doi: 10.1387/ijdb.082719mk

THE INTERNATIONAL JOURNAL OF DEVELOPMENTAL BIOLOGY
www.intjdevbiol.com

SUPPLEMENTARY MATERIAL

corresponding to:

Proteomics analysis of regenerating amphibian limbs: changes during the onset of regeneration

MICHAEL W. KING*, ANTON W. NEFF and ANTHONY L. MESCHER

MS and data analysis

Peptides prepared from each pool of tissue were run on a Surveyor HPLC system (Thermo Electron) with a Zorbax 300SB-C18 column (1mm x 5cm). Each peptide pool (20 µg) was injected twice for a total of 20 independent injections. Samples were injected in random order. Peptides were eluted with a linear gradient from 5% to 45% acetonitrile developed over 120 min at a flow rate of 50 µL/min and the effluent was electro-sprayed into the Thermo Electron LTQ mass spectrometer. The source was set in positive ion mode with 4.8 kV electrospray potential, a sheath gas flow of 20 arbitrary units, and a capillary temperature of 225°C. The source lenses voltages were set by maximizing the ion current for the 2+ charge state of angiotensin. The data were collected in the "Data-Dependent Triple Play" (MS scan, Zoom scan, and MS/MS scan) mode. Dynamic exclusion was set to a repeat count of one, exclusion list duration of two minutes and rejection widths of –0.75 *m/z* and +2.0 *m/z*. Peak lists from raw MS/MS data were generated using Xcalibur 2.0. The acquired data were filtered and analyzed by a proprietary algorithm developed by Higgs *et al.* (2005, 2007). In all instances where data was analyzed using proprietary software developed by Higgs *et al.* (2005, 2007), this software was licensed to Monarch Life Sciences, LLC from Eli Lilly and Co.

Protein quantitation was carried out using a proprietary protein quantitation algorithm (Higgs *et al.*, 2005; Higgs *et al.*, 2007). Briefly, once the raw files were acquired from the LTQ, all extracted ion chromatograms (XIC) were aligned by retention time. To be used in the protein quantification procedure, each aligned peak must match parent ion, charge state, daughter ions (MS/MS data), and retention time (within a 1 min window). After alignment, the area-under-the-curve (AUC) for each individually aligned peak from each sample was measured, normalized, and compared for relative abundance. All peak intensities were transformed to a log scale before quantile normalization (Bolstadt *et al.*, 2003).

Protein identification

Database searches were carried out using the non-redundant (NR) NCBI protein database. SEQUEST and X!Tandem database search algorithms were used for peptide sequence identification. Each algorithm compares the observed peptide MS/MS spectrum and a theoretically derived spectrum from the database to assign quality scores (*XCorr* in SEQUEST and *E-Score* in X!Tandem). These quality scores and other important predictors are combined in a proprietary algorithm that assigns an overall score, %ID confidence, to each peptide (Higgs *et al.*, 2005; Higgs *et al.*, 2007). The assignment is based on a random forest recursive partition supervised learning algorithm. The %ID confidence cutoff in our study was 75%. The %ID confidence score is calibrated so that approximately X% of the peptides with %ID confidence >X% are correctly identified (Higgs *et al.*, 2005). It is also intuitively understood that our confidence in protein identification is increased with the number of distinct amino acid sequences identified. Therefore we also categorize proteins depending on whether they have one or multiple sequences of the required confidence. A protein is classified as 'YES' in the 'Multiple Sequences' column if it has at least two distinct amino acid sequences with the required ID confidence; otherwise it is classified as 'NO' (see Table 1).

The peptide ID confidence assigns a protein into a 'HIGH' or 'LOW' classification. This is based on the peptide with the highest peptide ID confidence (the best peptide). Proteins with best peptide having a confidence between 90-100% are assigned to the 'HIGH' category. Proteins with best peptide having a confidence between 75-89% are assigned to the 'LOW' category. All peptides with confidence less than 75% were pre-filtered out.

Proteins assigned with HIGH confidence were further divided into two categories (see Table 1) depending on whether multiple unique peptides were identified for the same protein (category 1) or only a single peptide was found for the identified protein (category 2). Table 1 gives the number of proteins with significant changes for each category. The threshold for significance is set to control the False Discovery Rate (FDR) for each two group comparisons at 5% (Reiner *et al.*, 2003). The FDR is estimated by the q-value which is an adjusted p-value. The FDR is the proportion of significant changes that are false positives. If proteins with a q-value ≤ 0.05 are declared significant it is expected that 5% of the declared changes will be false positives. It is a misconception that the p-value estimates the FDR. The p-value estimates the False Positive Rate (FPR) which is the proportion of false positives among the proteins that in reality did not change. The FPR = 1- specificity and FDR = 1 – positive predictive value in the language of medical diagnostics. (Note: the p-value to q-value adjustment is done separately for category 1, category 2 and the LOW confidence categories). The percent coefficient of variance (%CV) values are derived from the standard deviation divided by the mean on a % scale. Shown in Table 1 is the median %CV for each category. Variance components for the CV calculation were computed on the log scale and then converted to a CV on the arithmetic scale corresponding to the original AUC or 'area under the curve'. This transformation was done assuming the original scale has the log normal distribution. Therefore, ln(2) converts scale from log base 2 to natural log; exp is the exponential function (Limpert *et al.*, 2001).

Protein quantification, quality assurance and statistical analysis

All procedures for quantification of proteins, assurance of quality of the results and statistical analysis were carried out according to the detailed steps outlined by Fitzpatrick *et al.* (2007). Briefly, every peptide quantified has an intensity measurement for every sample. The intensity measurement is a relative quantity giving the area under the curve (AUC) from the extracted ion chromatogram (XIC) after background noise removal. The AUC is measured at the same retention time for each sample after the sample chromatograms have been aligned (Higgs *et al.*, 2005). The intensities are then transformed to the log base 2 scale which serves two purposes. First, relative changes in protein expression are best described by ratios. However ratios are difficult to model statistically and the log transformation converts a ratio to a difference which is easier to model. Second, as is frequently the case in biology, the data better approximate the normal distribution on a log scale (Limpert *et al.*, 2001), which is important because normality is an assumption of the ANOVA models used to analyze this data. The base of the log transform is arbitrary with base 2 the most common with genomic data. Log base 2 is popular because a 2-fold change (or doubling, or 100% increase) yielding an expression ratio of 2 is transformed to 1 on a log base 2 scale (i.e. a 2-fold change is a unit change on the log base 2 scale). After log transformation the data are then quantile normalized (Bolstadt *et al.*, 2003). This normalization removes trends introduced by sample handling, sample preparation, and possible total protein differences, as well as changes in instrument sensitivity while running multiple samples.

If multiple peptides have the same protein identification then their quantile normalized log base 2 intensities are averaged to obtain log base 2 protein intensities. The average of the normalized log peptide intensities is a weighted average. A peptide is weighted proportional to the peptide ID confidence. The log base 2 protein intensity is the final quantity that is fit by a separate ANOVA statistical model for each protein.

In this study, all injections were performed using the same microbore column and other instrumentation. In order to assess the stability of the column and instrument, chicken lysozyme was spiked into every sample at a constant amount before tryptic digestion. Since a constant amount of chicken lysozyme was spiked into each of the samples, it should show no significant change between groups. Therefore, we interpret significant changes in other proteins whose absolute FC is less than that for chicken lysozyme to be due to random non-specific effects. All proteins assigned to categories 1 and 2 in Tables 2 and 3 have an FC value higher than that for chicken lysozyme whose average FC was 1.085. FC is computed by dividing the mean treated group (3dPA) by the mean control group (0dPA) or the reciprocal for peptides higher at 0dPA. An FC of 1 means there is no change.

Additional references

BOLSTAD, B.M., IRIZARRY, R.A., ASTRAND, M., AND SPEED, T.P. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* 19:185-193.

FITZPATRICK, D.P.G., YOU, J.S., BEMIS, K.G., WERY, J.P., LUDWIG, J.R., AND WANG, M. (2007). Searching for potential biomarkers of cisplatin resistance in human overian cancer using a label-free LC/MS-based protein qualtification method. *Proteomics - Clinical Applications* 1: 246-263.

LIMPERT, E., STAHEL, W.A., AND ABBT, M. (2001). Log-normal distributions across the sciences: Keys and clues. *Bioscience* 51: 341-352.

REINER, A., YEKUTIELI, D., AND BENJAMINI, Y. (2003). Identifying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics* 19: 368-375.