Title Options:

Sequencing and characterization of the Pacific oyster (Crassostrea gigas) gill proteome using shotgun proteomics

Shotgun proteomics as an effective means to describe the oyster proteome

Beyond the genome

Evaluating the value of shotgun proteomics in the Pacific oyster

Evaluating feasibility of gel-free proteomics analyses in the oyster

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) throughout the experiment. At the end of 6 weeks, gill tissue was removed from 4 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 4 times with a probe sonicator, 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 diothiolthreitol (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 used to enzymatically digest proteins. Trypsin was solubilized in a trypsin dilution buffer (20 µl) (Promega, Madison, WI) and 3 µl of this solution was added to each sample. The samples were incubated overnight at 37°C. The next day, the samples were evaporated on the speed vac to near dryness.

Desalting of the samples was running samples through a pre-prepared column. Specifically, MacroSpin columns (The Nest Group, Southborough, MA, USA) were prepared by adding Solvent A (80% acetonitrile with 0.1% trifluoroacetic acid in water, 200 µl) and centrifuging at 2,000 rpm for 3 minutes, repeating 3 additional times. The same steps were followed with Solvent B (5% acetonitrile with 0.1% trifluoroacetic acid in water) for a total of 3 Solvent B additions. To prepare the samples for loading onto the columns, Solvent A (200 µl) was added. The entire volume of protein digest (200 µl) was added to the column and spun down at 3,000 rpm for 3 minutes, collecting flow-through and repeating once. The columns were then washed with Solvent B (200 µl), centrifuging at 3,000 rpm for 3 minutes, repeating 2 more times. The columns were placed in clean collection tubes and Solvent A (100 µl) was added in 2 additions. The columns were centrifuged at 3,000 rpm for 3 minutes and effluent was collected. The 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). 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). 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 scans were acquired in the Orbitrap at 400-2000 m/z. Scans were acquired at 60,000 resolution followed by 6 data dependent MS/MS scans in the Linear ion trap (LTQ). Ion selection threshold was 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 8 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). Search results were then analyzed in Protein Prophet to create protein and peptide groupings and probability statistics (Nesvizhskii et al. 2003). Only proteins with a probability of at least 0.9 (estimated false discovery rate of 0.6%) and had at least 4 total spectra assignments 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 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 spectra for each protein. Comparisons of proteins identified across biological replicates were visualized using Venny (Oliveros 2007).

RESULTS

*Tandem Mass Spectrometry*

A combined total of 175,818 MS/MS spectra were generated across all biological (4) and technical (3) replicates using the Orbitrap platform, with an average number of MS/MS spectra per injection (12) of 14,652 (standard deviation of 756). A total of 54,521 unique peptides contributed to the identification of 2,850 proteins (probability score threshold of 0.9, [Supplemental File](https://www.dropbox.com/s/3m8ododnmeb6nko/Supplemental%20File_Annotations%20of%20Sequenced%20Proteins.xlsx)). 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 Files).

 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 3 technical replicates. Using spectral counts as a proxy for expression, 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). 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 number of proteins identified in only 1 technical replicate or in 2 or all technical replicates are also given. The last row shows how many proteins in each biological replicate were identified across all technical replicates.***



***FIGURE 1. Technical replicates for oyster D. Technical replicates are shown in different colors (black, blue, and green). Total independent spectral counts are plotted on the y-axis for each protein. For the most part, 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. 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 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 4 oysters.

***TABLE 2. Most abundant proteins identified. Protein ID is given, as well as protein description from SwissProt/UniProt-KB, SwissProt ID (SPID), and the oysters where the protein is present***



Of the 1,671 proteins expressed across all samples, 1,496 were annotated using the UnitProt/Swiss-Prot 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. Gill Proteome, sequenced on Orbitrap, 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 Slim terms) in the gill proteome compared to the full proteome as deduced from the C. gigas genome.***

