BIOINFORMATICS DISCOVERY NOTE Vol. 26 no. 24 2010, pages 3007–3011
doi:10.1093/bioinformatics/btq541

*Data and text mining*

Advance Access publication November 9, 2010

# The necessity of adjusting tests of protein category enrichment in discovery proteomics

Brenton Louie[1,2,3], Roger Higdon[1,2,3] and Eugene Kolker[1,2,3,4,*]

[1]Bioinformatics and High-throughput Analysis Laboratory, [2]High-throughput Analysis Core, Seattle Children's Research Institute, Seattle, WA 98101, [3]Predictive Analytics, Seattle Children's Hospital, Seattle, WA 98145 and [4]Division of Biomedical and Health Informatics, Department of Medical Education and Biomedical Informatics, University of Washington Medical School, Seattle, WA 98195, USA

Associate Editor: Dmitrij Frishman

**ABSTRACT**

**Motivation:** Enrichment tests are used in high-throughput experimentation to measure the association between gene or protein expression and membership in groups or pathways. The Fisher's exact test is commonly used. We specifically examined the associations produced by the Fisher test between protein identification by mass spectrometry discovery proteomics, and their Gene Ontology (GO) term assignments in a large yeast dataset. We found that direct application of the Fisher test is misleading in proteomics due to the bias in mass spectrometry to preferentially identify proteins based on their biochemical properties. False inference about associations can be made if this bias is not corrected. Our method adjusts Fisher tests for these biases and produces associations more directly attributable to protein expression rather than experimental bias.

**Results:** Using logistic regression, we modeled the association between protein identification and GO term assignments while adjusting for identification bias in mass spectrometry. The model accounts for five biochemical properties of peptides: (i) hydrophobicity, (ii) molecular weight, (iii) transfer energy, (iv) beta turn frequency and (v) isoelectric point. The model was fit on 181 060 peptides from 2678 proteins identified in 24 yeast proteomics datasets with a 1% false discovery rate. In analyzing the association between protein identification and their GO term assignments, we found that 25% (134 out of 544) of Fisher tests that showed significant association ($q$-value $\leq$0.05) were non-significant after adjustment using our model. Simulations generating yeast protein sets enriched for identification propensity show that unadjusted enrichment tests were biased while our approach worked well.

**Contact:** eugene.kolker@seattlechildrens.org

**Supplementary information:** Supplementary data are available at *Bioinformatics* online.

Received on June 29, 2010; revised on August 26, 2010; accepted on September 18, 2010

## 1 INTRODUCTION

Enrichment tests are commonly performed in high-throughput discovery research such as mass spectrometry-based proteomics. They are used, for instance, to determine the association between whether a protein is expressed in a particular condition and membership in a particular biologically relevant category such as Gene Ontology (GO) term assignment. Discovery of significant associations, such as GO terms with over- or underrepresentation of expressed proteins, then provide new hypotheses and directions for further research. These enrichment tests are commonly carried out by using Fisher's exact or chi-square tests (Curtis *et al.*, 2005). These are the most basic of the enrichment tests; others have been proposed such as gene set enrichment analysis (Subramanian *et al.*, 2005). However, given their clear-cut nature Fisher's exact tests are still often applied and are considered robust enough to be used as a guide or ranking mechanism for biological categories of proteins in discovery research.

Mass spectrometry has now become the standard tool in high-throughput discovery proteomics. It enables measurement of expressed proteins (as opposed to gene sequences) that control and enable most biological processes (Kolker *et al.*, 2006; Pandey and Mann, 2000). Research studies will often compare the distribution of expressed proteins in a given set of conditions or between organisms (Schrimpf *et al.*, 2009). The problem with analyzing proteomics data, however, is the bias in how proteins are identified; which is a function of their biochemical properties (Braisted *et al.*, 2008; Mallick *et al.*, 2007). These biases can affect tests of enrichment for expressed proteins within proteins sets and produce misleading biological conclusions in discovery proteomics studies if they are not accounted for.

To address this, we demonstrate an approach to adjust Fisher's exact tests specifically for protein identification bias. Our approach is similar to 'propensity scores', developed for observational studies in epidemiology, to adjust comparisons due to confounding variables possibly associated with grouping and outcome variables (Rosenbaum and Rubin, 1983). This helps avoid finding spurious relationships between study groups. Propensity scores are used by building a model to predict the probability that an object has a given outcome based on potentially confounding variables (covariates). They are then used to create strata that are balanced by their covariates. In our case, the study groups are the protein categories (GO terms) and the response variable is protein identification by mass spectrometry. We used logistic regression to model the relationship between protein identification, protein GO term assignment and propensity scores of proteins assigned to the GO term (see Section 2). Our method is different from those which adjust enrichment $P$-values by accounting for the hierarchical relationships

*To whom correspondence should be addressed.

between GO terms (Alexa *et al.*, 2006; Falcon and Gentleman, 2007) and not for protein propensities.

When we analyzed a large yeast proteomics dataset, we found that many significant associations between GO terms and protein identifications found using the Fisher's exact test were actually not significant by our method. In addition, some non-significant Fisher's exact tests became significant by our method, but the number was smaller. Given these results we believe that our method produces associations which are more directly attributable to true differences in protein expression between categories rather than the inherent experimental bias in discovery proteomics.

## 2 METHODS

### 2.1 Mass spectrometry discovery datasets

A total of 24 mass spectrometry discovery datasets in yeast were obtained from the Open Proteomics Database, Peptidome and PeptideAtlas and searched against the Saccharomyces Genome Database (SGD) (Hong *et al.*, 2008). Datasets were analyzed as in our previous publication (Higdon *et al.*, 2010). The data were input to the SPIRE proteomics pipeline with false discovery rates (FDRs) estimated on a randomized database approach using isotonic regression (Hather *et al.*, 2010). Proteins with FDRs fixed at 1% were combined across experiments.

Of 6717 proteins (open reading frames) from the SGD there were 2678 identified at 1% FDR, within the 24 datasets. We generated fully tryptic peptides from these 2678 proteins for a total of 273 270 peptides. A total of 35 properties (e.g. hydrophobicity) were then determined for the peptides (Braisted *et al.*, 2008). Peptides were then filtered to only retain peptides with molecular weights between 530 and 3326 Da, which were the minimum and maximum molecular weights actually identified in the experiments. This resulted in 181 068 final peptides which were separated into equal size test and training sets by random sampling.

### 2.2 Peptide and protein identification model

We developed a logistic model to represent the bias in mass spectrometry to preferentially identify peptides with particular biochemical properties. Our model for peptide identification is:

$$\mathrm{id}_{\mathrm{pep}} \sim \mathrm{mw}_{\mathrm{pep}} + \mathrm{hyd}_{\mathrm{pep}} + \mathrm{pI}_{\mathrm{pep}} + \mathrm{bt}_{\mathrm{pep}} + \mathrm{te}_{\mathrm{pep}} \quad (1)$$

where $\mathrm{id}_{\mathrm{pep}}$ is the probability (logit) of a peptide identification based on the molecular weight (mw), hydrophobicity (hyd), pI (isoelectric point), frequency of beta turn (bt) and transfer energy (te) of the peptide. These were determined from the AAindex (Kawashima and Kanehisa, 2000).

The five predictors were selected based on hierarchical clustering (R hclust function). Many of the possible 35 predictors were highly correlated (Supplementary Material), which left us with the choice of the remaining five. A mixed model with proteins as a random effect was evaluated to estimate possible differences in peptide properties while accounting for differences in protein concentrations. There was little difference in the fits between the mixed model and a model ignoring protein membership, indicating little variation in the peptide properties of proteins across different concentrations.

Protein propensities were determined by aggregating the propensities of their tryptic peptides using the following formula:

$$\mathrm{propensity}_{\mathrm{protein}} = -\log\left(\prod 1 - \mathrm{id}_{\mathrm{peptides}}\right) \quad (2)$$

where $\mathrm{id}_{\mathrm{peptides}}$ are the peptide identification probabilities associated with a given protein determined from the above peptide identification model. We considered a number of different summarization methods but chose this approach because it was most highly associated with protein identification and produced the most normally distributed protein propensity scores. All models were fit using the glm function in the R statistical package.

### 2.3 Unadjusted enrichment tests

Unadjusted enrichment tests to determine odds ratios (ORs) and *P*-values for annotation categories were performed using two-sided Fisher's exact tests as reviewed in (Curtis *et al.*, 2005). Categories correspond to GO terms assigned to proteins as determined by the SGD. Note the basic equivalence between Fisher's exact, chi-square and likelihood ratio tests based on logistic regression for large sample sizes (McCullagh and Nelder, 1989). The null hypothesis for these tests is that the proportions (distributions) between two variables (e.g. has GO term or not) are the same. Also note the hierarchy of the GO should be accounted for (Falcon and Gentleman, 2007). To account for the GO hierarchy, children of parent terms were included in counts to satisfy the 'true path rule' of the GO.

### 2.4 Adjusted enrichment tests

We developed a second logistic model that represents our method for adjusting enrichment tests. The formula for the model is:

$$\mathrm{id}_{\mathrm{prot}} \sim \mathrm{category}_{\mathrm{prot}} + \mathrm{propensity}_{\mathrm{prot}} \quad (3)$$

where $\mathrm{id}_{\mathrm{prot}}$ is the probability (logit) that a protein is identified, the coefficient for the *category* term corresponds to the adjusted (log) OR of the protein category (GO term) after adjusting for the propensity of the proteins associated with that term to be identified (the *propensity* term). A *P*-value is associated with the OR of the category term. The model was fit using the glm function in R.

### 2.5 Adjusting for multiple testing

*P*-values for unadjusted and adjusted enrichment tests were converted to *q*-values using the R *q*-value package (Storey and Tibshirani, 2003). *Q*-values $\leq 0.05$ were considered significant. Differences in ORs between adjusted and unadjusted tests were also calculated. Enrichment tests where the number of proteins (identified or not) under consideration were $<10$ were excluded.

## 3 RESULTS

The peptide model developed on the training set indicated that mass spectrometry indeed preferentially identifies peptides with particular properties, with molecular weight and hydrophobicity having the largest influence. Model selection on the test dataset indicated that a model with quadratic terms (with the exception of hydrophobicity) provided superior performance over a model with linear terms by receiver operator characteristic (ROC) curve analysis (Supplementary Material). This is to be expected given the relationship between many of the peptide properties and the probability of peptide identification (Fig. 1). As a validity check of our peptide model, predicted probabilities from our model were compared with a model using a 'random forest' approach, the enhanced signature peptide predictor (ESP) (Fusaro *et al.*, 2009). ROC curve analysis indicated that our logistic model outperformed the random forest model (area under the curve 76.5% versus 66.9%). Possible explanations could be overfitting or high correlation between peptide properties, indicating non-independence between variables. Also, the ESP model was not developed specifically for shotgun proteomics and is possibly inappropriate in our scenario. Developing and training our own model was therefore a logical choice. Another advantage of our logistic model it that it is easier to deduce that molecular weight and hydrophobicity have by far the most influence on the probability of peptide identification by mass spectrometry.

For protein propensity scores, the propensity for each protein was determined by aggregating the predicted probabilities of its tryptic

peptides (see Section 2). Statistical analysis (using another logistic model) indicated a 6.2% increase in the odds of identifying a protein per unit increase in their propensity score.

To compare unadjusted versus adjusted tests, $q$-values for all GO terms for identified and not identified proteins were determined (see Section 2). Terms were from all three GO ontologies: GO cellular component (CC), GO biological process (BP) and GO molecular function (MF). The incorrect assumption in unadjusted enrichment tests is that a significant result reflects the true difference in protein expression for proteins assigned a particular GO term. We know that is not true since we have seen that mass spectrometry preferentially identifies certain proteins (Fig. 1). In Figure 2, we illustrate the difference in $q$-values between unadjusted and adjusted enrichment tests for all applicable GO terms in our study as well as the absolute percent change in the ORs for GO terms. Out of 1587 terms (326 CC, 904 BP, 357 MF) 29 had significant enrichment tests ($q$-value $\leq$0.05) in the adjusted test and not significant in the unadjusted test, and 134 terms were not significant in the adjusted test and significant in the unadjusted test. In all, 134 out 544 (25%) originally significant unadjusted tests became non-significant when adjusting for protein propensity. Overall, the tendency was for enrichment tests to become non-significant after adjustment (34.3% versus 27.7% significant). Note, however, that the decrease in the number of significant tests would tend to increase $q$-values in general since the denominator in the FDR would tend to be smaller. All test results are provided as Supplementary Material.

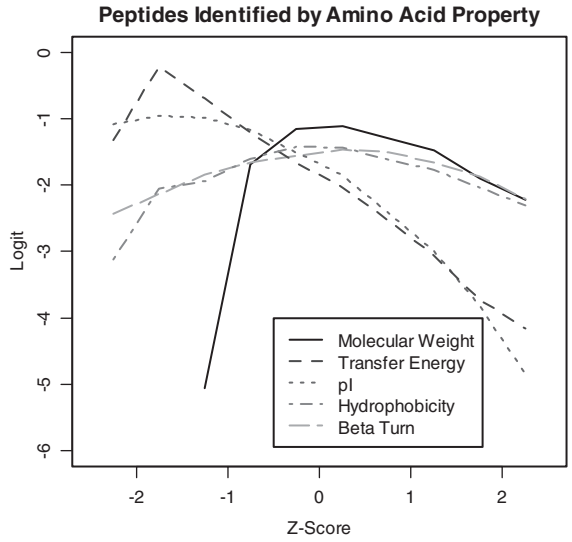


**Fig. 1.** Relationships between the proportion of identified peptides (logit transformation) and standard scores ($Z$-scores) of five peptide properties. The relationship between molecular weight and identification, for instance, suggests a quadratic relationship in the peptide identification model, unlike hydrophobicity that is a linear relationship.

Specific examples of GO terms with $P$ and $q$-values which changed from significant to non-significant and vice versa can be found in Table 1. All terms are associated with proteins with much higher or lower propensity to be identified by mass spectrometry than average. For instance, GO:0005815 (*mitochondrial organizing center*) contains GO terms that have lower propensity scores than average ($Z$-score = −2.55). Our method produces a significant $P$-value for this term versus a non-significant $P$-value from a Fisher's exact test (0.01 versus 0.07). In contrast, for GO:0006418 (*tRNA aminoacylation for protein translation…*) a Fisher's exact test produces a significant result versus a non-significant result with our method (0.02 versus 0.10). Proteins associated with this GO term have higher propensity scores than average ($Z$-score = 0.93) and are more likely to be identified. See Table 1 for examples of unadjusted versus adjusted tests for GO terms from all three ontologies.

Adjusting is clearly more necessary when protein sets are enriched in terms of identification propensity. We simulated

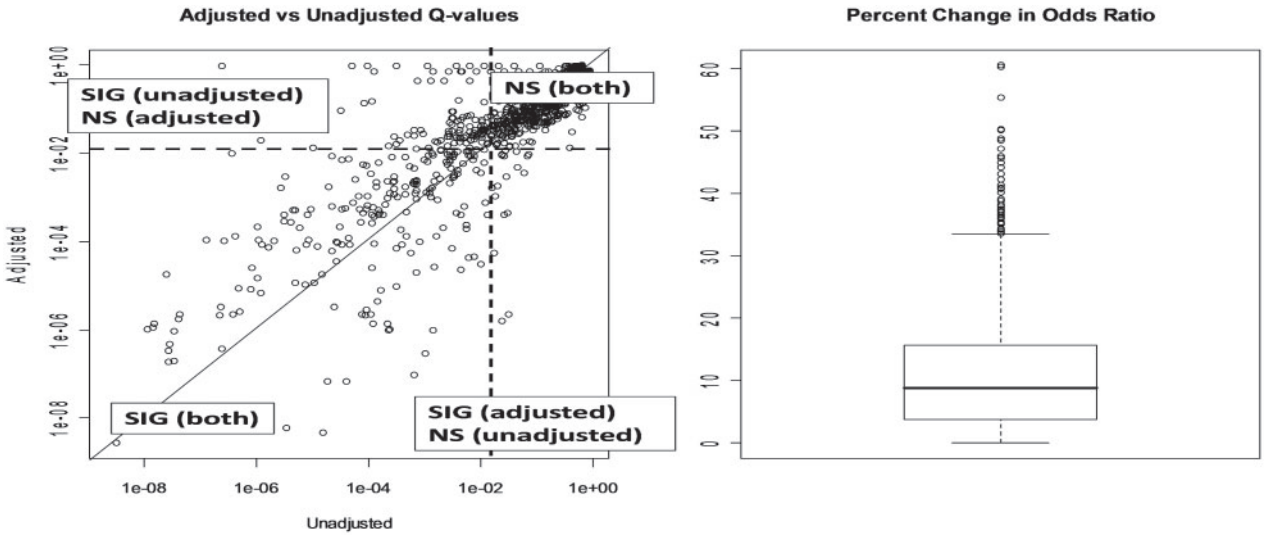


**Fig. 2.** Pairwise plot of $q$-values from unadjusted and adjusted enrichment tests of GO terms (log transformed) for identified versus not identified proteins. Dashed lines indicate significance cutoffs ($q$-value $\leq$0.05) for both unadjusted and adjusted tests. The 163 enrichment tests either changed from not significant (NS) to significant (SIG) or vice versa (indicated by boxes with SIG and NS). Also shown is a boxplot of the absolute percent change in ORs between unadjusted and adjusted tests. The average percent change is about 11% (8.8% median), with outliers over 30%.

**Table 1.** Examples of GO terms with a practical change in their *P*-values between unadjusted and adjusted tests

| GO term | *P*/*q*-values unadjusted | OR unadjusted | *P*/*q*-values adjusted | OR adjusted | *Z*-score | Description |
|---|---|---|---|---|---|---|
| GO:0005815 | 0.07/0.09 | 0.65 | 0.01/0.04 | 0.54 | −2.55 | Microtubule organizing center |
| GO:0005740 | 0.24/0.22 | 1.41 | 0.03/0.08 | 1.29 | 2.24 | Mitochondrial envelope |
| GO:0051276 | 0.29/0.41 | 0.90 | 0.01/0.03 | 0.76 | −2.61 | Chromosome organization and biogenesis |
| GO:0000003 | 0.25/0.38 | 0.89 | 0.003/0.01 | 0.74 | 0.65 | Reproduction |
| GO:0006418 | 0.02/0.05 | 2.29 | 0.10/0.18 | 1.77 | 0.93 | tRNA aminoacylation for protein translation… |
| GO:0015075 | 0.11/0.13 | 0.69 | 0.05/0.04 | 0.61 | −2.11 | Chromatin binding |
| GO:0032559 | $1.2e^{-05}$/0.14 | 1.46 | 0.17/$8.4e^{-05}$ | 1.14 | 1.11 | Adenyl ribonucleotide binding |

GO terms associated with proteins which have lower propensity to be identified have *Z*-scores less than 0.0. *Z*-scores greater than 0.0 indicate GO terms associated with proteins with higher propensity to be identified. Adjusted and unadjusted *q*-values and ORs are also included in the table.

protein datasets that were enriched for identification propensity by randomly selecting proteins with higher than average propensity. We performed 1000 simulations based on protein sets of size 25, 50 and 100. The sets on average did not have increased protein expression as measure by tandem affinity purification and green fluorescent protein (Ghaemmaghami *et al.*, 2003) as the correlation between propensities and concentration was small (actually slightly negative at −0.07). Unadjusted enrichment tests were severely biased toward being statistically significant with 12–44% having *P*-values <0.05. Adjusted enrichment tests have very near nominal levels of *P*-values (4–6% with *P*-values <0.05). Similar results were obtained for below average propensity datasets (see Supplementary Material for full details).

## 4 DISCUSSION

We introduce here a discovery proteomics data analysis method for performing accurate tests of enrichment for expressed proteins by adjusting for the bias in mass spectrometry to identify proteins with particular biochemical properties. The fact that mass spectrometry instruments preferentially identify proteins has been studied previously (Braisted *et al.*, 2008; Fusaro *et al.*, 2009). To our knowledge, however, no one before has used this information to adjust enrichment tests to more accurately determine which categories of proteins are significantly associated with over- or underrepresentation of expressed proteins. Our approach theoretically works on any type of protein category: GO terms, KEGG pathways or otherwise. It is certainly logical from a biological standpoint that proteins with similar biochemical properties associate in similar categories. Very hydrophobic proteins for example often tend to be associated with cell membranes (e.g. GO:0015075 in Table 1). It is these biochemical properties that translate into varying propensities for identification in mass spectrometry and make adjusting enrichment tests necessary to avoid incorrect conclusions of biological significance in discovery proteomics experiments.

We emphasize again that previous methods to adjust the significance of enrichment tests on GO terms do so in an entirely different manner than our approach. These other methods work by 'de-correlating' related GO terms in various ways by accounting for the relationships between GO terms (Alexa *et al.*, 2006; Falcon and Gentleman, 2007). Unlike their approach, we adjust for the nature of the data itself by considering the propensity of some proteins to be preferentially identified in mass-spectrometry. Also, the functionality of these previous methods is largely confined to the GO (or produces trivial results on non-hierarchical categories), whereas our method is category independent. For instance, it works on KEGG pathways that are not hierarchically structured. Although a caveat with our method may be that it is specifically designed for mass spectrometry data, mass spectrometry has basically become the standard method in discovery proteomics. Also, despite the fact that we have applied our method to protein identification, it is relevant for comparing quantitative expression. This is due to the fact that the inherent bias in how proteins are identified remains true in how they are quantified (e.g. using spectral counts or total ion current for determining their concentrations). Adjusting for identification bias is obviously less relevant for relative expression since comparisons are made between identical sequences across conditions. However, adjustment for identification bias may improve precision in measurements of protein expression by more optimally weighting individual peptides thereby improving the power of tests for relative expression.

So, it is apparent that naïve application of enrichment tests is problematic. Enrichment tests are intended to measure the true difference in expression between protein sets not bias in how proteins are measured. We would like to state here that assessing the areas of possible variation and biases in data is an important analytical exercise. Transparent statistical models in research such as the logistic model that we have employed herein can more clearly illuminate areas of variability and possible bias (Higdon and Kolker, 2007; Higdon *et al.*, 2008; Louie *et al.*, 2009). Also, all instrumentation methods certainly contain biases in measurement; therefore, adaptations of our approach could be applied to other high-throughput analysis approaches. Our intent is to open the door to the possibilities of adjusting enrichment tests for other sorts of data gathered by a myriad of various means.

## ACKNOWLEDGEMENTS

We thank Gerald van Belle and Elizabeth Stewart for their feedback and insightful comments.

*Funding*: We appreciate the support of National Institutes of Health 5R01 GM076680-02 and UO1 DK072473, National Science Foundation DBI-0544757 and NSF-07140 and Seattle Children's Research Institute internal funds.

*Conflict of Interest*: none declared.

## REFERENCES


Alexa,A. *et al.* (2006) Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. *Bioinformatics*, **22**, 1600–1607.

Braisted,J. *et al.* (2008) The APEX quantitative proteomics tool: generating protein quantitation estimates from LC-MS/MS proteomics results. *BMC Bioinformatics*, **9**, 529.

Curtis,R.K. *et al.* (2005) Pathways to the analysis of microarray data. *Trends Biotechnol.*, **23**, 429–435.

Falcon,S. and Gentleman,R. (2007) Using GOstats to test gene lists for GO term association. *Bioinformatics*, **23**, 257–258.

Fusaro,V.A. *et al.* (2009) Prediction of high-responding peptides for targeted protein assays by mass spectrometry. *Nat. Biotechnol.*, **27**, 190–198.

Ghaemmaghami,S. *et al.* (2003) Global analysis of protein expression in yeast. *Nature*, **425**, 737–741.

Hather,G. *et al.* (2010) Estimating false discovery rates for peptide and protein identification using randomized databases. *Proteomics*, **10**, 2369–2376.

Higdon,R. and Kolker,E. (2007) A predictive model for identifying proteins by a single peptide match. *Bioinformatics*, **23**, 277–280.

Higdon,R. *et al.* (2008) A note on the false discovery rate and inconsistent comparisons between experiments. *Bioinformatics*, **24**, 1225–1228.

Higdon,R. *et al.* (2010) Meta-analysis for protein identification: a case study on yeast data. *OMICS*, **14**, 309–314.

Hong,E.L. *et al.* (2008) Gene Ontology annotations at SGD: new data sources and annotation methods. *Nucleic Acids Res.*, **36**, D577–D581.

Kawashima,S. and Kanehisa,M. (2000) AAindex: amino acid index database. *Nucleic Acids Res.*, **28**, 374.

Kolker,E. *et al.* (2006) Protein identification and expression analysis using mass spectrometry. *Trends Microbiol.*, **14**, 229–235.

Louie,B. *et al.* (2009) A statistical model of protein sequence similarity and function similarity reveals overly-specific function predictions. *PLoS ONE*, **4**, e7546.

Mallick,P. *et al.* (2007) Computational prediction of proteotypic peptides for quantitative proteomics. *Nat. Biotechnol.*, **25**, 125–131.

McCullagh,P. and Nelder,J.A. (1989) *Generalized Linear Models*, 2nd edn. Chapman & Hall/CRC, London.

Pandey,A. and Mann,M. (2000) Proteomics to study genes and genomes. *Nature*, **405**, 837–846.

Rosenbaum,P.R. and Rubin,D.B. (1983) The central role of the propensity score in observational studies for causal effects. *Biometrika*, **70**, 41–55.

Schrimpf,S.P. *et al.* (2009) Comparative functional analysis of the Caenorhabditis elegans and Drosophila melanogaster proteomes. *PLoS Biol.*, **7**, e48.

Storey,J.D. and Tibshirani,R. (2003) Statistical significance for genomewide studies. *Proc. Natl Acad. Sci. USA*, **100**, 9440–9445.

Subramanian,A. *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA*, **102**, 15545–15550.