Exploring dinoflagellate biology with high-throughput proteomics

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

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

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

RUNNING HEAD

Dinoflagellate proteomics

KEYWORDS

Dinoflagellates, gene expression, MS-sequencing, proteomics

Originality and significance

This manuscript provides a timely overview of current proteomic technologies, with an emphasis on how such techniques could be used to advance our understanding of dinoflagellate biology.

Conflict of interest

The authors declare no conflicts of interest.

Summary

Dinoflagellates are notorious for their ability to form the harmful algal blooms known as "red tides," yet the mechanisms underlying bloom formation remain poorly understood. Despite recent advances in nucleic acid sequencing, which have generated transcriptomes from a wide range of species exposed to a variety of different conditions, measuring changes in RNA levels have not generally produced great insights into dinoflagellate cell biology or environmental physiology, nor do we have a thorough grasp on the molecular events underpinning bloom formation. Herein we discuss the application of high-throughput proteomics to the study of dinoflagellate biology. By profiling the cellular protein complement (the proteome) instead of mRNA (the transcriptome), the biomolecular events that underlie the changes of phenotypes can be more readily evaluated as proteins directly determine the structure and the function of the cell. Recent advances in proteomics have seen this technique become a high-throughput method that is now able to provide a perspective different from the more commonly employed nucleic acid sequencing. We suggest that the time is ripe to exploit these new technologies in addressing the many mysteries of dinoflagellate biology, such as the adaptive changes of coral with their symbiotic dinoflagellates, as well as the initiation mechanisms of harmful algal blooms. Furthermore, while proteomics may be particularly useful for the dinoflagellates, whereas transcriptional responses are often muted, the techniques are readily applicable to other organisms.

Biology of dinoflagellates

Dinoflagellates are unicellular, generally marine protists with a variety of lifestyles (Gomez, 2012; Taylor et al., 2008). The roughly 2000 species that are distinguishable by morphology contain members that are about half heterotrophic, and half plastid-containing (which can include phototrophs, mixotrophs, and species that have simply ingested the photosynthetic apparatus from prey). The vast majority of dinoflagellates are free-living, though roughly 10% and 1% are parasitic and symbiotic species respectively; the latter are disproportionally important given that their endosymbiotic association with reef-building scleractinians serve as the basis of all coral reefs. However, the number of morphologically distinct species is most likely an underrepresentation of the number of species, as roughly 8000 operational taxonomic units (OTU) were uncovered by the Tara Oceans rRNA sequencing project (de Vargas et al., 2015). Based on this new study, dinoflagellates actually have a much greater species richness than any other group of marine protozoan. While it is difficult to ascribe a definite percentage to the contribution of dinoflagellates to global primary production, the oceans account for roughly half the global total, and dinoflagellates and diatoms together account for a large proportion of the marine contribution (Field et al., 1998).

Dinoflagellates are also known for their ability to form the harmful algal blooms (HABs) called red tides. While they occur naturally, their frequency has increased over the past 30-40 years, possibly as a result of nutrient runoff from agriculture (van Dolah, 2000). Many of the HAB-forming species can release toxins into the environment, of which the most well-known are probably the sodium channel blocker saxitoxin (which leads to paralytic shellfish poisoning), brevotoxin (which results in neurotoxic shellfish poisoning), ciguatoxin (resulting in ciguatera fish poisoning) and the protein phosphatase inhibitor okadaic acid (which causes diarrheic shellfish poisoning) (Wang, 2008). These toxins are complex secondary metabolites, generally polyketides, with the cyclic perhydropurine saxitoxin being an important exception. Why dinoflagellates synthesize such a wide array of potent toxins is currently unknown. However, it was found that toxin-producing *Karlodinium veneficum* changes the swimming pattern of *Storeatula major*, a cryptophyte prey (Sheng et al., 2010). In this case, the toxin acts as an aid to feeding.

Dinoflagellates also display a dazzling array of unusual cytological features (Fig 1) (Spector, 1984). For example, some species are able to bioluminesce and form specialized organelles called scintillons for this purpose (Fig 1B) (Fogel et al., 1972). These scintillons contain the bioluminescence reaction catalyst (luciferase) and a substrate (luciferin) binding protein (LBP) (Desjardins and Morse, 1993), and both proteins are found at higher levels during the night when the cells produce light than during the day when they do not (Johnson et al., 1984; Morse et al., 1989). In addition, dinoflagellate plastids are different from those found in higher plants or green algae as they are derived from either secondary or tertiary endosymbiotic events (Keeling, 2010). The plastid genome is generally in the form of small plasmid-like circles (Zhang et al., 1999), and only a small number of genes are expressed (Wang and Morse, 2006). The most common form of plastid contains (1) an unusual reddish carotenoid called peridinin, (2) a novel peridinin-chlorophyll-binding protein (PCP) as the major light harvesting protein (LHC) (Hofmann et al., 1996), and (3) a particularly oxygen-sensitive form II Rubisco as a carbon fixing enzyme (Morse et al., 1995; Rowan et al., 1996; Siegelman et al., 1977). The subcellular location of Rubisco changes during the day to form regions of concentrated enzyme termed pyrenoids (Fig 1C) (Nassoury et al., 2001). Lastly, the dinoflagellate nucleus contains chromatin that is permanently condensed throughout the cell cycle (Fig 1D), and the nuclear membrane remains intact during mitosis. The amount of basic protein in the nucleus is unusually low, with a ten-fold lower protein to DNA ratio than what is found in other eukaryotes (Rizzo et al., 1982). It seems likely that both RNA and divalent cations are involved in maintaining chromatin structure. There are no observable nucleosomes, and indeed, despite the finding of conserved histone genes in the transcriptomes of several species (Bayer et al., 2012; Roy and Morse, 2012), levels of histone expression at the protein level are extremely low (Roy and Morse, 2012). Instead of histones, dinoflagellate nuclei possess both histone-like proteins and dinoflagellate/viral nuclear proteins (DVNP) (Chan and Wong, 2007; Gornik et al., 2012; Janouskovec et al., 2017). Other species have a light sensitive eyespot, called an ocelloid, which is formed from a combination of mitochondria and plastids (Gavelis et al., 2015).

Gene expression in dinoflagellates

In order to address these many unusual and important facets of dinoflagellate biology, changes in gene expression under different conditions have been extensively studied. Typically these studies use microarray analysis or high-throughput nucleic acid sequencing to look at changes in transcript levels. Studies on many other species of algae have indicated that changes in gene expression in response to changes in the environment can provide valuable insight into the mechanisms allowing them to cope (Mayfield et al., 2014a; Mayfield et al., 2014b). However, and perhaps in keeping with their unusual nuclear chromatin architecture, the degree to which dinoflagellates use transcriptional responses to alter mRNA expression levels appears limited. For example, an examination by RNA-Seq of the transcriptome in *Lingulodinium polyedrum* over the course of an 12:12 light dark cycle and during 24 hours of constant light detected no significant change in expression of any of the transcripts (Roy et al., 2014). There are thus no transcriptional responses due to changes in light intensity, or over the cell cycle, or brought about by the circadian (daily) clock in this species.

In contrast, transcriptional responses have been reported for other species of dinoflagellates, although for the most part these are of low amplitude. One study of *rbcL* transcript levels in the symbiotic dinoflagellate *Symbiodinium* showed robust circadian changes when cells in culture were examined (Mayfield et al., 2014a). There are also changes in gene expression in *Symbiodinium* in response to thermal stress, widely investigated because of its potential importance in the phenomenon of coral bleaching (Barshis et al., 2014; Krueger et al., 2015; Leggat et al., 2011; Mayfield et al., 2012; Mayfield et al., 2014b; Rosic et al., 2011), although the majority of stresses experienced by *Symbiodinium* fail to induce a transcriptional response (Baumgarten et al., 2013; Mayfield et al., 2011). In *Amphidinium carterae* low light was reported to induce PCP and other LHC-protein transcripts (ten Lohuis and Miller, 1998), and phosphate limitation was found to increase alkaline phosphatase transcript levels (Lin et al., 2011). *Karenia brevis* has both diurnal and circadian regulated genes (van Dolah et al., 2007), as well as a transcriptional response to nitrogen and phosphorus depletion (Morey et al., 2011), while *Prorocentrum* has been reported to have a diurnal difference in the abundance of transcripts encoding rhodopsin and Rubisco (Shi et al., 2013). Metals change the expression level of superoxide dismutase in *Gonyaulax* (now *Lingulodinium*) (Okamoto et al., 2001), and a number of genes were observed to be regulated by redox state in *Pyrocystis* (Okamoto and Hastings, 2003). Gene expression in *Alexandrium minutum* was found to be affected by growth stage and

nutrient status (Yang et al., 2010). In another *Alexandrium* species, a transcriptional response to the presence of bacteria in the media was noted (Moustafa et al., 2010). Lastly, there has also been a report of cell cycle-regulated changes in levels of transcripts encoding the translational initiation factor 5A in the heterotrophic species *Crypthecodinium* (Chan et al., 2002a).

Dinoflagellates also regulate gene expression at a post-transcriptional level. This has been extensively studied in *L. polyedrum* with respect to control over protein synthesis by the circadian clock (Akimoto et al., 2005; Fagan et al., 1999; Milos et al., 1990; Morse et al., 1989). These changes are regulated at a post-transcriptional level as no changes in the amounts of the corresponding mRNAs have been found by RNA Seq (Roy et al., 2014). Furthermore, post-transcriptional regulation of S-phase gene expression in *Karenia brevis* was reported (Brunelle and van Dolah, 2011). The finding that there is an extensive family of sequences encoding the translational initiator (cap-binding protein) eIF4E provide additional support for the importance of translational regulation in the dinoflagellates (Jones et al., 2015).

It is important to stress that while studies of post-transcriptional changes in gene expression are more difficult to perform than those using high-throughput sequencing techniques to measure RNA levels, there are many cellular mechanisms that act at a post-transcriptional level. These include protein-mediated regulation of RNA structure, sequestration or translation (Castello et al., 2013), as well as miRNA mediated effects on translation (Iwakawa and Tomari, 2015). In general, the correlation between protein levels and RNA levels is not very good. For example, only 2 of 167 regulated *Symbiodinium* proteins varied in concert with their transcripts (Mayfield et al., 2016). Another study demonstrated a decrease in the amount of Rubisco (RBCL) protein despite constant *rbcL* mRNA levels in response to elevated temperatures (Putnam et al., 2013). A more global evaluation of the relationship between dinoflagellate RNA and protein levels was carried out with over 3000 *Lingulodinium* proteins (Tse et al., 2018). The correction was poor (Pearson's $r = 0.46$), even lower than what was reported in a similar study in mammalian cells (Pearson's $r = 0.54$ using over 1000 genes) (Vogel et al., 2010). As a consequence, transcriptome profiling may be of limited use in addressing significant biological problems in dinoflagellates, as shown in a comparative study of mRNA expressions and the corresponding protein expressions of *Symbiodinium* (Mayfield et al., 2016). One exception may be a recent comparison

of the transcriptomes of toxic and non-toxic *Alexandrium* strains, which revealed that among the 35 differentially expressed sequences, a putative saxitoxin pathway enzyme *sxtA* was up regulated in the toxic strain (Zhang et al., 2014). However, it must also be noted that the net effect of the decrease in this one sequence is unclear, as *Alexandrium* species can contain up to 12 different *sxtA* sequences (Hackett et al., 2013).

Proteomics

As a general rule, measurements of protein levels are likely to provide a more accurate representation of the catalytic capability of the cells than are measurements of transcript levels as proteins are more directly related to an observed cellular phenotype than are RNA levels. For the dinoflagellates, proteomic studies are thus of particular interest. Traditional explorations of the cellular proteome typically involved analysis of one protein at a time, using gel electrophoresis and Western blotting techniques to evaluate the levels and potential changes that the protein was undergoing in different conditions. The development of two-dimensional electrophoresis (2D-E) allowed many hundreds of proteins to be examined at a time, although these techniques were difficult to master. A recent review of dinoflagellate proteomics makes clear that the majority of work in this area has indeed used two-dimensional gel electrophoresis to separate and visualize differentially expressed proteins (Wang et al., 2014) before identification using mass spectrometry (MS) techniques. Proteomics has wide application in dinoflagellates, as it has been used to look for proteins involved in symbiosis (Weston et al., 2012) and to look for proteins that may be involved in toxin production (Wang et al., 2013). It has been employed as a means of determining surface proteins (Chan et al., 2012) as well as a way to rapidly identify the species in a natural harmful algal bloom (Lee et al., 2011). The proteomics of cellular chronobiology has also been examined (Akimoto et al., 2004).

Now, however, advances in high-throughput protein sequencing have allowed datasets of many thousands of proteins to be generated routinely. In the most recent application to the dinoflagellates, high-throughput MS techniques coupled with tags to aid protein quantification (isobaric tags for relative and absolute quantification, [iTRAQ]; described in more detail below) have identified 3488 different proteins, setting the current record for number of dinoflagellate proteins identified (Zhang et al., 2015b). High-throughput proteomics has the immense

advantage that, like high-throughput DNA or RNA sequencing, the large datasets that are created are not hypothesis driven, but rather represent a catalog that can be pored over to associate sequences, whose nature was not previously suspected to be involved with the phenomenon or treatment under study, with a particular phenotype. While not yet at the levels of nucleic acid sequencing, the number of proteins that can be quantitated by top-end MS is still quite remarkable. Furthermore, the ease of sample preparation and price of sample analysis is now bringing the technique to the domain of the regular researcher. For example, while labeling strategies designed to aid protein quantitation do require technical expertise, in what is called "label-free proteomics" thousands of proteins can be measured in a sample prepared simply by precipitating the proteins in a cell lysate with acetone. The number of proteins identified in model systems, the high end of the technology to date, has reached ~10,000 proteins in a study of a single human cell type (Nagaraj et al., 2011) by a simple fractionation of a trypsin digested crude protein extract into six parts by ion exchange chromatography. More recently, a single injection of a yeast total protein digest identified 34,000 peptides and almost 4,000 proteins (Hebert et al., 2014); this represents almost 90% of the yeast proteome (Ghaemmaghami et al., 2003). These examples underscore the degree to which MS-based proteomics is technology driven, and it must be stressed that advances in the field show no signs of abating (Padula et al., 2017).

Despite the high resolving power of LC-MS/MS, protein extraction procedures are often the first obstacle which researchers must overcome when trying to use proteomics technologies with dinoflagellates. In part this is due to the tough and complex cell cortex, containing internal flattened membrane bound structures called alveolae which can often contain cellulosic plaques. Some species have a fibrous layer under the alveolae called the pellicle, which may contain sporopollenin. High endogenous levels of salts, nucleic acids, phenolic compounds and pigments can also interfere with the proteomic analyses. The general solution to the cell breakage problem is application of sufficient mechanical force, and techniques such as ultra-sonic disruption of the cells, vigorous shaking in a bead-beater with bead sizes appropriate to the size of the cell, or grinding in liquid nitrogen have been used with success (Tse et al., 2018; Tse and Lo, 2017b). The problems associated with chemical contaminants were also encountered during preparation of samples for 2D-E, and solution involved a number of lengthy and tedious procedures (Chan et al., 2004; Chan et al., 2002b; Wang et al., 2009) or toxic chemicals like

Trizol (Lee and Lo, 2008; Zhang et al., 2015c). Trizol was also used for gel-free shotgun proteomics of field and laboratory-cultured diatoms (Zhang et al., 2015a), whereas a simple trichloroacetic acid/acetone precipitation step has been successfully used with the dinoflagellate species *Scrippsiella acuminata* (Tse and Lo, 2017a; Tse and Lo, 2017b) and *Lingulodinium polyedra* (Tse et al., 2017).

During what is variously termed "bottom-up," "discovery," or "shotgun" proteomics, complex protein mixtures are digested with a protease (usually trypsin) that cleaves the protein at specific amino acids (the basic amino acids lysine and arginine in the case of trypsin) thus producing a series of peptides that are the same for a protein and different for other proteins. The tryptic peptides are then separated by liquid chromatography (LC) and analyzed by ionization of the peptides and injection into an electrospray-tandem mass spectrometer (ESI-MS/MS). For a full description of the linear trap quadrupole (LTQ)-Orbitrap (Thermo Scientific, USA), one of the mass analyzers with the highest resolution and mass accuracy for proteomics works, an excellent recent review covers the technical details (Eliuk and Makarov, 2015). Briefly, the tryptic peptides partially resolved by a liquid chromatography column are injected into an electrospray (ESI) ion source where they are ionized, allowing the mass-to-charge (m/z) ratio of the peptides to be determined by the mass spectrometer subsequently. A typical LC profile shows a complex series of peaks (Fig 2A), and at each time point on the chromatograph, several mass peaks are typically seen in the precursor mass spectrum (MS^1) (Fig 2B).

Peptides are generally observed as precursor ions with two positive charges (+2 ions) on the m/z axis. These precursor ions are easily identified as having a +2 charge, as when the m/z axis is expanded, they form a series of equally spaced peaks whose mass differs by 0.5 m/z. These different peaks are due to a random incorporation of the carbon-13 isotope instead of the normal carbon-12 isotope (a difference of 1 Da); ^{13}C is a naturally occurring stable isotope which makes up 1.1% of the total carbon on the planet. The exact pattern of the series of peaks varies depending on how many ^{13}C atoms are found in the peptide (Fig 2B, insert), but it is the spacing between the peaks that defines the charge on the ion.

Suitable precursor peptides ions detected by the MS^1 are then selected to go through the secondary mass spectrometry (MS^2), where the precursor ions selected collide with atoms of an inert gas and fragment into pieces in what is called collision induced dissociation (CID). Precursor peaks in the MS^1 are deemed suitable for CID when they meet a number of predefined

characteristics (what is called data-dependent analysis), including being +2 charged ions, as well as having a peak intensity higher than a preset threshold. It should be mentioned that intensity, or current flow, is proportional to the amount of the precursor ion, and thus of the original amount of the peptide. When fragmentation occurs, it generally occurs at the peptide bond. A single fragmentation event for a single peptide molecule produces two fragments, one containing the original N terminal end and the other containing the original C terminal end. If fragmentation was completely random, then a population containing equal amounts of all possible peptide fragments would be observed, producing a series of peaks whose mass differs by only a single amino acid. The peptide sequence can actually be read twice in these fragmented peptide mass spectra, once by looking at the mass difference between peptide fragments as they increase in size from the N terminal end (called the b-series) and the other when they increase in size from the C terminal end (called the y-series). However, in practice, fragmentation is rarely completely random, and thus only a sub-section of the expected mass peaks is observed.

When the genome (or, for the dinoflagellates, most likely a transcriptome) of the organism under study is known, the sequence of a given peptide can be identified by a comparison of the experimentally obtained pattern of peptide fragment masses with a computer predicted pattern of peptide fragment masses for every tryptic peptide sequence predicted by the genome (or transcriptome) sequence. For a single peptide from the PCP protein of *L. polyedrum* (Fig 2C), the predicted masses of the b and y series (vertical green and blue lines, respectively) are compared to the actual m/z patterns (red lines). The analysis also provides a statistical estimate of the validity of the identification, typically being calculated as a peptide "score" when the analysis is performed by Mascot Server (Matrix Science, United Kingdom), and this represents the probability that the actual pattern of m/z profile corresponds to the predicted m/z profile of a particular peptide. The score depends on a number of factors, including how many peaks are seen in the fragmented peptide, the signal-to-noise ratio of the peaks, and the accuracy of the mass determination for the peaks. In general, the highest scoring identification above a preset threshold is retained. Experiments are typically performed in what is called data dependent analysis mode, which simply means that MS1 peaks are selected for MS2 fragmentation based on their intensity. This has the disadvantage of biasing the analysis against low abundance proteins. The advantage is that because the experiment is performed without any preconceived idea of the desired target proteins, unexpected candidates can be discovered.

However, there are some confounding problems when transcriptomes are used as the sequencing database for protein identity searches. These include the fact that the transcriptome is likely incomplete, that the transcripts may have been incompletely assembled, or that the sequences contain rearrangements or frame shift mutations, although some of these problems can be mitigated by using MS data to refine the sequences present in the transcriptome (Tse et al., 2018). Furthermore, some peptides may feature post-translational modifications which will change the mass of some of the fragments, thus hindering identification. However, for the most part these problems are relatively minor, and it is the resolution, sensitivity and accuracy of the instrument itself which limits the number of proteins identified.

It is often convenient to think of the mass spectra produced from a complex mixture of tryptic peptides as analogous to the patterns of proteins observed after two-dimensional gel electrophoresis. In this case, however, the two axes are not isoelectric point and size but rather retention time on the LC column and the mass-to-charge ratio (Fig 2D). This view is an easy way to grasp the essential features of the process, but obscures the wealth of detail actually present. If blown up, each of the "spots" will be seen to be a series of mass peaks (due to random incorporation of C^{13}), and in many cases, these spots will be linked to an MS^2 spectrum which does not appear in this view. The intensity of the different spots does not typically provide a good visual estimate of the amounts of different peptides, which can vary by up to five orders of magnitude. As one example, a small region of the spectra in the region containing a peptide derived from the bioluminescence protein LBP is shown (boxed in red in Fig 2E). This protein is known to change in amount by up to 10-fold over the course of the light-dark (LD) cycle (Morse et al., 1989) yet visual examination of the spectra shows only a slight change.

Identification of proteins

One of the earliest techniques developed to identify proteins was peptide mass fingerprinting, in which the masses of all peptide fragments produced by protease digestion of a given protein were measured and compared to the predicted pattern of masses from all proteins in a sequence database (Pappin et al., 1993). This powerful technique does not require excessively sophisticated equipment and can also identify regions of a protein that have undergone post-

translational modification. However, the disadvantages include a requirement for complete protein sequences in the database and the fact that it is not suitable for analysis of complex mixtures of proteins.

Proteins can also be identified by determining the sequence of a single peptide, or where more stringent identification is required by two or more peptides from the same protein. Identification is less sure with single peptides there may be many matches for short peptide sequences and because there is unavoidable "noise" in low-m/z range of typical MS^2 spectra. The sequences of peptides from two or more regions in the same protein provide more confidence in the identification (Tse et al., 2018; Tse and Lo, 2017b; Zhang et al., 2015b). The former would be chosen when the goal is to glean the most information possible from the protein sample at the expense of confidence in the identification, while latter is more likely to reflect the true identity of the transcript and will deliver a more faithful estimate of the quantity of the protein. There are always potential errors in identification, a part of which is due to the large number of sequences in the DNA database as indicated above; errors in identification (or false positives) increase along with the increasing number of peptides identified. Another contributing factor is that two amino acids have identical masses (leucine and isoleucine are said to be isobaric.) while another pair is near-isobaric (lysine and glutamine have very similar masses.), and these can cause errors in the identification of sequence from peptide mass patterns. A last issue is that a given peptide may be found in more than one protein. For example, one peptide identified may be found in proteins A and B while a second may be found in proteins A and C, and it is not possible to experimentally determine if both peptides are derived from A or if an equal mixture of B and C is present. In both cases, when peptide identification scores are combined to yield protein identification scores, both peptides will be deemed to be derived from A.

Fractionation of complex mixtures

While the performance of instruments and their ability to distinguish a large number of peptides in a complex mixture are constantly increasing, there are still many reasons to fractionate complex mixtures. The most compelling of these is that the complexity of the mixture decreases, thus making it easier to characterize the partial proteome more thoroughly. However, in many

other cases fractionation is used to uncover a specific subclass of protein sequences. For example, phosphorylated proteins can be identified by first enriching a crude protein preparation on phosphate binding columns. As another example, MS analysis of purified organelles can provide an inventory of proteins in a specific sub-cellular location (Peng et al., 2011).

The two most generally used fractionation methods include SDS-PAGE for proteins and ion exchange chromatography for peptides. In the former, proteins are resolved using standard electrophoretic techniques, and the gels are consequently stained with Coomassie blue and cut into small pieces. Proteins in each gel piece are then digested *in situ* with trypsin and the peptides extracted and sequenced. In the latter, proteins are first digested with trypsin, and the resulting tryptic peptides then separated chromatographically, generally with cation exchangers (Kong et al., 2013).

Other fractionation methods can be used to enrich phosphoproteins or phosphopeptides (Fig 3). In the former, protein samples are fractionated on commercial (and proprietary) chromatographic resins, e.g. the phosphoprotein enrichment kits by Pierce (Thermo Scientific, USA). In the latter, proteins are first digested with trypsin, and the phosphorylated peptides are separated by metal oxide affinity chromatography (MOAC, typically using TiO_2 (Ibáñez-Vea et al., 2017; Iliuk et al., 2010; Melo-Braga et al., 2015; Thingholm and Larsen, 2016c; Yu and Veenstra, 2013)), by using immobilized metal affinity chromatography (IMAC, typically using Fe^{+3} NTA, a nitrilotriacetic acid derivative) (Thingholm and Larsen, 2016a, 2016b; Yan et al., 2013), or by binding to antibodies recognizing phosphorylated amino acids (Blaydes et al., 2000; Foy et al., 2007). The advantage to the former is that many peptides from the same protein can be identified, thus facilitating identification and quantification, whereas that of the latter is that a larger number of phosphorylated peptides are identified, thus facilitating the characterization of the phosphosites that are actually phosphorylated in the sample. Phosphorylation is important as it is the most frequently found post translational modification, and it is used extensively to regulate protein structure and activity (Cieśla et al., 2011; Fischer, 2013). Some estimates indicate the number of phosphorylated proteins in eukaryotes could be as high as 30% (Mann et al., 2002), underscoring its pervasive nature. The extensive use of phosphorylation to modify proteins is mirrored by an abundance of protein kinases, which number about 500 in the human

genome (Manning et al., 2002), whereas the transcriptome of the dinoflagellate *Lingulodinium* has just over 600 (Roy and Morse, 2014).

Another fractionation option involves purification of sub-cellular structures such as organelles (Yates et al., 2005). This typically involves centrifugation of cell lysates using differential or isopycnic (density gradient), or a combination of both. Unfortunately, sub-cellular fractionation protocols are limited for the dinoflagellates, and there are two potential issues that must be overcome for this to be useful. First, the organelles must remain intact after breaking the cellulosic cell wall which often requires considerable destructive force, and second, a means of following the organelles through the procedure and determining their purity must be available. Plastids in *Lingulodinium* were found to be unstable and cannot be purified intact (Wang et al., 2005b), although a cell wall preparation (Bertomeu et al., 2003) and the bioluminescent organelles termed scintillons (Desjardins and Morse, 1993) have been prepared from this species. Purifications of nuclei have been reported from a number of dinoflagellate species (Bhaud et al., 1999; Mendiola et al., 1966; Rizzo et al., 1982; Sun et al., 2012). The proteomics of dinoflagellate nuclei is a potential gold mine of information given their many unusual features, and proteomics has begun to provide insight into how the dinoflagellate chromatin may be organized by cataloging the proteins identified in purified chromatin preparations (Beauchemin and Morse, 2017). For example, this latter study found a surprisingly high proportion of RNA binding proteins in the chromatin, suggesting the high levels of RNA reported in the dinoflagellate nucleus (Soyer and Haapala, 1974) may play a structural role.

Finally, proteins and their binding partners can be analyzed using affinity purification (Dunham et al., 2012). There are many versions of this but, due to the lack of widespread transformation technologies that would allow proteins carrying a tag to be expressed in the cell, not all are applicable to the dinoflagellates. The techniques that remain may still be powerful, however, as they include antibodies that are able to recognize an epitope of a naturally occurring protein. Furthermore, this would also include the use of natural tags such as phosphate groups on protein or peptide residues whose purification has been discussed above. However, apart from purification of phosphorylated proteins, affinity purification has not been widely used in study of dinoflagellate proteins. An affinity column, $p13^{suc1}$ that has been used to purify cyclin-dependent

kinases in yeast, was tested in *Lingulodinium* but was not successful in binding the dinoflagellate enzyme (Bertomeu et al., 2007).

Quantitative comparisons

Quantitation of the amount of each peptide identified in the LC-MS/MS is conceptually easy but technically challenging (Bantscheff et al., 2012), and a number of different techniques have been developed to determine protein levels. Isotopic labeling is currently the gold standard in protein quantitation, and it generally involves mixing two samples, one of which has been labeled with a stable isotope that changes the mass of a peptide, and analyzing the ratio of the labeled and unlabeled peptides in the mixture (Figure 4). The isotopically labeled version of the peptide is chemically identical to the label normal version and thus has an identical LC retention time, and differs only in mass. This technique has two variants, one in which proteins in the cell are labeled (so called metabolic labeling) and one where proteins (or peptides) are labeled after extraction of the cells. A second class of techniques, termed label-free methods, mostly estimate the amount of the protein from either the intensity of the peak in the MS^1 spectrum, or the number of MS^2 spectra that can be identified as peptides belonging to a given protein.

Isotopic labeling in living cells can be performed in two ways. In one, a non-radioactive isotopically labeled substance is administered for a short time period, essentially a pulse-labeling approach similar to what has been often used with radioactive isotopes. If the substance chosen is an amino acid and the labeling time is short, then only a small proportion of a protein pool will be labeled with the proportion depending on the speed of synthesis. While radio-labeled amino acids and 2D gels have been used to study protein synthesis *in vivo* in dinoflagellate cultures (Morse et al., 1989; Wang et al., 2005a), stable isotope-labeled amino acids have so far only been used to to test for incorporation of an amino acid into the toxin molecule in a dinoflagellate toxin biosynthesis study (Macpherson et al., 2003). As an alternative to the pulse labeling approach, cultures can be grown in the presence of a heavy isotope such as ^{15}N-$NaNO_3$ for many generations, which ensures all proteins in the cell are uniformly and completely labeled. The labeled proteins are mixed in with unlabeled proteins, and the mixture then hydrolyzed and

analyzed. This technique has been shown in plants to be highly reproducible for quantification (Nelson et al., 2007) and has recently been employed in the dinoflagellates (Tse and Lo, 2017a).

Isotope-coded affinity tags (ICAT) contain (1) reactive groups that can be used to label specific amino acids in proteins (typically cysteine residues), (2) linkers with different mass tags, and (3) an affinity tag (generally biotin) that can be used to purify labeled from unlabeled peptides (Gygi et al., 1999). Because different protein samples can carry different mass tags (heavy or light), two different protein samples can be mixed, processed and analyzed concurrently, and the relative amounts of the two peptides can be determined simply from the peak intensities for the different mass-tagged peptides in the precursor ion spectra. A variant of this, termed isobaric (meaning equal weights) mass tags, also have a reactive group that are covalently linked to amino acids. However, in contrast to ICAT, in which the linkers are with different masses, linkers in isobaric mass tags have the same mass in MS^1 analyses. Instead they are distinguished when they enter the second MS, as they fragment to produce tags whose masses differ in the MS^2 spectra, and the ratios between different mass forms are again used for quantitation. These tags can be used either with intact proteins or with tryptic peptides. Two types of labels exploiting this technology are iTRAQ; (Ross et al., 2004) and tandem mass tags (TMT; (Thompson et al., 2003)), and they are sold with eight or ten different mass forms, respectively. This means that either 8 or 10 different samples can be labeled separately and subsequently being mixed and analyzed together, facilitating quantitation between samples. TMT Labeling of proteins prior to digestion has been used to examine the response of coral symbionts (genus *Symbiodinium*) to thermal stress (Weston et al., 2012). iTRAQ Labeling of peptides following digestion has been used to compare the protein complement of toxin producing and non-producing strains of *Alexandrium* (Zhang et al., 2015b).

In label-free methods, the height of the peak in the MS^1 spectrum is conceptually the easiest way to estimate and compare protein abundances. However, there are several potential problems associated with this approach. One of which lies in ensuring that peptides of very similar but actually different masses can be adequately resolved in the MS^1. In this case, peak intensity measurements could reflect a contribution from multiple peptides. A second problem is that a peptide is spread out along the time axis of the chromatograph during the LC step. Therefore, a

particular peptide will be found in multiple MS^1 spectra, and the intensities in each must be collated and summed together. A third important issue is that of standardizing different samples to accommodate differences in the total amount of sample loaded, and here a general solution is to select peptides whose levels are not expected to change between the different samples and to use them to adjust the amount of signal obtained for all the peptides in the sample. In high resolution LC-MS/MS, such as the Orbitrap Fusion Lumos (Thermo Scienctific), most of these difficulties are reduced by the high performance of the instrument with higher consistency of intensities of mass peaks over samples. Several different (and free) computer programs are available for normalizing and quantifying peptides using MS intensity data, including MaxQuant (Tyanova et al., 2016) and Skyline (Schilling et al., 2012). MaxQuant is designed to work with large numbers of peptides, while Skyline is better suited for detailed analysis of a smaller number of peptides.

A second method for measuring peptide abundances is called spectral counting, where the number of MS^2 corresponding to a given peptide sequence are simply counted (Zybailov et al., 2005). This is based on the principle that an abundant peptide will constitute a larger peak in the LC chromatogram, and in consequence, will have levels above the preset threshold in a larger number of MS^1 spectra and will be selected more often to generate MS^2 spectra. One problem associated with this measure is that there is a decreased dynamic range compared to peptide intensity measurements, although the correlation in-between the two approaches are seems to be strong (Zybailov et al., 2005). Furthermore, the number of MS^1 spectra from a given peptide that are actually selected to produce an MS^2 spectrum has been found to be influenced by the other peaks in the MS^1 spectrum, suggesting that the quantitation of a given peptide may differ depending on the sample in which it is found.

Moving up from quantitation of individual peptides to proteins, one measure of protein quantity can be made on the basis of the number of peaks among the MS1 spectra that can be identified to correspond to a given protein; this measure is termed the exponentially modified protein abundance index (emPAI) (Ishihama et al., 2005). Similar to the spectral counting method, the idea is that a more abundant protein will produce more above-threshold value peptides that will be selected for to produce an MS^2 spectrum. This method is rooted in the fact that different

peptides usually ionize to different degrees, making greater levels of the original protein capable of producing more of a difficult-to-ionize peptide and thus an increase in the number of peptides identified for the protein. Similar to spectral counting, this measurement has a reduced dynamic range compared to MS^1 peak intensities, as there are a finite and discrete number of peptides that can be produced from a given protein. Other approaches, such as using weighted averages of the amount of the individual peptides making up a given protein, are also available (ProteoProfile; http://www.thibault.iric.ca/proteoprofile/files/TechnicalGuide.pdf) for calculating protein concentrations (Beauchemin and Morse, 2017).

Conclusions

Dinoflagellate proteomes are vastly understudied and deserve much wider attention. What can we learn from proteomics? Proteins are more relevant indicators of cellular behavior than are RNA levels, so proteomics techniques allow the biochemical abilities of a cell to be probed more directly than ever before, thus have the potential to address a number of outstanding issues in dinoflagellate biology (Table 1). Comparisons of the types and amounts of proteins found in toxic and non-toxic strains can allow uncovering how such toxins are synthesized. Furthermore, levels of proteins and phosphoproteins monitored over the course of the cell cycle can provide valuable clues into mechanisms of cell division and how this process is regulated. Proteomics has become a force to be reckoned with first because the wide spread availability of transcriptome sequencing can easily produce the databases on which protein identification relies, and second because the continuing advances in LC-MS/MS technologies allow identification and quantification of an ever-increasing number of proteins. The current state of the art has reached a stage where it now handily outperforms 2D gel electrophoresis for analysis of complex mixtures, and the use of simple and rapid fractionation protocols prior to analysis increases the number of proteins identified substantially. Prices of the service in academic institutions are typically around $100- $150 US per sample, which can be either a single purified protein or a complex mixture from which several thousand proteins can be identified. While these analyses do not include mass-tagging or fractionation, they often do include data analysis (peptide identification and quantification). This is in the same range as the price for outsourcing 2D gels, although unlike 2D gels, the cost of purchasing high-end MS technology for an individual investigator

would be prohibitive. It must also be kept in mind that biological triplicates (at least) must be prepared for each sample to ensure reliable quantitation, which thus triples the cost per sample. Proteomics promises to be very useful for the dinoflagellates, for which transcriptional control may not be the primary means of regulating gene expression. However, the techniques lend themselves to a plethora of other microorganisms, opening the doors to extensive analyses of protein levels and post translational modifications under different environmental conditions for entire ecosystems.

Acknowledgements

We are indebted to Dr. AB Mayfield for his thoughtful and thorough critique of the manuscript. We gratefully acknowledge financial support by the National Science and Engineering Research Council of Canada (Grant number 171382-03 to DM) and the Hong Kong Polytechnic University (grant number 4-ZZEG to SCLL).

Figure Legends

Figure 1 Some unusual features of the dinoflagellate *Lingulodinium polyedra*. (A) A confocal micrograph of a night-phase cell shows numerous blue scintillons (sc) and pale orange chloroplasts (p). The location of the nucleus (n) appears black because plastids are excluded from this compartment. The entire cell is 35-40 µm in diameter. (B) A scintillon observed by transmission electron microscopy (TEM) is seen as an electron dense sphere in the cytoplasm in contact with the vacuole. (C) Chloroplasts in day-phase cells have two compartments, one rich in thylakoid membranes (p) and the other, termed a pyrenoid, which is enriched in Rubisco (py). This section was labeled with an anti-Rubisco, and the antibody visualized using 20 nm gold-conjugated secondary antibodies. (D) A part of the nucleus (n) showing the permanently condensed chromosomes.

Figure 2 Peptide identification by MS fragmentation. (A) A typical liquid chromatogram (LC) profile for a complex mixture of tryptic peptides prepared from a dinoflagellate protein extract. At the time shown by the dotted line, the sample is fed into the first mass spectrometer for precursor fragment detection (MS^1). (B) The MS spectrum of the components found at the time indicated in A, shows three main mass peaks. The peak with the highest m/z ratio has, at higher resolution, a pattern of masses consistent with a +2 ion (inset). (C) The peak containing the +2 ion in B is then fragmented in the second MS, and the masses of the different observed fragments (red lines) are compared with the masses predicted for a peptide from the peridinin chlorophyll *a* protein (PCP). The masses predicted for the y series and the b series are shown by the vertical blue and green lines, respectively. (D) A graphical display of peptides in a manner analogous to that of a two-dimensional electrophoresis gel. Tools developed for picking and comparing spots between different gels can be applied to these m/z vs. retention time plots to uncover peptide precursor peaks and hence protein identification and quantitation. Each spot, if examined at higher magnification, would have a series of mass peaks similar to those in the inset in B. (E) A region containing a peptide derived from the bioluminescence-related protein LBP (boxed in red), examined at four times over a light-dark (LD) cycle. LBP levels are low during the day (LD 0 and LD 6) and high at night (LD 12 and LD 18).

Figure 3 Phosphoprotein and phosphopeptide enrichments.

Phosphopeptides can be enriched in a sample destined for MS analysis either by enriching phosphorylated peptides from a tryptic digest (left-hand side) or by first enriching phosphorylated proteins from which tryptic peptides are produced (right-hand side).

Figure 4 Labeling strategies for quantitative proteomics.

Cells in culture can be labeled (red) as a pulse, such as with labeled amino acids, in which case only newly synthesized proteins become labeled. The relative amounts of labeled and unlabeled peptides is indicative of protein synthesis rates. Cells can be labeled for many generations, such as with $^{15}NO_3^-$, in which case all the proteins are uniformly labeled. The amount of the labeled peptides can be measured relative to a second unlabeled sample that is mixed in with the first before analysis. Unlabeled proteins can also be labeled following digestion, and the use of different mass tags permits up to ten different samples to be analyzed concurrently and their abundances compared.

610 **References**

611

612 Akimoto, H., Kinumi, T., Ohmiya, Y., 2005. Circadian rhythm of a TCA cycle enzyme is apparently
613 regulated at the translational level in the dinoflagellate Lingulodinium polyedrum. Journal of
614 biological rhythms 20(6), 479-489.
615 Akimoto, H., Wu, C., Kinumi, T., Ohmiya, Y., 2004. Biological rhythmicity in expressed proteins
616 of the marine dinoflagellate *Lingulodinium polyedrum* demonstrated by chronological
617 proteomics. Biochem Biophys Res Commun 315(2), 306-312.
618 Bantscheff, M., Lemeer, S., Savitski, M.M., Kuster, B., 2012. Quantitative mass spectrometry in
619 proteomics: critical review update from 2007 to the present. Anal Bioanal Chem 404(4), 939-
620 965.
621 Barshis, D.J., Ladner, J.T., Oliver, T.A., Palumbi, S.R., 2014. Lineage-specific transcriptional
622 profiles of Symbiodinium spp. unaltered by heat stress in a coral host. Mol Biol Evol 31(6), 1343-
623 1352.
624 Baumgarten, S., Bayer, T., Aranda, M., Liew, Y.J., Carr, A., Micklem, G., Voolstra, C.R., 2013.
625 Integrating microRNA and mRNA expression profiling in Symbiodinium microadriaticum, a
626 dinoflagellate symbiont of reef-building corals. BMC Genomics 14, 704.
627 Bayer, T., Aranda, M., Sunagawa, S., Yum, L.K., Desalvo, M.K., Lindquist, E., Coffroth, M.A.,
628 Voolstra, C.R., Medina, M., 2012. Symbiodinium transcriptomes: genome insights into the
629 dinoflagellate symbionts of reef-building corals. PLoS One 7(4), e35269.
630 Beauchemin, M., Morse, D., 2017. A proteomic portrait of dinoflagellate chromatin reveals
631 abundant RNA binding proteins. Chromosoma in press.
632 Bertomeu, T., Hastings, J.W., Morse, D., 2003. Vectorial labeling of dinoflagellate cell surface
633 proteins. J. Phycol. 39, 1254-1260.
634 Bertomeu, T., Rivoal, J., Morse, D., 2007. A dinoflagellate CDK5-like cyclin-dependent kinase.
635 Biol Cell 99(9), 531-540.
636 Bhaud, Y., Geraud, M.L., Ausseil, J., Soyer-Gobillard, M.O., Moreau, H., 1999. Cyclic expression
637 of a nuclear protein in a dinoflagellate. The Journal of eukaryotic microbiology 46(3), 259-267.
638 Blaydes, J.P., Vojtesek, B., Bloomberg, G.B., Hupp, T.R., 2000. The Development and Use of
639 Phospho-Specific Antibodies to Study Protein Phosphorylation, pp. 177-189.
640 Brunelle, S.A., van Dolah, F.M., 2011. Post-transcriptional regulation of S-phase genes in the
641 dinoflagellate, Karenia brevis. The Journal of eukaryotic microbiology 58(4), 373-382.
642 Castello, A., Fischer, B., Hentze, M.W., Preiss, T., 2013. RNA-binding proteins in Mendelian
643 disease. Trends Genet 29(5), 318-327.
644 Chan, K.L., New, D., Ghandhi, S., Wong, F., Lam, C.M., Wong, J.T., 2002a. Transcript levels of the
645 eukaryotic translation initiation factor 5A gene peak at early G(1) phase of the cell cycle in the
646 dinoflagellate Crypthecodinium cohnii. Appl Environ Microbiol 68(5), 2278-2284.
647 Chan, L.L., Hodgkiss, I.J., Wan, J.M.F., Lum, J.H.K., Mak, A.S.C., Sit, W.H., Lo, S.C.L., 2004.
648 Proteomic study of a model causative agent of harmful algal blooms, Prorocentrum triestinum
649 II: the use of differentially expressed protein profiles under different growth phases and growth
650 conditions for bloom prediction. Proteomics 4(10), 3214-3226.

Chan, L.L., Li, X.Y., Sit, W.H., Lam, P.K.S., Leung, K.M.Y., 2012. Development of theca specific antisera for the profiling of cell surface proteins in the marine toxic dinoflagellate genus Alexandrium Halim. Harmful Algae 16, 58-62.

Chan, L.L., Lo, S.C.L., Hodgkiss, I.J., 2002b. Proteomic study of a model causative agent of harmful red tide, Prorocentrum triestinum I: Optimization of sample preparation methodologies for analyzing with two-dimensional electrophoresis. Proteomics 2(9), 1169-1186.

Chan, Y.H., Wong, J.T., 2007. Concentration-dependent organization of DNA by the dinoflagellate histone-like protein HCc3. Nucleic Acids Res 35(8), 2573-2583.

Cieśla, J., Frączyk, T., Rode, W., 2011. Phosphorylation of basic amino acid residues in proteins: important but easily missed. Acta Biochimica Polonica 58(2), 137-148.

de Vargas, C., Audic, S., Henry, N., Decelle, J., Mahe, F., Logares, R., Lara, E., Berney, C., Le Bescot, N., Probert, I., Carmichael, M., Poulain, J., Romac, S., Colin, S., Aury, J.M., Bittner, L., Chaffron, S., Dunthorn, M., Engelen, S., Flegontova, O., Guidi, L., Horak, A., Jaillon, O., Lima-Mendez, G., Lukes, J., Malviya, S., Morard, R., Mulot, M., Scalco, E., Siano, R., Vincent, F., Zingone, A., Dimier, C., Picheral, M., Searson, S., Kandels-Lewis, S., Tara Oceans, C., Acinas, S.G., Bork, P., Bowler, C., Gorsky, G., Grimsley, N., Hingamp, P., Iudicone, D., Not, F., Ogata, H., Pesant, S., Raes, J., Sieracki, M.E., Speich, S., Stemmann, L., Sunagawa, S., Weissenbach, J., Wincker, P., Karsenti, E., 2015. Ocean plankton. Eukaryotic plankton diversity in the sunlit ocean. Science 348(6237), 1261605.

Desjardins, M., Morse, D., 1993. The polypeptide components of scintillons, the bioluminescence organelles of the dinoflagellate *Gonyaulax polyedra*. Biochem. Cell. Biol. 71(3-4), 176-182.

Dunham, W.H., Mullin, M., Gingras, A.C., 2012. Affinity-purification coupled to mass spectrometry: basic principles and strategies. Proteomics 12(10), 1576-1590.

Eliuk, S., Makarov, A., 2015. Evolution of Orbitrap Mass Spectrometry Instrumentation. Annu Rev Anal Chem (Palo Alto Calif) 8, 61-80.

Fagan, T., Morse, D., Hastings, J., 1999. Circadian synthesis of a nuclear encoded chloroplast Glyceraldehyde-3-phosphate dehydrogenase in the dinoflagellate *Gonyaulax polyedra* is translationally controlled. Biochemistry 38, 7689-7695.

Field, C.B., Behrenfeld, M.J., Randerson, J.T., Falkowski, P., 1998. Primary production of the biosphere: integrating terrestrial and oceanic components. Science 281(5374), 237-240.

Fischer, E.H., 2013. Cellular regulation by protein phosphorylation. Biochemical and biophysical research communications 430(2), 865-867.

Fogel, M., Schmitter, R.E., Hastings, J.W., 1972. On the physical identity of scintillons: bioluminescent particles in Gonyaulax polyedra. J Cell Sci 11(1), 305-317.

Foy, M., Harney, D.F., Wynne, K., Maguire, P.B., 2007. Enrichment of Phosphotyrosine Proteome of Human Platelets by Immunoprecipitation, pp. 313-318.

Gavelis, G.S., Hayakawa, S., White, R.A., 3rd, Gojobori, T., Suttle, C.A., Keeling, P.J., Leander, B.S., 2015. Eye-like ocelloids are built from different endosymbiotically acquired components. Nature 523(7559), 204-207.

Ghaemmaghami, S., Huh, W.K., Bower, K., Howson, R.W., Belle, A., Dephoure, N., O'Shea, E.K., Weissman, J.S., 2003. Global analysis of protein expression in yeast. Nature 425(6959), 737-741.

694 Gomez, F., 2012. A quantitative review of the lifestyle, habitat and trophic diversity of
695 dinoflagellates (Dinoflagellata, Alveolata). Systematics and Biodiversity 10, 267-275.
696 Gornik, S.G., Ford, K.L., Mulhern, T.D., Bacic, A., McFadden, G.I., Waller, R.F., 2012. Loss of
697 nucleosomal DNA condensation coincides with appearance of a novel nuclear protein in
698 dinoflagellates. Curr Biol 22(24), 2303-2312.
699 Gygi, S.P., Rist, B., Gerber, S.A., Turecek, F., Gelb, M.H., Aebersold, R., 1999. Quantitative
700 analysis of complex protein mixtures using isotope-coded affinity tags. Nat Biotechnol 17(10),
701 994-999.
702 Hackett, J.D., Wisecaver, J.H., Brosnahan, M.L., Kulis, D.M., Anderson, D.M., Bhattacharya, D.,
703 Plumley, F.G., Erdner, D.L., 2013. Evolution of saxitoxin synthesis in cyanobacteria and
704 dinoflagellates. Mol Biol Evol 30(1), 70-78.
705 Hebert, A.S., Richards, A.L., Bailey, D.J., Ulbrich, A., Coughlin, E.E., Westphall, M.S., Coon, J.J.,
706 2014. The one hour yeast proteome. Mol Cell Proteomics 13(1), 339-347.
707 Hofmann, E., Wrench, P.M., Sharples, F.P., Hiller, R.G., Welte, W., Diederichs, K., 1996.
708 Structural basis of light harvesting by carotenoids: peridinin- chlorophyll-protein from
709 *Amphidinium carterae*. Science 272(5269), 1788-1791.
710 Ibáñez-Vea, M., Kempf, S.J., Larsen, M.R., 2017. Characterization of the Phosphoproteome and
711 Sialoproteome in Brain Tissues by Mass Spectrometry #, T Current Proteomic Approaches
712 Applied to Brain Function, pp. 191-206.
713 Iliuk, A.B., Martin, V.A., Alicie, B.M., Geahlen, R.L., Tao, W.A., 2010. In-depth analyses of kinase-
714 dependent tyrosine phosphoproteomes based on metal ion-functionalized soluble
715 nanopolymers. Molecular & cellular proteomics 9(10), 2162-2172.
716 Ishihama, Y., Oda, Y., Tabata, T., Sato, T., Nagasu, T., Rappsilber, J., Mann, M., 2005.
717 Exponentially modified protein abundance index (emPAI) for estimation of absolute protein
718 amount in proteomics by the number of sequenced peptides per protein. Molecular & Cellular
719 Proteomics 4(9), 1265-1272.
720 Iwakawa, H.O., Tomari, Y., 2015. The Functions of MicroRNAs: mRNA Decay and Translational
721 Repression. Trends Cell Biol 25(11), 651-665.
722 Janouskovec, J., Gavelis, G.S., Burki, F., Dinh, D., Bachvaroff, T.R., Gornik, S.G., Bright, K.J.,
723 Imanian, B., Strom, S.L., Delwiche, C.F., Waller, R.F., Fensome, R.A., Leander, B.S., Rohwer, F.L.,
724 Saldarriaga, J.F., 2017. Major transitions in dinoflagellate evolution unveiled by
725 phylotranscriptomics. Proc Natl Acad Sci U S A 114(2), E171-E180.
726 Johnson, C.H., Roeber, J.F., Hastings, J.W., 1984. Circadian changes in enzyme concentration
727 account for rhythm of enzyme activity in *Gonyaulax*. Science 223, 1428-1430.
728 Jones, G.D., Williams, E.P., Place, A.R., Jagus, R., Bachvaroff, T.R., 2015. The alveolate
729 translation initiation factor 4E family reveals a custom toolkit for translational control in core
730 dinoflagellates. BMC Evol Biol 15, 14.
731 Keeling, P.J., 2010. The endosymbiotic origin, diversification and fate of plastids. Philosophical
732 transactions of the Royal Society of London. Series B, Biological sciences 365(1541), 729-748.
733 Kong, H.K., Wong, M.H., Chan, H.M., Lo, S.C., 2013. Chronic exposure of adult rats to low doses
734 of methylmercury induced a state of metabolic deficit in the somatosensory cortex. J Proteome
735 Res 12(11), 5233-5245.

736 Krueger, T., Fisher, P.L., Becker, S., Pontasch, S., Dove, S., Hoegh-Guldberg, O., Leggat, W., Davy,
737 S.K., 2015. Transcriptomic characterization of the enzymatic antioxidants FeSOD, MnSOD, APX
738 and KatG in the dinoflagellate genus Symbiodinium. BMC Evol Biol 15, 48.
739 Lee, F.W.-F., Lo, S.C.-L., 2008. The use of Trizol reagent (phenol/guanidine isothiocyanate) for
740 producing high quality two-dimensional gel electrophoretograms (2-DE) of dinoflagellates.
741 Journal of microbiological methods 73(1), 26-32.
742 Lee, F.W.F., Ho, K.C., Mak, Y.L., Lo, S.C.L., 2011. Authentication of the proteins expression
743 profiles (PEPs) identification methodology in a bloom of Karenia digitata, the most damaging
744 harmful algal bloom causative agent in the history of Hong Kong. Harmful Algae 12, 1-10.
745 Leggat, W., Seneca, F., Wasmund, K., Ukani, L., Yellowlees, D., Ainsworth, T.D., 2011.
746 Differential responses of the coral host and their algal symbiont to thermal stress. PLoS One
747 6(10), e26687.
748 Lin, X., Zhang, H., Huang, B., Lin, S., 2011. Alkaline Phosphatase Gene Sequence And
749 Transcriptional Regulation By Phosphate Limitation In Amphidinium Carterae (Dinophyceae)(1).
750 J Phycol 47(5), 1110-1120.
751 Macpherson, G.R., Burton, I.W., LeBlanc, P., Walter, J.A., Wright, J.L., 2003. Studies of the
752 biosynthesis of DTX-5a and DTX-5b by the dinoflagellate Prorocentrum maculosum:
753 regiospecificity of the putative Baeyer-Villigerase and insertion of a single amino acid in a
754 polyketide chain. J Org Chem 68(5), 1659-1664.
755 Mann, M., Ong, S.E., Gronborg, M., Steen, H., Jensen, O.N., Pandey, A., 2002. Analysis of protein
756 phosphorylation using mass spectrometry: deciphering the phosphoproteome. Trends
757 Biotechnol 20(6), 261-268.
758 Manning, G., Whyte, D.B., Martinez, R., Hunter, T., Sudarsanam, S., 2002. The protein kinase
759 complement of the human genome. Science 298(5600), 1912-1934.
760 Mayfield, A.B., Chan, P.H., Putnam, H.M., Chen, C.S., Fan, T.Y., 2012. The effects of a variable
761 temperature regime on the physiology of the reef-building coral Seriatopora hystrix: results
762 from a laboratory-based reciprocal transplant. J Exp Biol 215(Pt 23), 4183-4195.
763 Mayfield, A.B., Hsiao, Y.Y., Chen, H.K., Chen, C.S., 2014a. Rubisco Expression in the
764 Dinoflagellate *Symbiodinium* sp. Is Influenced by Both Photoperiod and Endosymbiotic Lifestyle.
765 Mar Biotechnol 16, 371-384.
766 Mayfield, A.B., Wang, L.H., Tang, P.C., Fan, T.Y., Hsiao, Y.Y., Tsai, C.L., Chen, C.S., 2011. Assessing
767 the impacts of experimentally elevated temperature on the biological composition and
768 molecular chaperone gene expression of a reef coral. PLoS One 6(10), e26529.
769 Mayfield, A.B., Wang, Y.B., Chen, C.S., Chen, S.H., Lin, C.Y., 2016. Dual-compartmental
770 transcriptomic+ proteomic analysis of a marine endosymbiosis exposed to environmental
771 change. Molecular ecology 25(23), 5944-5958.
772 Mayfield, A.B., Wang, Y.B., Chen, C.S., Lin, C.Y., Chen, S.H., 2014b. Compartment-specific
773 transcriptomics in a reef-building coral exposed to elevated temperatures. Mol Ecol 23(23),
774 5816-5830.
775 Melo-Braga, M.N., Ibáñez-Vea, M., Larsen, M.R., Kulej, K., 2015. Comprehensive protocol to
776 simultaneously study protein phosphorylation, acetylation, and N-linked sialylated
777 glycosylation. Proteomic Profiling: Methods and Protocols, 275-292.
778 Mendiola, L.R., Price, C.A., Guillard, R.R., 1966. Isolation of nuclei from a marine dinoflagellate.
779 Science 153(3744), 1661-1663.

780 Milos, P., Morse, D., Hastings, J.W., 1990. Circadian control over synthesis of many *Gonyaulax*
781 proteins is at a translational level. Naturwiss. 77, 87-89.
782 Morey, J.S., Monroe, E.A., Kinney, A.L., Beal, M., Johnson, J.G., Hitchcock, G.L., van Dolah, F.M.,
783 2011. Transcriptomic response of the red tide dinoflagellate, Karenia brevis, to nitrogen and
784 phosphorus depletion and addition. BMC Genomics 12, 346.
785 Morse, D., Milos, P.M., Roux, E., Hastings, J.W., 1989. Circadian regulation of bioluminescence
786 in *Gonyaulax* involves translational control. Proc. Natl. Acad. Sci. U.S.A. 86(1), 172-176.
787 Morse, D., Salois, P., Markovic, P., Hastings, J.W., 1995. A nuclear encoded form II rubisco in
788 dinoflagellates. Science 268, 1622-1624.
789 Moustafa, A., Evans, A.N., Kulis, D.M., Hackett, J.D., Erdner, D.L., Anderson, D.M., Bhattacharya,
790 D., 2010. Transcriptome profiling of a toxic dinoflagellate reveals a gene-rich protist and a
791 potential impact on gene expression due to bacterial presence. PLoS One 5(3), e9688.
792 Nagaraj, N., Wisniewski, J.R., Geiger, T., Cox, J., Kircher, M., Kelso, J., Paabo, S., Mann, M., 2011.
793 Deep proteome and transcriptome mapping of a human cancer cell line. Molecular systems
794 biology 7, 548.
795 Nassoury, N., Fritz, L., Morse, D., 2001. Circadian changes in ribulose-1,5-bisphosphate
796 carboxylase/oxygenase distribution inside individual chloroplasts can account for the rhythm in
797 dinoflagellate carbon fixation. Plant Cell 13(4), 923-934.
798 Nelson, C.J., Huttlin, E.L., Hegeman, A.D., Harms, A.C., Sussman, M.R., 2007. Implications of
799 15N-metabolic labeling for automated peptide identification in Arabidopsis thaliana.
800 Proteomics 7(8), 1279-1292.
801 Okamoto, O., Robertson, D., Fagan, T., Hastings, J., Colepicolo, P., 2001. Different regulatory
802 mechanisms modulate the expression of a dinoflagellate iron-superoxide dismutase. J. Biol.
803 Chem. 276, 19989-19993.
804 Okamoto, O.K., Hastings, J.W., 2003. Novel dinoflagellate circadian-clock genes identified
805 through microarray analysis of a phase shifted clock. J. Phycology 39, 1-9.
806 Padula, M.P., Berry, I.J., MB, O.R., Raymond, B.B., Santos, J., Djordjevic, S.P., 2017. A
807 Comprehensive Guide for Performing Sample Preparation and Top-Down Protein Analysis.
808 Proteomes 5(2).
809 Pappin, D.J., Hojrup, P., Bleasby, A.J., 1993. Rapid identification of proteins by peptide-mass
810 fingerprinting. Curr Biol 3(6), 327-332.
811 Peng, S.E., Chen, W.N.U., Chen, H.K., Lu, C.Y., Mayfield, A.B., Fang, L.S., Chen, C.S., 2011. Lipid
812 bodies in coral–dinoflagellate endosymbiosis: proteomic and ultrastructural studies. Proteomics
813 11, 3540-3555.
814 Putnam, H.M., Mayfield, A.B., Fan, T.Y., Chen, C.S., Gates, R.D., 2013. The physiological and
815 molecular responses of larvae from the reef-building coral *Pocillopora damicornis* exposed to
816 near-future increases in temperature and pCO2. Mar Biol 160, 2157-2173.
817 Rizzo, P.J., Jones, M., Ray, S.M., 1982. Isolation and properties of isolated nuclei from the
818 Florida red tide dinoflagellate Gymnodinium breve (Davis). The Journal of protozoology 29(2),
819 217-222.
820 Rosic, N.N., Pernice, M., Dove, S., Dunn, S., Hoegh-Guldberg, O., 2011. Gene expression profiles
821 of cytosolic heat shock proteins Hsp70 and Hsp90 from symbiotic dinoflagellates in response to
822 thermal stress: possible implications for coral bleaching. Cell Stress Chaperones 16(1), 69-80.

Ross, P.L., Huang, Y.N., Marchese, J.N., Williamson, B., Parker, K., Hattan, S., Khainovski, N., Pillai, S., Dey, S., Daniels, S., Purkayastha, S., Juhasz, P., Martin, S., Bartlet-Jones, M., He, F., Jacobson, A., Pappin, D.J., 2004. Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents. Mol Cell Proteomics 3(12), 1154-1169.

Rowan, R., Whitney, S.M., Fowler, A., Yellowlees, D., 1996. Rubisco in marine symbiotic dinoflagellates: form II enzymes in eukaryotic oxygenic phototrophs encoded by a nuclear multigene family. Plant Cell 8(3), 539-553.

Roy, S., Beauchemin, M., Dagenais-Bellefeuille, S., Letourneau, L., Cappadocia, M., Morse, D., 2014. The *Lingulodinium* circadian system lacks rhythmic changes in transcript abundance. BMC biology 12(1), 107.

Roy, S., Morse, D., 2012. A full suite of histone and histone modifying genes are transcribed in the dinoflagellate *Lingulodinium*, PLoS One.

Roy, S., Morse, D., 2014. The dinoflagellate *Lingulodinium* has predicted casein kinase 2 sites in many RNA binding proteins. Protist 165(3), 330-342.

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., 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(5), 202-214.

Sheng, J., Malkiel, E., Katz, J., Adolf, J.E., Place, A.R., 2010. A dinoflagellate exploits toxins to immobilize prey prior to ingestion. Proc Natl Acad Sci U S A 107(5), 2082-2087.

Shi, X., Zhang, H., Lin, S., 2013. Tandem repeats, high copy number and remarkable diel expression rhythm of form II RuBisCO in Prorocentrum donghaiense (Dinophyceae). PLoS One 8(8), e71232.

Siegelman, H., Kycia, J., Haxo, F., 1977. Peridinin chlorophyll a proteins of dinoflagellate algae. Brookhaven Nat. Symp. 28, 162-169.

Soyer, M.O., Haapala, O.K., 1974. Electron microscopy of RNA in dinoflagellate chromosomes. Histochemistry 42, 239-246.

Spector, D., 1984. Dinoflagellates. Academic Press, New York, p. 545.

Sun, S., Wong, J.T., Liu, M., Dong, F., 2012. Counterion-mediated decompaction of liquid crystalline chromosomes. DNA Cell Biol 31(12), 1657-1664.

Taylor, F.J.R., Hoppenrath, M., Saldarriaga, J.F., 2008. Dinoflagellate diversity and distribution. Biodivers Conserv 17, 407-418.

ten Lohuis, M.R., Miller, D.J., 1998. Light-regulated transcription of genes encoding peridinin chlorophyll a proteins and the major intrinsic light-harvesting complex proteins in the dinoflagellate amphidinium carterae hulburt (Dinophycae). Changes In cytosine methylation accompany photoadaptation. Plant Physiol 117(1), 189-196.

Thingholm, T.E., Larsen, M.R., 2016a. Phosphopeptide enrichment by immobilized metal affinity chromatography. Phospho-Proteomics: Methods and Protocols, 123-133.

Thingholm, T.E., Larsen, M.R., 2016b. Sequential elution from IMAC (SIMAC): An efficient method for enrichment and separation of mono-and multi-phosphorylated peptides. Phospho-Proteomics: Methods and Protocols, 147-160.

Thingholm, T.E., Larsen, M.R., 2016c. The use of titanium dioxide for selective enrichment of phosphorylated peptides. Phospho-Proteomics: Methods and Protocols, 135-146.

867 Thompson, A., Schafer, J., Kuhn, K., Kienle, S., Schwarz, J., Schmidt, G., Neumann, T., Johnstone,
868 R., Mohammed, A.K., Hamon, C., 2003. Tandem mass tags: a novel quantification strategy for
869 comparative analysis of complex protein mixtures by MS/MS. Anal Chem 75(8), 1895-1904.
870 Tse, S., PK, Lo, S.C., 2017a. Comparative proteomic studies of a Scrippsiella acuminata bloom
871 with its laboratory-grown culture using a 15 N-metabolic labeling approach. Harmful Algae 67,
872 26-35.
873 Tse, S.P., Beauchemin, M., Morse, D., Lo, S.C., 2017. Refining Transcriptome Gene Catalogs by
874 MS-Validation of Expressed Proteins. Proteomics.
875 Tse, S.P.K., Beauchemin, M., Morse, D., Lo, S.C.L., 2018. Refining transcriptome gene catalogs by
876 MS-validation of expressed proteins. Protemics 18, DOI: 10.1002/pmic.201700271.
877 Tse, S.P.K., Lo, S.C.L., 2017b. Comparative proteomic studies of a Scrippsiella acuminata bloom
878 with its laboratory-grown culture using a 15 N-metabolic labeling approach. Harmful Algae 67,
879 26-35.
880 Tyanova, S., Temu, T., Cox, J., 2016. The MaxQuant computational platform for mass
881 spectrometry-based shotgun proteomics. Nat Protoc 11(12), 2301-2319.
882 van Dolah, F.M., 2000. Marine algal toxins: origins, health effects, and their increased
883 occurrence. Environ Health Perspect 108 Suppl 1, 133-141.
884 van Dolah, F.M., Lidie, K.B., Morey, J.S., Brunelle, S.A., Ryan, J.C., Monroe, E.A., Haynes, B.L.,
885 2007. Microarray analysis of diurnal and circadian regulated genes in the florida red-tide
886 dinoflagellate *Karenia brevis* (Dinophyceae). J. Phycol. 43, 741-752.
887 Vogel, C., Abreu Rde, S., Ko, D., Le, S.Y., Shapiro, B.A., Burns, S.C., Sandhu, D., Boutz, D.R.,
888 Marcotte, E.M., Penalva, L.O., 2010. Sequence signatures and mRNA concentration can explain
889 two-thirds of protein abundance variation in a human cell line. Molecular systems biology 6,
890 400.
891 Wang, D.-Z., Lin, L., Chan, L.L., Hong, H.-S., 2009. Comparative studies of four protein
892 preparation methods for proteomic study of the dinoflagellate Alexandrium sp. using two-
893 dimensional electrophoresis. Harmful Algae 8(5), 685-691.
894 Wang, D.Z., 2008. Neurotoxins from marine dinoflagellates: a brief review. Mar Drugs 6(2), 349-
895 371.
896 Wang, D.Z., Gao, Y., Lin, L., Hong, H.S., 2013. Comparative proteomic analysis reveals proteins
897 putatively involved in toxin biosynthesis in the marine dinoflagellate Alexandrium catenella.
898 Mar Drugs 11(1), 213-232.
899 Wang, D.Z., Zhang, H., Zhang, Y., Zhang, S.F., 2014. Marine dinoflagellate proteomics: current
900 status and future perspectives. J Proteomics 105, 121-132.
901 Wang, Y., Jensen, L., Hojrup, P., Morse, D., 2005a. Synthesis and degradation of dinoflagellate
902 plastid-encoded psbA proteins are light-regulated, not circadian-regulated. Proc. Natl. Acad. Sci.
903 U.S.A. 102(8), 2844-2849.
904 Wang, Y., MacKenzie, T., Morse, D., 2005b. Purification of plastids from the dinoflagellate
905 *Lingulodinium*. Mar Biotechnol (NY) 7(6), 659-668.
906 Wang, Y., Morse, D., 2006. Rampant polyuridylylation of plastid gene transcripts in the
907 dinoflagellate Lingulodinium. Nucleic Acids Res 34(2), 613-619.
908 Weston, A.J., Dunlap, W.C., Shick, J.M., Klueter, A., Iglic, K., Vukelic, A., Starcevic, A., Ward, M.,
909 Wells, M.L., Trick, C.G., Long, P.F., 2012. A profile of an endosymbiont-enriched fraction of the

910 coral Stylophora pistillata reveals proteins relevant to microbial-host interactions. Mol Cell
911 Proteomics 11(6), M111 015487.
912 Yan, Y., Zheng, Z., Deng, C., Li, Y., Zhang, X., Yang, P., 2013. Hydrophilic polydopamine-coated
913 graphene for metal ion immobilization as a novel immobilized metal ion affinity
914 chromatography platform for phosphoproteome analysis. Analytical chemistry 85(18), 8483-
915 8487.
916 Yang, I., John, U., Beszteri, S., Glockner, G., Krock, B., Goesmann, A., Cembella, A.D., 2010.
917 Comparative gene expression in toxic versus non-toxic strains of the marine dinoflagellate
918 Alexandrium minutum. BMC Genomics 11, 248.
919 Yates, J.R., 3rd, Gilchrist, A., Howell, K.E., Bergeron, J.J., 2005. Proteomics of organelles and
920 large cellular structures. Nat Rev Mol Cell Biol 6(9), 702-714.
921 Yu, L.-R., Veenstra, T., 2013. Phosphopeptide enrichment using offline titanium dioxide columns
922 for phosphoproteomics. Proteomics for Biomarker Discovery, 93-103.
923 Zhang, H., Wang, D.-Z., Xie, Z.-X., Zhang, S.-F., Wang, M.-H., Lin, L., 2015a. Comparative
924 proteomics reveals highly and differentially expressed proteins in field-collected and
925 laboratory-cultured blooming cells of the diatom Skeletonema costatum. Environmental
926 Microbiology 17(10), 3976-3991.
927 Zhang, S.F., Zhang, Y., Xie, Z.X., Zhang, H., Lin, L., Wang, D.Z., 2015b. iTRAQ-based quantitative
928 proteomic analysis of a toxigenic dinoflagellate Alexandrium catenella and its non-toxic mutant.
929 Proteomics 15(23-24), 4041-4050.
930 Zhang, Y., Zhang, S.F., Lin, L., Wang, D.Z., 2014. Comparative transcriptome analysis of a toxin-
931 producing dinoflagellate Alexandrium catenella and its non-toxic mutant. Mar Drugs 12(11),
932 5698-5718.
933 Zhang, Y.J., Zhang, S.F., He, Z.P., Lin, L., Wang, D.Z., 2015c. Proteomic analysis provides new
934 insights into the adaptive response of a dinoflagellate Prorocentrum donghaiense to changing
935 ambient nitrogen. Plant, cell & environment 38(10), 2128-2142.
936 Zhang, Z., Green, B.R., Cavalier-Smith, T., 1999. Single gene circles in dinoflagellate chloroplast
937 genomes. Nature 400(6740), 155-159.
938 Zybailov, B., Coleman, M.K., Florens, L., Washburn, M.P., 2005. Correlation of relative
939 abundance ratios derived from peptide ion chromatograms and spectrum counting for
940 quantitative proteomic analysis using stable isotope labeling. Analytical chemistry 77(19), 6218-
941 6224.
942

Figure

Click here to download high resolution image

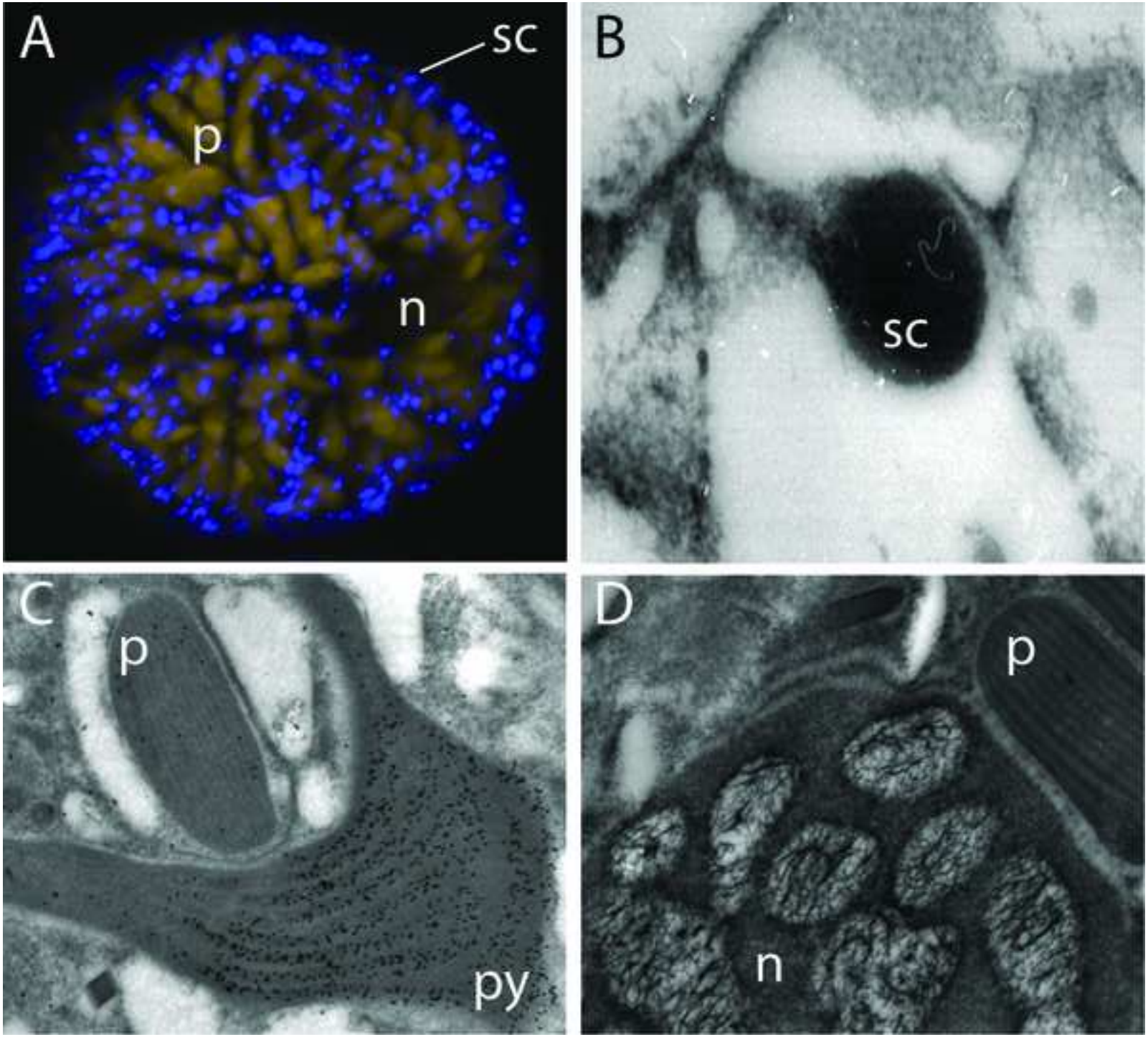

Figure 2

Click here to download high resolution image

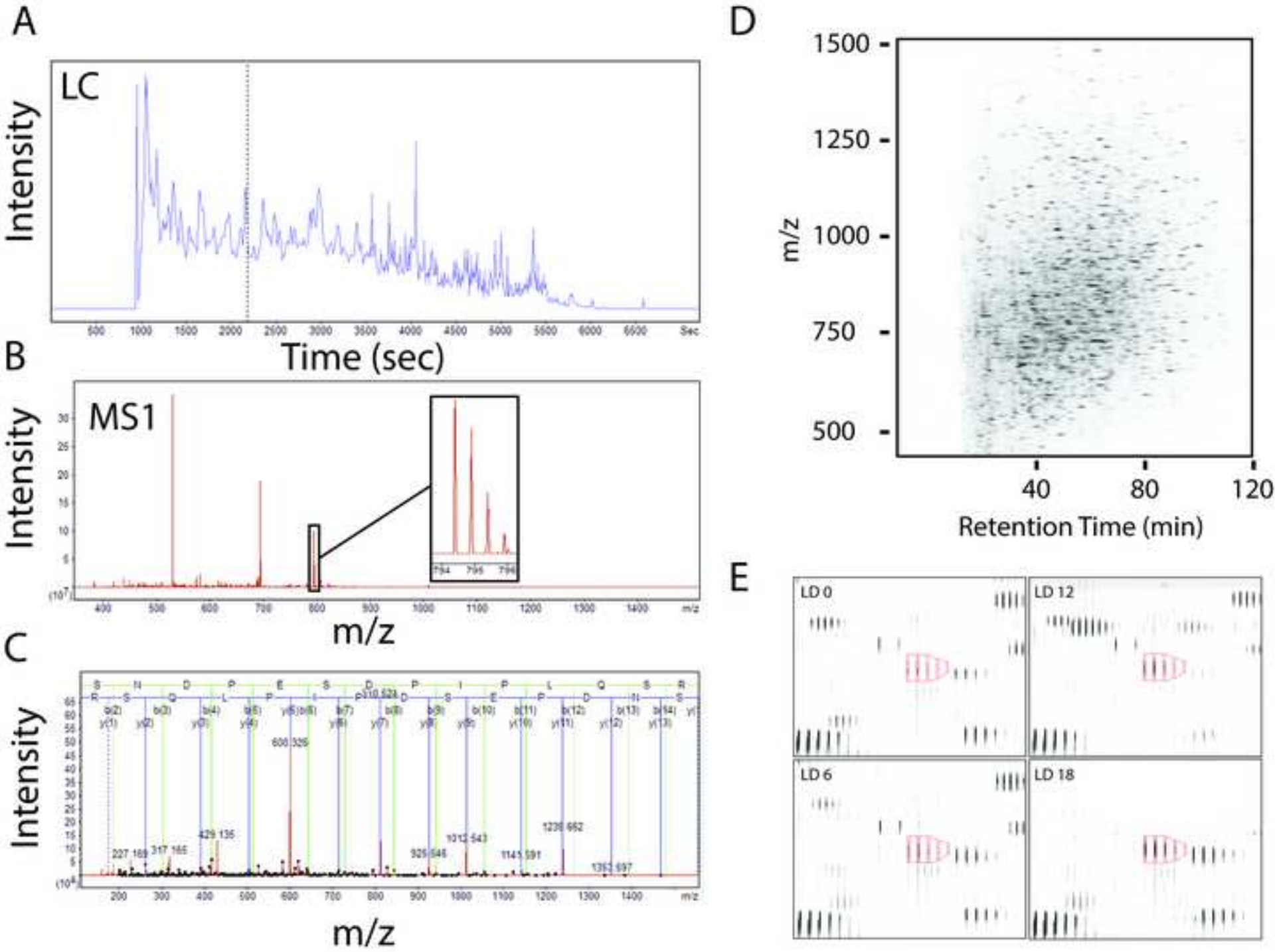

Figure 3

Click here to download high resolution image

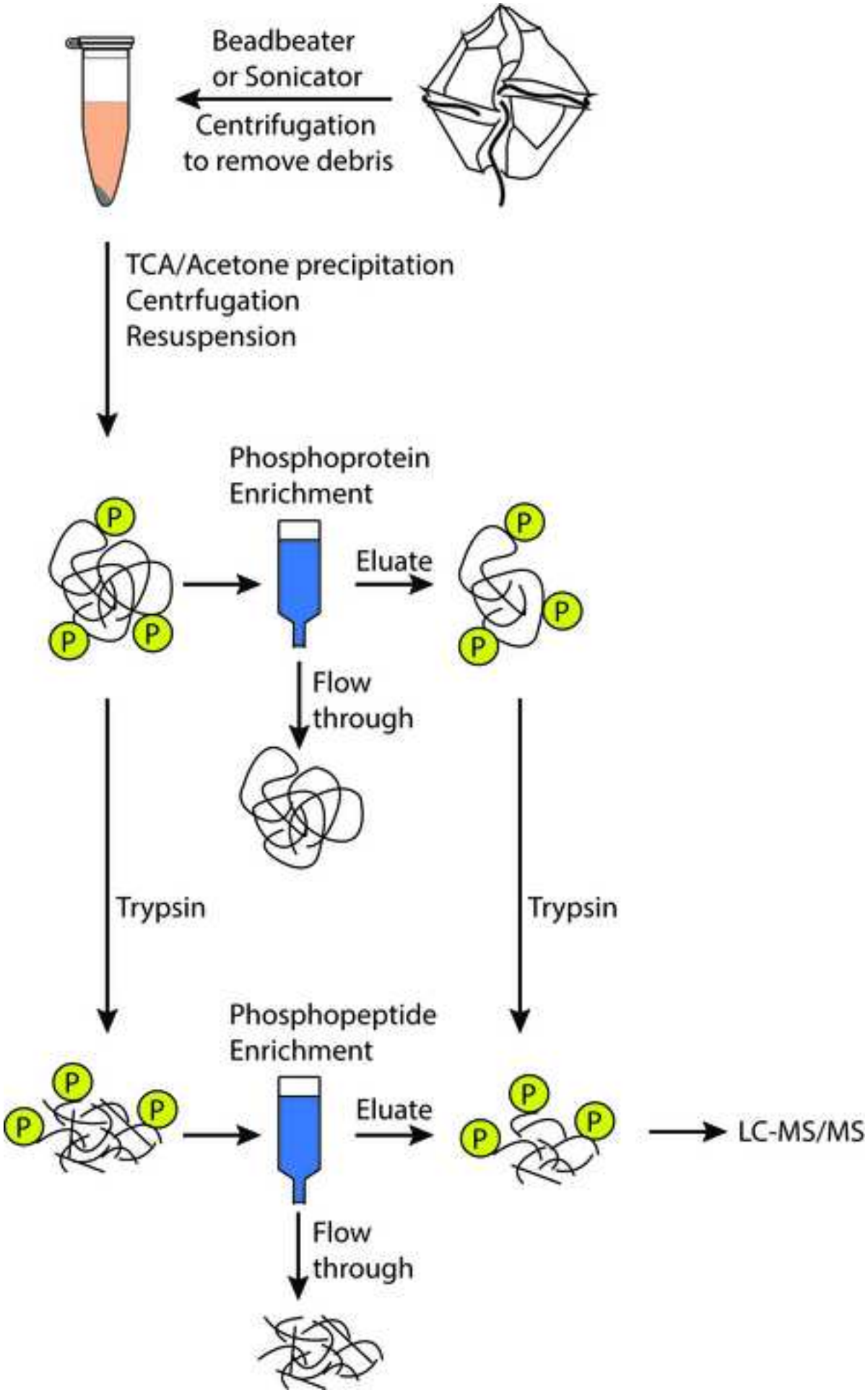

Figure 4

Click here to download high resolution image

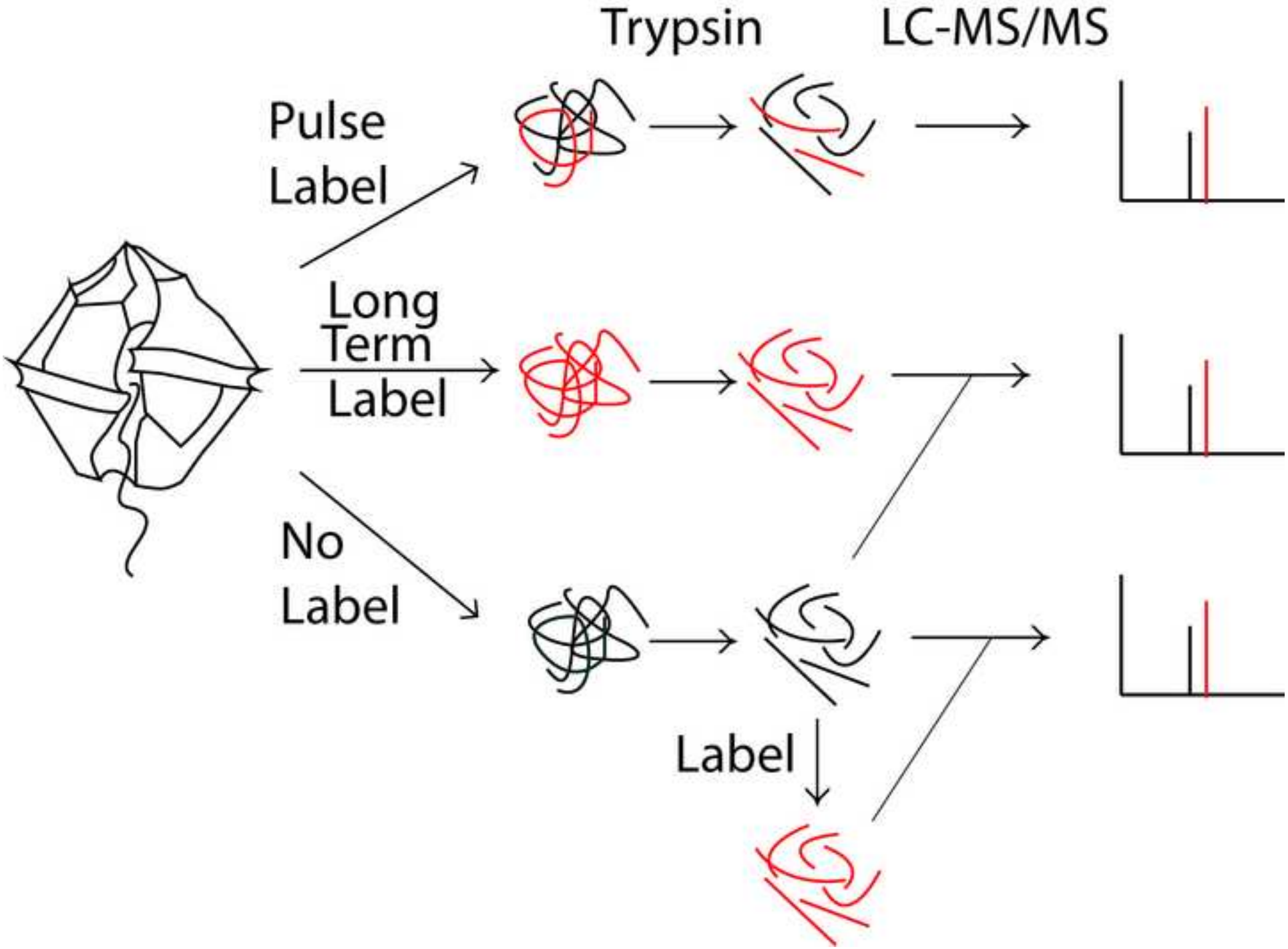

Table 1: Key questions in dinoflagellate biology and the potential of proteomic technologies of solving the corresponding questions.

Questions	Proteomic strategies	Targets
Bloom initiation and progression	Fractionation shotgun proteomics; Metaproteomics of field samples	Protein changes during bloom progression Differences between field and culture samples
Toxin production and role	Fractionation shotgun proteomics; Posttranslational modifications	Differences between toxic and nontoxic strains Protein changes correlated with differences in toxin production
Circadian rhythms	Fractionation shotgun proteomics; Protein targeted proteomics; Posttranslational modifications	Differences over a 24h cycle in light-dark and constant conditions
Species identification In field blooms	Peptide mass fingerprinting	Housekeeping protein fingerprints