Refining transcriptome gene catalogs by MS-validation of expressed proteins

Sirius P.K Tse[1], Mathieu Beauchemin[2], David Morse[2] and Samuel C.L. Lo[1]

[1] Shenzhen Key Laboratory of Food Biological Safety Control, AND
Department of Applied Biology and Chemical Technology, The Hong Kong Polytechnic University

[2] Institut de Recherche en biologie Végétale, Département de Sciences Biologiques, Université de Montréal

KEYWORDS
MS-sequencing, Proteomics, Transcriptome, Dinoflagellate

RUNNING TITLE
MS validated protein sequence databases

ABSTRACT

Protein sequencing by tandem mass spectroscopy (LC-MS/MS) identifies thousands of protein sequences even in complex mixtures, and provides valuable insight into the biological functions of different cells. For non-model organisms, transcriptomes are generally used to allow peptide identification, an important addition to their use as a gene catalog allowing the potential metabolic activities of cells to be determined. Here, we used LC-MS/MS data to identify which of the six possible reading frames in the transcriptome was actually used by the cell to make protein, and asked whether this would have an impact on downstream analyses using the transcriptome. We first compiled a list of 6628 translated nucleic acid sequences that contained the peptide matches to a 74,655-sequence transcriptome from the dinoflagellate *Lingulodinium polyedra*. When compared with BLASTx analyses of the DNA sequences, the MS-validated protein sequences analysed BLASTp showed differences in gene ontology, had more identified BLAST hits and contained more KEGG pathway enzymes. The MS-validated protein sequences also differ from datasets containing longest ORF protein sequences. We also note a poor correlation between the levels of protein and mRNA abundance, a comparison not previously performed for dinoflagellates. We suggest use of MS-validated protein sequences instead of the DNA sequence directly may provide a more accurate representation of cellular capacity.

INTRODUCTION

Recent advances in high throughput Mass Spectroscopy-based protein sequencing have allowed an unprecedented examination of the biochemical potential of numerous organisms and are particularly useful for non-model organisms (1). The template sequences used to identify the peptides can be derived from the genome, in which case information on gene order, pseudogenes and regulatory elements is found in addition to the gene complement. Template sequences can also be derived from the transcriptome, which provides information on the types of genes that are present as well as their expression levels under defined conditions. As with other eukaryotes, only a small fraction of the genome is transcribed in dinoflagellates (2), typically on the order of several percent. Transcriptome sequences are thus easier to analyse and provide a rapid as well as a cost-effective means to explore the metabolic potential of cells. A transcriptome can also provide insight into what reactions a cell is able to catalyse by determining the best BLAST hit (the most similar sequence) for each translated sequence in the transcriptome. These types of results are conveniently summarized by categorizing the sequences identified by gene ontology (GO) (3).

Identification of proteins in a transcriptome by BLAST searches, developed in the 1990s (4), is still the accepted standard for sequence characterisation. However, some potential confounding aspects can be readily imagined for transcriptome assemblies. First, since transcriptome assemblies often have difficulty in completely assembling a given transcript, a given gene may be spread across several entries in the transcriptome. Clearly, when a transcriptome sequence is incomplete, identification of a particular functional domain in the sequence will not provide a complete portrait of its true functional role. Second, when transcriptomes are prepared without regard for strand specificity, there is no means of distinguishing which of the one six possible reading frames constitutes that actually used. Lastly, assembly errors can create chimeric sequences whose deduced functions may be erroneously assigned because of the presence of inappropriate protein domains.

Gene catalogs, derived from either genomic DNA or transcripts, are essential for bioinformatic interpretation of the mass spectrums obtained during protein sequencing. Top-end tandem mass

spectrometers (MS/MS) coupled with a liquid chromatographic column can now identify several thousands of peptide sequences in a single sample (5, 6). The bottom-up sequencing method involves digestion of a protein sample with an endopeptidase (usually trypsin), separation of the digested peptides by liquid chromatography (LC), a determination in the first MS of the mass of the each of the peptides that is separated, and finally a fragmentation of the peptide in the second MS and a determination of all the masses in the ladder-like pattern of fragments. Two successive mass peaks in each ladder differ by one amino acid, so in principal the sequence of amino acids which initially defined the peptide could be simply read off. In practise, incomplete and non-random fragmentation of the original peptide means certain fragment peaks are of low abundance, so the sequence of a peptide is identified by comparing the experimental pattern of peaks in the ladder with a virtual peak ladder produced by computer from every sequence in a genome or a transcriptome. Computational methods also exist to assess the intensity of peaks in the MS^1 mass spectrum whose calculated carbon isotope ratio and MS^2 spectrum peaks agree with that predicted for a given peptide in the database. These intensity values can then be used to estimate amounts of a protein (7, 8).

Despite the importance of transcriptomes in estimating the functional characteristics of cells, few studies have examined the consequences of using experimentally determined peptide sequences to refine the transcriptome sequences. We have sequenced protein extracts from the marine dinoflagellate *Lingulodinium polyedra*, a non-model organism for which a transcriptome (9) but no genome sequence is available, and extracted a dataset containing all the nucleic acid sequences that contained one or more peptide sequences. These nucleic acids sequences were translated and the reading frames encoding the MS-derived peptides were used to obtain what we term an MS-validated protein dataset. We find these proteins sequences differ markedly from those obtained by simply translating the longest ORF. We also find significantly differences in some Gene Ontology categories when nucleic acid and protein sequence lists were compared. We suggest that the interpretation of transcriptomes in non-model organisms could be enhanced by MS-validated sequences. The protein sequences also provide considerable time saving when used to identify peptide mass spectra.

METHODS

Cell Culture and proteomic analyses

Culture growth conditions and protein extraction methods for MS analysis from the dinoflagellate *Lingulodinium polyedra* (strain 1936 from the National Center for Marine Algae, East Boothbay, Maine) have been reported previously (10). Briefly, cells were harvested by filtration on Whatman 541 paper, and the filtered cells resuspended in extraction buffer (25 mM MES pH 6, 1M NaCl, 0.25% CHAPS with added protease and phosphatase inhibitors, Qiagen phosphoprotein preparation kit). After breaking the cells with two minutes vigorous shaking in a beadbeater (BioSpec Products), the extract was clarified by centrifugation at 13,000xg for 10 minutes at 4°C, and the protein precipitated in 80% acetone at -20°C overnight. Protein was recovered by centrifugation at 13,000xg for 30 minutes in the cold, and the pellets washed twice with 80% cold acetone and left to air dry for 15 min. The cell pellets were resuspended in lysis buffer (6M Urea, 50 mM DTT, 10 mM Tris) to a final concentration of roughly 10 mg/ml, and the protein concentration measured using the Bradford reagent. For each sample to be analyzed, 100 µg protein in a final volume of 10 µL was reduced by incubation at 60°C for 45 minutes after addition of 20 µL 1.5 mg/ml fresh DTT, then alkylated by incubation in the dark at room temperature for 30 min after addition of 20 µL 10 mg/ml fresh IAA. The proteins were again precipitated overnight at -20°C by addition of 250 µl cold acetone. Protein was recovered by centrifugation as above, and the air-dried pellet resuspended in 10 µL lysis buffer. The sample was then diluted with 200 µL 25 mM NH_4HCO_3 and digested overnight with 5 µg trypsin. Peptides were purified from the mixture after acidification to pH < 4 by addition of 5% TCA using a C18 ZipTip (Millipore) and by following the manufacturer's instructions. Samples were dried in a Speedvac, resuspended in 20 µL 0.1% FA, and transferred to an HPLC vial for injection into the MS.

For ion exchange fractionation of peptides, 100 µg protein was digested with trypsin as above. However, instead of using a ZipTip to isolate peptides from the digest, the sample was diluted to less than 0.5 M urea with 25 mM NH_4HCO_3 and loaded onto an strong cation exchanger (SCX, Millipore). The flow-though was collected as a 25 mM NH_4HCO_3 fraction, and five additional fractions (50 mM, 100 mM, 150 mM, 200 mM and 400 mM NH_4HCO_3) were also collected. All

fractions were dried in a Speedvac and resuspended in 5% TFA ensuring a pH < 4 before purifying the peptides using ZipTips as above.

For SDS-PAGE fractionation, 100 µg of protein was loaded and run on a 12% SDS PAGE (11). The gel was lightly stained with Coomassie blue, and the gel cut into 12 slices each containing roughly similar amount of stain. The gel slices were chopped with a razor blade and washed with 400 µl 25mM and 1:1 solution 50mM ammonium bicarbonate:ACN for at least 4 times until the gel cubes became colorless. The gel cubes were dehydrated (by addition of 200 µl of ACN for 10 minutes), rehydrated using 200 µL of 1:1 solution and dehydrated again with 200 µl of ACN and left to air dry. Proteins were reduced, alkylated and digested with trypsin as above. To extract peptides from the gel pieces, 40 µl of 0.1% TFA was added and the gel cubes sonicated for 10 minutes in a water bath sonicator. The liquid was removed and the extraction repeated three times before combining the supernatants and drying them in a Speedvac. Samples were re-dissolved in 10 µL 0.1% FA and transferred to an HPLC tube for MS acquisition.

Mass spectroscopy was performed using two different instruments, an LTQ-Fusion Lumos (Thermo Scientific, USA) and a 6600 Triple-TOF (AB Sciex, USA). For the 6600 triple-TOF, 2 µL peptides were resolved by a 15 cm nanoflow C18 column (ABSciex, USA) with the gradient set from 6% to 30% of ACN in 0.1% of FA for 90 minutes. The eluents were introduced into the 6600 triple-TOF with settings described previously {Tse and Lo, 2017, in press). For the LTQ-Fusion Lumos, peptides were first resolved by a 15 cm nanoflow C18 column (LC packings, Netherland) using an Ultimate 3000 nanoflow liquid chromatography (Thermo Scientific, USA) with the same gradient described above. Eluents were introduced into an electrospray (ESI) where peptides were ionized by a nozzle potential of 2300V in positive mode. The temperature of the ESI was kept in 150°C. The mass spectrometer was operated in data-dependant acquisition (DDA) mode. Precursor ions were first introduced to an Orbitrap mass analyzer for precursor mass acquisition (MS^1) with the mass-per-charge range of 350-1500 and a resolution of 60,000. Fragment mass acquisitions (MS^2) were then performed in a linear iontrap mass analyzer.

Data analysis

Sequences were identified using Mascot Distiller 2.5 (Matrix Science) with an Mascot Server 2.5 using a previously described *Lingulodinium* transcriptome with 74,655 entries assembled using Velvet (9) (available in Genbank under the accession numbers JO692619-JO767447). Carbamidomethylation on the cysteine residue was set as the fixed modification, whereas oxidation on the methionine residue was set as the variable modification. One missed cleavage was permitted. For the search of data acquired from the 6600 triple-TOF, precursor ion tolerance and fragment ion tolerance were set at 10 ppm and 0.1 Da respectively, whereas 10 ppm and 0.5 Da was set respectively for spectra acquired from the Fusion Lumos. Global identity false discovery rate (FDR) was kept under 1% by searching the data against a decoy database made by reversing the sequence of the *Lingulodinium* transcriptome library.

Blast2Go (3) was used to determine sequence identities and establish GO categories, and either tBLASTn (for DNA sequences) or BLASTp (for the MS-validated protein sequences) were used at their default settings. The number of sequences assigned to the different KEGG pathways was counted manually for each pathway. The number of sequences for the different categories in the Biological Process, Cell Component or Molecular Function lists were tested using the BLAST results determined with the DNA sequences and the MS-validated protein sequences separately. Statistical significance was determined by first calculating a z-score (as $(X - Y) / (X + Y)^{1/2}$), where X and Y are the number of sequences in the category determined using DNA sequence or protein sequence, respectively, then calculating a p value from the z-score using the Norm.S.Dist function in Microsoft Excel. Enrichment profiles for the MS-validated test set were determined using Fisher's exact test with the entire Velvet transcriptome as a reference.

The deduced protein sequences in the Velvet transcriptome assembly that corresponded to experimentally determined peptide sequences was determined using Geneious (12). The 6628 sequence DNA dataset was first translated in all 6 potential reading frames, and each of the 6 translated reading frames queried separately for matches (100% sequence identity and 100% coverage) with the list of 21,040 MS determined peptides. Comparisons between the MS-validated protein sequences and the longest ORFs for each sequence in the nucleic acid dataset were also made using Geneious. The longest ORF for each DNA sequence in the dataset was determined using Galaxy (13).

To assess relative protein levels, raw MS data files were used for protein quantitation using Progenesis QI for proteomics (Waters). Three technical replicates were averaged to obtain a relative value for each peptide. Correlations to RNA levels determined previously using RNA Seq data (14) were evaluated using the ggpubr package in R.

RESULTS

Peptide sequences from a total of four different experiments were used to recover a final list of 6,628 sequences from a Velvet assembled transcriptome (Table 1). In terms of the efficiency of protein sequencing, we found that the Fusion Lumos delivered more peptide sequences than the 6600 Triple-TOF. We also found that fractionation increased the number of peptides obtained, with SCX fractionation of tryptic peptides performing markedly better than fractionation of the proteins prior to digestion using SDS-PAGE. Almost 6,000 proteins can be identified when both a non-fractionated sample and an SCX fractionated sample are analyzed by the Fusion Lumos, indicating that this combination of protocols should produce the greatest number of peptides for the least investment of time and money. A hybrid strategy was used in the Fusion Lumos, where an orbitrap was used for precursor acquisition and an iontrap was used for fragment ion acquisition. This approach enabled the fragment acquisition of the current cycle and the precursor acquisition of the next cycle happened simultaneously. The sequences identified by MS appear highly dependent on the abundance of the peptides in the sample. For example, the Biological Process group of GO categories shows significant enrichment in basic sugar and amino acid metabolism (Figure 1), and enzymes involved in basic metabolism might be expected to be more abundant and thus more likely to be detected. Only one GO category, protein phosphorylation, was found to be under represented.

In order to determine which deduced protein sequences in our Velvet transcriptome assembly corresponded to experimentally determined peptide sequences, peptide sequences were used to select protein sequences from the 6628 sequences translated in all 6 potential reading frames. All translated sequences with a match to any peptide were then combined to form a single MS-validated sequence dataset. Interestingly, 94 sequences showed a match to peptides in more than

one reading frame. These were examined manually by comparing the Velvet assembly sequence to almost identical sequences in two other datasets, a Trinity assembly of our data (15) and a *L. polyedra* (strain CCMP1738) dataset from the Community for Advanced Microbial Ecology Research and Analysis (CAMERA) (http://imicrobe.us/) assembled using BPA (16). The majority of the sequences where peptide matches were seen in more than one reading frame were found to be assembly artifacts, with 39 containing a tail-to-tail duplication, 12 containing a head-to-tail duplication and 23 appearing to be chimeras formed from two separate sequences. An additional 11 nucleotide sequences had a single frame shift mutation, while the remaining 9 sequences appeared to have resulted from a false positive match, as the peptides matches were in reading frames surrounded by stop codons and thus unlikely to represent a *bone fide* peptide.

To test if analyses using MS-validated proteins sequences differed from those using the DNA sequences in the transcriptome, we first compared the results of BLAST searches using either BLASTp with the MS-validated protein sequences and tBLASTn with the DNA sequences in the transcriptome. Interestingly, the number of sequences with a BLAST hit was greater when the protein sequences were used (Figure 2). After the BLAST searches were completed, the proteins identified were classified by Gene Ontogeny, and the number of proteins in the different categories determined for each of the two searches (Supplementary Table 1). We found that the molecular process classification did not change markedly between the two methods, and biological process categories were also very similar. However, cellular component categories differed markedly between the two BLAST search results. We also tested for the degree to which proteins identified by the two searches could be assigned functions in the KEGG pathway maps (Table 2). In a total of 19 pathways, 97 enzymes were assigned after tBLASTn searches, while 139 enzymes were assigned after BLASTp searches. This represents an increase of over 40% in the number of pathway enzymes represented using MS-validated protein sequences.

We next compared the translated MS-validated transcriptome with the proteins constituting the longest ORF for all the DNA sequence (Figure 3). Roughly two-thirds of these latter (4190 sequences) were 100% identical to the MS-validated sequences, but the sequence similarity in the remaining third decreased rapidly. Thus, simply using longest ORFs does not produce a good yield of authentic protein sequences.

We were also curious in the degree to which RNA and protein levels were correlated in the dinoflagellate *Lingulodinium*, as this species has daily changes in protein synthesis rates (17) but does not alter RNA levels to accomplish this (14). We thus predicted that that the correlation between the two might be on the low side compared to what has been observed in other systems. Using the total number of proteins identified from all runs (6628) to interpret the raw data for the unfractionated sample run on an Orbitrap (Table 1), 3199 proteins were quantified using Progenesis. Relative protein levels were then compared to transcript levels as determined previously by RNA Seq (14) (Figure 4). The correlation between the levels of protein and RNA (Pearson r = 0.46, $p<0.0001$; Spearman rho = 0.33; Kendall tau = 0.23) does indeed appear to be lower than what has been observed in a range of other species (Spearman rho between 0.5 – 0.73) (18).

DISCUSSION

Transcriptomes are an invaluable aid to understanding the biological processes that can be carried out by organisms, and are especially important for non-model organisms where genome sequencing projects are not likely to be undertaken in the foreseeable future. However, transcriptomes have some disadvantages, especially when sequencing efforts are not strand specific, as each DNA sequence has six different reading frames that potentially encode the protein sequence. One often used method to infer the correct reading frame is to simply choose the longest open reading frame (ORF) under the assumption that reading frames will be subject to random mutations that introduce stop codons unless selective pressure acts to conserve the sequence important for the cell. We have tested this using a subset of 6628 sequences from the 74,655 sequences in the transcriptome of the dinoflagellate *Lingulodinium*, the subset being defined by the experimental MS-based identification of at least one peptide in each of the different sequences. However, the longest ORF agrees with the MS-validated sequences only two thirds of the time, suggesting longest ORFs do not always appear to be faithful representations of the encoded proteins. Determining the correct reading frame is important for the dinoflagellates where many of the genes cannot be identified by BLAST searches. As an example, almost a third of the sequences in the *Lingulodinium* transcriptome have no match to

sequences in GenBank (9), so homology cannot be used to infer the proteins that are actually expressed.

This MS-validated protein dataset represents the largest number of dinoflagellate protein sequences reported to date, and validates the use of untargeted bottom-up proteomics with the dinoflagellates. High throughput proteomics appears to have almost completely supplanted the use of 2D electrophoresis in protein analysis. Recently, a similar high throughput approach was used to compare protein levels in toxic and non-toxic *Alexandrium catanella* using iTRAQ (isobaric tags for relative and absolute quantification) mass tags. This study identified 3488 proteins (19) of which 185 had different levels in the two strains. While none of the known toxin biosynthesis enzymes were among them, this important proof of principal clearly shows the importance of untargeted proteomics in assessing how toxins are made in dinoflagellates. This study described here used a label-free approach, which has the advantage that little sample manipulation is required. However, greater care must be taken during analysis of unlabeled samples to ensure correct normalization compared to the iTRAQ technique.

Determining the correct translation products from a transcriptome by incorporating data from experimentally sequenced peptides also influences the results of GO analysis and assignment to KEGG pathways, and this is one aspect that has not previously been observed. In fact, we have found few reports in the literature that have used protein sequence to validate virtual translation products. In one, expressed protein sequences were used to compare protein content as determined by translation of the transcriptome with that obtained by genome annotation (20). Another proposed a software package that could be used to validate genome sequence protein predictions (21), and the package was subsequently used to validate isoforms generated by splicing (22).

The construction of a validated protein database from peptide sequences as described here is for the most part an automated procedure, with only 1.5 % of the sequences requiring manual curation. This seems a worthwhile investment to allow a more efficient interpretation of future MS data from non-model organisms. The elimination of five of the six possible reading frames is likely to reduce the number of false positives when performing database searches, since false

discovery rates are measured as percentage values (typically FDR <1%). A six-fold reduction of the number of sequences would thus reduce the number of false positives by a similar ratio. In addition, search times should also be reduced, thus reducing analysis times for large datasets.

REFERENCES

1. Aebersold, R., and Mann, M. (2016) Mass-spectrometric exploration of proteome structure and function. *Nature* 537, 347-355
2. Moustafa, A., Evans, A. N., Kulis, D. M., Hackett, J. D., Erdner, D. L., Anderson, D. M., and Bhattacharya, D. (2010) Transcriptome profiling of a toxic dinoflagellate reveals a gene-rich protist and a potential impact on gene expression due to bacterial presence. *PLoS One* 5, e9688
3. Conesa, A., Gotz, S., Garcia-Gomez, J. M., Terol, J., Talon, M., and Robles, M. (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* 21, 3674-3676
4. Altschul, S. F., Gish, W., Miller, E. W., and Lipman, D. J. (1990) Basic local alignment search tool. *Journal of molecular biology* 215, 403-410
5. Nagaraj, N., Wisniewski, J. R., Geiger, T., Cox, J., Kircher, M., Kelso, J., Paabo, S., and Mann, M. (2011) Deep proteome and transcriptome mapping of a human cancer cell line. *Molecular systems biology* 7, 548
6. Hebert, A. S., Richards, A. L., Bailey, D. J., Ulbrich, A., Coughlin, E. E., Westphall, M. S., and Coon, J. J. (2014) The one hour yeast proteome. *Mol Cell Proteomics* 13, 339-347
7. Schilling, B., Rardin, M. J., MacLean, B. X., Zawadzka, A. M., Frewen, B. E., Cusack, M. P., Sorensen, D. J., Bereman, M. S., Jing, E., Wu, C. C., Verdin, E., Kahn, C. R., Maccoss, M. J., and Gibson, B. W. (2012) Platform-independent and label-free quantitation of proteomic data using MS1 extracted ion chromatograms in skyline: application to protein acetylation and phosphorylation. *Mol Cell Proteomics* 11, 202-214
8. Tyanova, S., Temu, T., and Cox, J. (2016) The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. *Nat Protoc* 11, 2301-2319
9. Beauchemin, M., Roy, S., Daoust, P., Dagenais-Bellefeuille, S., Bertomeu, T., Letourneau, L., Lang, B. F., and Morse, D. (2012) Dinoflagellate tandem array gene transcripts are highly conserved and not polycistronic. *Proc Natl Acad Sci U S A* 109, 15793-15798
10. Roy, S., and Morse, D. (2014) The dinoflagellate *Lingulodinium* has predicted casein kinase 2 sites in many RNA binding proteins. *Protist* 165, 330-342
11. Kong, H. K., Wong, M. H., Chan, H. M., and Lo, S. C. (2013) Chronic exposure of adult rats to low doses of methylmercury induced a state of metabolic deficit in the somatosensory cortex. *J Proteome Res* 12, 5233-5245
12. Drummond, A., Ashton, B., Buxton, S., Cheung, M., Cooper, A., Duran, C., Field, M., Heled, J., Kearse, M., Markowitz, S., Moi, r. R., Stones-Havas, S., Sturrock, S., Thierer, T., and Wilson, A. (2011) Geneious v5.4 Ed.
13. Afgan, E., Baker, D., van den Beek, M., Blankenberg, D., Bouvier, D., Cech, M., Chilton, J., Clements, D., Coraor, N., Eberhard, C., Gruning, B., Guerler, A., Hillman-Jackson, J., Von Kuster, G., Rasche, E., Soranzo, N., Turaga, N., Taylor, J., Nekrutenko, A., and Goecks, J. (2016) The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. *Nucleic Acids Res* 44, W3-W10
14. Roy, S., Beauchemin, M., Dagenais-Bellefeuille, S., Letourneau, L., Cappadocia, M., and Morse, D. (2014) The *Lingulodinium* circadian system lacks rhythmic changes in transcript abundance. *BMC biology* 12, 107

15. Roy, S., Letourneau, L., and Morse, D. (2014) Cold-induced cysts of the photosynthetic dinoflagellate *Lingulodinium polyedra* have an arrested circadian bioluminescence rhythm and lower levels of protein phosphorylation. *Plant Physiol* 164, 966-977
16. Keeling, P. J., Burki, F., Wilcox, H. M., Allam, B., Allen, E. E., Amaral-Zettler, L. A., Armbrust, E. V., Archibald, J. M., Bharti, A. K., Bell, C. J., Beszteri, B., Bidle, K. D., Cameron, C. T., Campbell, L., Caron, D. A., Cattolico, R. A., Collier, J. L., Coyne, K., Davy, S. K., Deschamps, P., Dyhrman, S. T., Edvardsen, B., Gates, R. D., Gobler, C. J., Greenwood, S. J., Guida, S. M., Jacobi, J. L., Jakobsen, K. S., James, E. R., Jenkins, B., John, U., Johnson, M. D., Juhl, A. R., Kamp, A., Katz, L. A., Kiene, R., Kudryavtsev, A., Leander, B. S., Lin, S., Lovejoy, C., Lynn, D., Marchetti, A., McManus, G., Nedelcu, A. M., Menden-Deuer, S., Miceli, C., Mock, T., Montresor, M., Moran, M. A., Murray, S., Nadathur, G., Nagai, S., Ngam, P. B., Palenik, B., Pawlowski, J., Petroni, G., Piganeau, G., Posewitz, M. C., Rengefors, K., Romano, G., Rumpho, M. E., Rynearson, T., Schilling, K. B., Schroeder, D. C., Simpson, A. G., Slamovits, C. H., Smith, D. R., Smith, G. J., Smith, S. R., Sosik, H. M., Stief, P., Theriot, E., Twary, S. N., Umale, P. E., Vaulot, D., Wawrik, B., Wheeler, G. L., Wilson, W. H., Xu, Y., Zingone, A., and Worden, A. Z. (2014) The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP): illuminating the functional diversity of eukaryotic life in the oceans through transcriptome sequencing. *PLoS biology* 12, e1001889
17. Milos, P., Morse, D., and Hastings, J. W. (1990) Circadian control over synthesis of many *Gonyaulax* proteins is at a translational level. *Naturwiss.* 77, 87-89
18. Maier, T., Guell, M., and Serrano, L. (2009) Correlation of mRNA and protein in complex biological samples. *FEBS Lett* 583, 3966-3973
19. Zhang, S. F., Zhang, Y., Xie, Z. X., Zhang, H., Lin, L., and Wang, D. Z. (2015) iTRAQ-based quantitative proteomic analysis of a toxigenic dinoflagellate Alexandrium catenella and its non-toxic mutant. *Proteomics* 15, 4041-4050
20. Adamidi, C., Wang, Y., Gruen, D., Mastrobuoni, G., You, X., Tolle, D., Dodt, M., Mackowiak, S. D., Gogol-Doering, A., Oenal, P., Rybak, A., Ross, E., Sanchez Alvarado, A., Kempa, S., Dieterich, C., Rajewsky, N., and Chen, W. (2011) De novo assembly and validation of planaria transcriptome by massive parallel sequencing and shotgun proteomics. *Genome Res* 21, 1193-1200
21. Pang, C. N., Tay, A. P., Aya, C., Twine, N. A., Harkness, L., Hart-Smith, G., Chia, S. Z., Chen, Z., Deshpande, N. P., Kaakoush, N. O., Mitchell, H. M., Kassem, M., and Wilkins, M. R. (2014) Tools to covisualize and coanalyze proteomic data with genomes and transcriptomes: validation of genes and alternative mRNA splicing. *J Proteome Res* 13, 84-98
22. Tay, A. P., Pang, C. N., Twine, N. A., Hart-Smith, G., Harkness, L., Kassem, M., and Wilkins, M. R. (2015) Proteomic Validation of Transcript Isoforms, Including Those Assembled from RNA-Seq Data. *J Proteome Res* 14, 3541-3554

Table 1 Peptides and proteins determined by MS analysis

Experiment	# unique Peptides	# unique proteins	# proteins ≥ 2 peptides	# proteins 1 peptide
1 - SCX fractionation	14,208	5,115	2,978	2,137
2 - SDS fractionation	7,836	3,509	1,870	1,639
3 - non fractionated (OB)	6,107	3,046	1,432	1,614
4 - non fractionated (SciEx)	1,731	1,220	449	771
TOTAL (1+2+3+4)	21,040	6,628	3525	2,889
TOTAL (1+3)	18,037	5,958	3,286	2,672

Table 2 Number of entries in KEGG pathways

KEGG map	DNA sequences	Protein Sequences
Glycolysis	7	7
TCA cycle	9	10
Carbon fixation	9	10
Ox. Phosphorylation	4	5
Purine biosynthesis	14	20
Pyrimidine biosynthesis	3	5
Fatty acid metabolism	3	4
F,Y,W biosynthesis	1	1
S,G,T biosynthesis	4	9
R, P biosynthesis	4	6
A,D,N,E,Q biosynthesis	4	7
C,M biosynthesis	7	11
V,L,I metabolism	7	11
K biosynthesis	1	2
H biosynthesis	3	3
Pyruvate metabolism	6	9
Nitrogen metabolism	2	6
Sulfur metabolism	5	7
Methane metabolism	4	6

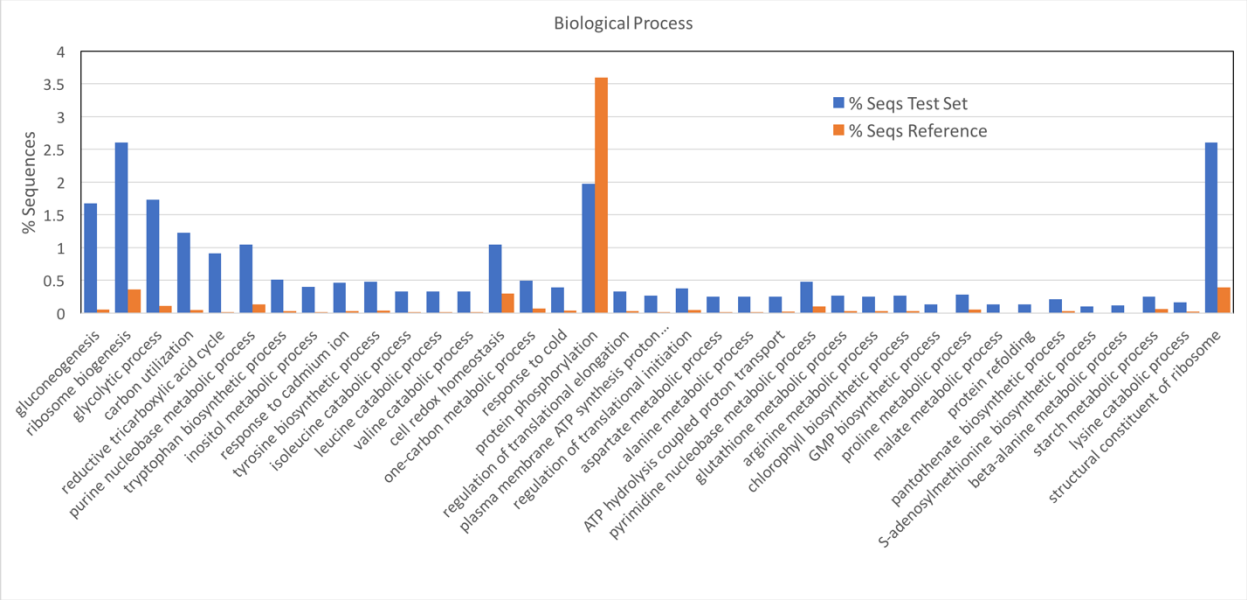

Figure 1 Biological Process GO categories with significantly enriched sequences among the MS identified proteins. MS-validated protein sequences were used as the test set and the entire Velvet transcriptome was used as a reference set to determine significantly enriched categories with Fisher's exact test. The only category found to be underrepresented in the test set is protein phosphorylation. All categories shown are significant with $p < 0.00001$.

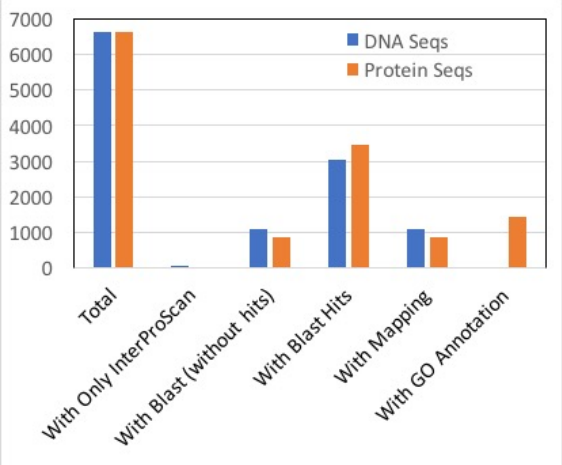

Figure 2 Use of an MS-validated transcriptome catalog increases the number of BLAST hits. The statistics of BLAST searches is shown for the DNA sequences used directly (upper panel) and for the MS-validated protein sequences (lower panel).

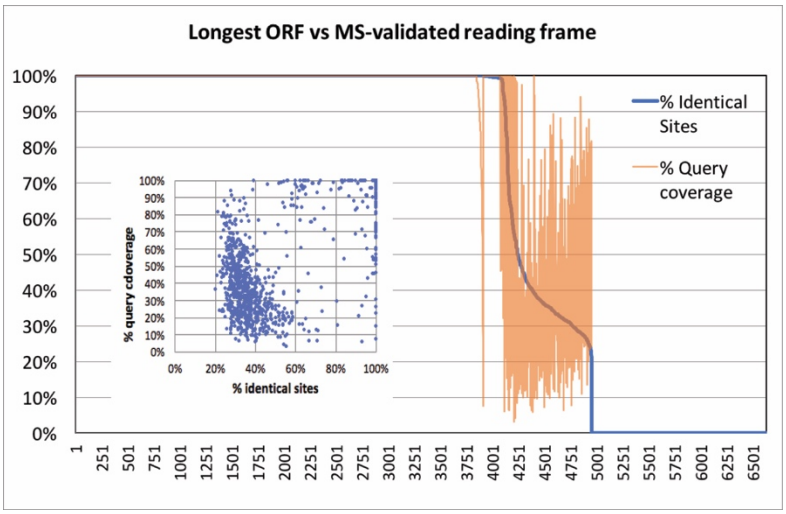

Figure 3 Longest ORF predictions show poor agreement to the MS-validated protein sequence. The protein sequences determined from our MS-validated dataset were compared to protein sequences derived from the longest ORF for each of the DNA sequences (e-value cut-off set to 1). The % sequence coverage is shown as a function of the % identical sites for all sequence pairs (inset).

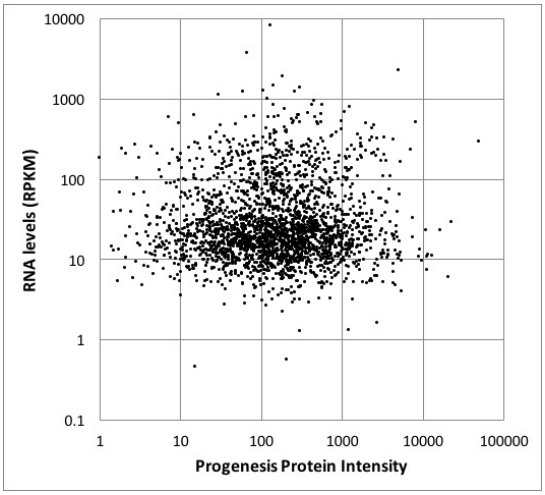

Figure 4 Correlation between Protein and RNA levels is poor. Average log protein intensities (relative values) were plotted as a function of log RNA levels (in reads per kilobase per million, RPKM).

Supplementary Table 1 Number of sequences in different GO categories using BLAST2GO with DNA sequences or with MS-validated protein sequences. Significant difference between the two ($p < 0.05$) are shown in red.

	DNA sequences	Protein Sequences	p value
Molecular Function	1152	1267	0.02593
Catalytic Activity	626	688	0.09240
Transferase	133	153	0.19825
Hydrolase	190	179	0.33861
Oxidoreductase	190	197	0.37447
Lyase	58	69	0.24775
transferase	133	153	0.19825
kinase	74	74	0.39894
ligase	64	60	0.37402
Binding	585	624	0.21268
Nucleic acid	59	71	0.22929
RNA	51	63	0.21214
Protein	27	35	0.23810
Ion binding	371	377	0.38946
Structural	97	98	0.39792
Ribosomal	82	83	0.39774
Cellular Component	480	660	0.00000
Cell	407	555	0.00000
Cytoplasm	275	395	0.00001
Intracellular organelle	241	352	0.00001
Membrane bound	140	234	0.00000
Plastid	80	101	0.11799
Nucleus	25	75	0.00000
Mitochondrial	26	49	0.01173
Non membrane bound	109	122	0.27672
Ribosome	89	87	0.39443
Macromolecular complex	163	188	0.16378
Protein	74	101	0.04970
Ribonucleioprotein	89	87	0.39443
Membrane	7	20	0.01745
Biological Process	809	875	0.10945
Metabolic process	550	613	0.07242
Cellular Process	736	603	0.00054
Response to stimuli	73	79	0.35439
Signaling	43	35	0.26469
Localization	37	60	0.02610
Biological regulation	69	45	0.03190
Celllar component organisation	16	49	0.00009
Small Molecules	257	266	0.36921
Biosynthesis	242	261	0.27865
Catabolism	85	106	0.12576
N metabolism	323	368	0.09216
Photosyntehsis	71	67	0.37647
Gene exp@ression	110	128	0.20198
Translation	107	110	0.39075
Protein modification	81	98	0.17796