Genetic Variation Underlying Protein Expression in Eggs of the Marine Mussel *Mytilus edulis*†

Angel P. Diz‡, Edward Dudley§, Barry W. MacDonald¶, Benjamin Piña∥, Ellen L. R. Kenchington¶, Eleftherios Zouros**, and David O. F. Skibinski‡ †‡

Study of the genetic basis of gene expression variation is central to attempts to understand the causes of evolutionary change. Although there are many transcriptomics studies estimating genetic variance and heritability in model organisms such as humans there is a lack of equivalent proteomics studies. In the present study, the heritability underlying egg protein expression was estimated in the marine mussel *Mytilus*. We believe this to be the first such measurement of genetic variation for gene expression in eggs of any organism. The study of eggs is important in evolutionary theory and life history analysis because maternal effects might have profound effects on the rate of evolution of offspring traits. Evidence is presented that the egg proteome varies significantly between individual females and that heritability of protein expression in mussel eggs is moderate to high suggesting abundant genetic variation on which natural selection might act. The study of the mussel egg proteome is also important because of the unusual system of mitochondrial DNA inheritance in mussels whereby different mitochondrial genomes are transmitted independently through female and male lineages (doubly uniparental inheritance). It is likely that the mechanism underlying this system involves the interaction of specific egg factors with sperm mitochondria following fertilization, and its elucidation might be advanced by study of the proteome in females having different progeny sex ratios. Putative identifications are presented here for egg proteins using MS/MS in *Mytilus* lines differing in sex ratio. Ontology terms relating to stress response and protein folding occur more frequently for proteins showing large expression differences between the lines. The distribution of ontology terms in mussel eggs was compared with those for previous mussel proteomics studies (using other tissues) and with mammal eggs. Significant differences were observed between mussel eggs and mussel tissues but not between the two types of eggs. *Molecular & Cellular Proteomics* 8:132–144, 2009.

The potential of proteomics approaches in population and evolutionary biology is gradually being recognized (1, 2). The analysis of the functional significance of variation in gene expression between individuals and populations is an active area of investigation (3). Knowledge of the genetic and environmental components of such variation is of central importance for assessment of the roles of selection, mutation, and genetic drift in causing evolutionary change in gene expression profiles (4–6). A number of studies have demonstrated significant variation in gene expression patterns within and between species (for a review, see Ref. 7). For example in a study of the marine fish *Fundulus heteroclitus* using transcriptomics techniques, 18% of 907 genes showed significant variation between individuals within populations with a much smaller percentage showing statistically significant differences between populations (8).

The heritability of gene expression has been measured directly using quantitative genetics approaches. For example, in a study of heritability of gene expression in 15 human families using lymphoblastoid cell lines, 31% of 2340 genes had significant heritability, and for 25% of these the heritability was greater than 40% (9). Other studies have partitioned genetic variation further. For example, of 8131 transcripts studied in two strains of mice and their reciprocal F1 crosses, 18% showed a heritability of gene expression of >50%, and 20% showed evidence of dominance effects. In addition, about 4% of the transcripts differed in expression between the reciprocal F1 crosses, indicating maternal effects (10). Non-additive variation for gene expression including over-dominance was observed for about 50% of transcripts in an analysis of strains and hybrids of the Pacific oyster (11).

Maternal effects, which can be genetically or environmentally determined, including those mediated through the properties of eggs can have substantial and complex effects on the rate of evolution of offspring traits (12–14). The trait egg
size is of particular interest in life history analysis. A trade-off has been suggested between egg size and number (15, 16) such that in a specific environment there is a single optimal egg size that maximizes progeny fitness. Egg quality characteristics and protein expression patterns might similarly be of high relevance to maternal effects and their evolutionary and ecological consequences.

Although the vast majority of genomics studies are on somatic tissues, both transcriptomics (17) and proteomics (18, 19) approaches have been used to study gene expression during oocyte maturation in mammals. Such work may be of practical importance for the improvement of oocyte selection during assisted reproduction (20). Proteomics might have advantages over transcriptomics for oocyte study because the level of accumulated mRNA in eggs might not reflect that of the corresponding proteins (21). In addition, abundant housekeeping proteins, easily detectable through proteomics, might have a variety of important functions during oocyte maturation (22). Among non-mammals, proteomics approaches have been used to identify egg proteins in silkworms (23). The above studies on oocytes, however, have not attempted to measure genetic and environmental components of phenotypic variation in gene expression. This is one aim of the current study of the marine mussel *Mytilus*.

The mussel *Mytilus* is an important model organism in studies of evolution and ecology in the marine environment. For example mussels are important for biomonitoring (24) and have been widely used in studies of population genetics and speciation (25, 26). Given the increasing interest in the use of proteomics techniques in marine biology (27, 28), mussels are thus positioned to play an important future role in this area. The technology for transcriptomics in *Mytilus* is being developed (29–31), and several studies also indicate the feasibility of pursuing successful proteomics work in mussels and the identification of proteins differentially expressed between species (32–35) and between different environments (36–38). Adaptive rationales were given in these studies for some of the differentially expressed proteins identified by mass spectrometry. For example it was suggested that higher expression of heat shock protein in intertidal mussels might be related to the heat stress that these mussels experience at low tide (36).

Marine mussels (*Mytilus* spp.) are a particularly interesting model system because of their unusual mechanism of mitochondrial DNA (mtDNA) inheritance. Separate mtDNA genomes are passed through both the female and male lines of descent, a mechanism known as doubly uniparental inheritance (DUI) (39–42). Females are usually homoplasmic for F-type mtDNA, which they transmit to both daughters and sons. Males are usually heteroplasmic for F-type and a male specific M-type mtDNA. The M-type is transmitted from fathers to offspring through the sperm. DUI provides a unique opportunity to study the evolutionary and functional consequences of different modes of mtDNA inheritance (43). The mechanism of DUI requires that sperm mtDNA have a different fate during development in mussels destined to become males and females. Studies with fluorescence microscopy suggest that sperm mitochondria disperse randomly in female embryos but remain aggregated at least until the D-stage larvae in male embryos (44). The domination of male gonad by the M-type and egg and somatic tissue by the F-type (45, 46) might be explained by this difference in behavior, but the precise molecular mechanism underlying DUI and sex determination is unknown. One possibility is that egg proteins are involved, and thus a proteomics study to test for differential protein expression in eggs destined to become male and female could be illuminating.

In the present study, a proteomics comparison was made of pairs of sibs from two lines of the mussel *Mytilus edulis* established in a breeding program. This provides the first example of the measurement of the genetic component underlying egg protein expression variation. The results are discussed in light of previous studies of the genetic determination of egg traits and their evolutionary and ecological relevance. Proteins identified by mass spectrometry in mussel eggs are compared with those identified in mammalian egg proteomics studies, and the relevance of these results to future work on the mechanism of DUI are considered.

### MATERIALS AND METHODS

**Animals and Egg Collection**—*M. edulis* female mussels belonging to two lines from a breeding program were used. The lines differed in sex ratio with one having high frequency of males (Line M, with 91% males) and the other producing females (Line F, 100% females). The lines were established from wild mussels recently collected from Mahone Bay, Nova Scotia, Canada (44.44° north, 64.38° west). Mussels were reared in a hatchery where they were kept together in the same holding tank for at least 2 years. Because egg formation and development occur on an annual cycle, egg proteome differences between individuals could thus not easily be attributed to the pre-hatchery environment. For spawning, mussels were removed from the holding tank and placed in individual plastic cups on the day of spawning. The water in these containers was from the same source as that used in the holding tanks. Thus all mussels were reared and spawned under identical environmental conditions. Samples of eggs were collected from one pair of sisters from each line, (98Ax98WM8)b and (98Ax98WM8)c from Line M and X102H and X102K from Line F. After spawning, 15 µl of eggs, obtained by centrifugation for 3 min at 13,000 × g, were resuspended in preservation medium (10% glycerol in 0.9 M NaCl) for each of the four mussels. The eggs in preservation medium were snap frozen in a dry ice/ethanol bath and stored at −80 °C prior to analysis.

**Protein Extraction**—Proteins were extracted from eggs in 0.5 ml of lysis buffer (7 M urea, 2 M thiourea, 4% CHAPS, 2% DTT, and 1% IPG) and solubilized with a sonicator (Branson Digital Sonifer 250) using 12 blasts of 15% amplitude and 5 s each with 10-s breaks. This was done on ice to avoid protein burning. After centrifugation for 30 min at

---

1 The abbreviations used are: mtDNA, mitochondrial DNA; ANOVA, analysis of variance; 2-DE, two-dimensional electrophoresis; df, degrees of freedom; DUI, doubly uniparental inheritance; FDR, false discovery rate; NCBI, National Center for Biotechnology Information.
Proteome of Mussel Eggs

21,000 × g at 4 °C, the pellet was discarded, and protein supernatant was stored at −80 °C until electrophoresis. Protein concentration was measured with the Protein 2-D Quant kit (GE Healthcare), and cleaning was performed with a 2-D Clean-Up kit (GE Healthcare) to remove interfering substances (salt or charged detergents) for the first dimension IEF.

Two-dimensional Electrophoresis (2-DE)—Egg proteins from the four female mussels were analyzed by two-dimensional electrophoresis with three gels (technical replicates) per mussel. Approximately 90 μg (analytical gels) or 400 μg (preparative gels) of total protein was used for each gel. The first dimension (IEF) electrophoresis was carried out on immobilized pH gradient strips (pH 3–10 nonlinear, 24 cm; GE Healthcare) with a horizontal electrophoresis apparatus (Ettan IPGphor, GE Healthcare). IPG dry strips were rehydrated overnight directly with sample in DeStreak Rehydration solution (GE Healthcare). After two steps of strip equilibration (2 × 15 min) with DTT and iodoacetamide, the second dimension of gel electrophoresis was carried out with precast 12.5% polyacrylamide gels (22 × 27 × 0.1 cm²; GE Healthcare) using an Ettan Daltix II electrophoresis system (GE Healthcare). Electrophoresis was carried out using the Multi Temp III (GE Healthcare) at 25 °C at a maximum of 100 mA for ~5 h until the Bromphenol blue front reached the bottom of the gel. Protein spots were visualized by silver staining using either an image analysis method (47) or a fully compatible mass spectrometry method (48). Co-migrating broad range standards (Bio-Rad) were used in the second dimension to allow estimation of molecular masses.

Analysis of 2-DE Patterns—Silver-stained gels were scanned to TIFF files using an Image Scanner (GE Healthcare). The TT900 S2S (Nonlinear Dynamics Ltd.) software was used initially for semiautomatic alignment of gels. The Progenesis PG240 software (Nonlinear Dynamics Ltd.) “Samespots” was then used for spot detection, filtering, and volume measurement. Filtering incorporated visual checking, and spots showing clear anomalies or resulting from artifacts such as specks on the gels were discarded. The lowest on boundary background subtraction method was used as recommended by Nonlinear Dynamics Ltd. The method ensures that zero or negative spot volume values are not returned by the software, a convenience for subsequent statistical analysis. The number of spots detected per gel varied depending mainly on the spot detection parameters settings.

Variable Transformation, Normalization, and Standardization—For analysis, 261 protein spots were used. Absolute spot volume was normalized for each gel by dividing each spot volume by the gel total over the 261 spots. The 2-DE procedure, incorporating normalization, has the consequence that spot volume differences between individuals for a specific protein spot are differences relative to total spot volume over the spots used. A higher normalized volume for a specific protein spot in individual A compared with B should also reflect a higher relative amount, or relative concentration, of this protein in the volume of tissue extracted. This conclusion still applies if the relationship between spot volume and protein amount, although increasing, is nonlinear. For analytical convenience, the normalized spot volumes were then multiplied by the gel total for the gel with the lowest total (resulting variable = NormVol). NormVol was transformed by taking the log to base 2, and the resulting values were then renormalized for each gel on the log scale by subtracting each value from the arithmetic average of the 261 log₂ values for that gel. Twenty outlying values (of 12 gels × 261 = 3132 values) identified using Grubb’s test (49) were removed from the data set and replaced with values imputed using the regression method, implemented in SPSS 13.0 for Windows (Statistical Package for the Social Sciences (SPSS Inc.)), with added random residuals. This was followed by renormalization by subtraction (resulting variable = log₂final). The variable values obtained directly after imputation were also standardized by dividing each value by the standard deviation of the 12 values (4 mussels × 3 technical replicates) for each spot and then renormalized by subtraction. This was iterated twice until convergence of values (resulting variable = log₂std). Thus for log₂std each spot has equal weighting, that is the same variance over the 12 values. Other possible normalization and transformation procedures were investigated and tested. The method described above is consistent with approaches used in the quantitative proteomics literature and has the advantage that it results in variables that do not deviate markedly from the normal distribution permitting parametric statistics.

Global Analysis on Entire Data Set—A fully factorial two-way analysis of variance (ANOVA; model II, type III sums of squares) was carried out on the entire data set for the variables log₂final and log₂std using SPSS (Statistical Package for the Social Sciences (SPSS Inc.)). There are two Lines, four Mussel (two per line), three technical Replicates per mussel, and 261 Spots. The factors Lines and Mussel were treated as fixed, and the other factors were treated as random; although for computation of variance components, all factors were treated as random. To test for differential gene expression between lines, a nested ANOVA was constructed with the mean square for Lines × Spots tested against (Mussels within Lines) × Spots. To test for biological variation, the mean square for (Mussels within Lines) × Spots was tested against the Residual mean square (between technical replicates within Mussels × Spots). Three added variance components (s²) were calculated from these means squares, for Lines × Spots (s²_L), (Mussels within Lines) × Spots (s²_M), and the residual technical variation (s²_R). Two intraclass correlations (t₁ and t₂) were calculated from these variance components, t₁ = s²_L/(s²_L + s²_M) and t₂ = s²_M/(s²_L + s²_M). The correlation t₂ measures the relative amount of genetic variation; t₁ measures the relative amount of biological variation. Although both variables log₂final and log₂std show a good visual fit to the normal distribution over the whole data set, an acceptable fit to a straight line in the Q-Q plot, and a non-significant Kolmogorov-Smirnov test result, the Levene test for homogeneity of variance is significant. Thus confidence limits (95%) for selected ANOVA parameters were determined by bootstrapping in Excel using sampling functions from PopTools (50). For each bootstrap replicate, a new data set was constructed by selecting 261 spots with replacement and log₂final or log₂std recalculated followed by ANOVA on the bootstrapped data set. This was repeated 10,000 times. Bias in the bootstrapped estimates was corrected using the bias-corrected percentile method (51). Supporting analyses not based on ANOVA were also carried out. The Manhattan distance (absolute difference in spot volume between two mussels) and the squared Euclidean distance (square of difference in spot volume between two mussels) averaged over spots were calculated for log₂final and log₂std with confidence intervals determined by bootstrapping over spots. Finally a nonparametric rank test was performed with log₂final. The number of spots in which the spot volume for both mussels in Line M either exceeded or were less than both mussels from Line F was counted. The fraction of such spots expected by chance, the null hypothesis, based on the possible permutations of four mussels and assuming no difference between lines is ½.

Protein Identifications and Mass Spectrometry Analysis—The spots of interest were excised using an Etten Spot Picker (Amersham Biosciences) into a microplate using Milli-Q H₂O. Trypsin digestion was carried out using an Etten Digester (Amersham Biosciences) following the manufacturer’s instructions. The excised spots were destained and washed using 15 mM potassium ferricyanide, 50 mM sodium thiosulfate, Milli-Q H₂O, and 75% acetonitrile. The trypsin digestion was performed by adding 10 μl of trypsin (20 ng/μl) in 20 mM ammonium bicarbonate solution (Promega sequencing grade) to each spot and incubating at 37 °C overnight. The peptides were recovered and transferred to a new microplate using 0.1% TFA, 50% acetonitrile and
finally dried under a nitrogen stream using a TurboVap96 (Zymark). Protein identifications were attempted with MALDI-TOF with poor results in part because of the paucity of Mytilus database information. With HPLC-MS/MS, the number of positive identifications increased in number and quality. HPLC-MS/MS was carried out using a nano-ESI ion trap, the LCQ DECA XP (ThermoFinnigan, Hemel Hempstead, UK), equipped with a nanospray interface. For the HPLC separation of the peptide sample, 10 μl (dissolved in 0.1% formic acid in water) was injected using a FAMOS autosampler (Dionex) onto an in-house prepared fused silica C18 PepMap column (15 cm × 75 μm) with a pulled tip that formed the electrospray needle thereby eliminating any dead volume after the column. The sample was injected into a mobile phase of 2% acetonitrile, 98% water (0.1% formic acid) (mobile phase A) with mobile phase B representing 0.1% formic acid in acetonitrile. After loading the sample for 5 min, a gradient between 100% A and 60% B over 45 min was performed followed by an increase to 96% B over 15 min, washing of the column with 95% B for 10 min before returning to 100% A mobile phase composition, and column re-equilibration over 10 and 15 min, respectively. The mobile phase was delivered by an Ultimate pump system with a nanoflow cartridge split system (Dionex). A spray voltage of 1.6 kV was applied to the sample at a liquid junction just prior to the column, and the spray passed into a heated capillary at 165 °C and a capillary voltage of 10 V. Data were acquired in a positive, data-dependent acquisition mode in which the mass spectrometer first acquires a full scan mass spectrum between 475 and 2000 Da. The MS/MS spectra of the three most abundant ions in the spectrum were then obtained before another full scan spectrum was again monitored. This process was repeated throughout the HPLC-MS/MS run with dynamic exclusion parameters excluding any ions whose MS/MS spectra were obtained three times from further analysis for 3 min. The proteins in the spots were identified from selected peptide spectra obtained by nano-ESI-ion trap MS/MS using Bioworks Browser version 3.2 (ThermoElectron) which compared the spectra produced against a non-redundant database of all organisms. Protein identifications were improved by combining the peptide information obtained from the triplicate HPLC-MS/MS analyses of each replicate protein spot. The recommended default settings of Bioworks were used. These assumed that the protein had been digested with trypsin, allowed one missed cleavage during digestion, had no fixed or variable modifications, and had a precursor ion mass tolerance of 1.4 amu and a fragment ion tolerance of 1.0 amu. The database searched was the nr.fasta database acquired and indexed on August 16, 2006. No species were defined, and so the entire database was searched (2,481,719 proteins). A protein was considered a positive match when the XCorr for the peptide exceeded the threshold set for each charge state (XCorr = 1.8 for z = 1, 2.2 for z = 2, and 2.8 for z = 3). These thresholds had been recommended in a personal communication with ThermoElectron and had been shown to be efficient in allowing correct protein identification using standard proteins (data not included here). The molecular weights of the identified proteins, obtained from the database, were also compared with their spot positions on the two-dimensional gel with a tolerance of 10 kDa allowed between the observed and theoretical molecular mass. To further validate protein identifications, peptide sequences were used to search a local Mytilus cDNA sequence database and Mytilus sequences in NCBI databases using TBlast.

RESULTS

2-DE Gels—The number of protein spots observed on the gels exceeded 1000 per gel, but after the filtering process, 261 spots were retained for further analysis. Most spots were present on all gels; others were present on all technical replicates of at least one of the mussels analyzed. Silver staining has a linear sensitivity range of 3 orders of magnitude (52, 53). In the present study the weakest and strongest spots fell within this range. It is likely that those selected were dominated by housekeeping proteins, which tend to have high expression levels (54). Representative examples of 2-DE gels from the two lines are shown in Fig. 1.

Technical Error—The proportion of variance in NormVol explained statistically by Spots was calculated by ANOVA for each mussel; the residual variation is the technical variance between replicates. The mean proportion is 0.92 (range, 0.90–0.96). This value is higher than previously reported values in
the range 0.83–0.91 (35) in part because of stringent data filtering and possibly the use of precast polyacrylamide gels. The coefficient of variation was computed from the three technical replicates and averaged over spots for each mussel. The mean is 19.1% (range, 15.5%–23.3%) lower or comparable to values in previous reports (e.g. Molloy et al. (55), who quote 20–30%, and Asirvatham et al. (56), who quote a mean of 16.2%).

Global Analysis on Entire Data Set—Nested ANOVA parameters are given for log2final and log2std in Table I. For both variables, both the difference between lines and the biological variation are highly significant. ANOVA statistics with bootstrapped confidence limits are given in Table II. The results are consistent with those of Table I. The confidence limits of the F statistics do not overlap unity, and those of the variance components and intraclass correlation coefficients do not overlap zero, suggesting significant genetic and biological variation consistent with the results of Table I.

The Manhattan and Euclidean distances for log2final and log2std were significantly greater between lines than between mussels within lines (data not shown), supporting the conclusions from ANOVA. In the nonparametric rank test the observed fraction of spots that completely differentiated mussels from the two lines is 114/147, and the expected fraction under the null hypothesis is 87/174. This is significant (χ² = 12.569, df = 1, p = 0.0004), once again confirming that there are overall expression differences between the lines.

Mean Expression Considered on a Spot by Spot Basis—An ANOVA was carried out on log2final separately for each spot. The factors are Lines (df = 1), (Mussels within Lines) (df = 2), and Residual (df = 8). (Mussels within Lines) was tested against the Residual. Lines was tested first against (Mussels within Lines) and second against Pooled error ((Mussels within Lines) + Residual) for those spots where (Mussels within Lines) was not itself significant. The Levene test was used to test for homogeneity of variance within mussels. Significant spots using false discovery rates (FDRs) of 5, 20, and 50% (57, 58) are shown in Table III together with the number showing a significant Levene test result (in parentheses). The number of protein spots significant at FDRs of 5, 20, and 50% together with the number showing a significant Levene test result (in parentheses) is shown in Table III.
spots show significant differences between lines and between mussels within lines.

Of the 261 spots, 50 were expressed at \( \geq 1.5 \)-fold difference between lines (26 higher in Line F and 24 higher in Line M). This \( k \)-fold difference level has some advantages as a threshold in relation to power and efficiency and is typically of interest to researchers \((59-61)\). Twenty-six spots were expressed with a \( \geq 1.5 \)-fold difference (10 higher in Line F and 16 higher in Line M), 17 spots were expressed with a \( 2.0-3.0 \)-fold difference (11 in Line F and six in Line M), and seven were expressed with a \( >3.0 \)-fold difference (five in Line F and two in Line M). FDR values for spots identified by mass spectrometry (see below) are given in Table IV.

**Protein Identifications by Mass Spectrometry Analysis (MS/MS)**—Protein identifications by mass spectrometry were attempted on the 50 spots with a \( \geq 1.5 \)-fold difference in expression between Lines. Of these, 17 were identified by MS/MS on the basis of the charge state threshold parameters (Table IV and supplemental Table 1), six were identified in Line M, and 11 were identified in Line F. Two proteins were identified with two and four peptides, respectively; the others were identified with one peptide. Identifications were attempted on a further 40 spots with \(<1.5\)-fold expression difference between Lines. Of these, 26 were identified (Table IV and supplemental Table 1). Of these, 14 proteins were identified with from two to six peptide sequences; the others were identified with one peptide (supplemental Table 1). Many of the spots identified were nominally for the same protein. These might be different isoforms, allelic charge variants, or different splice variants or reflect post-translational modification.

The spots identified in a proteomics study can be assigned ontology terms relating to protein identity, function, and process. Three such classification systems were used in the present study. The first was simply the protein identity corresponding with the NCBI record. The second (Function/Process) used ontology terms provided by The Universal Protein Resource (UniProt) \((62)\). The third (Overall Process) was arrived at using best judgment and focused on process. This resulted in a smaller number of terms that have some advantages for statistical analysis. Some difficulties arise in such classifications. For example DNA polymerase III (Table IV) is a prokaryotic enzyme, so it cannot be related to nuclear DNA metabolism. Whether this particular protein could be related to mitochondrial DNA polymerases is still to be elucidated. Similarly albumin-like proteins are unknown in molluscs, so the identification albumin precursor-1-ho in mussel tissue (supplemental Table 2) is uncertain. The P4hb protein is a subunit of proline 4-hydroxylase and therefore could be classified as a chaperone. However, this enzyme is also involved in an essential step in the synthesis of collagen. In bivalves it is important for byssal synthesis, so it has been assigned the term structural element for Overall Process. For Function/Process, some similar ontology terms were pooled for the analyses described below.

The frequency distribution of terms for the spots of Table IV are shown in Table V for the three classifications (Protein Identity, Function/Process, and Overall Process). Statistical analyses of such distributions, dependent on the usual assumptions regarding random selection and independence, could be a useful tool in data exploration. Clearly the chance observation of a single ontology term \( (e.g. \text{gelsolin}) \) in a sample of spots is not informative of function. However, significant multiple occurrences of a particular term might be regarded as a characterizing feature of the underlying population of spots. This can be tested statistically with a \( \chi^2 \) analysis against the null hypothesis that all terms within the sample have equal frequency. Because of small numbers for some terms, a dedicated Monte Carlo bootstrapping program written in FORTRAN language was used with 30,000 repetitions to determine significance overall and for individual terms against this null hypothesis. There is significance overall for Protein Identity terms \( (p = 0.013) \), and heat shock protein 60 is significant individually at the 5% level \( \text{a posteriori} \) taking account of number of terms in the table. For Function/Process terms, there is significance overall \( (p = 0.000) \) and individual significance \( \text{a posteriori} \) for term “stress response/protein folding.” The term “cellular motility/cytoskeleton” is significant \( \text{a priori} \) \( (p = 0.025) \) but not in the \( \text{a posteriori} \) test.

The terms were also sorted into two groups according to four criteria: (i) expression higher in Line M or Line F, (ii) significant or non-significant by the FDR levels applied in Table IV when tested against the Mussels within Lines effect as error, (iii) significant or non-significant when tested against the Pooled error, and (iv) \( \text{-fold} \) expression difference between lines \( \geq 1.5 \) or \(<1.5\). This allows construction of two-way contingency tables \( (k \times m) \) where \( k = 2 \) (for two groups) and \( m \) is the number of terms. A dedicated Monte Carlo program was again used to assess significance of \( \chi^2 \) in a model III contingency test. There is significance only for (iv) for Protein terms \( (p = 0.005) \), Function/Process terms \( (p = 0.033) \), and Overall Process \( (p = 0.003) \), and the individual terms “heat shock protein 60,” “stress response/protein folding,” and “chaperones/protein folding and degradation” are significant \( \text{a posteriori} \).

The frequencies of ontology terms were also determined for other published studies on mussel tissues (excluding eggs) \((32, 33, 36-38) \) and mammal oocytes or eggs \((18, 19, 21, 22) \). The data for individual proteins are given in supplemental Table 2 and summarized for Function/Process and Overall Process terms in Fig. 2. Because the number of spots identified for a particular term was not available in some studies, the three data groups (referred to as Mussel Egg, Mammal Egg, and Mussel Tissue) are reduced so that the occurrence of a particular protein term was assigned a count of one regardless of the number of spots for this term. Thus only the function or process terms will have multiple occurrences within data groups. Contingency tests were carried out to compare the distribution of terms in Mussel Egg with that in Mammal Egg and Mussel Tissue. The
| Spot no. | -Fold changes | p (Lines/Mussels) | FDR | p (Lines/Pool) | FDR | Protein name | Function/Process | Overall Process |
|---------|---------------|------------------|-----|----------------|-----|--------------|------------------|-----------------|
| 6       | 4.2 F         | 0.066            | 0.066 | 50            |      | Putative α-tubulin | Cellular motility/cytoskeleton | Structural element |
| 12      | 2.6 F         | 0.072            | 0.001 | 5             |      | RACK1        | Intracellular receptor | RNA/DNA metabolism/regulation |
| 30      | 2.4 M         | 0.003            | 0.000 | 5             |      | Acyl-CoA dehydrogenase | Fatty acid/lipid metabolism | Metabolism |
| 28      | 2.4 M         | 0.004            | 0.000 | 5             |      | Isocitrate dehydrogenases | Glycolysis/tricarboxylic acid cycle | Metabolism |
| 25      | 2.2 F         | 0.101            | 0.002 | 5             |      | Homologous to chaperonin protein | Stress response/protein folding | Chaperones/protein folding and degradation |
| 14      | 2.0 M         | 0.194            | 0.194 | 5             |      | DNA topoisomerase II, putative | DNA replication | RNA/DNA metabolism/regulation |
| 109     | 1.5 M         | 0.107            | 0.040 | 20            |      | Proteasome subunit 1b (proteasome δ chain) | Protein catabolism | Chaperones/protein folding and degradation |
| 41      | 1.5 M         | 0.200            | 0.225 | 5             |      | Chaperonin subunit 6b (g) | Stress response/protein folding | Chaperones/protein folding and degradation |
| 75      | 1.4 M         | 0.001            | 0.000 | 5             |      | Glycerate dehydrogenase | Amino acid synthesis/oxidoreductase | Metabolism |
| 138     | 1.4 F         | 0.016            | 0.013 | 20            |      | P4hb protein | Isomerase/protein binding | Structural element |
| 177     | 1.4 F         | 0.010            | 0.000 | 5             |      | Gelsolin     | Actin-binding protein | Structural element |
| 175     | 1.3 F         | 0.032            | 0.055 | 50            |      | Homologous to chaperonin protein | Stress response/protein folding | Chaperones/protein folding and degradation |
| 227     | 1.2 M         | 0.030            | 0.059 | 50            |      | DNA polymerase III α chain | DNA replication | RNA/DNA metabolism/regulation |
| 218     | 1.2 F         | 0.153            | 0.277 | 5             |      | Caireticulin | Protein folding/calcium ion binding | Chaperones/protein folding and degradation |
| 141     | 1.2 F         | 0.396            | 0.289 | 5             |      | Heat shock protein Hsp70 | Stress response/protein folding | Chaperones/protein folding and degradation |
| 245     | 1.1 M         | 0.163            | 0.087 | 50            |      | Tubulin α | Cellular motility/cytoskeleton | Structural element |
| 244     | 1.1 M         | 0.100            | 0.119 | 50            |      | α-Tubulin | Amino acid synthesis/oxidoreductase | Metabolism |
| 257     | 1.1 M         | 0.284            | 0.284 | 50            |      | Tubulin α | Cellular motility/cytoskeleton | Structural element |
| 161     | 1.1 M         | 0.656            | 0.436 | 5             |      | ATP synthase F1 subunit α | Ion transportation | Metabolism |
| 239     | 1.1 M         | 0.245            | 0.444 | 5             |      | NADPH-specific isocitrate dehydrogenase | Glycolysis/tricarboxylic acid cycle | Metabolism |
comparison of Mussel Egg with Mussel Tissue was significant for Overall Process (p < 0.001) with the term “RNA/DNA metabolism/regulation” at significantly higher frequency in Mussel Egg and the term “stress response/defense” at significantly higher frequency in Mussel Tissue in a posteriori tests.

**DISCUSSION**

**Genetic Variation in Protein Expression in Mussel Eggs**—The results provide evidence for significant variation in gene expression between Lines as judged by the global ANOVA, a consideration of Euclidean and Manhattan distances, and a non-parametric analysis applied to the collection of spots as a whole. The intraclass correlation ($r_{GL}$) varies between 0.394 (for standardized) and 0.496 (for unstandardized) data. In relation to estimation of genetic variance, the full sib mating design is most appropriate, although application of other models (e.g., for inbred lines) would not alter the general conclusion that there is appreciable genetic variation underlying expression. The main contributors to the covariance of full sibs are the additive ($V_A$) and dominance ($V_D$) variance and the variance due to common environment ($V_{Ec}$). Then covariance $V_A = \frac{1}{2} V_{Ec} = r_{GL}^2 V_{Ec}$ (63). In the present study, the mussels used were raised in the same environment prior to and during spawning. This justifies the conclusion that com-

---

**TABLE IV—continued**

| Spot no. | -Fold changes$^a$ | p (Lines/Mussels)$^b$ | FDR | p (Lines/Pool)$^c$ | FDR | Protein name$^d$ | Function/Process | Overall process |
|----------|-------------------|----------------------|-----|-------------------|-----|-----------------|-----------------|-----------------|
| 77       | 1.1 M             | 0.463                | 0.463| Glycerate dehydrogenase | Amino acid synthesis/oxidoreductase | Metabolism |
| 73       | 1.1 M             | 0.507                | 0.507| FGF1 ATP synthase subunit α | Ion transportation | Metabolism |
| 63       | 1.1 F             | 0.610                | 0.559| ATP synthase β subunit | Ion transportation | Metabolism |
| 129      | 1.1 F             | 0.754                | 0.754| Tubulin β | Cellular motility/cytoskeleton | Structural element |
| 246      | 1.1 F             | 0.693                | 0.797| Similar to cytoplasmic β-actin | Structural constituent of cytoskeleton | Structural element |
| 197      | 1                 | 0.630                | 0.439| Isocitrate dehydrogenases | Glycolysis/tricarboxylic acid cycle | Metabolism |
| 253      | 1                 | 0.227                | 0.456| Cytoplasmic actin Cyll | Structural constituent of cytoskeleton | Structural element |
| 145      | 1                 | 0.702                | 0.496| Heat shock protein 60 | Stress response/protein folding | Chaperones/protein folding and degradation |
| 236      | 1                 | 0.602                | 0.583| Isocitrate dehydrogenases | Glycolysis/tricarboxylic acid cycle | Metabolism |
| 260      | 1                 | 0.622                | 0.591| Isocitrate dehydrogenase DNA polymerase III α chain | DNA replication | RNA/DNA metabolism/regulation |
| 261      | 1                 | 0.242                | 0.614| α-Tubulin | Cellular motility/cytoskeleton | Structural element |
| 155      | 1                 | 0.890                | 0.843| Isocitrate dehydrogenase | Glycolysis/tricarboxylic acid cycle | Metabolism |
| 67       | 1                 | 0.965                | 0.965| Likely mitochondrial malate dehydrogenase Malate dehydrogenase | Glycolysis/tricarboxylic acid cycle | Metabolism |

---

$^a$ F = higher expression in Line F; M = higher expression in line M.

$^b$ Values in italics have significant value for the Levene test for homogeneity of variance.

$^c$ Denominator is Pooled error when (Mussels within Lines) not significant; otherwise denominator is (Mussels within Lines).

$^d$ Protein names are the definitions for the NCBI entries corresponding to the accession numbers for the different peptide sequences given in supplemental Table 1.

$^e$ CCT, chaperone-containing T-complex polypeptide 1.
Proteome of Mussel Eggs

| Table V | Protein identity, function, and process ontology terms for 43 protein spots identified by MS/MS in Mytilus eggs |
|---------|---------------------------------------------------------------------------------------------------------|
| Protein identity                  | Number of spots |
| Heat shock protein 60 (chaperonin) | 9 |
| Isocitrate dehydrogenase           | 5 |
| α-Tubulin                          | 4 |
| ATP synthase                       | 3 |
| Glycerate dehydrogenase            | 3 |
| DNA polymerase III α chain         | 2 |
| Heat shock protein 70              | 2 |
| Pama2 protein                      | 2 |
| Terms occurring only once pooled   | 13 |
| Function/Process                   | |
| Stress response/protein folding    | 12 |
| Cellular motility/cytoskeleton     | 8 |
| Glycolysis/tricarboxylic acid cycle| 6 |
| Ion transportation                 | 4 |
| DNA replication/synthesis          | 3 |
| Amino acid synthesis/oxidoreductase| 3 |
| Protein metabolism, transport      | 2 |
| Terms occurring only once pooled   | 5 |
| Overall Process                    | |
| Chaperones/protein folding and degradation | 14 |
| Metabolism                         | 13 |
| Structural element                 | 9 |
| RNA/DNA metabolism/regulation      | 7 |

Table V: Protein identity, function, and process ontology terms for 43 protein spots identified by MS/MS in Mytilus eggs.

Common environment does not contribute to the resemblance between sibs, and thus $t_{G} \geq (0.5 \times \text{heritability} \times 100\%)$. This sets the average broad sense heritability of gene expression in this study at about 50% or greater. For a single spot, a different sample of individual mussels could have resulted in values of $t_{G}$ by chance higher or lower depending on the similarity of sibs. Treating the different spots as independent characters to obtain an average heritability might in part be equivalent to increasing the number of mussels analyzed.

Evolutionary and Ecological Significance of Expression Variation in Eggs—Although many studies have been made on egg characters in relation to life history or evolutionary theory, egg proteome variation has not yet been considered in this context. Knowledge of the phenotypic and genetic variation for fitness-related egg traits is important for the prediction of how they might change in evolution (64). Egg size is known to affect larval fitness in marine invertebrates (65). Thus whereas there is a fecundity benefit to small eggs, larger eggs, although more costly, might shorten the time of planktonic development by reducing dependence on external food sources. Such considerations may be less relevant for some animals, for example mammals where the embryo develops inside the mother and is supplied with nutrients by her (66). In the mussel Mytilus californianus, considerable phenotypic variation in egg volume and energy content was reported between individual females within populations (67). This demonstration of phenotypic variation for egg traits is consistent with the present study on M. edulis where significant biological variation ($t_{G}$) in protein expression is reported.

From the evolutionary viewpoint it is more pertinent to know the extent to which this phenotypic variation has a genetic basis. There are numerous examples of moderate to high heritability estimates for egg traits (e.g., Refs. 64 and 68–70). This suggests that egg traits have good potential to respond to natural selection. The present study is possibly the first to make an estimate of broad sense heritability for proteome profiles in eggs in a natural population. The value obtained suggests a moderate to high value for heritability, a result generally in line with the estimates for other egg traits. It is important to note that although the methods of quantitative proteomics by 2-DE allow conclusions about differences between individuals in the amount of a particular protein, they are always relative to the total amount for the spots studied.

Although a substantial heritability indicates good potential for evolutionary change, a fairly well established empirical observation is that traits closely related to fitness, for example reproductive traits, tend to have lower heritability than those less closely related (63, 71). The consequences of this for the rate of natural selection are unclear (72–74). The evolutionary significance of moderate to high heritability values for egg traits including protein expression thus remains an open question and justifies further work on the functional significance of egg proteins and their fitness-related variation.

Protein Identifications in Mussel Eggs Compared with Other Tissues—Comparative analysis is a fundamental tool in the identification of functional differences and adaptation (75). Comparison of expression differences of specific proteins between diverse tissues is difficult in proteome studies. However, the presence or absence of specific protein or function/process terms or their frequencies can be analyzed bearing in mind the need for a statistical framework and the assumptions of specific statistical tests.

It might be expected that a comparison between mussel and mammal eggs would reveal a difference in proteomic profile, reflecting the differences in how nutrients are delivered to the developing embryo (66) in these organisms. The external fertilization and development of mussel eggs in a highly variable and uncertain environment might require emphasis on proteins involved in resistance to heat stress such as chaperonins or in energy production during extended planktonic development. However, contingency analysis provides no evidence of significant differences in the distribution of ontology terms between mussel and mammal eggs. In fact, in both types of eggs, terms associated with stress response and protein folding and in cellular motility and the cytoskeleton have the highest and similar frequencies (Fig. 2). The cause of the significantly higher frequency of the Overall Process term RNA/DNA metabolism/regulation in mussel eggs compared with mussel tissue can be noted, but its cause is unclear. The Overall Process term stress response/defense has a significantly higher frequency in mussel tissue than in...
eggs. In this classification, proteins involved in chemical, oxidative, or parasitic stress responses are classified separately from those involved in protein folding and degradation. The former group is especially abundant in mussel tissue perhaps because the corresponding studies were more focused on chemical pollution or other forms of stress. However, many of these proteins were identified from a peroxisomal prefractation process used in mussel tissue studies that may bias toward this group.

The protein spots identified by MS in many proteomics studies on non-model organisms are likely to be biased toward those proteins that are abundant because they are visible on the gel and those that are easily identifiable in the databases. In the present study, for each line, log₂final was compared by ANOVA between three groups of spots. These are spots analyzed by MS/MS and identified, those analyzed but not identified, and those not analyzed. For both lines the spots identified by MS had significantly higher spot volume (p = 0.003, Line F; p = 0.000, Line M) than those in the other two groups, which were homogenous using the Tukey honestly significant difference test. Thus the spots identified here by MS/MS have significantly higher spot volume than those not identified. These high or medium abundance proteins are thus likely to be biased toward housekeeping genes that might be conserved between species and thus also facilitate database identification. However, even housekeeping proteins differing in expression between tissues cannot be excluded as candidates for functional analysis particularly because of the evidence that many proteins can have multiple functions (76).

Study of the Identified Proteins Using Text Mining—The protein terms identified in the present and the mammal egg and mussel tissue studies were also investigated using PubMed and text mining, a bioinformatics approach in which there is increas-

![Pie diagrams illustrating distribution of Function/Process and Overall Process ontology terms for Mussel Egg, Mammal Egg, and Mussel Tissue. The data are derived from supplemental Table 2.](image-url)
ing interest (77, 78), for example in relation to the detection of protein-protein interactions (79, 80). The potential functional importance of an interaction between two terms can be gauged by the frequency of co-occurrence of these two terms in PubMed abstracts. For example this is the basis of available programs such as EBIMed (81).

The number of PubMed hits were recorded for each protein term when searching in the title or abstract (supplemental Table 2, column C). The searches were then repeated but including terms relating to egg or oocyte (supplemental Table 2, column D). Supplemental Table 2 also gives an example of the search query format. A high number of hits in the second search (D) or a high value for the ratio (D/C) would indicate a high relative frequency of co-occurrence of the protein and egg/oocyte terms in the literature. The nonparametric Kruskal-Wallis test was then used to analyze the mean ranking of the values of D and the ratio D/C between three sets, x, y, and z. In the first test, x = proteins unique to Mussel Egg, y = proteins unique to Mammal Egg, and z = proteins shared by Mussel Egg and Mammal Egg. In the second test this was repeated substituting Mussel Tissue for Mammal Egg. In all cases, the ranking for those proteins shared by the two groups was higher. These proteins are actin, β-actin, α-tubulin, Hsp70, and calreticulin shared by Mussel Egg and Mammal Egg and actin, β-tubulin, Hsp70, calreticulin, mitochondrial malate dehydrogenase, ATP synthase, and calreticulin shared by Mussel Egg and Mussel Tissue. However, only for Mussel Egg and Mammal Egg are the ranking differences significant. For D the mean ranks are x = 24.5, y = 18.0, and z = 40.8 (p = 0.001), and for D/C these are x = 20.3, y = 22.0, and z = 36.7 (p = 0.040). All the shared proteins are high in the rankings. These are chaperonins and proteins important in cellular structure and motility corresponding with the function and process term proteins at highest frequency in both types of egg (Fig. 2). Thus these two analyses provide consistent and mutually supportive results that the shared terms at high-frequency co-occurrence of these two terms might have some causal link with the sex determination mechanism. One suggested mechanism involves the differential molecular labeling of mitochondria derived from eggs and sperm and interaction of these labels with recognition factors, which might or might not be proteins, to determine mitochondrial fate (83). Proteins involved in intracellular movement and the cytoskeleton might seem plausible candidates a priori for involvement in the differential behavior of mitochondria in Mytilus male embryos (where they remain aggregated) and female embryos (where they are dispersed) (44).

In the present study, terms relating to stress response and protein folding and perhaps cellular motility and the cytoskeleton were significantly more frequent in mussel eggs than were other terms (Table V). Of comparisons made between groups of terms sorted according to different criteria, terms relating to chaperones and stress response proteins have a higher frequency in the group having ≳1.5-fold difference in expression between lines, whereas proteins related to the core of cell metabolism (“housekeeping” functions, such as cytoskeleton and energy production) have higher frequency in the alternative group. The individual spots showing the greatest differential expression between lines (Table IV) might well be candidates for further investigation into the causes of DUI. As a further remark, it should be borne in mind that differences in sampling and identification of spots might reveal different pictures. The difficulties of interpretation are illustrated by α-tubulin. It is observed three times with -fold differences between lines of 4.2 (spot 6 greater in F), 1.1 (spot 245 greater in M), and 1.1 (spot 257 greater in M). Clearly any conclusions drawn about the functional significance of α-tubulin would be different if spot 6 were by chance omitted from the list compared with chance omission of spots 245 and 257.

Ideally future studies would compare the proteome of eggs from random samples of females from the population identified retrospectively as producing either mainly female or male offspring. It is hoped that a continuation of such studies will eventually allow egg proteins involved in DUI to be identified. Even if such proteins have only an indirect or consequential association with DUI they might provide important clues to the functional mechanism.

* This work was supported by a Marine Genomics Europe Network grant (European Union FP6 Contract GOCE-CT-2004-505403) and postdoctoral fellowship (to A. P. D.). The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked “advertisement” in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

†† To whom correspondence should be addressed. Tel.: 44-1792-295390; Fax: 44-1792-513430; E-mail: d.o.f.skibinski@swansea.ac.uk.

REFERENCES

1. Navas, A., and Albar, J. P. (2004) Application of proteomics in phylogenetic and evolutionary studies. Proteomics 4, 299–302
2. Biron, D. G., Loxdale, H. D., Ponton, F., Moura, H., Marché, L., Brugidou, C., and Thomas, F. (2006) Population proteomics: an emerging discipline to study metapopulation ecology. Proteomics 6, 1712–1715
3. Vasemägi, A., and Primmer, C. R. (2005) Challenges for identifying functionally important genetic variation: the promise of combining complementary research strategies. Mol. Ecol. 14, 3623–3642
4. Nuzhdin, S. V., Wayne, M. L., Harmon, K. L., and McIntyre, L. M. (2004) Common pattern of evolution of gene expression level and protein sequence in Drosophila. Mol. Biol. Evol. 21, 1308–1317
5. Whitehead, A., and Crawford, D. L. (2006) Neutral and adaptive variation in

Proteome of Mussel Eggs
