarities of the life cycle. Diatom cell size within a population is not constant due to the method of vegetative reproduction, which leads to a progressive cell size reduction of the population (Crawford 1981), followed by restoration of cell size via a sexual event. For this reason, different cell sizes can be observed in the same diatom population, either in pure cultures of (*e.g.* in the marine diatom *Thalassiosira weissflogii* Grunow: Armbrust & Chisholm 1992) or in environmental populations (*e.g.* the freshwater species *Sellaphora pupula* (Kützing) Mereschk: (Mann, Chepurnov & Droop 1999). However, although the range of cell sizes within a given diatom population may vary by a factor of 2 to 5 in the environment (Hense & Beckmann 2015), natural populations usually have a rather narrow range of sizes and larger cells form a negligible fraction of the population (Mann 2011). Furthermore, the distribution of cell size within environmental populations is often close to being normal (Mann, Chepurnov & Droop 1999; Spaulding *et al.* 2012). The balance between small and big individuals in the same population will therefore limit errors associated with the use of a mean biovolume. Hence, we propose to use the mean of biovolume to calculate CFs; without considering other potential HTS quantification biases, its application to DNA reads of environmental material should allow a good correction of their proportions.

Acknowledgments

The authors declare no conflict of interest. Funding provided by the French National Agency for Water and Aquatic Environments (ONEMA-AFB) and supported by the European COST action DNAqua-Net (CA 15219). A special thanks to David G. Mann for the constructive discussions that helped to improve the manuscript.

Data accessibility

All PGM raw sequence data are available for the 15 libraries, corresponding to the 5 DNA mock communities with 3 replicates, on the Zenodo repository website (http://doi.org/10.5281/zenodo.807178).

Author contributions

V.V., A.B., F.R., S.J., M.C., K.T., I.D contributed to the study designed. V.V., M.C and S.J. conducted the laboratory work. V.V. analyzed the data and wrote the manuscript. All the authors contributed to the discussions and to manuscript editing.

References

Amend, A.S., Seifert, K.A. & Bruns, T.D. (2010). Quantifying microbial communities with 454 pyrosequencing: does read abundance count? *Molecular Ecology*, **19**, 5555–5565.

Angly, F.E., Dennis, P.G., Skarshewski, A., Vanwonterghem, I., Hugenholtz, P. & Tyson, G.W. (2014). CopyRighter: a rapid tool for improving the accuracy of microbial community profiles through lineage-specific gene copy number correction. *Microbiome*, **2**, 11.

Apothéloz-Perret-Gentil, L., Cordonier, A., Straub, F., Iseli, J., Esling, P. & Pawlowski, J. (2017). Taxonomy-free molecular diatom index for high-throughput eDNA biomonitoring. *Molecular ecology resources*, **(in press)**.

Armbrust, E.V. & Chisholm, S.W. (1992). Patterns of cell size change in a marine centric diatom: variability evolving from clonal isolates. *Journal of Phycology*, **28**, 146–156.

Bendich, A.J. (1987). Why do chloroplasts and mitochondria contain so many copies of their genome? *BioEssays*, **6**, 279–282.

Crawford, R.M. (1981). The Siliceous Components of the Diatom Cell Wall and Their Morphological Variation. *Silicon and Siliceous Structures in Biological Systems*, pp. 129–156. Springer New York, New York, NY.

von Dassow, P., Petersen, T.W., Chepurnov, V.A. & Virginia Armbrust, E. (2008). Inter- and Intraspecific relationships between nuclear DNA content and cell size in selected members members of the centric diatom genus Thalassiosira (Bacillariophyceae). *Journal of Phycology*, **44**, 335–349.

Deagle, B.E., Thomas, A.C., Shaffer, A.K., Trites, A.W. & Jarman, S.N. (2013). Quantifying sequence proportions in a DNA-based diet study using Ion Torrent amplicon sequencing: which counts count? *Molecular Ecology Resources*, **13**, 620–633.

Eberhard, S., Drapier, D. & Wollman, F.-A. (2002). Searching limiting steps in the expression of chloroplast-encoded proteins: relations between gene copy number, transcription, transcript abundance and translation rate in the chloroplast of Chlamydomonas reinhardtii. *The Plant Journal*, **31**, 149–160.

Elbrecht, V. & Leese, F. (2015). Can DNA-Based Ecosystem Assessments Quantify Species Abundance? Testing Primer Bias and Biomass—Sequence Relationships with an Innovative Metabarcoding Protocol (M. Hajibabaei, Ed.). *Plos One*, **10**, e0130324.

Elbrecht, V., Peinert, B. & Leese, F. (2017). Sorting things out: Assessing effects of unequal specimen biomass on DNA metabarcoding. *Ecology and Evolution*, **7**, 6918–6926.

Ersland, D.R., Aldrich, J. & Cattolico, R. a. (1981). Kinetic Complexity, Homogeneity, and Copy Number of Chloroplast DNA from the Marine Alga Olisthodiscus luteus. *Plant Phisiology*, **68**, 1468–1473.

European Committee for Standardization (CEN). (2006). EN 15204 - Water quality - Guidance standard on the enumeration of phytoplankton using inverted microscopy (Utermöhl technique). *European Standard*, 1–42.

European Council. (2000). Directive 2000/60/EC of the European Parliament and of the Council of 23 October 2000 Establishing a Framework for Community Action in the Field of Water Policy. *Office for official publications of the European Communities, Brussels*.

Evans, N.T., Olds, B.P., Renshaw, M.A., Turner, C.R., Li, Y., Jerde, C.L., Mahon, A.R., Pfrender, M.E., Lamberti, G.A. & Lodge, D.M. (2016). Quantification of mesocosm fish and amphibian species diversity via environmental DNA metabarcoding. *Molecular Ecology Resources*, **16**, 29–41.

Godhe, A., Asplund, M.E., Härnström, K., Saravanan, V., Tyagi, A. & Karunasagar, I. (2008). Quantification of diatom and dinoflagellate biomasses in coastal marine seawater samples by real-time PCR. *Applied and environmental microbiology*, **74**, 7174–82.

Gruber, A. (2008). *Molecular Characterisation of Diatom Plastids (PhD thesis)*. University of Konstanz.

Gutteridge, S. & Gatenby, A. (1995). Rubisco Synthesis, Assembly, Mechanism, and Regulation. *The Plant Cell Online*, **7**, 809–819.

Hense, I. & Beckmann, A. (2015). A theoretical investigation of the diatom cell size reduction–restitution cycle. *Ecological Modelling*, **317**, 66–82.

Keck, F., Vasselon, V., Tapolczai, K., Rimet, F. & Bouchez, A. (2017). Freshwater biomonitoring in the Information Age. *Frontiers in Ecology and the Environment*, **15**, 266–274.

Kermarrec, L., Franc, A., Rimet, F., Chaumeil, P., Frigerio, J.-M., Humbert, J. & Bouchez, A. (2014). A next-generation sequencing approach to river biomonitoring using benthic diatoms. *Freshwater Science*, **33**, 349–363.

Kermarrec, L., Franc, A., Rimet, F., Chaumeil, P., Humbert, J.F. & Bouchez, A. (2013). Next-generation sequencing to inventory taxonomic diversity in eukaryotic communities: a test for freshwater diatoms. *Molecular Ecology Resources*, **13**, 607–619.

Koop, H.-U., Herz, S., Golds, T.J. & Nickelsen, J. (2007). The genetic transformation of plastids. *The genetic transformation of plastids. In R. Bock (Ed.) Cell and molecular biology of plastids. Topics in current genetics (Vol. 19). Berlin, Heidelberg: Springer*.

Lecointe, C., Coste, M. & Prygiel, J. (1993). 'Omnidia': software for taxonomy, calculation of diatom indices and inventories management. *Hydrobiologia*, **269–270**, 509–513.

Lommer, M., Specht, M., Roy, A.-S., Kraemer, L., Andreson, R., Gutowska, M.A., Wolf, J., Bergner, S. V, Schilhabel, M.B., Klostermeier, U.C., Beiko, R.G., Rosenstiel, P., Hippler, M. & LaRoche, J. (2012). Genome and low-iron response of an oceanic diatom adapted to chronic iron limitation. *Genome biology*, **13**, R66.

Mann, D.G., Chepurnov, V.A. & Droop, S.J.M. (1999). Sexuality, incompatibility, size variation, and preferential polyandry in natural populations and clones of Sellaphora pupula (Bacillariophyceae). *Journal of Phycology*, **35**, 152–170.

Mann, D.G. (2011). Size and Sex. *The Diatom World* (ed E. J Seckbach & JP Kociolek), pp. 145–166. Springer, Dordrecht.

Okie, J.G., Smith, V.H. & Martin-Cereceda, M. (2016). Major evolutionary transitions of life, metabolic scaling and the number and size of mitochondria and chloroplasts. *Proceedings. Biological sciences*, **283**, 20160611.

Pandey, L.K., Bergey, E.A., Lyu, J., Park, J., Choi, S., Lee, H., Depuydt, S., Oh, Y.T., Lee, S.M. & Han, T. (2017).

The use of diatoms in ecotoxicology and bioassessment: Insights, advances and challenges. *Water Research*, **118**, 39–58.

Pawlowski, J., Lejzerowicz, F., Apotheloz-Perret-Gentil, L., Visco, J. & Esling, P. (2016). Protist metabarcoding and environmental biomonitoring: Time for change. *European Journal of Protistology*, **55**, 12–25.

R Development core team. (2013). R: a language and environment for statistical computing. *R Foundation for Statistical Computing, Vienna, Austria.*

Rauwolf, U., Golczyk, H., Greiner, S. & Herrmann, R.G. (2010). Variable amounts of DNA related to the size of chloroplasts III. Biochemical determinations of DNA amounts per organelle. *Molecular Genetics and Genomics*, **283**, 35–47.

Rimet, F. (2012). Recent views on river pollution and diatoms. *Hydrobiologia*, **683**, 1–24.

Rimet, F., Chaumeil, P., Keck, F., Kermarrec, L., Vasselon, V., Kahlert, M., Franc, A. & Bouchez, A. (2016). R-Syst::diatom: an open-access and curated barcode database for diatoms and freshwater monitoring. *Database*, **2016**, baw016.

Rimet, F., Trobajo, R., Mann, D.G., Kermarrec, L., Franc, A., Domaizon, I. & Bouchez, A. (2014). When is Sampling Complete? The Effects of Geographical Range and Marker Choice on Perceived Diversity in Nitzschia palea (Bacillariophyta). *Protist*, **165**, 245–259.

Round, F.E., Crawford, R.M. & Mann, D.G. (1990). *Diatoms: Biology and Morphology of the Genera* (Cambridge University Press, Ed.).

Sabir, J.S.M., Yu, M., Ashworth, M.P., Baeshen, N.A., Baeshen, M.N., Bahieldin, A., Theriot, E.C. & Jansen, R.K. (2014). Conserved gene order and expanded inverted repeats characterize plastid genomes of Thalassiosirales. *Plos One*, **9**, e107854.

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., Sahl, J.W., Stres, B., Thallinger, G.G., Van Horn, D.J. & Weber, C.F. (2009). Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and Environmental Microbiology*, **75**, 7537–7541.

Snoeijs, P., Busse, S. & Potapova, M. (2002). The importance of diatom cell size in community analysis. *Journal of Phycology*, **38**, 265–281.

Spaulding, S. a., Jewson, D.H., Bixby, R.J., Nelson, H. & McKnight, D.M. (2012). Automated measurement of diatom size. *Limnology and Oceanography: Methods*, **10**, 882–890.

Stein, E.D., Martinez, M.C., Stiles, S., Miller, P.E. & Zakharov, E. V. (2014). Is DNA barcoding actually cheaper and faster than traditional morphological methods: results from a survey of freshwater bioassessment efforts in the United States? (M. Casiraghi, Ed.). *Plos One*, **9**, e95525.

Sun, J. & Liu, D. (2003). Geometric models for calculating cell biovolume and surface area for phytoplankton. *Journal of Plankton Research*, **25**, 1331–1346.

Taberlet, P., Coissac, E., Hajibabaei, M. & Rieseberg, L.H. (2012). Environmental DNA. *Molecular Ecology*, **21**, 1789–1793.

Tan, B., Ng, C., Nshimyimana, J.P., Loh, L.L., Gin, K.Y.-H. & Thompson, J.R. (2015). Next-generation sequencing (NGS) for assessment of microbial water quality: current progress, challenges, and future opportunities. *Frontiers in microbiology*, **6**, 1027.

Thomas, A.C., Deagle, B.E., Eveson, J.P., Harsch, C.H. & Trites, A.W. (2016). Quantitative DNA metabarcoding: improved estimates of species proportional biomass using correction factors derived

492 from control material. *Molecular Ecology Resources*, **16**, 714–726.

493 Treusch, A.H., Demir-Hilton, E., Vergin, K.L., Worden, A.Z., Carlson, C.A., Donatz, M.G., Burton, R.M. &
494 Giovannoni, S.J. (2012). Phytoplankton distribution patterns in the northwestern Sargasso Sea
495 revealed by small subunit rRNA genes from plastids. *The ISME Journal*, **6**, 481–492.

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

499 Vasselon, V., Rimet, F., Tapolczai, K. & Bouchez, A. (2017b). Assessing ecological status with diatoms DNA
500 metabarcoding: Scaling-up on a WFD monitoring network (Mayotte island, France). *Ecological*
501 *Indicators*, **82**, 1–12.

502 Visco, J.A., Apothéloz-Perret-Gentil, L., Cordonier, A., Esling, P., Pillet, L. & Pawlowski, J. (2015).
503 Environmental Monitoring: Inferring the Diatom Index from Next-Generation Sequencing Data.
504 *Environmental Science & Technology*, **49**, 7597–7605.

505 Vivien, R., Lejzerowicz, F. & Pawlowski, J. (2016). Next-generation sequencing of aquatic oligochaetes:
506 Comparison of experimental communities. *Plos One*, **11**, 1–14.

507 Weber, A. a-T. & Pawlowski, J. (2013). Can abundance of protists be inferred from sequence data: a case
508 study of foraminifera. *PloS one*, **8**, e56739.

509 Zetsche, E.-M. & Meysman, F.J.R. (2012). Dead or alive? Viability assessment of micro- and mesoplankton.
510 *Journal of Plankton Research*, **34**, 493–509.

511 Zimmermann, J., Glöckner, G., Jahn, R., Enke, N. & Gemeinholzer, B. (2015). Metabarcoding vs.
512 morphological identification to assess diatom diversity in environmental studies. *Molecular Ecology*
513 *Resources*, **15**, 526–542.

514 Znachor, P., Rychtecký, P., Nedoma, J. & Visocká, V. (2015). Factors affecting growth and viability of natural
515 diatom populations in the meso-eutrophic Římov Reservoir (Czech Republic). *Hydrobiologia*, **762**,
516 253–265.

517

518 ## Tables

519 **Table 1** Characteristics of the 8 diatom species selected in the Thonon Culture Collection (TCC) and used in
520 this study.

521 **Table 2** – CF calculated for the 8 diatom species using their respective cell biovolume (Table 1) and the
522 linear equation between the *rbc*L copy number and the cell biovolume (Fig. 3).

523

524 ## Figures

525 **Figure 1** Experimental design applied to the 8 diatom species. After the inoculation of 21 flasks containing
526 40mL of DV media, diatom culture growth was followed at 7 sampling time (from T0 to T6) and analysis
527 was performed in triplicate (3 flasks per sampling time).

528 **Figure 2** Estimation of the *rbc*L copy number per diatom cell for the 8 diatom species. Mean values
529 calculated using the gene and the diatom cell concentrations obtained respectively by qPCR and inverted
530 microscopy at T0, T1 and T2 sampling points (n = 9).

531 **Figure 3** Correlation between the diatom cell biovolume and the *rbc*L gene copy number per cell after
532 log(x+1) transformation.

Figure 4 Relative abundances of the 8 diatom species in the 5 DNA mock communities based (A) on mean of HTS DNA reads without (left) and with (right) correcting quantification using the biovolume correction factor and (B) on mean of morphological counts from inverted microscopy.

Figure 5 Dominant taxa (relative abundance > 0.5 %) obtained in HTS Mayotte molecular inventories without (left) and with (right) application of the biovolume correction factor. All samples (n=80) are considered.

Figure 6 Distribution of the differences between the molecular and the morphological SPI (∆SPI) for all Mayotte samples using original molecular SPI values (left) and new molecular SPI values based on molecular inventories corrected with the biovolume CF (right).

Supporting Information

Table S1 *rbc*L primers, qPCR reactions mix and condition used for the qPCR assays. Information is provided for 1 reaction in a final volume of 25µL.

Table S2 Estimation of the diatom cell concentration and the live/dead cell proportion per mL of media, based on microscopy counts, for the 8 diatom species at each sampling time and for the 3 replicates (A, B, C). Mean values of cell concentration per mL of media, which only take into account living cells, is provided and used for the calculation of *rbc*L copy number per diatom cell (bold values).

Table S3 Estimation of the *rbc*L copy number per mL of media determined by qPCR for the 8 diatom species at each sampling time and for the 3 replicates (A, B, C). Mean values of *rbc*L concentration per mL of media is provided and used for the calculation of *rbc*L copy number per diatom cell (bold values).

Table S4 CF calculated for the 84 diatom taxa detected in Mayotte environmental samples. Calculation performed using the respective cell biovolume of each taxa (available in the Rsyst::diatom library) and the linear equation between the *rbc*L copy number and the cell biovolume produced in the Fig. 3.

Table S5 Number of DNA reads assigned to the 8 species in each of the 5 DNA mock communities. A, B, and C represent the 3 replicates.

Figure S1 Correlation between the diatom cell biovolume and the *rbc*L gene copy number per cell after log(x+1) transformation based on live (black) or live/dead (grey) microscopical counts. Linear equation of the model and the Pearson correlation coefficient (r) with is associated p-value are indicated.

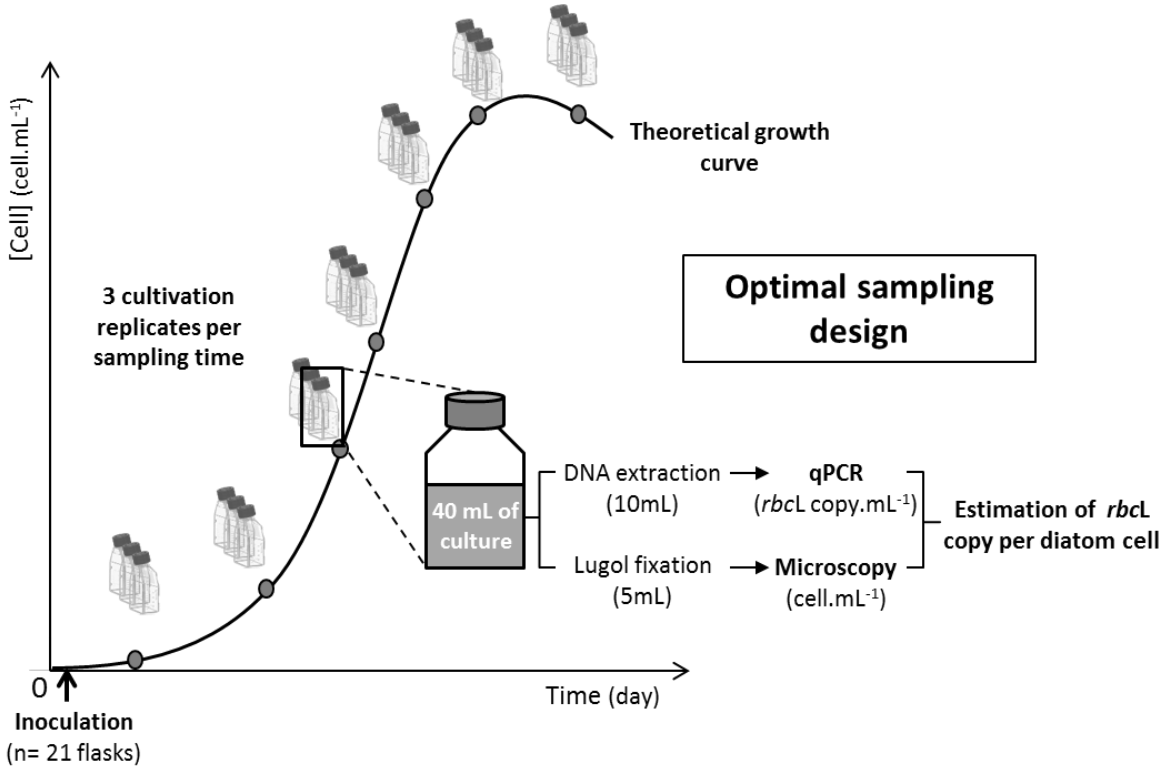
Species	TCC code	Chloroplast nb./cell	Length (µm)	Width (µm)	Thickness (µm)	Biovolume (µm^3)
Achnanthidium minutissimum (Kützing) Czarnecki	TCC667	1	7.1	3.2	2.5	45
Nitzschia palea (Kützing) W.Smith	TCC139-1	2	22.7	4,0	4,0	183
Ulnaria ulna (Nitzsch) Compère	TCC670	2	54.6	7.9	9.5	4087
Pinnularia viridiformis (Nitzsch) Ehrenberg	TCC890	2	51.4	14.3	17.8	10282
Diatoma tenuis Kützing	TCC861	≈ 8	42.4	4.8	4.8	769
Nitzschia inconspicua Grunow	TCC488	2	8.1	4.3	3.6	98
Fragilaria perminuta (Grunow) Lange-Bertalot	TCC753	2	11.1	4.2	3.7	135
Cyclotella meneghiniana Kützing	TCC690	≈ 20	12.1		4.7	539

Table 1 – Characteristics of the 8 diatom species selected in the Thonon Culture Collection (TCC) and used in this study.

Species	Calculated CF
A. minutissimum	0.6
N. inconspicua	1.7
N. palea	3.3
P. viridiformis	78.5
D. tenuis	11.1
F. perminuta	2.4
U. ulna	39.6
C. meneghiniana	8.3

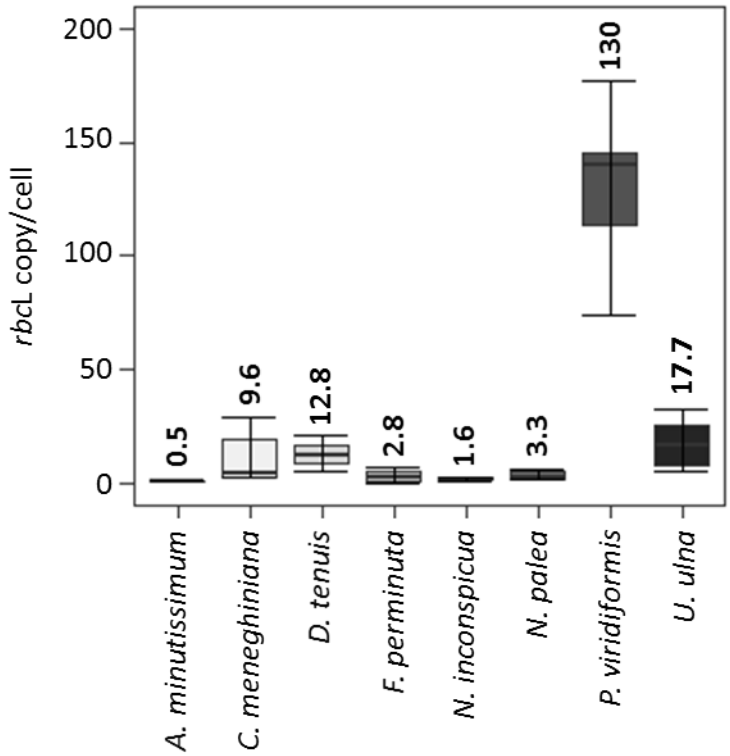
565

566 **Table 2** – CF calculated for the 8 diatom species using their respective cell biovolume (Table 1) and the linear
567 equation between the *rbc*L copy number and the cell biovolume (Fig. 3).



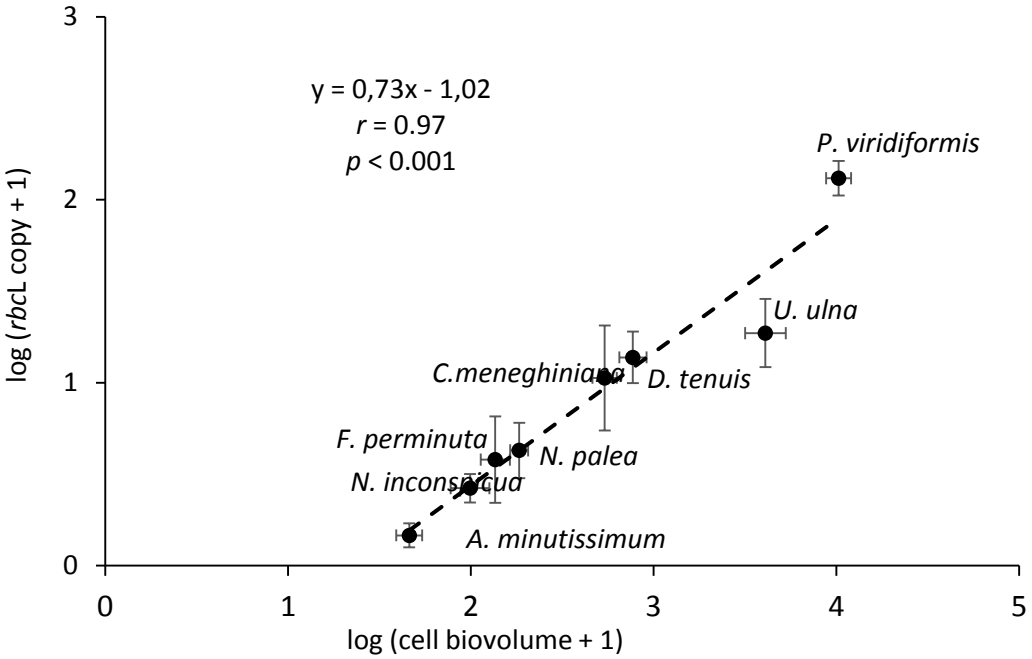
568

569 **Figure 1** – Experimental design applied to the 8 diatom species. After the inoculation of 21 flasks containing 40mL of
570 DV media, diatom culture growth was followed at 7 sampling time (from T0 to T6) and analysis was performed in
571 triplicate (3 flasks per sampling time).



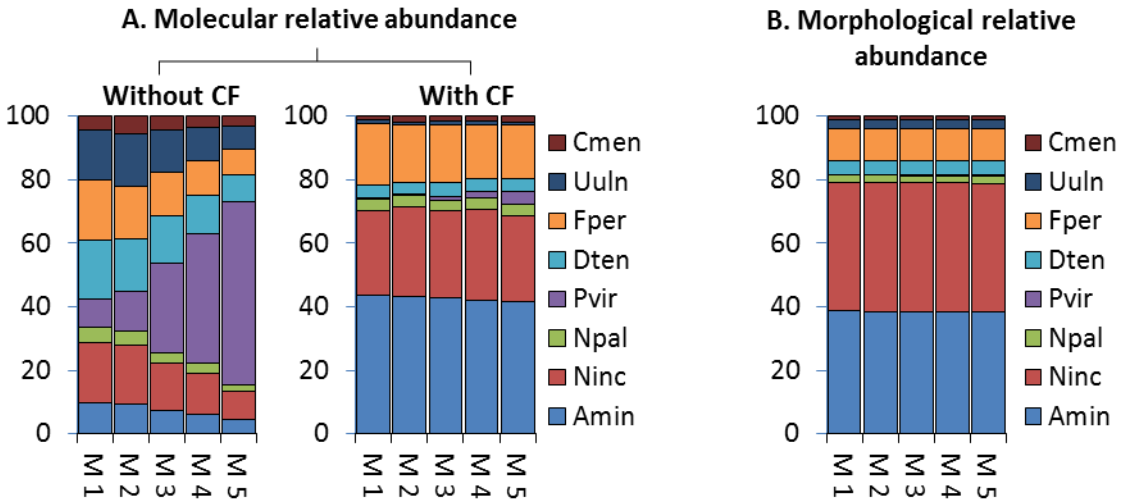
572

573 **Figure 2** – Estimation of the *rbc*L copy number per diatom cell for the 8 diatom species. Mean values calculated using
574 the gene and the diatom cell concentrations obtained respectively by qPCR and inverted microscopy at T0, T1 and T2
575 sampling points (n = 9).



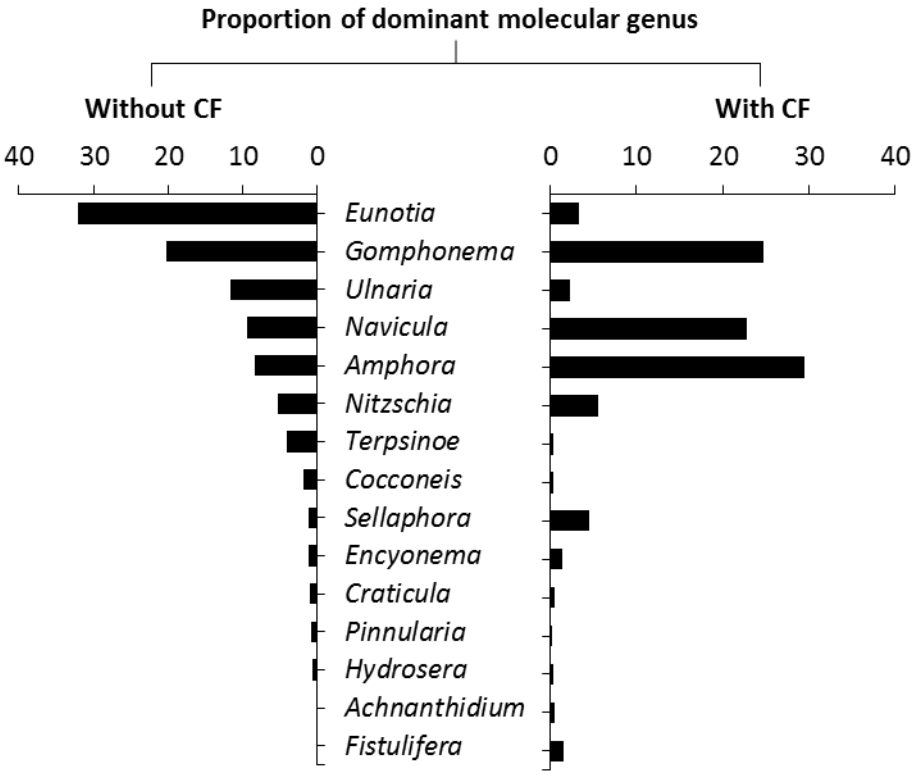
576

577 **Figure 3** – Correlation between the diatom cell biovolume and the *rbc*L gene copy number per cell after log(x+1)
578 transformation.



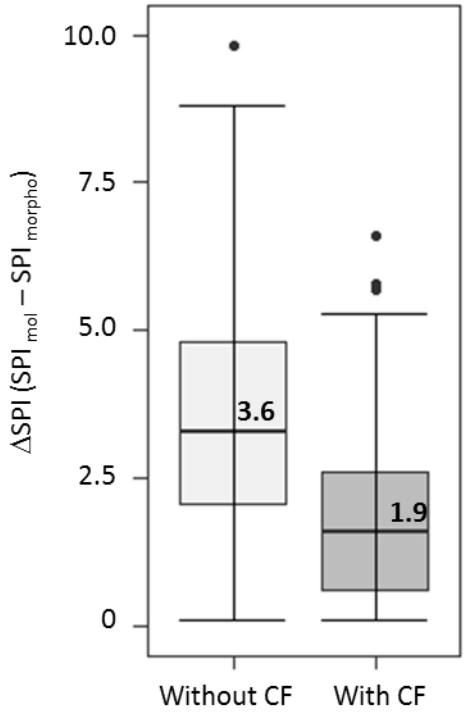
579

580 **Figure 4** – Relative abundances of the 8 diatom species in the 5 DNA mock communities based (A) on mean of HTS
581 DNA reads without (left) and with (right) correcting quantification using the biovolume correction factor and (B) on
582 mean of morphological counts from inverted microscopy.



583

584 **Figure 5** – Dominant taxa (relative abundance > 0.5 %) obtained in HTS Mayotte molecular inventories without (left)
585 and with (right) application of the biovolume correction factor. All samples (n=80) are considered.



586

587 **Figure 6** – Distribution of the differences between the molecular and the morphological SPI (ΔSPI) for all Mayotte
588 samples using original molecular SPI values (left) and new molecular SPI values based on molecular inventories
589 corrected with the biovolume CF (right).

Table S1 – *rbc*L primers, qPCR reactions mix and condition used for the qPCR assays. Information is provided for 1 reaction in a final volume of 25µL.

Primer name		**Primer sequence (5' - 3')**	**Length (bp)**
Forward	Diat_rbcL_708F_1	AGGTGAAGTAAAAGGTTCWTACTTAAA	27
	Diat_rbcL_708F_2	AGGTGAAGTTAAAGGTTCWTAYTTAAA	27
	Diat_rbcL_708F_3	AGGTGAAACTAAAGGTTCWTACTTAAA	27
Reverse	R3_1	CCTTCTAATTTACCWACWACTG	22
	R3_2	CCTTCTAATTTACCWACAACAG	22

Reagents	**Initial conc.**	**Final conc.**	**Volume (µL)**
Sybr MIX	2X	1X	12.5
H_2O molecular grade	-	-	6.75
Forward (Diat_rbcL_708F_1 + _2 + _3)	10 µM	0.5 µM	1.25
Reverse (R3_1 + R3_2)	10 µM	0.5 µM	1.25
Bovine Serum Albumin (BSA)	10 mg/mL	0.5 mg/mL	1.25
DNA	25 ng/µL	2 ng/µL	2

Step	**Time (s)**	**Temperature (°C)**	**Cycles**
1	900	95	
2	45	95	X 40
3	45	55	
4	45	72	
5	1° every 5s	60 to 95	

Table S2 – Estimation of the diatom cell concentration and the live/dead cell proportion per mL of media, based on microscopy counts, for the 8 diatom species at each sampling time and for the 3 replicates (A, B, C). Mean values of cell concentration per mL of media, which only take into account living cells, is provided and used for the calculation of *rbc*L copy number per diatom cell (bold values).

Species	**Sampling time**	**Days after inoculation**	**[cell.mL^{-1}] per replicate**			**% of dead cell**			**Mean (living cells)** (cell.mL-1)
			A	**B**	**C**	**A**	**B**	**C**	
Cmen	T0	5	3.7E+02	4.1E+02	4.2E+02	9.8	3.8	4.0	**3.8E+02**
	T1	10	8.2E+03	6.0E+03	8.5E+03	6.1	14.5	11.9	**6.8E+03**
	T2	13	1.2E+04	1.1E+04	2.0E+04	13.5	16.1	12.0	**1.2E+04**
	T3	20	1.2E+05	5.7E+04	1.3E+05	20.2	13.9	19.6	**8.1E+04**
	T4	25	2.6E+05	4.1E+05	3.2E+05	53.9	51.2	56.3	**1.5E+05**
	T5	31	2.0E+05	2.4E+05	2.3E+05	59.9	50.4	48.2	**1.0E+05**
	T6	38	4.6E+05	5.5E+05	2.6E+05	59.1	55.1	57.1	**1.8E+05**
Npal	T0	5	1.8E+04	2.1E+04	3.9E+04	0.0	0.0	1.0	**2.6E+04**
	T1	10	6.1E+05	4.9E+05	4.1E+05	0.0	0.0	0.0	**5.1E+05**
	T2	13	4.6E+05	4.8E+05	5.2E+05	0.9	1.9	1.0	**4.8E+05**
	T3	17	4.8E+05	3.9E+05	6.2E+05	5.9	4.0	6.7	**4.7E+05**
	T4	25	4.1E+05	4.3E+05	9.4E+05	15.4	9.9	8.1	**5.3E+05**
	T5	34	6.2E+05	7.2E+05	7.0E+05	23.5	30.3	25.0	**5.0E+05**
	T6	40	1.3E+06	1.1E+06	6.4E+05	46.6	38.5	54.1	**5.6E+05**
Uuln	T0	5	8.2E+03	7.9E+03	1.5E+04	3.8	2.8	0.7	**1.0E+04**
	T1	10	1.3E+04	1.2E+04	1.5E+04	5.4	7.8	7.2	**1.2E+04**
	T2	13	1.2E+04	3.4E+04	7.9E+03	14.3	13.6	10.5	**1.5E+04**
	T3	20	1.8E+04	1.6E+04	2.6E+04	27.1	23.8	23.9	**1.5E+04**
	T4	31	1.6E+04	1.1E+04	9.2E+03	83.3	74.4	63.9	**3.0E+03**
	T5	38	8.6E+03	9.2E+03	3.6E+04	82.8	84.8	82.2	**3.1E+03**
Ninc	T0	5	3.9E+03	8.7E+03	5.8E+03	0.5	1.0	0.0	**6.1E+03**
	T1	10	3.5E+05	3.9E+05	4.3E+05	0.0	0.2	0.2	**3.9E+05**
	T2	12	4.3E+05	2.6E+05	1.1E+06	0.6	0.2	0.7	**5.9E+05**
	T3	17	4.1E+05	6.4E+05	1.1E+06	6.9	7.0	5.1	**6.8E+05**
	T4	25	1.7E+06	1.4E+06	9.9E+05	11.6	10.4	12.6	**1.2E+06**
	T5	34	1.6E+06	1.3E+06	1.4E+06	7.2	9.9	6.7	**1.3E+06**
	T6	40	1.3E+06	1.9E+06	1.7E+06	21.4	28.9	30.8	**1.2E+06**
Dten	T0	12	1.2E+04	4.3E+04	2.6E+04	0.2	0.0	0.3	**2.7E+04**
	T1	17	1.1E+05	9.4E+04	1.0E+05	5.7	7.0	5.5	**9.6E+04**
	T2	20	1.8E+05	2.2E+05	1.3E+05	6.9	5.7	5.5	**1.7E+05**
	T3	25	4.9E+05	2.3E+05	1.4E+05	6.3	8.8	8.2	**2.7E+05**
	T4	34	2.7E+05	2.0E+05	2.1E+05	26.5	35.8	43.1	**1.5E+05**
	T5	38	4.1E+05	2.4E+05	1.6E+05	48.3	49.3	45.5	**1.4E+05**
Pvir	T0	13	6.0E+02	4.7E+02	4.1E+02	8.0	7.5	11.7	**4.5E+02**
	T1	20	9.6E+02	7.2E+02	1.1E+03	12.0	9.3	7.3	**8.3E+02**
	T2	31	1.5E+03	1.7E+03	3.1E+03	14.3	17.3	18.5	**1.8E+03**
	T3	34	2.0E+03	2.0E+03	2.4E+03	16.5	23.5	29.6	**1.6E+03**
	T4	40	2.7E+03	2.2E+03	3.6E+03	26.1	22.6	26.7	**2.1E+03**
	T5	73	4.9E+03	2.7E+03	2.6E+03	83.7	75.8	66.8	**7.7E+02**
Fper	T0	12	6.0E+04	3.4E+04	3.3E+04	0.7	0.7	1.3	**4.2E+04**
	T1	17	2.7E+05	1.1E+05	1.7E+05	14.6	12.6	11.6	**1.6E+05**
	T2	20	2.2E+05	1.6E+05	1.2E+05	23.4	24.1	19.7	**1.3E+05**
	T3	25	1.5E+05	1.8E+05	1.6E+05	62.2	65.4	62.3	**6.0E+04**
	T4	31	6.6E+05	3.0E+06	4.4E+05	69.3	73.8	65.5	**3.8E+05**

	T5	34	1.2E+06	3.2E+05	2.6E+05	78.5	74.8	76.8	**1.3E+05**
	T6	40	2.9E+05	5.8E+05	5.4E+05	82.5	75.8	73.5	**1.1E+05**
Amin	T0	12	1.8E+03	6.2E+03	3.7E+03	0.7	1.7	1.0	**3.9E+03**
	T1	17	3.0E+04	7.4E+04	8.4E+04	4.1	3.7	3.0	**6.0E+04**
	T2	25	5.5E+05	4.0E+05	1.4E+06	4.7	7.7	4.6	**7.5E+05**
	T3	31	1.3E+06	1.0E+06	5.2E+05	13.1	13.1	10.2	**8.3E+05**
	T4	34	2.1E+06	2.9E+06	6.7E+05	11.6	10.5	13.8	**1.7E+06**
	T5	38	2.7E+06	1.2E+06	5.6E+05	15.2	11.4	16.9	**1.3E+06**
	T6	40	2.8E+06	2.7E+06	1.7E+06	16.2	11.5	17.5	**2.0E+06**

Table S3 – Estimation of the *rbc*L copy number per mL of media determined by qPCR for the 8 diatom species at each sampling time and for the 3 replicates (A, B, C). Mean values of *rbc*L concentration per mL of media is provided and used for the calculation of *rbc*L copy number per diatom cell (bold values).

Species	Sampling time	Days after inoculation	[*rbc*L] (copy.mL^{-1})			Mean (copy.mL^{-1})
			A	B	C	
Cmen	T0	5	7.1E+03	7.3E+03	1.1E+04	**8.4E+03**
	T1	10	4.8E+04	2.2E+04	1.3E+04	**2.8E+04**
	T2	13	4.6E+04	2.6E+04	2.4E+04	**3.2E+04**
	T3	20	1.2E+05	1.1E+05	1.9E+05	**1.4E+05**
	T4	25	6.1E+05	6.2E+05	7.6E+05	**6.6E+05**
	T5	31	4.3E+05	2.3E+06	5.3E+05	**4.8E+05**
	T6	38	9.4E+05	1.0E+06	7.3E+05	**9.1E+05**
Npal	T0	5	3.8E+04	3.4E+04	7.0E+04	**4.7E+04**
	T1	10	1.3E+06	1.5E+06	9.1E+05	**1.3E+06**
	T2	13	2.5E+06	2.8E+06	2.7E+06	**2.6E+06**
	T3	17	2.5E+06	2.7E+06	2.6E+06	**2.6E+06**
	T4	25	3.3E+06	2.3E+06	3.0E+06	**2.9E+06**
	T5	34	1.7E+06	1.9E+06	2.2E+06	**2.0E+06**
	T6	40	1.1E+06	1.3E+06	8.4E+05	**1.1E+06**
Uuln	T0	5	1.4E+05	2.5E+05	1.6E+05	**1.8E+05**
	T1	10	4.0E+05	3.3E+05	3.1E+05	**3.5E+05**
	T2	13	1.2E+05	1.1E+05	7.5E+04	**1.0E+05**
	T3	20	4.9E+05	1.8E+05	2.6E+05	**3.1E+05**
	T4	31	1.2E+05	1.4E+05	2.2E+05	**1.6E+05**
	T5	38	7.5E+04	5.6E+04	5.7E+04	**6.3E+04**
Ninc	T0	5	1.1E+04	1.3E+04	1.5E+04	**1.3E+04**
	T1	10	3.5E+05	7.2E+05	5.2E+05	**5.3E+05**
	T2	12	8.1E+05	6.3E+05	1.1E+06	**8.3E+05**
	T3	17	9.9E+06	8.6E+06	7.5E+06	**8.7E+06**
	T4	25	4.7E+06	5.1E+06	6.3E+06	**5.4E+06**
	T5	34	7.3E+06	7.8E+06	8.1E+06	**7.7E+06**
	T6	40	4.8E+06	4.5E+06	3.2E+06	**4.2E+06**
Dten	T0	12	4.7E+05	2.1E+05	3.0E+05	**3.3E+05**
	T1	17	1.5E+06	2.0E+06	1.1E+06	**1.6E+06**
	T2	20	7.6E+05	1.4E+06	2.8E+06	**1.7E+06**
	T3	25	1.3E+06	5.0E+05	4.6E+05	**7.5E+05**
	T4	34	4.3E+05	2.3E+05	5.3E+05	**4.0E+05**
	T5	38	3.2E+05	4.5E+05	2.3E+05	**3.4E+05**
Pvir	T0	13	7.9E+04	4.6E+04	7.1E+04	**6.6E+04**
	T1	20	1.2E+05	1.2E+05	1.2E+05	**1.2E+05**
	T2	31	2.0E+05	1.3E+05	2.0E+05	**1.8E+05**
	T3	34	1.6E+05	2.6E+05	3.0E+05	**2.4E+05**
	T4	40	2.6E+05	2.2E+05	3.8E+05	**2.9E+05**
	T5	73	3.1E+05	5.5E+05	4.8E+05	**4.5E+05**
Fper	T0	12	1.4E+04	3.4E+03	9.4E+03	**9.0E+03**
	T1	17	3.0E+05	2.4E+05	3.6E+05	**3.0E+05**
	T2	20	8.3E+05	7.1E+05	9.2E+05	**8.2E+05**
	T3	25	8.1E+05	4.8E+05	1.4E+06	**8.8E+05**
	T4	31	4.4E+05	4.8E+05	4.4E+05	**4.5E+05**
	T5	34	6.7E+05	6.8E+05	1.0E+06	**8.0E+05**

	T6	40	4.0E+05	4.7E+05	4.5E+05	**4.4E+05**
Amin	T0	12	1.8E+03	2.8E+03	3.6E+03	**2.7E+03**
	T1	17	3.4E+04	1.5E+04	2.7E+04	**2.6E+04**
	T2	25	1.9E+05	1.7E+05	2.2E+05	**1.9E+05**
	T3	31	1.2E+05	1.6E+05	1.6E+05	**1.5E+05**
	T4	34	2.6E+05	3.6E+05	3.1E+05	**3.1E+05**
	T5	38	2.8E+05	3.2E+05	2.6E+05	**2.9E+05**
	T6	40	5.1E+05	3.6E+05	2.0E+05	**3.6E+05**

603

Table S4 – CF calculated for the 84 diatom taxa detected in Mayotte environmental samples. Calculation performed using the respective cell biovolume of each taxa (available in the Rsyst::diatom library) and the linear equation between the *rbc*L copy number and the cell biovolume produced in the Fig. 3.

Diatom taxa	Biovolume (µm^3)	Calculated CF
Achnanthes_coarctata	53	0.7
Achnanthidium_helveticum	316	5.3
Achnanthidium_minutissimum	76	1.3
Achnanthidium_sp.	76	1.3
Amphora_pediculus	72	1.2
*Amphora_*sp.	20096	128.3
Caloneis_silicula	1994	23.1
*Caloneis_*sp.	523	8.1
Cocconeis_placentula	2963	31.2
Craticula_cuspidata	2850	30.3
Craticula_molestiformis	119	2.1
*Cyclotella_*sp.	328	5.5
Cymbella_excisa	520	8.1
Cymbella_heterogibbosa	5817	51.5
*Cymbella_*sp.	520	8.1
Cymbopleura_naviculiformis	1148	15.1
Encyonema_minutum	213	3.8
Encyonema_muelleri	12784	92.1
Encyonema_silesiacum	821	11.7
*Encyonema_*sp.	213	3.8
Eolimna_subminuscula	112	2.0
*Epithemia_*sp.	5967	52.5
Eunotia_bilunaris	617	9.3
Eunotia_minor	755	10.9
Eunotia_pectinalis	4219	40.6
*Eunotia_*sp.	15700	107.1
Fallacia_pygmaea	1229	16.0
Fistulifera_saprophila	14	0.03
*Fragilaria_*sp.	294	5.0
Frustulia_vulgaris	1625	19.8
*Frustulia_*sp.	1625	19.8
Gomphonema_acuminatum	1860	21.9
Gomphonema_affine	926	12.8
Gomphonema_bourbonense	270	4.6
Gomphonema_clevei	484	7.6
Gomphonema_parvulum	331	5.5
*Gomphonema_*sp.	510	8.0
Halamphora_montana	161	2.9
*Halamphora_*sp.	161	2.9
*Hydrosera_*sp.	500	7.8
Lemnicola_hungarica	436	7.0
Luticola_sparsipunctata	176	3.1
Mayamaea_permitis	66	1.0
Navicula_cryptocephala	431	6.9
Navicula_cryptotenella	386	6.3
Navicula_lanceolata	1227	15.9
Navicula_radiosa	1852	21.9
Navicula_rostellata	854	12.0
*Navicula_*sp.	88	1.5
Navicula_symmetrica	818	11.6
Navicula_tripunctata	966	13.2
Navicula_veneta	279	4.8

*Neidium*_sp.	240	4.2
Nitzschia_amphibia	334	5.6
Nitzschia_filiformis	737	10.7
Nitzschia_fonticola	344	5.7
Nitzschia_inconspicua	89	1.5
Nitzschia_lorenziana	1362	17.3
Nitzschia_palea	391	6.4
*Nitzschia*_sp.	307	5.2
Nitzschia_tubicola	336	5.6
Pinnularia_divergens	3908	38.3
Pinnularia_subanglica	1188	15.6
Pinnularia_subgibba	3454	35.0
*Pinnularia*_sp.	1258	16.3
Placoneis_clementis	1123	14.9
Placoneis_elginensis	1266	16.3
*Planothidium*_sp.	267	4.6
Rhopalodia_gibba	185472	649.8
*Rhopalodia*_sp.	185472	649.8
Sellaphora_minima	88	1.5
Sellaphora_pupula	1183	15.5
Sellaphora_seminulum	69	1.1
*Sellaphora*_sp.	88	1.5
Seminavis_robusta	5308	48.1
Staurosira_elliptica	29	0.1
*Staurosira*_sp.	315	5.3
Stephanodiscus_hantzschii	670	9.9
*Surirella*_sp.	1034	14.0
Tabellaria_flocculosa	500	7.8
Terpsinoe_musica	10563	80.0
*Tryblionella*_sp.	655	9.7
Ulnaria_ulna	4724	44.1
*Ulnaria*_sp.	5260	47.8

607

Table S5 – Number of DNA reads assigned to the 8 species in each of the 5 DNA mock communities. A, B, and C represent the 3 replicates.

	Mock 1			**Mock 2**			**Mock 3**			**Mock 4**			**Mock 5**		
Species	*A*	*B*	*C*	*A*	*B*	*C*	*A*	*B*	*C*	*A*	*B*	*C*	*A*	*B*	*C*
A. minutissimum	2828	1934	2410	1785	2129	2109	1837	1900	1882	2025	1342	1683	1202	1273	1332
N. inconspicua	5480	3484	4648	3673	4533	4083	3777	3622	3824	3920	3074	3741	2462	2588	2571
N. palea	1452	1059	1126	912	850	1037	718	896	904	899	715	888	695	567	634
P. viridiformis	2573	1966	2066	2372	2823	2999	6440	7861	7461	11586	10430	11722	18424	16703	14159
D. tenuis	5311	3423	4552	3286	4461	3172	4578	3377	3376	4013	2679	3442	2206	2861	2522
F. perminuta	5817	3796	4452	3484	3844	3549	3492	3569	3341	3427	2449	3083	2117	2318	2226
U.ulna	4486	3037	3863	3303	3893	3561	3259	3343	3449	3321	2412	2897	2395	2053	1992
C. meneghiniana	1360	844	984	1202	1344	1204	1129	1113	1235	1137	807	1126	994	869	779

Figure S1 – Correlation between the diatom cell biovolume and the *rbc*L gene copy number per cell after log(x+1) transformation based on live (black) or live/dead (grey) microscopical counts. Linear equation of the model and the Pearson correlation coefficient (r) with is associated p-value are indicated.