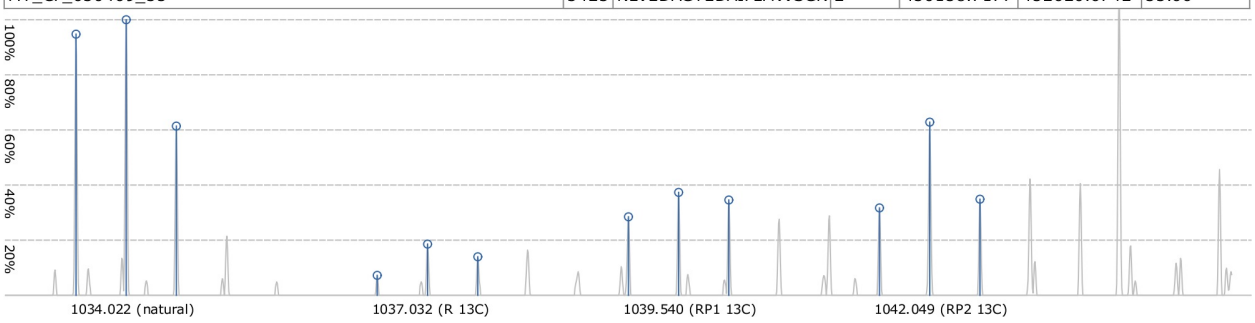
Supplementary notes: qTrace algorithmic details

In addition to the outline of the qTrace algorithm provided in the main text, a detailed explanation of label definitions including variable labels and an exemplary comparison of the *fixed peak count* and the *isotope envelope fitting* modes follow.

Variable labels

In the following example, the effect of defining a variable label is shown. A fixed number of 3 required peaks per isotope envelope has been chosen, and the label has been defined as `RP* 13C`, indicating a ^{13}C label in every arginine residue, and also variably in every proline residue, as indicated by the star symbol.

Filename	Scan	Peptide	Charge	Amount light	Amount heavy	Retention time
MT_CP_030409_35	3423	NLVEDAGTEDAIPLPNVSGR	2	430158.7177	452620.6742	33.66



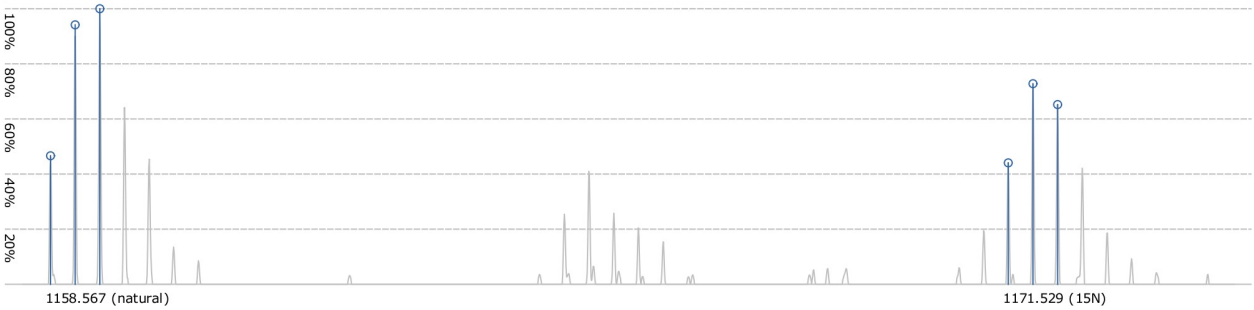
The quantified peptide NLVEDAGTEDAIPLPNVSGR contains 2 proline residues, thus resulting in 2 additional heavy isotope envelopes for 1 or 2 labeled proline residues respectively. It can be seen from the example figure that the heavy sister peptide amount would be dramatically underestimated if the additional heavy proline isotope envelopes would not have been taken into account.

Isotope envelope fitting

As an alternative to the *fixed peak count* mode, qTrace provides the *isotope envelope fitting* mode, which obviates the need to define a somewhat arbitrary number of isotope peaks per isotope envelope. In

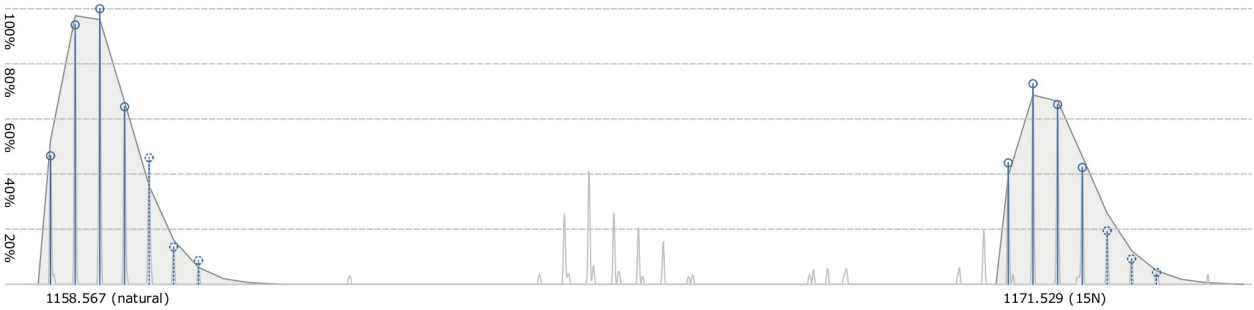
this mode, the actual number of isotope peaks depends on the theoretical isotope envelope of a target peptide and is defined indirectly by the relative peak height in the theoretical isotope envelope. The following figure shows an example sister peptide pair as quantified with a fixed peak count of 3.

Filename	Scan	Peptide	Charge	Amount light	Amount heavy	Retention time
RH_chlmh_20091105_39	5059	ANGGPLGEGLDPLHPGGAFDPLGLADDPDTFAELK	3	930744.9751	703786.1248	45.29



It can be seen that some peaks are not taken into consideration, although they obviously stem from the same peptide and should contribute to the total amount. For the following figure, isotope envelope fitting has been used with a *required peak abundance* of 40% and a *considered peak abundance* of 1%. The amounts are estimated by determining the area under the fitted theoretical isotope envelope, indicated by grey shading.

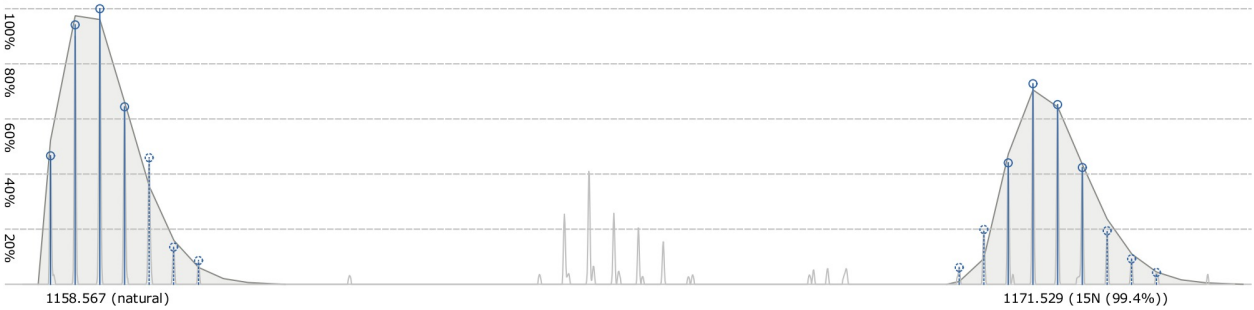
Filename	Scan	Peptide	Charge	Amount light	Amount heavy	Retention time
RH_chlmh_20091105_39	5059	ANGGPLGEGLDPLHPGGAFDPLGLADDPDTFAELK	3	1438828.6791	1029314.6113	45.29



Although the fitted isotope envelopes seem to match quite well, two additional peaks are visible on the left hand side of the heavy isotope envelope, indicating that the ^{15}N label was not completely incorporated and a small amount of ^{14}N atoms is still present in the heavy sample. Specifying a labeling efficiency of 99.4% leads to a slightly shifted isotope envelope which captures the aforementioned two

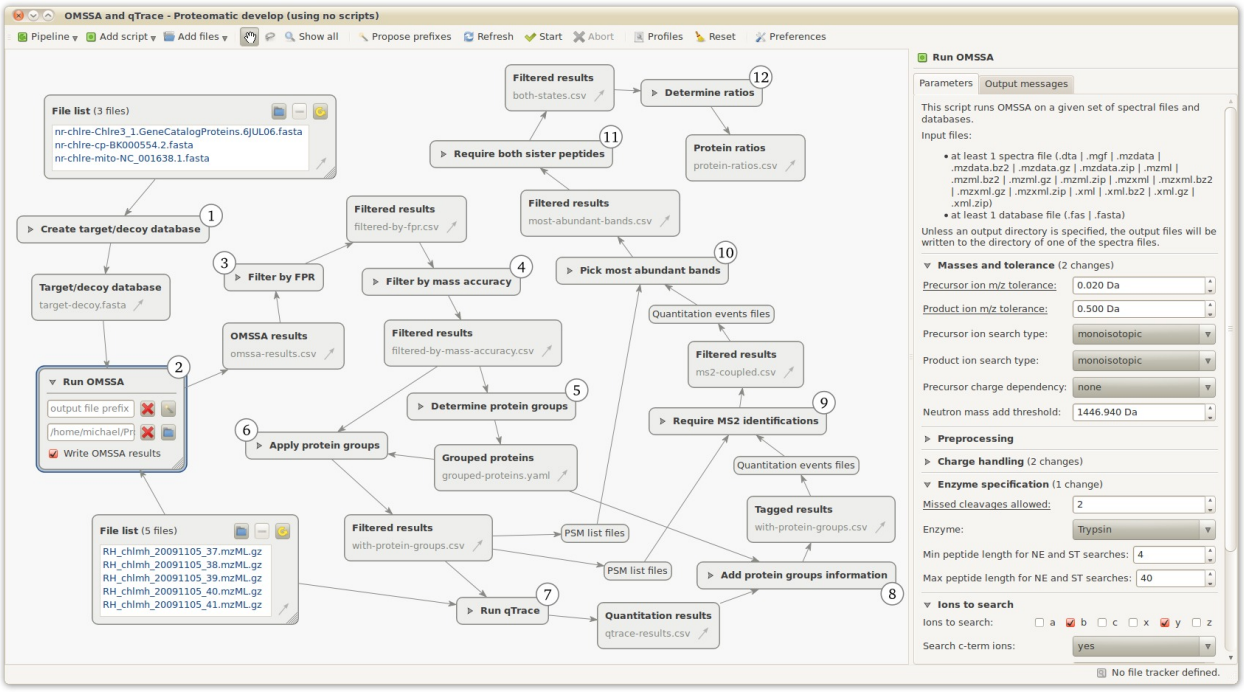
peaks, as demonstrated in the following figure.

Filename	Scan	Peptide	Charge	Amount light	Amount heavy	Retention time
RH_chlmh_20091105_39	5059	ANGGPLGEGLDPLHPGGAFDPLGLADDPDTFAELK	3	1438828.6791	1072173.9272	45.29



Supplementary Figure S1: A Proteomatic pipeline for protein identification and quantitation explained in detail

In order to clarify the processing pipeline depicted in Fig. 3, a detailed description of the individual steps follows, in topological order of the processing steps. Please note that the workflow can be easily changed by re-arranging arrows, adding new processing steps, or adjusting the parameters in the right hand pane of the window. The pipeline was set up for a mass spectrometric experiment that dealt with a sample containing differentially labeled $^{15}N/^{14}N$ mixed iron-sufficient and -deficient *C. reinhardtii* chloroplasts that were separated by one-dimensional SDS-PAGE. Individual protein bands were excised, digested with trypsin and analyzed by Nano-LC-coupled mass spectrometry as described[4].



1. **Create target/decoy database.** A target/decoy database is assembled from the *C. reinhardtii* gene models deduced from nuclear, chloroplast, and mitochondrial genome databases[2], with decoys being created by reversing the input sequences.

2. **Run OMSSA.** OMSSA is used to perform an *in silico* tryptical digestion of the input target/decoy

protein database and match the resulting tryptic peptides to the input fragmentation scans.

3. **Filter by FPR.** The resulting *true positive* (targets) and *false positive* (decoys) peptide/spectral matches (PSM) are sorted by score and an adaptive score threshold is determined in such a way that the estimated FPR is not higher than 1%, according to the formula $FPR = \frac{2 \cdot decoys}{targets + decoys}$. Matches to peptides from decoy proteins are discarded.

4. **Filter by mass accuracy.** The remaining PSM are filtered by a 5 ppm precursor mass accuracy filter, compensating for the fact that OMSSA only handles absolute mass tolerances.

5. **Determine protein groups.** Because peptides that appear in multiple proteins cannot be unambiguously assigned, the affected proteins are grouped as far as possible in this step[3]: two proteins A and B are joined into a protein group if the identified peptides of protein A are a subset of the identified peptides of protein B. This way, protein families sharing equal peptides may be identified and later quantified as a protein group.

6. **Apply protein groups.** The protein groups are applied to the filtered OMSSA results by replacing the protein entry of every PSM by the appropriate protein group.

7. **Run qTrace.** The peptides identified by OMSSA and the spectral files are passed to qTrace, which searches for unlabeled and ^{15}N labeled sister peptides in MS1 full scans, as depicted in Fig. 4. At this time, the resulting quantitation events (QE) are available on the peptide level only.

8. **Add protein groups information.** The quantitation events are shifted to the protein level by adding the appropriate protein group for every quantified peptide, discarding quantitation events from peptides that appear in multiple protein groups.

9. **Require MS2 identifications.** In order to increase the confidence that the observed isotope envelopes in the MS1 full scans are really the target peptides, an MS2 fragment scan identification is required within a retention time difference of no more than 1 minute. The necessary information

is taken from the final filtered OMSSA result file.

Note: The input boxes labeled *PSM list files* and *Quantitation events files* are necessary in the pipeline to resolve filename extension ambiguities. Whenever possible, Proteomatic determines the correct files for an input file group by their filename extension (e. g. *Run OMSSA* requires two input file groups *spectra* and *protein databases*, and input files can be distinguished automatically because the filename extensions differ: `.fasta` for protein databases, `.mzML` etc. for spectra files). However, because both *Run OMSSA* and *Run qTrace* use the CSV format for output files, the user must tell Proteomatic which files should be assigned to which input group by connecting the appropriate files to the input file group boxes instead of the script box.

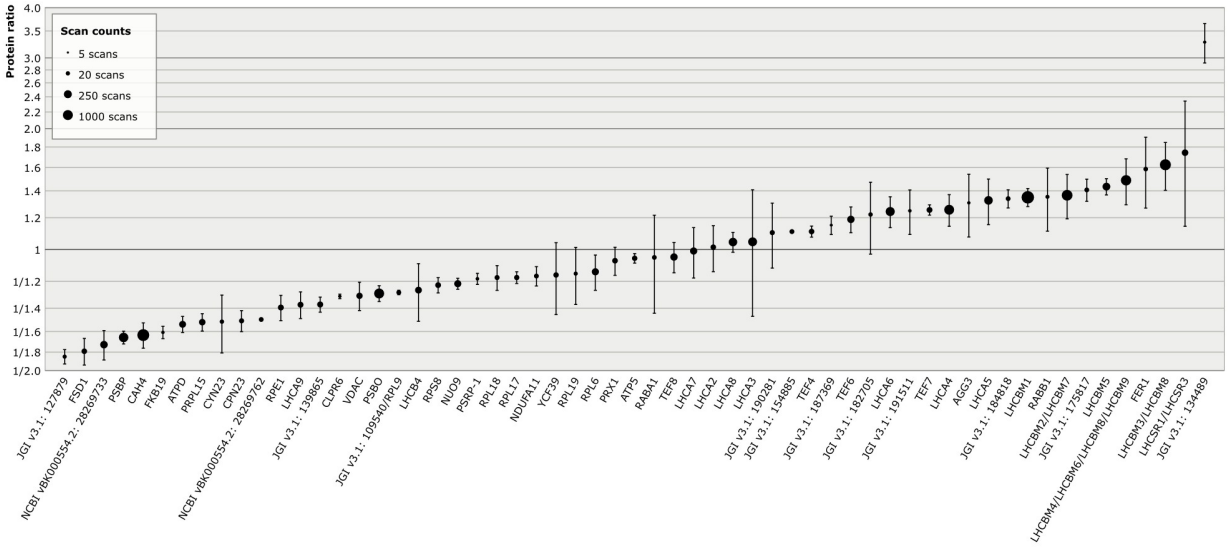
10. **Pick most abundant SDS-PAGE bands.** To further reduce the risk of falsely assigned quantitation events, the OMSSA identification results are used to determine the SDS-PAGE band where a protein (or protein group) was most abundantly identified in, and only quantitation events from this band (± 1 band) are retained.

11. **Require both sister peptides.** As a final filtering step, all quantitation events in which only one of the two sister peptides is observed are discarded.

12. **Determine ratios.** Until here, only abundances have been estimated for both light and heavy sister peptides, and no ratios have been calculated yet. Following the observation that peptides are generally visible in multiple subsequent MS1 full scans while they elute from the HPLC, we group the quantitation events of any peptide by its combination of charge state and SDS-PAGE band, thereby capturing the elution process of any peptide. For each of these groups, all light and heavy abundances are added and divided to yield a ratio per peptide/band/charge (PBC) combination. The final protein (or protein group) ratio is obtained from calculating the mean and standard deviation of the individual PBC combination ratios.

Supplementary Figure S2: qTrace metabolic labeling demonstration

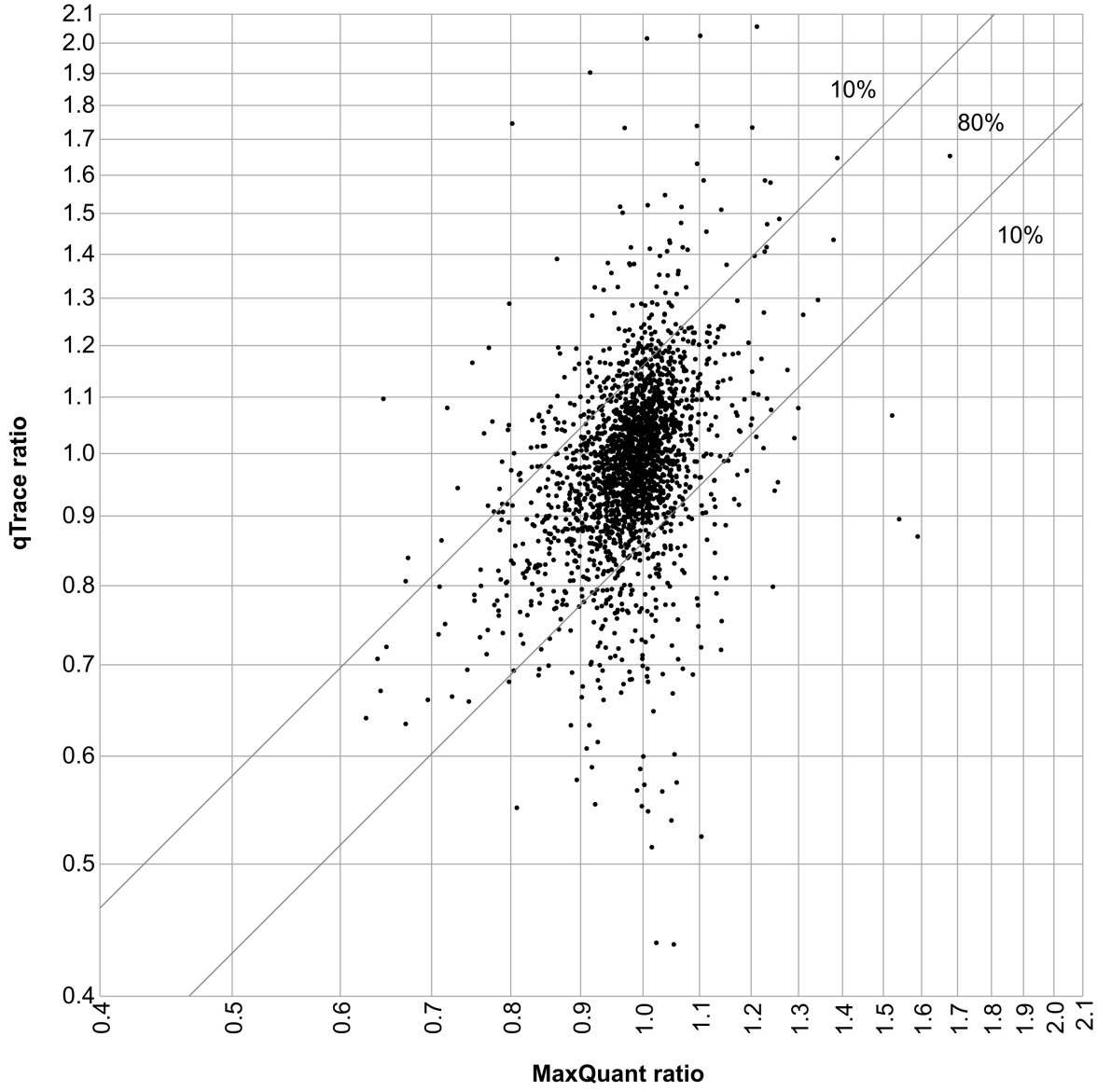
In order to assess the performance of qTrace, isolated chloroplast proteins from *Chlamydomonas reinhardtii* cultures grown under iron-deficient and iron-sufficient conditions were mixed and subsequently separated via SDS-PAGE. A selection of five consecutive SDS-PAGE bands known to contain light-harvesting proteins were excised and tryptically digested and the resulting peptides further separated by an HPLC which was coupled on-line to a Thermo Scientific LTQ Orbitrap mass spectrometer, measuring each band for one hour, yielding a total of $\sim 11,500$ full scans and $\sim 18,000$ fragmentation scans. OMSSA was used to identify a total of 2,754 distinct peptides at an estimated FPR of 1% and at a precursor mass accuracy of 5 ppm. The identified peptides were subsequently passed to qTrace, which was used to search for the individual unlabeled and ^{15}N labeled isotope envelopes in all full scans. The resulting quantitation events were further filtered as described in **Supplementary Figure 1**. After adding protein information and determining protein ratios, we were able to quantify a total of 59 proteins or protein groups, including many light harvesting proteins.



Supplementary Figure S3: qTrace performance assessment

In order to assess the performance of qTrace, we have conducted a comparison between the well-established protein quantitation program MaxQuant[1] and qTrace. In the MaxQuant paper, a set of 24 measurements with three repeats was used for protein identification (using Mascot) and quantitation (using MaxQuant), leading to a total of $\sim 4,100$ proteins quantified, $\sim 3,800$ of which could be quantified with non-zero amounts in both light and heavy states. We performed the comparison based on these results only because quantitation results of zero or infinity must be considered less reliable due to the fact that one of the sister peptides might have escaped detection resulting from an unrecognized post-translational modification.

Using the same HeLa sample spectral files as in the MaxQuant paper (which are publicly available via Tranche) and the same IPI human protein database (version 3.48), we were able to quantify a total of $\sim 2,200$ proteins (also with non-zero amounts for both light and heavy states). The fact that we could only identify 58% of the proteins in comparison to the MaxQuant example likely results from the use of different search engines (Mascot vs. OMSSA) and, most of all, differences in the subsequent handling of peptide identification results. Regardless of the number of quantified proteins, the resulting protein ratios are consistent: 80% of the results are within an acceptable difference range in which, for example, a MaxQuant ratio of 1.0 corresponds to a qTrace ratio of about 0.85 to 1.15, as shown in the comparison figure.


qTrace ratio
MaxQuant ratio
10%
80%
10%
2.1
2.0
1.9
1.8
1.7
1.6
1.5
1.4
1.3
1.2
1.1
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.4
0.5
0.6
0.7
0.8
0.9
1.0
1.1
1.2
1.3
1.4
1.5
1.6
1.7
1.8
1.9
2.0
2.1

Supplemental references

[1] Cox, J. and Mann, M. (2008). MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat Biotechnol*, **26**(12), 1367–1372. PMID: 19029910.

[2] Merchant, S. S., Prochnik, S. E., Vallon, O., Harris, E. H., Karpowicz, S. J., Witman, G. B., Terry, A., Salamov, A., Fritz-Laylin, L. K., Maréchal-Drouard, L., Marshall, W. F., Qu, L., Nelson, D. R., Sanderfoot, A. A., Spalding, M. H., Kapitonov, V. V., Ren, Q., Ferris, P., Lindquist, E., Shapiro, H., Lucas, S. M., Grimwood, J., Schmutz, J., Cardol, P., Cerutti, H., Chanfreau, G., Chen, C., Cognat, V., Croft, M. T., Dent, R., Dutcher, S., Fernández, E., Fukuzawa, H., González-Ballester, D., González-Halphen, D., Hallmann, A., Hanikenne, M., Hippler, M., Inwood, W., Jabbari, K., Kalanon, M., Kuras, R., Lefebvre, P. A., Lemaire, S. D., Lobanov, A. V., Lohr, M., Manuell, A., Meier, I., Mets, L., Mittag, M., Mittelmeier, T., Moroney, J. V., Moseley, J., Napoli, C., Nedelcu, A. M., Niyogi, K., Novoselov, S. V., Paulsen, I. T., Pazour, G., Purton, S., Ral, J., Riaño-Pachón, D. M., Riekhof, W., Rymarquis, L., Schroda, M., Stern, D., Umen, J., Willows, R., Wilson, N., Zimmer, S. L., Allmer, J., Balk, J., Bisova, K., Chen, C., Elias, M., Gendler, K., Hauser, C., Lamb, M. R., Ledford, H., Long, J. C., Minagawa, J., Page, M. D., Pan, J., Pootakham, W., Roje, S., Rose, A., Stahlberg, E., Terauchi, A. M., Yang, P., Ball, S., Bowler, C., Dieckmann, C. L., Gladyshev, V. N., Green, P., Jorgensen, R., Mayfield, S., Mueller-Roeber, B., Rajamani, S., Sayre, R. T., Brokstein, P., Dubchak, I., Goodstein, D., Hornick, L., Huang, Y. W., Jhaveri, J., Luo, Y., Martínez, D., Ngau, W. C. A., Otillar, B., Poliakov, A., Porter, A., Szajkowski, L., Werner, G., Zhou, K., Grigoriev, I. V., Rokhsar, D. S., and Grossman, A. R. (2007). The chlamydomonas genome reveals the evolution of key animal and plant functions. *Science (New York, N.Y.)*, **318**(5848), 245–250. PMID: 17932292.

[3] Nesvizhskii, A. I. and Aebersold, R. (2005). Interpretation of shotgun proteomic data: the protein inference problem. *Mol Cell Proteomics*, **4**(10), 1419–1440. PMID: 16009968.

[4] Terashima, M., Specht, M., Naumann, B., and Hippler, M. (2010). Characterizing the anaerobic response of chlamydomonas reinhardtii by quantitative proteomics. *Mol Cell Proteomics*, **9**(7), 1514–1532.