Shotgun proteomics as an effective means to describe the oyster gill proteome

ABSTRACT

(to be written this weekend)

INTRODUCTION

Fluctuation in gene and protein expression patterns can be sensitive and specific approaches to provide fundamental information on biological processes. At the transcript level, several methodologies can be used to characterize expression from the gene-centric to systems level, including qPCR (e.g. Griffitt et al. 2006; Stumpp et al. 2011), microarrays (e.g. Todgham and Hofmann 2009; Lockwood et al. 2010), and high-throughput sequencing (HTS) (e.g. Polato et al. 2011; Philipp et al. 2012). The use of HTS technology has exponentially increased the available DNA and transcript information for non-model organisms in recent years. While these results give us an accurate portrayal of changes at the molecular level, it is common that proteins have a more direct role in physiological processes including homeostasis maintenance.

Historically there have been several technical and analytical challenges in characterizing global protein expression. One challenge is the need to have sufficient genomic resources available to describe proteins of interest. Specifically, in most instances protein sequencing will produce a ~20 amino acid fragment that requires a known corresponding gene to aid in annotation. However, the lack of genomic resources have not totally hampered proteomic studies. For example researchers characterized the putative physiological response of *Gillichthys mirabilis* gill tissue exposed to osmotic and temperature stress using two dimensional gel electrophoresis without sequencing any proteins (Kültz & Somero 1996). In another study, researchers used Surface Enhanced Laser Desorption/Ionisation and identified 11 differentially expressed proteins in the gill tissue of *Oncorhynchus mykiss* exposed to zinc stress (Hogstrand et al. 2002). Four proteins were identified based on a combination of their physical properties (i.e. mass and binding) coupled with sequence similarity comparisons with limited teleost protein sequences in the SwissProt database (Hogstrand et al. 2002).

Using predicted protein sequences in closely related species can assist in annotation, but species specific information will provide more accurate results. This is evident in a study on protein expression in pea (*Pisum sativum*) chloroplasts where concurrent cDNA sequencing facilitated a greater number of protein identifications compared to identifications through homology searches with closely related model species (Bräutigam et al. 2008). The reason that species specific information provides such an advantage is due to how modern day protein sequence identification is carried out. The vast majority of high-throughput mass spectrometry proteomics is accomplished by matching observed peptide fragmentation patterns (tandem mass spectra) to theoretical spectra. This is possible because peptides fragment in a predictable manner allowing for theoretical tandem mass spectra to be created *in silico* from a given protein sequence, stressing the importance of the database used. These correlation-based algorithms require the peptide mass (precursor mass) and peptide fragmentation (tandem mass spectrum). Even when employing databases of closely related species a large number of peptides viable tandem mass spectra might not accurately be assigned to a protein since a single amino acid mutation could significantly alter the peptide mass and resulting fragmentation pattern.

As technological advances have continued to increase accessibility of whole transcriptomes and genomes to researchers working on ecological species, there is increasing interest in leveraging these data to carry out proteomic studies. Recently, the Pacific oyster genome was sequenced. Given this resource our object was to quantify the level information and variation across samples in the oyster. There have been a several prior studies examining protein expression in oysters using liquid chromatography coupled with tandem mass spectrometry (MS) with samples separated by two-dimensional gel electrophoresis (2-DE) beforehand. These proteomic techniques have identified specific proteins that are responsible for early developmental changes in *C. gigas* (Huan et al. 2012) and larval *C. gigas* response to elevated pCO2 (ocean acidification) (Dineshram et al. 2012). Similarly, 2-DE methods have been used to identify and sequence proteins that are differentially regulated in a range of physiological situations in adult oyster species. Antioxidant proteins are up-regulated in adult *C. virginica* mantle tissue upon exposure to ocean acidification (Tomanek et al. 2011), high quality oocytes in *C. gigas* have specific protein expression profiles (Corporeau et al. 2012), *C. hongkongensis* expresses both general and specific responses to metal contamination (Liu & Wang 2012), and *Saccostrea glomerata* has differential proteome regulation in response to acid sulfate runoff (Amaral et al. 2012).

 An alternative to 2-DE approaches is to perform shotgun proteomics. Shotgun proteomics is the sequencing of a complex mixture of peptides using liquid chromatography and tandem MS without prior separation (i.e. 2-DE). One of the main advantages of using 2-DE methods is that information on the proteins’ physical properties (mass, isoelectric point) can be used in the protein identification, whereas these empirical data are lost in the strictly tandem MS approaches. However, tandem MS has significantly greater data efficiency than gel-based approaches, even considering the greater cost. The use of shotgun proteomics allows for a greater number of proteins to be rapidly identified from a single cell state providing a more complete metabolic picture of cellular function and acclimation. With the exception of Muralidharan et al. (2012) who used shotgun proteomics to uncover *S. glomerata* hemocyte proteomic responses to metal contamination, there are limited examples in marine invertebrates that have made use of this high-throughput tandem MS approach.

In this study, we used shotgun proteomics to sequence the gill proteome in the Pacific oyster, *Crassostrea gigas*. The objective was to determine the effectiveness of a shotgun proteomics approach using the *C. gigas* genome as a database for protein identification. The proteome was sequenced in four biological replicates, with three technical replicates per oyster. This is the first published shotgun sequencing of the proteome in this species.

METHODS

*Oysters*

 Pacific oysters (*Crassostrea gigas*, 18 months old) were collected in Shelton, Washington, United States. Oysters were transferred to Friday Harbor Laboratories (Friday Harbor, WA) into a flow-through system at 13°C for 6 weeks. Eight 4-L vessels containing 6 oysters each were kept in a water bath with seawater flowing through at 57.5 mL/min. Vessels were cleaned every other day with fresh water and salt water rinses. Oysters were fed Shellfish Diet 1800 (Reed Mariculture, Campbell, CA, USA) throughout the experiment. At the end of 6 weeks, gill tissue was removed from four oysters and immediately flash frozen in liquid nitrogen for proteomic analysis.

*Protein Digestion and Desalting*

 Gill tissue samples (50-100 mg) were homogenized in 50 mM NH4HCO3 (100ul) using RNAse-free plastic pestles. Each homogenized gill sample was sonicated four times with a probe sonicator and stored on dry ice between sonications. After sonication, protein concentrations were measured using the Bradford assay, following the manufacturer’s protocol (Pierce, Thermo Fisher Scientific, Rockford, IL, USA). Urea (36mg) was added to each sample (for a total concentration of 6M) to stabilize peptides. Next, 1.5 M Tris (pH 8.8) (6.6 µl) was added followed by 200 mM TCEP (2.5 µl). Samples were incubated for 1 hour at 37°C on a shaker. To alkylate the proteins, 200 mM iodoacetamide (IAM) (20 µl) was added. Samples were then vortexed, and incubated for 1 hour at room temperature in the dark. To absorb excess IAM, 200 mM dithiolthreitol (20 µl) was added, then samples were vortexed and incubated at room temperature for 1 hour. A volume equal to approximately 100 µg was removed and the remainder was discarded. NH4HCO3 (200 µl of 25 mM) was added to dilute the urea and then HPLC grade MeOH (50 µl) was added to each tube. Trypsin was solubilized in a trypsin dilution buffer (20 µl) (Promega, Madison, WI, USA) and 3 µl of this solution was added to each sample to enzymatically digest the proteins. The samples were incubated overnight at 37°C. The next day, the samples were evaporated on the speed vac to near dryness.

Samples were desalted by passage through a pre-prepared MacroSpin column following the manufacturer’s specifications (The Nest Group, Southborough, MA, USA). After desalting, the remaining solvent was evaporated using a speed vac.

*Liquid Chromatography and Tandem Mass Spectrometry (LC-MS/MS)*

 Mass spectrometry was performed at the University of Washington Proteomics Resource (Seattle, WA, USA). Samples were resuspended in 2% acetonitrile and 0.1% fluoroacetic acid in water (100 µL). Samples were then vortexed to mix and spun down at 15,000 rpm for 10 minutes. The supernatant was aliquoted to autosampler vials. Nano LC separation was performed with a nanoACUITY system (Waters, Milford, MA, USA) interfaced to an LTQ Orbitrap XL mass spectrometer (Thermo Scientific, San Jose, CA, USA). Peptides were trapped on a 100 µm i.d. x 20 mm long pre-column packed with 200 angstrom (5 µm) Magic C18 particles (C18AQ; Michrom, Auburn, CA, USA). For separation, a 75 µm i.d. x 250 mm long analytical column with a laser pulled emitter tip packed with 100 angstrom (5 µm) Magic C18 particles (C18Q; Michrom) was used and analyzed in positive ion mode. For each LC-MS/MS analysis, an estimated amount of 0.5 µg of peptides was loaded onto the pre-column at 2 µL/min in water/acetonitrile (98%/2%) with 0.1% (v/v) formic acid. Peptides were eluted using an acetonitrile gradient flowing at 240 nL/min using mobile phase consisting of the following: Solvent C (water, 0.1% formic acid) and Solvent D (acetonitrile, 0.1% formic acid). The gradient program was as follows: 0 - 1 minute: Solvent C (98%) Solvent D (2%); 1 minute: Solvent C (90%) Solvent D (10%); 90 minutes: Solvent C (65%) Solvent D (35%); 91 - 101 minutes: Solvent C (20%) Solvent D (80%); 102 - 120 minutes, Solvent C (98%) Solvent D (2%). Peptide spectra were acquired by scans in the Orbitrap followed by the ion trap.

*Data acquisition*

 High resolution full precursor ion scans were acquired at 60,000 resolution in the Orbitrap over 400-2000 m/z while six consecutive tandem mass spectra (MS/MS) were acquired by collision induced dissociation (CID) in the Linear ion trap (LTQ). Data dependent ion threshold was set at 5,000 counts for MS/MS, and the maximum allowed ion accumulation times were 400 ms for full scans and 100 ms for MS/MS measurements. The number of ions accumulated was set to 1 million for Orbitrap scans and 10,000 for linear ion trap MS/MS scans. An angiotensin and neurotensin standard was run after every eight injections. Each sample was injected in triplicate in a novel randomized order.

*Protein Identification and Data Analysis*

 Peptide sequence and corresponding protein identification for all mass spectra were carried out using SEQUEST (Eng et al. 1994) and the Crassostrea gigas proteome version 9 (Zhang et al. 2012, <http://dx.doi.org/10.5524/100030>). A DECOY database was created by reversing the *C. gigas* proteome and adding it to the forward database. This was completed in order to determine false positive matches and yielded a false discovery rate of ~0.6%. Search parameters included trypsin as the assigned enzyme and a precursor mass accuracy of ± 3 Da. SEQUEST results were then analyzed using PeptideProphet and ProteinProphet in order to statistically evaluate peptide matches and assign protein probabilities (Nesvizhskii et al. 2003). Only proteins with a probability of ≥0.9 (estimated false discovery rate of 0.6%) and a minimum of four total tandem mass spectral assignments in all replicates were used in further characterizations described below.

In order to annotate corresponding proteins, the *C. gigas* proteome (version 9) was compared to the UniProtKB/Swiss-Prot database (www.uniprot.org) using Blastp with an e-value limit of 1E-10. Associated Gene Ontology terms were used to classify sequences based on biological process as well as categorize genes into parent categories (GO Slim). Enrichment analysis was used to identify overrepresented biological processes in the gill proteome compared to the entire proteome as downloaded from the *Crassostrea gigas* proteome version 9 (Database for Annotation, Visualization and Integrated Discovery (DAVID, v. 6.7; Huang et al. 2009a and 2009b, <http://david.abcc.ncifcrf.gov/>)). In order to assess biological variation of expressed proteins, technical replicates were pooled by taking the sum of total independent tandem mass spectra for each protein. Comparisons of proteins identified across biological replicates were visualized using Venny (Oliveros 2007).

The minimum number of peptides needed to be sequenced to optimize unique protein identifications was determined. A list was constructed of all sequenced peptides and their matching protein identification. Redundancies were maintained in this list, so that if a certain peptide was sequenced multiple times it was included multiple times in the list. The list had 72,387 peptide entries Using the sample function in R, randomized subsets of this list were generated using the sample function. The number of hypothetically sequenced peptides in these lists ranged from 500-70000.

RESULTS

*Tandem Mass Spectrometry*

A combined total of 175,818 tandem MS spectra were generated across all four biological and three technical replicates using the Orbitrap mass spectrometer. A total of 54,521 unique peptides contributed to the identification of 2,850 proteins with probability score threshold of 0.9 ([Supplemental File](https://www.dropbox.com/s/3m8ododnmeb6nko/Supplemental%20File_Annotations%20of%20Sequenced%20Proteins.xlsx)). 1,671 of these proteins had at least four tandem mass spectral hits across all replicates. The mean amino acid coverage across all proteins was 13.3%. Protein identifications for each injection including protein probability scores, number of total and unique spectra, and peptide sequences are supplied in ([Supplemental File](https://www.dropbox.com/s/n1u8u3h04937b9r/Supplemental%20File_All%20replicates.xlsx)).

 For all biological samples, the number of proteins identified in each technical replicate was consistent with minimal standard deviation (standard deviation ranged from 2.2-3.5% of average across technical replicates). Likewise, for each oyster sample 56-58% of the identified proteins were present in all three technical replicates. Using spectral counts as a proxy for relative expression, it was observed that relative expression levels were consistent across technical replicates (see Figure 1).

***TABLE 1 (***[***https://docs.google.com/spreadsheet/ccc?key=0An4PXFyBBnDEdFNmeE1BMmtMX3pZd1VqeDNUbFBSbnc***](https://docs.google.com/spreadsheet/ccc?key=0An4PXFyBBnDEdFNmeE1BMmtMX3pZd1VqeDNUbFBSbnc)***) . Summary of technical replicates for each biological replicate (oyster). Numbers are for total proteins identified, regardless of SwissProt annotation. Overall number of proteins identified in each oyster is in the first row, followed by how many proteins were identified in each technical replicate. The last row shows how many proteins in each biological replicate were identified across all technical replicates.***



***FIGURE 1. Technical replicates for oyster A. Technical replicates are shown in different colors (black, blue, and green). Total independent spectral counts are plotted on the y-axis for each protein. Technical replicates were consistent and similar patterns were seen for the other 3 oysters (data not shown).***



 The number of proteins identified in each oyster (after pooling technical replicates, see Methods) was 1,558, 1,554, 1,519, and 1,500 (Table 1). Most proteins (n=1,249) were identified across all biological replicates (Figure 2).

***FIGURE 2. Venn diagram of proteins identified among biological replicates (with a minimum 4 tandem mass spectral counts across replicates). Proteins identified in oyster A are in the blue ellipse, B are in yellow, C are in green, and oyster D proteins are in red.***



In order to evaluate general protein expression and assess sample variability, the ten most highly expressed proteins in each oyster were identified. These proteins represent core cell structure and function, such as maintenance of membrane charge and cellular attachment, cytoskeleton structure, muscle components, turnover of intracellular proteins, and protection against oxidative stress. Eight out of ten of these proteins (basement membrane-specific heparan sulfate proteoglycan core protein, rootletin, actin, probable cytosol aminopeptidase, myosin heavy chain, filamin-C, extracellular superoxide dismutase) were identified in all four oysters.

***TABLE 2. Ten most abundant proteins identified across oysters. Protein ID is given, as well as protein description from UniProt-KB/SwissProt, SwissProt Accession Number, and the oysters where the protein is present***



Of the 1,671 proteins expressed across all samples, 1,496 were annotated using the UnitProt-KB/SwissProt database. Of the annotated proteins, 1,368 were associated with Gene Ontology classifications. A majority of proteins were associated with the biological process of transport (n=175), followed by protein metabolism (n=124), and RNA metabolism (n=115) (Figure 3).

***Figure 3. Expressed gill proteome, represented by component GO Slim categories.***



Enrichment analysis was carried out to determine which biological processes were overrepresented in gill tissue compared to the entire proteome. The proteins associated with processes corresponding to GO Slim terms cell organization and biogenesis, developmental processes, and protein metabolism were found to be enriched in the gill proteome (Figure 4).

***Figure 4. Enrichment of processes (GO terms) in the gill proteome compared to the full proteome as deduced from the C. gigas genome. Processes are arranged in semantic space with color corresponding to significance (see p-value scale) and circle size corresponding to relative contribution of process to the enriched dataset.***



 The number of unique proteins identified with different hypothetical numbers of sequenced peptides created an exponential curve (Figure 5). The plateau began around 30,000-40,000 sequenced peptides with a total of 2,400-2,516 unique peptides identified. New unique peptides were still identified in larger sample sizes of peptides, but the return per sequenced peptide began to diminish.

***FIGURE 5. Unique proteins identified for varying sizes of hypothetical datasets of sequenced peptides.***



DISCUSSION

 Challenges in protein identification in non model species has resulted in limited focus on quantitative proteomics approaches in ecological studies. Given the recent technological advances in the proteomics field (Yates et al. 2009) and the recent release of the Pacific oyster genome, we set out to assess the practical use of quantitative proteomics in this model species. Approximately 1,000 proteins can be identified per sample injection with consistency across technical replicates, suggesting this approach will be a valuable means of protein discovery. For all biological samples a majority of the proteins identified (56-58%) were present in all technical replicates. Given that some proteins were not identified in all technical replicates, proteins with limited expression, at least in the oyster gill tissue, might not be detected and/or expression levels might not be accurately reflected with too few technical replicates. By sequencing three technical replicates per oyster, we were able to capture the majority of the proteome since novel unique protein identifications became less frequent in a random sampling of different numbers of peptides (Figure 5). The accuracy of our methodology was supported by the high degree of similarity in protein identification across biological replicates: more than 75% of the proteins were identified across all four oysters and 99% were identified in at least two oysters.

The number of proteins identified and subsequently annotated can vary tremendously based on experimental design, target tissue, match thresholds, and genomic resources available. In a study of European whitefish, *Coregonus lavaretus*, proteomics on fish larvae yielded sequencing of peptides corresponding to 1,500 proteins (Papakostas et al. 2012). The high number of protein identifications in whitefish is most likely due to the complexity of the tissue analyzed (whole body) and the availability of a large protein database from a closely related species. In a metaproteomics study of marine microbes, 2,273 distinct proteins were identified across 10 samples (Morris et al. 2010). The large number of proteins identified in Morris et al. (2010) is evidence of the large number of organisms and ecological niches that were sampled in their study. Only 49 proteins were identified in the hemolymph of Sydney rock oysters, *Saccostrea glomerata*, due in large part to the simplicity of the biological material analyzed and the lack of a genome from which a database could be constructed (Muralidharan et al. 2012). The relatively high number of proteins identified in the oyster gill proteome (1,671) supports that this tissue is the location of a variety of cells and physiological processes.

 In addition to assessing the feasibility of shotgun proteomics in the Pacific oyster we also were able to provide a functional characterization of the proteome. Some of the top GO Biological Processes represented in the gill proteome were transport, oxidation reduction, protein transport, mRNA processing, RNA splicing, multicellular organismal development, and cell adhesion. These GO terms depict a tissue where cellular signaling, transcription, and translation are important responses to potential environmental changes. These genes reflect the function of the oyster gill, which acts as an interface between the organism and its environment. Sequencing of the transcriptome in *C. gigas* gill tissue resulted in identification of genes involved in both signaling and the immune response (Gavery & Roberts 2012). Genes involved in the *C. gigas* gill tissue thermal response included those in the functional categories of cellular communication, cellular stress, transcription, and translation (Meisterzheim et al. 2007). These studies reinforce the dynamic capabilities of *C. gigas* gill tissue in response to environmental change.

Proteins involved in oxidation reduction and cell adhesion are expressed in the gill since the tissue is rich in hemocytes, the primary immune response cell in the oyster. Assays of gene expression support the role of gill tissue and its surrounding hemocytes in the immune and environmental stress responses. The transcript for the antioxidant protein peroxiredoxin 6 is expressed more highly in the gill tissue from oysters exposed to environmental contaminants (David et al. 2007). In *C. gigas* resistant to summer mortality, compared to susceptible oysters, gene expression profiles from the gill tissue demonstrate a greater capacity to respond to an immune challenge (Fleury & Huvet 2012).

 The oyster gill proteome is rich in proteins encompassing a wide variety of functions. The successful identification of these proteins from shotgun sequencing allows for further in-depth research into the biological function of gill tissue as well as possibilities for corroboration of gene and protein expression data. The success of the shotgun sequencing was due in part to the recent publication of the *C. gigas* genome, underlining that the dissemination of large genomic, transcriptomic, and proteomic datasets provides invaluable resources for the larger scientific community.

SUPPLEMENTAL FILES

Supplemental File 1: Protein identifications based on tandem mass spectra searched against the *C. gigas* proteome. The protein accession number is provided for each protein identified (n=2,850) as well as the SwissProt Accession Number (SPID) corresponding to the protein annotation, the e-value for the the BLASTp search, the gene description from SwissProt, and the total number of tandem mass spectra that corresponded to the protein entry across all replicates.

Supplemental File 2: ProteinProphet results for each replicate. Each worksheet in the spreadsheet corresponds to one replicate, i.e. B\_03 is the third technical replicate for oyster B. Each protein search result includes an entry number, one or more protein accession numbers, the probability that the protein assignment is correct, the percent coverage of the protein by the sequenced peptides, the number of unique peptides (non-redundant) used to identify the protein, the total number of peptides used to identify the protein, and the peptide sequences. Some of the entries have multiple protein accession numbers because both/all of the numbers correspond to proteins with the same or highly similar sequences.

REFERENCES

(to be finished this weekend)

Zhang, G; Fang, X; Guo, X; Li, L; Luo, R; Xu, F; Yang, P; Zhang, L; Wang, X; Qi, H; Zhu, Y; Yang, L; Huang, Z (2012): Genomic data from the Pacific oyster (*Crassostrea gigas*)