Label-Free Quantitative Shotgun Proteomics Using Normalized Spectral Abundance Factors

Article *in* Methods in molecular biology (Clifton, N.J.) · April 2013

DOI: 10.1007/978-1-62703-360-2_17 · Source: PubMed

5 authors, including:

Karlie A Neilson
Macquarie University
16 PUBLICATIONS **954** CITATIONS
SEE PROFILE

Tim Keighley
Macquarie University
25 PUBLICATIONS **383** CITATIONS
SEE PROFILE

Dana Pascovici
Macquarie University
86 PUBLICATIONS **966** CITATIONS
SEE PROFILE

Brett Cooke
Macquarie University
4 PUBLICATIONS **67** CITATIONS
SEE PROFILE

Chapter 17

Label-Free Quantitative Shotgun Proteomics Using Normalized Spectral Abundance Factors

Karlie A. Neilson, Tim Keighley, Dana Pascovici, Brett Cooke, and Paul A. Haynes

Abstract

In this chapter we describe the workflow used in our laboratory for label-free quantitative shotgun proteomics based on spectral counting. The main tools used are a series of R modules known collectively as the Scrappy program. We describe how to go from peptide to spectrum matching in a shotgun proteomics experiment using the XTandem algorithm, to simultaneous quantification of up to thousands of proteins, using normalized spectral abundance factors. The outputs of the software are described in detail, with illustrative examples provided for some of the graphical images generated. While it is not strictly within the scope of this chapter, some consideration is given to how best to extract meaningful biological information from quantitative shotgun proteomics data outputs.

Key words Shotgun proteomics, Label-free, Quantitative proteomics, Spectral counting, Normalized spectral abundance factors

1 Introduction

The field of shotgun proteomics has changed considerably in recent years as it has become less descriptive and more quantitative. Nowadays it has become increasingly common to see published studies which contain an abundance and richness of data which was once thought unattainable. There are many papers in the literature which include thousands of detailed individual protein measurements within a given cellular system, each of which includes the identity and relative amount of the protein in question. This can be performed for multiple samples, such as numerous points across a developmental or stress-imposition time course. The output of such experiments can be overwhelmingly large, but successful analysis of such data sets can reveal trends at the "big picture" level which are not discernible by other means.

In our laboratory we employ label-free quantitation using normalized spectral abundance factors (NSAFs). It must be emphasized

Ming Zhou and Timothy Veenstra (eds.), *Proteomics for Biomarker Discovery: Methods and Protocols*, Methods in Molecular Biology, vol. 1002, DOI 10.1007/978-1-62703-360-2_17,

that this is just one of many such techniques that can be used; we use this because it is simple, robust, and inexpensive, and relies on sound mathematical principles. Similarly, there are a myriad of possibilities for how to take biological samples of a given cell or tissue type, and transform them into a set of fractionated peptides or proteins suitable for mass spectrometric analysis. We present in this chapter one of the main techniques we use in our laboratory, which is SDS-PAGE (sodium dodecyl sulfate-polyacrylamide gel electrophoresis) fractionation of proteins prior to in-gel trypsin digestion. Again, we use this separation technique because it is simple, inexpensive, and robust. SDS-PAGE fractionation of proteins also has one advantage over many other techniques in that the SDS buffer is an excellent protein-solubilizing agent, especially in comparison to other, milder detergents.

There are two main ways of measuring changes in protein abundance without using metabolic or isotopic labels that involve measuring precursor ion intensity or counting spectra assigned to a particular protein (1–3). Approaches involving precursor ion intensity are based on the well-established analytical principle that the area under a chromatographic elution curve is proportional to the amount of eluting compound. This works very well for relatively simple mixtures but tends to be less accurate as peptide mixtures become increasingly complex. The approach relies on very accurate and reproducible chromatography, as peptide peaks from different chromatographic elution profiles need to be precisely aligned for the analysis to proceed.

The other main class of approach used in label-free quantitation involves counting spectra identified for the peptides in a protein, also known as spectral counting (4). This approach relies on the simple observation that as more of a digested protein is analyzed in a given mass spectrometric system, more peptides belonging to that protein will be identified. Hence, the number of spectra assigned to each protein present in a complex mixture can be used as a measure of relative abundance for each protein individually (5, 6).

A major conceptual advance in this field arose from the observation that the length of a protein affects the number of spectral counts; a longer protein will generate more identifiable peptides than the same molar amount of a smaller protein (3, 4). This will have an adverse effect when counting raw spectra to calculate abundance of a protein. The introduction of normalized spectral abundance factors (NSAFs) (4) provides an improved measure for relative abundance, by factoring the length of the protein into subsequent calculations (3). An NSAF value for a given protein is calculated by dividing the spectral counts (SpC) for a protein by its length (L). This value is then normalized by dividing by the sum of all SpC/L for all proteins identified in a complex mixture (7). NSAF values provide a measure of relative abundance

and the ability to compare the abundance of proteins within a sample (4). The dynamic range for NSAF values is approximately 3.6–3.8 orders of magnitude, allowing the measurement of abundance of a wide range of proteins present in a data set (8). When NSAF values are log-transformed they follow a normal distribution, facilitating analysis of statistically significant changes in expression (4). NSAF values have also been shown to have very similar statistical properties to comparable RNA transcript abundance values (8). This statistical comparability is important as it means that software and analysis tools developed for transcriptomics studies can also be applied to NSAF values in proteomic data sets.

It is important to emphasize the need for high-quality protein identification data when generating NSAF values. Protein and peptide data sets need to be filtered to a very low false discovery rate before meaningful NSAF values can be produced; only then can statistical significance be attached to changes in NSAF values of proteins observed in response to changes in a biological system. Also, it is important to optimize other experimental parameters in order to obtain worthwhile results. One detailed study has already demonstrated that the use of correct dynamic exclusion parameters in nanoLC-MS/MS has little or no effect on data quality, while the use of non-optimal dynamic exclusion parameters can cause distortions in quantitation (9).

Another important consideration is how to account for shared peptides between multiple proteins. It has been shown that the best approach is to apply NSAF values based on distributed spectral counts; shared spectral counts were distributed based on the number of additional spectral counts that belonged uniquely to each isoform (10). Other studies have shown that the far simpler approach of distributing multiple copies of a spectral count across shared protein sequences is a reasonably accurate approach to take (11, 12).

One of the early studies using NSAF-based quantitation involved the analysis of nuclear proteins from yeast. Nuclei were isolated and 2,674 proteins were identified and quantified. Low-abundance proteins associated with transcriptional regulation were identified and found to be present at low amounts, as expected. NSAF values have been used in a broad range of studies as a measure of relative abundance of identified proteins. Examples of such projects include peptide IPG-IEF profiling of rat liver membrane proteins (12), subcellular analysis of nuclear proteins in yeast (3), profiling temperature stress responses in rice (13), comparison of evolutionary adaptation of Pachycladon species (14), assembly of a probabilistic human protein interaction network (15), analysis of mouse renal cortex proteins (10), and characterization of the response of Sydney rock oysters to environmental heavy metal stresses (16).

The essential mathematical steps involved in transforming raw protein identification outputs into NSAF values can be performed in, for example, Excel spreadsheets. However, this is a laborious process and is constrained by the limited mathematical analysis tools available. Hence, numerous research groups have created software analysis packages suited to this purpose. One example is PepC, a program that identifies statistically significant differentially expressed proteins based on spectral counting (17). PepC is a Java-based program that can be used as web server module associated with the trans-proteomic pipeline (TPP) (18). The software statistically assesses spectral counting data based on a *G*-test to assess the difference in spectral counts across samples and a *t*-test to assess data reproducibility, but does not perform a data normalization step. Another example is Census, a software tool capable of processing most types of quantitative data, including both labelled and label-free proteomics experiments; the latter can be either area under the curve (AUC) or spectral counting methods (19). Census is able to quantitate data generated by both AUC and spectral counting methods, and employs an approach based on RelEx, an application previously released by the same group (20).

In this chapter we present details of the analysis pipeline we have used in a number of different publications and other ongoing projects. These details include peptide separation and analysis using nanoLC-MS/MS, peptide identification by peptide-to-sequence matching using the XTandem algorithm, quantitation of identified proteins using normalized spectral abundance factors, statistical analysis of proteins differentially expressed between samples using the Scrappy software package, and consideration of how to best extract biologically relevant information from such experiments.

2 Materials

Prepare all solutions using Milli-Q water or equivalent and the highest quality analytical grade reagents. Prepare and store all reagents at room temperature, unless otherwise indicated. All waste disposal regulations should be strictly adhered to when disposing of waste materials.

1. Zorbax C18 chromatography packing material (5 μm particle size: Agilent Technologies).
2. Readw.exe is available for free download from: http://sourceforge.net/projects/sashimi/files/.
3. The XTandem algorithm is available for free download from: http://www.thegpm.org/tandem/instructions.html.
4. A freely available version of the XTandem algorithm known as GPM-XE Tornado, which installs and runs locally on a

Windows PC, is available from https://proteomecommons.org/dataset.jsp?i=74059.

5. The Scrappy program is available as a series of R modules which can be download from: https://proteomecommons.org.

3 Methods

3.1 Shotgun Proteomics

The workflow described below is applicable to any type of label-free shotgun proteomics experiment. We routinely used SDS-PAGE gel slice shotgun experiments and gas phase fractionation, both of which have been described in detail elsewhere (13, 21–24). It is equally applicable to data produced from online or offline MudPIT experiments, peptide IPG-IEF fractionation, filter assisted sample preparation (FASP) (25), or any of the other myriad techniques commonly available.

The required features are that it is a shotgun data set comprising analysis of three biological replicates of at least two samples to be compared. Each of the individual replicate analyses typically contains hundreds of thousands of individual MS/MS spectra. For reasons of both clarity and brevity we have written this procedure focussing on a pairwise example experiment where the aim is a quantitative comparison of control versus stressed samples. For the figures in this chapter, we have used two data points ("control" and "48 cold") taken from a previously published experiment where rice plants were exposed to low temperature over a 4-day time period (24).

It is also possible to do this type of analysis with more than two samples, such as for a developmental time course or a comparison of varying degrees of temperature or water stress (13, 21, 23, 24). This requires different mathematical assumptions and models, and becomes much more difficult when comparing multiple samples without a defined reference point, such as in our study of five different New Zealand geographical isolates of Pachycladon, an endemic plant (22). That type of analysis becomes more about looking for broad trends in large amounts of data; in our case that was greatly facilitated by concurrent microarray and metabolite analysis which provided an information framework.

3.2 NanoLC-MS/MS

1. Sequentially analyze each of the peptide digest fractions using a nanoLC-MS/MS system, employing an LTQ-XL ion-trap mass spectrometer, Surveyor HPLC pump and Surveyor autosampler (Thermo, San Jose, CA).
2. Prepare an approximately 7 cm (100 μm i.d.) reversed phase columns using 100 Å, 5 mM Zorbax C18 resin (Agilent Technologies, CA, USA) in a fused silica capillary with an integrated electrospray tip (see Note 1).

3. Apply a 1.8 kV electrospray voltage to a gold-electrode liquid junction upstream of the C18 column.
4. Load each sample onto the C18 column followed by an initial wash step with buffer A (5 % (v/v) ACN, 0.1 % (v/v) formic acid) for 10 min at 1 μL/min.
5. Elute the peptides from the C18 column with 0–50 % buffer B (95 % (v/v) ACN, 0.1 % (v/v) formic acid) over a 30 min linear gradient min at 500 nL/min followed by 50–95 % buffer B over 5 min at 500 nL/min, and 5 min was with 95 % buffer B prior to column re-equilibration.
6. Direct the column eluate into the nanospray ionization source of the mass spectrometer (see Note 2).
7. Scan the spectra over the range 400–1,500 amu. Automated peak recognition, dynamic exclusion (90 s), and tandem MS of the top six most intense precursor ions at 40 % normalization collision energy were performed using Xcalibur software (Thermo) (see Note 3).

3.3 Protein and Peptide Identification

1. Acquire the set of data files from one experiment in the proprietary .Raw format. These are first converted to .mzxml format using the freeware Readw.exe program.
2. Place the set of .mzxml data files from a given sample into one directory, and peptide-to-spectrum matching is performed using the XTandem algorithm. We use the Global Proteome Machine software (26, 27), which is freely available and runs the XTandem Tornado version. Searching the set of .mzxml files stored in a directory enables the user to choose for a single combined summary output file to be created, in addition to all 16 individual result files (see Note 4).
3. Export the combined protein and peptide identification output file for all 16 gel slices to an Excel spreadsheet. This spreadsheet contains six columns of data, with the headers identifier, log(I), rI, log(e), pI, Mr (kDa), description, and annotated domains. It is necessary to remove the last column (annotated domains) prior to subsequent analysis as it interferes with subsequent data processing. The Excel file is then exported to comma-separated value format, which is then compatible with input into the Scrappy software.

3.4 Spectral Counting Reporting and Analysis Program (Scrappy): Uploading and Analyzing Data

The Scrappy program is an implementation of the R statistical analysis package, run from a simple web interface. It has a limited amount of variable input allowed, but performs a large number of calculations quickly and efficiently. The following steps are required:

1. Upload the csv files of XTandem protein identification outputs as described above. For a simple pairwise comparison of two

biological samples, it is designed to accept three files for each sample, representing three biological replicate analyses.

2. Define category names (e.g., "control" and "stressed") and upload all six files. It is also possible to upload more than three replicates in a category, or three replicates of any number of samples to be compared. For the purposes of this chapter we will mostly constrain it to the simpler version, a pairwise comparison comprising three biological replicates of each.
3. When finished entering the files, select parameters on the following screen, including the following: the minimum number of peptide identifications for a protein within one sample set to be considered a valid protein identification (default value is 5), whether to use untransformed or log-transformed data for the *t*-test analyses (default is log-transformed), the spectral fraction to be added to all counts for multigroup statistical analysis (default is 0.5), and which of the data categories are to be treated as the baseline for numerical comparisons (see Note 5).
4. Start the analysis. The calculations are performed over several minutes, depending on the size and number of the data files. A results folder is generated, with a series of files containing different data analysis results. At the end of the list the user has the option to download the results, with or without the initial data files.
5. The folder of results will contain a number of different analysis outputs, which are described below.

3.5 Spectral Counting Reporting and Analysis Program (Scrappy): Interpreting Results

The results output files can be grouped into five subheadings: data aggregation, data partitioning, data quality metrics, NSAF ratios, and ANOVA and clustering.

1. Data aggregation

   (a) *Output.csv*

   This file contains the combined data set, namely, the full set of reproducibly present proteins (proteins present in all replicates of at least one sample, having a total peptide count > minimum peptide level as set above), their description, spectral counts, logNSAF values, *t*-test statistic, and *p*-value.

   (b) *Up-regulated.csv*

   This file is the subset of the full data set containing only the up-regulated proteins: *p*-value <0.05 and ratio >1. The ratio is the mean of the two average NSAF values for the two groups, with the denominator being the first group in alphabetical order (so if the groups are stress and control, the ratio will be mean NSAF stress/mean NSAF control).

(c) *Down-regulated.csv*

As for up-regulated, but containing only the proteins with p-value <0.05 and ratio <1.

2. Data partitioning

(a) *DataCategoriesBarChart.png*

The data in the complete data set is partitioned based on reproducible presence and absence in the various experimental categories. For a two-group experiment (e.g., stress–control) this will simply show three bars: proteins present reproducibly (namely, present in all replicates) in stress only, proteins present reproducibly in control only, and proteins present in all samples. If the experiment has an arbitrary number of groups, n, there can be up to $2n-1$ separate bars. This image can be used to give a quick idea of which combinations of conditions are most prevalent in an experiment.

(b) *DataCategoriesTable*

The data categories table lists the precise numbers for the categories listed in the bar chart.

(c) *DataCategoriesPieChart.png*

This pie chart shows the numbers of proteins present in one group only, two groups, three groups, etc. For a two-group experiment it would only contain two "slices," proteins present in one group only and proteins present in both groups.

3. Data quality metrics

(a) *QQplot.png*

An image is shown in Fig. 1, displaying the quantile–quantile plot of the average logNSAF data for the control samples from the control and 48 h cold stress experiment referred to earlier. Such a plot shows the quantiles of the control category on the y-axis against the quantiles of a standard normal distribution (hence zero mean and unit standard deviation) on the x-axis. A separate plot is generated for each category. If the data distribution of the sample is relatively normal, then the points will lie approximately on the diagonal. A small departure from the diagonal is acceptable, but a large deviation may show that the data is not acceptable for further statistical analysis.

(b) *ED1densityPlot.png*

This plot is a kernel density plot of the logNSAF data distribution from each replicate overlaid; in essence it is like a set of smoothed histograms, one from each replicate, placed on top of each other. It shows visually whether the data is approximately normal, or indicates if there is any replicate that has a slightly unusual distribution when compared to

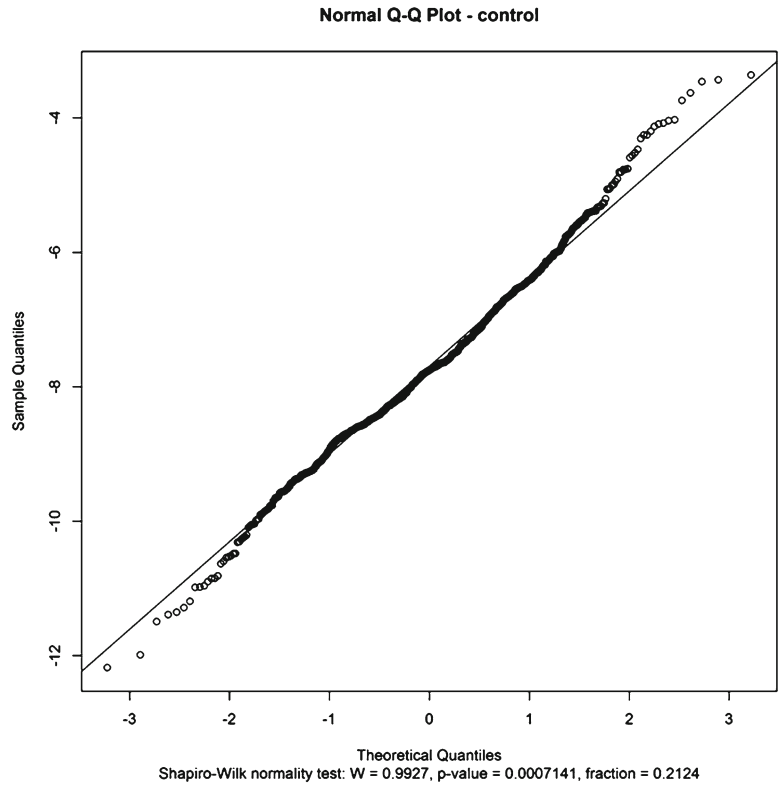

Fig. 1 Q-Q plot

the rest. If any replicate is visually very different from the rest, then the data quality should be examined.

(c) *ED2densityPlot*

This plot is similar to the previous density plot, but performed only for the proteins present reproducibly in all samples. An example is shown in Fig. 2, which includes three replicates each of the control and 48 h cold rice leaf samples referred to earlier. It is clear from this figure that all six samples overlay each other well with no major outliers.

4. Visualizing NSAF ratios

(a) *LogNSAF.png /LogNSAF.svg*

This graphic shows the logNSAF values for identified proteins, presented as logNSAF in the first specified category on the *x*-axis and logNSAF values in the second category on the *y*-axis. The dots are color coded, with light blue circles indicating that the logNSAF values are statistically unchanged between the two categories, while dark blue circles indicate those proteins with statistically significantly different logNSAF values between the two categories. Statistical significance is estimated using a student *t*-test on the log-transformed NSAF values from the original

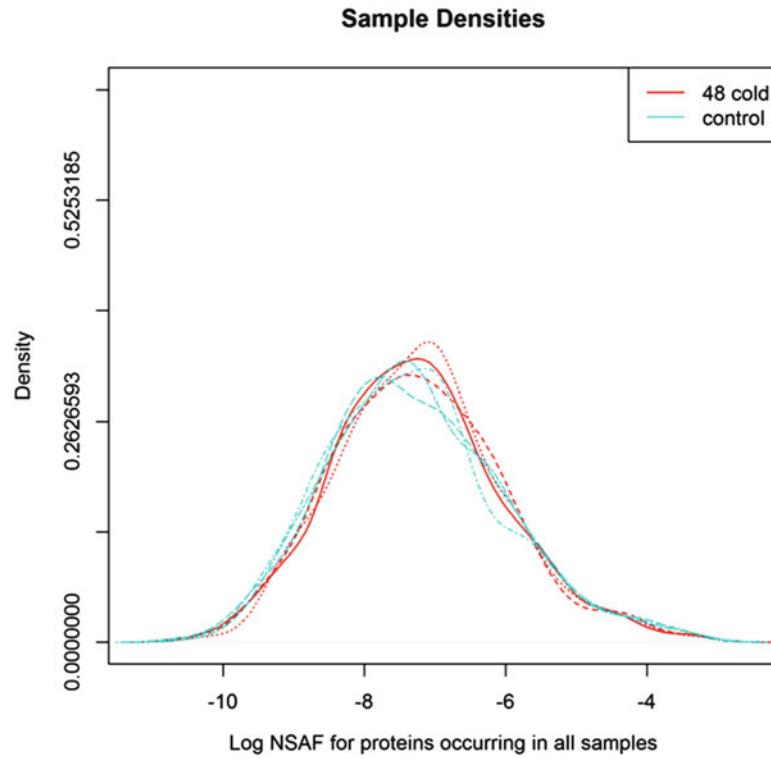

Fig. 2 ED2 density plot

biological triplicate experiments. In most experiments the majority of data points are clustered along the diagonal, with the lower values closer to the origin. The further from the diagonal, the greater the degree of differential expression.

Two different formats of this graph are generated, a png file and an interactive svg file (scalable vector graphics); hovering over the points in the interactive file will show additional information such as the protein identification (see Note 6). An example is shown in Fig. 3, which includes three replicates, each of the control and 48 h cold rice leaf samples referred to earlier.

(b) *RatioChart.png*

This plot shows the ratios of average NSAF in the two groups, ordered in increasing order of the *t*-test statistic. The size of the bars represents the average NSAF ratio for the up-regulated proteins, and the reciprocal (1/ratio) for the down-regulated proteins; therefore high bars show a big difference between groups. The up-regulated proteins are colored green, the down-regulated proteins are colored red, and the proteins showing no statistically significant difference are black. As above, there are two different formats of this image: one plain and one interactive.

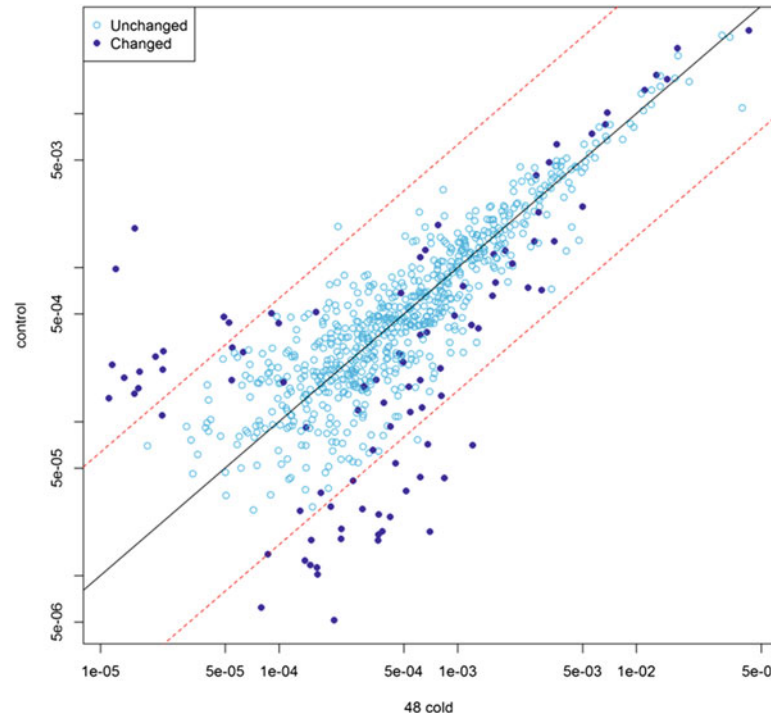

Fig. 3 LogNSAF ratio distribution

5. ANOVA and clustering

   The final category of analysis pertains to multiple group comparisons; hence they are not of immediate interest for a binary comparison. An analysis of variance is run on the subset of proteins present in all samples, separately for each protein. This approach is only of interest for more than two groups; in the two group case this is equivalent to the *t*-test already performed.

   (a) *expressionPatterns.png*

   This file contains the box plots of logNSAF values for the first 20 proteins identified as significantly changing by the analysis of variance (p-value <0.05), showing the pattern of change for those respective proteins. The proteins are ordered in increasing order of the p-value then plotted side by side. In the case of two groups this plot simply shows an up or down pattern. An example is shown in Fig. 4, which includes data from four time points of the rice leaf cold stress study referred to earlier: control, 48, 72, and 96 h cold stress.

   (b) *First20Genes.png*

   This file is very similar to the expression patterns, showing the pattern of change for the first 20 genes in

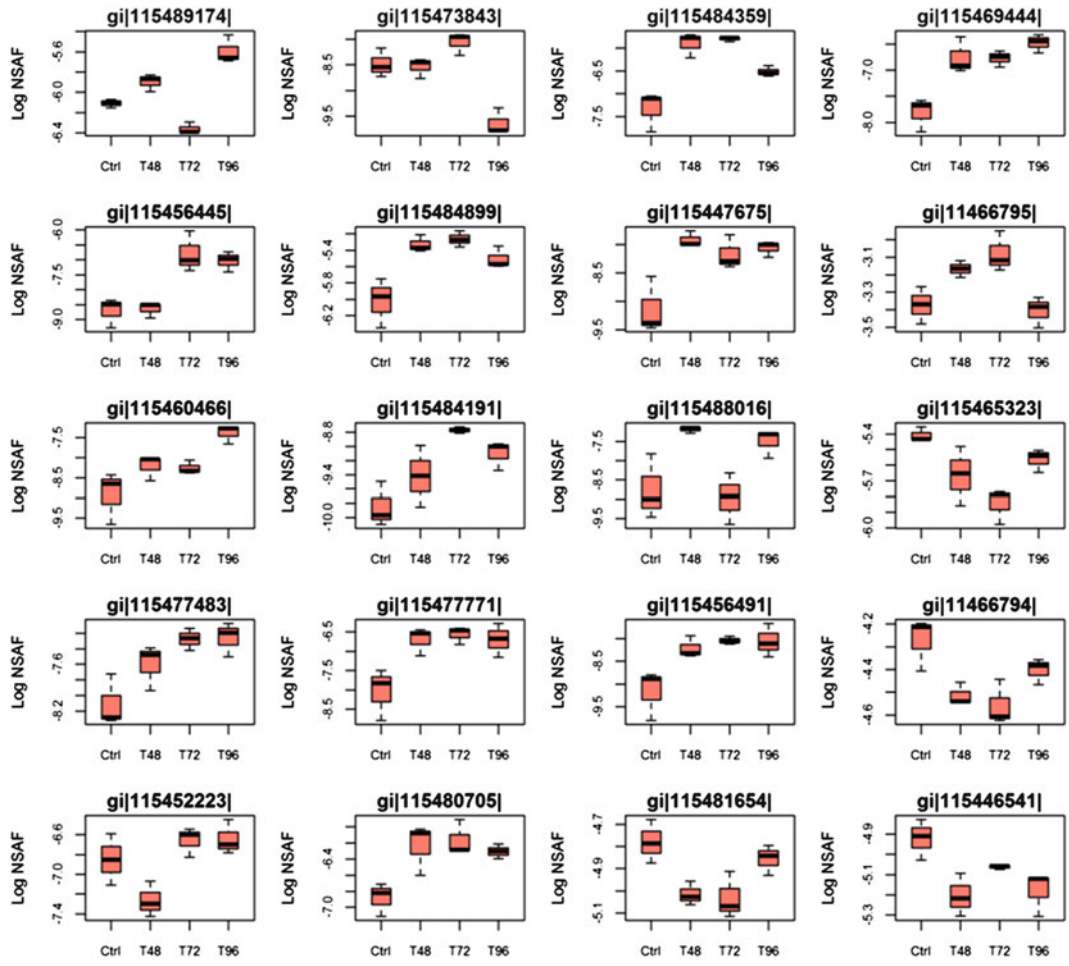

Fig. 4 Expression patterns

order of the ANOVA *p*-value, but in this instance they are plotted on the same scale, so differences in abundance between various proteins are apparent. An example is shown in Fig. 5, which includes data from four time points of the rice leaf cold stress study referred to earlier: control, 48, 72, and 96 h cold stress.

The proteins found to be changing by the ANOVA analysis are clustered on a heatmap and also by hierarchical clustering (complete linkage and Euclidean metric) and using the self-organizing maps algorithm; the resulting clusters are visualized in the images described below.

(c) *CLUST3heatmapCorrDist.png*

A heatmap is generated for all the proteins identified as significantly changing, using a correlation-based distance. A heatmap is a false color image of the logNSAF data;

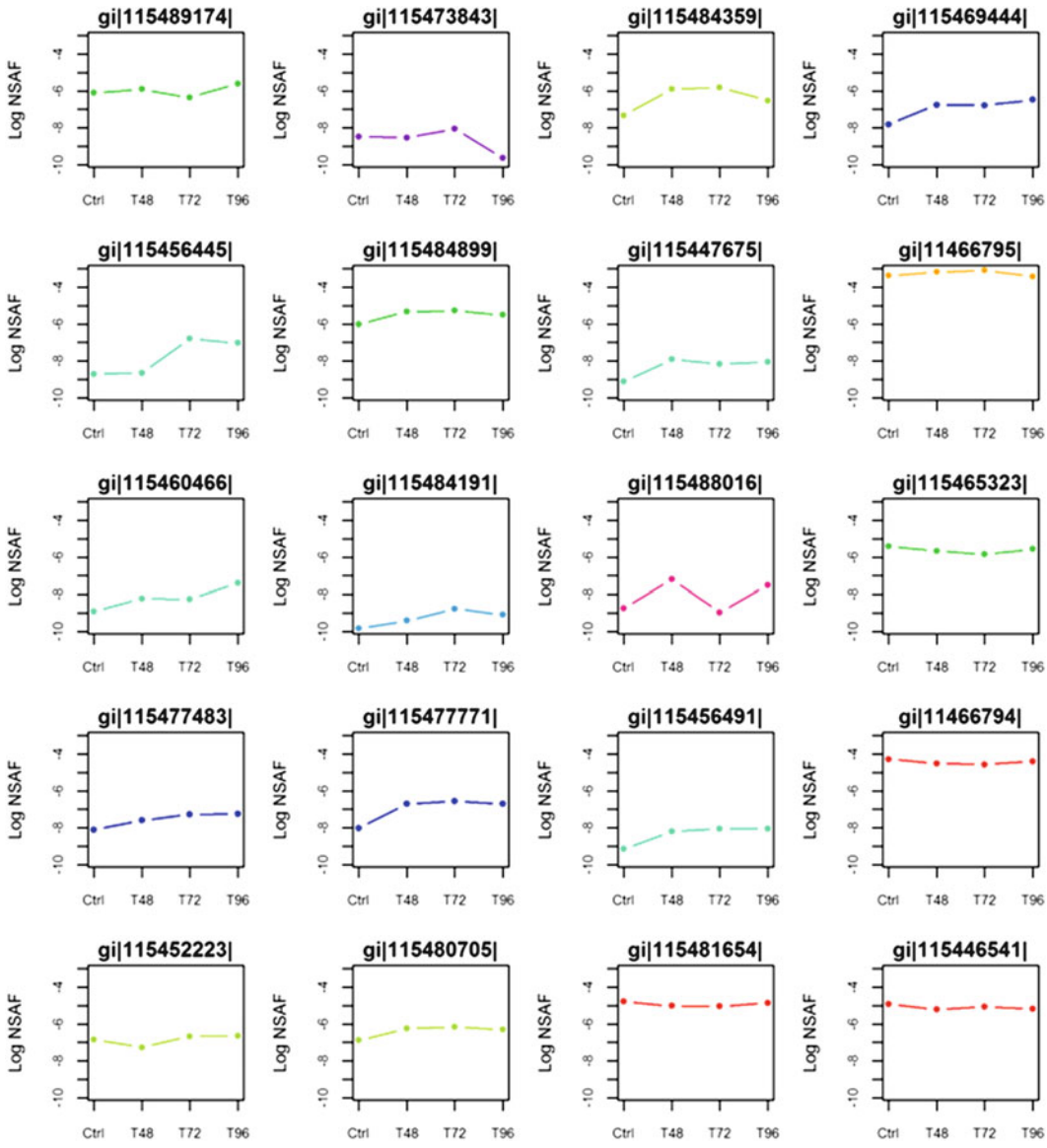

Fig. 5 First 20 genes

higher values appear in darker green. One column corresponds to one sample and one row corresponds to one protein. Rows and columns will be rearranged so that samples and respective proteins with more similar patterns are closer together. Replicate samples should be close together. An example is shown in Fig. 6, which includes data from four time points of the rice leaf cold stress study referred to earlier: control, 48, 72, and 96 h cold stress.

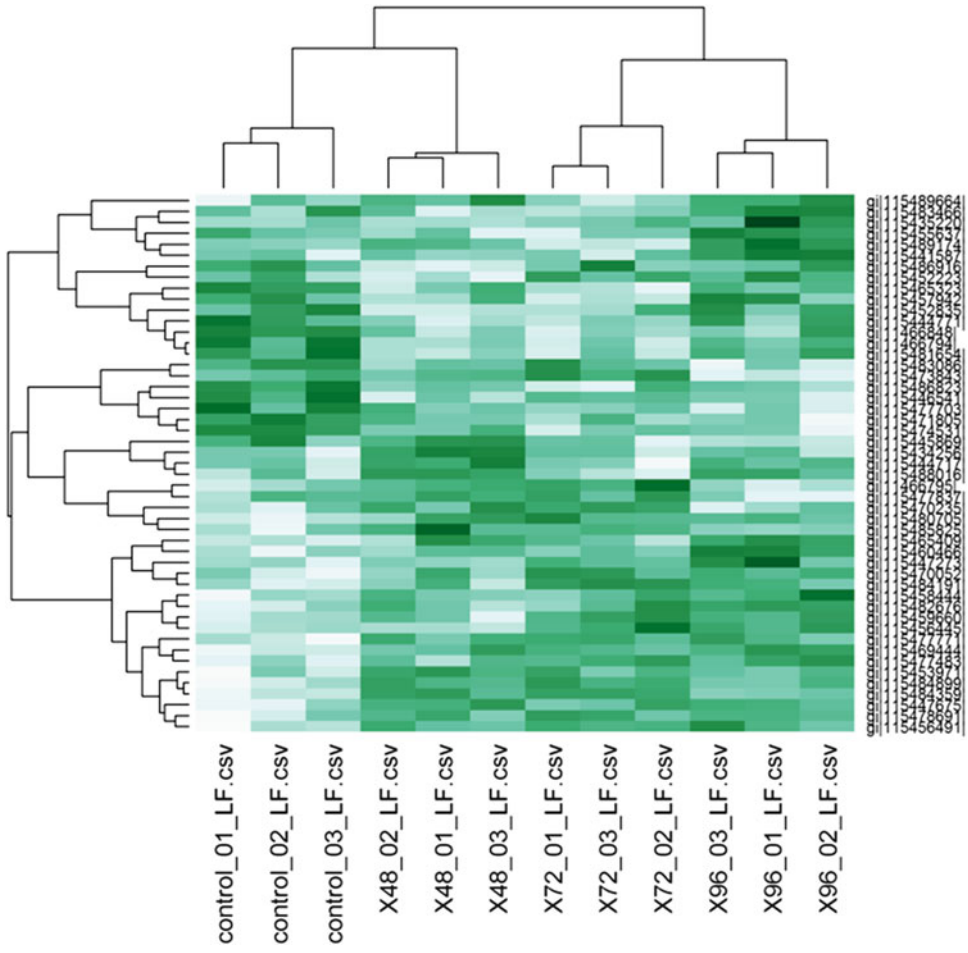

Fig. 6 Clust3 heatmap using a correlation based distance

(d) *CLUST1Genes.png*

This contains the cluster tree (dendrogram) which is the result of hierarchical clustering. There is no automatic choice of the right number of clusters—by cutting the tree at a particular height, one could end up with a certain number of clusters. Though there are ways of assessing the appropriate number of clusters, this is not done automatically for this analysis; instead an arbitrary choice of nine clusters is made, and they are colored with different colors.

(e) *ClusterExpressionProfiles.png*

The average logNSAF values for each experimental group are then plotted separately for each cluster identified in the dendrogram above. The individual protein averages are overlayed in light gray, to get an overall image of the patterns of change in protein abundance for each cluster.

(f) *SOM.png*

The self-organizing map algorithm (as implemented in the R *som* package) is also run for the same data, using an arbitrary *x* and *y* map dimension of 3. This algorithm will result into a partition of the NSAF data on a "map" of $9 = 3 \times 3$ cells, and the means and standard deviations for the average protein expressions in each of samples are visualized on this image. Thus each cell in the graph will show an average trend for the NSAF data in that cell.

(g) *PLGEM*

The power law global error model was shown to give a better estimate of NSAF data variability than that generated from the data alone, and can be used as an alternate to *t*-tests for determining up and down regulation between two conditions. Computationally this approach is made available in the *plgem* R package. In Scrappy it is set up to compare each of the conditions to a baseline, so in the case of two groups it is simply a comparison of, e.g., "stressed" versus "control." For more than two groups it will compare each of them in turn to the selected baseline group, e.g., "control." Unlike the *t*-tests, it is run only on the proteins present reproducibly in both samples (28). The *PLGEMFit.png* image describes the quality of the model fit, and the baselineComparisons.html lists the proteins found to be differentially expressed by this approach.

3.6 Finding Biological Meaning in the Results: Functional Annotation and Enrichment

The main output of Scrappy is the consolidated set of protein quantitation data, along with various subsets of proteins that are highlighted as potentially interesting; for example, proteins up-regulated in a drought-stressed condition, or the various subsets of proteins identified in a complex experiment with many groups. Although this is a richly detailed quantitative data set, extracting biologically interesting information can still be a daunting task.

The first option for proceeding further is via functional enrichment, mapping the identified proteins into categories such as cellular location, biological process, molecular function, or presence in a particular metabolic pathway, and then comparing the sets of interest to see which ones have more of a particular category than would be expected by chance. The tools used here depend very much on the organism studied and the number of groups considered in the experiment. For a binary comparison, such as control and stressed, there are many tools freely available online, with DAVID functional annotation platform (http://david.abcc.ncifcrf.gov/) and Blast2GO (http://www.blast2go.com/b2ghome) amongst the most popular. The researcher can

upload lists of up-regulated proteins and find, for example, biological processes or metabolic pathways overrepresented amongst the proteins up-regulated in the stressed condition as compared to the set of proteins at large. Commercial packages such as Ingenuity (http://www.ingenuity.com/) or GeneGO MetaCore (http://www.genego.com/metacore.php) can also provide a rapid and detailed analysis including functions, pathways, and networks, provided the license is purchased and the organism studied is supported in the software. The enrichment analysis may play a dual role: on the one hand it can help focus on a particular list of proteins involved in a biological category of interest, for example focus on signalling proteins up-regulated in the stress condition. On the other, it may help validate the set of proteins identified, provided the biological categories or pathways identified are meaningful in the context of the biology of the experiment.

For an experiment with multiple conditions, such as the five-way comparison of evolutionary adaptation of Pachycladon species mentioned earlier (14), the sheer number of numerical comparisons involved gets rapidly unwieldy: with five separate conditions, there could be ten separate binary comparisons to consider, each yielding up- and down-regulated and unchanged sets of proteins. For such experiments generating many subsets of proteins of interest we use the R-based software package PloGO, developed in our group (29), which allows us to categorize gene ontology information for various batches of proteins at once, compare the relative abundances of various gene ontology categories to a selected reference, and aggregate NSAF quantitation by GO category as well. Again, such an analysis may play multiple roles: firstly, focussing attention on the comparisons of conditions that produce functionally interesting sets of proteins, and secondly giving a list of biological processes or molecular functions that are overrepresented in particular categories of proteins, which can then be cross-correlated with the underlying biological paradigm of the experiment.

4 Notes

1. NanoLC columns can also be purchased from suppliers such as Michrom or New Objective.
2. It is not required to completely eliminate carryover between injections as all the samples in a given set will be combined for subsequent analysis.
3. Our standard experimental design is 16 SDS-PAGE gel slices from each of three biological replicates of a given sample. These samples can be completed in two and half days of mass spectrometric analysis time.

4. We use the "MudPIT combine" option in GPM searches to produce a unified output file from the individual search result files.
5. The default parameters in Scrappy are that a protein must be present in all three replicates with a minimum peptide count of 5. This parameter can be altered if necessary, but seems to work well for most of our experiments.
6. This feature is a very powerful (and popular), as it allows the user to present a single graphic to represent many thousands of data points in a manner that is easy to understand and interpret.

Acknowledgments

The authors acknowledge the funding support from the Macquarie University MQRES scholarship scheme (K.A.N.) and the Australian Research Council (P.A.H.). Aspects of this research were conducted at the Australian Proteome Analysis Facility funded by the Australian Government National Collaborative Research Infrastructure Scheme (NCRIS). P.A.H. acknowledges Robert Black for continued support and encouragement.

References

1. Anderson L, Hunter CL (2006) Quantitative mass spectrometric multiple reaction monitoring assays for major plasma proteins. Mol Cell Proteomics 5:573–588
2. Gao J, Opiteck GJ, Friedrichs MS, Dongre AR, Hefta SA (2003) Changes in the protein expression of yeast as a function of carbon source. J Proteome Res 2:643–649
3. Mosley AL, Florens L, Wen Z, Washburn MP (2009) A label free quantitative proteomic analysis of the Saccharomyces cerevisiae nucleus. J Proteomics 72:110–120
4. Zybailov BL, Florens L, Washburn MP (2007) Quantitative shotgun proteomics using a protease with broad specificity and normalized spectral abundance factors. Mol Biosyst 3:354–360
5. Liu H, Sadygov RG, Yates JR 3rd (2004) A model for random sampling and estimation of relative protein abundance in shotgun proteomics. Anal Chem 76:4193–4201
6. Zhang B, VerBerkmoes NC, Langston MA et al (2006) Detecting differential and correlated protein expression in label-free shotgun proteomics. J Proteome Res 5:2909–2918
7. Zybailov B, Mosley AL, Sardiu ME et al (2006) Statistical analysis of membrane proteome expression changes in *Saccharomyces cerevisiae*. J Proteome Res 5:2339–2347
8. Pavelka N, Fournier ML, Swanson SK, Pelizzola M, Ricciardi-Castagnoli P et al (2008) Statistical similarities between transcriptomics and quantitative shotgun proteomics data. Mol Cell Proteomics 7:631–644
9. Zhang Y, Wen Z, Washburn MP, Florens L (2009) Effect of dynamic exclusion duration on spectral count based quantitative proteomics. Anal Chem 81:6317–6326
10. Zhao Y, Denner L, Haidacher SJ, LeJeune WS, Tilton RG (2008) Comprehensive analysis of the mouse renal cortex using two-dimensional HPLC-tandem mass spectrometry. Proteome Sci 6:15
11. Chick JM, Haynes PA, Bjellqvist B, Baker MS (2008) A combination of immobilised pH gradients improves membrane proteomics. J Proteome Res 7:4974–4981
12. Chick JM, Haynes PA, Molloy MP et al (2008) Characterization of the rat liver membrane proteome using peptide immobilized pH gradient isoelectric focusing. J Proteome Res 7:1036–1045
13. Gammulla CG, Pascovici D, Atwell BJ, Haynes PA (2010) Differential metabolic response of cultured rice (Oryza sativa) cells exposed to high- and low-temperature stress. Proteomics 10:3001–3019

14. Voelckel C, Mirzaei M, Reichelt M et al (2010) Transcript and protein profiling identify candidate gene sets of potential adaptive significance in New Zealand Pachycladon. BMC Evol Biol 10:151
15. Sardiu ME, Cai Y, Jin J et al (2008) Probabilistic assembly of human protein interaction networks from label-free quantitative proteomics. Proc Natl Acad Sci USA 105:1454–1459
16. Muralidharan S, Thompson E, Girch G, Raftos D, Haynes PA (2011) Quantitative proteomics of heavy metal stress responses in Sydney rock oysters. Proteomics 12:906–921
17. Heinecke NL, Pratt BS, Vaisar T, Becker L (2010) PepC: proteomics software for identifying differentially expressed proteins based on spectral counting. Bioinformatics 26:1574–1575
18. Keller A, Eng J, Zhang N, Li X, Aebersold R (2005) A uniform proteomics MS/MS analysis platform utilizing open XML file formats. Mol Syst Biol 1:1–17
19. Park SK, Venable JD, Xu T, Yates JR 3rd (2008) A quantitative analysis software tool for mass spectrometry-based proteomics. Nat Methods 5:319–322
20. MacCoss MJ, Wu CC, Liu H, Sadygov R, Yates JR 3rd (2003) A correlation algorithm for the automated quantitative analysis of shotgun proteomics data. Anal Chem 75: 6912–6921
21. Gammulla CG, Pascovici D, Atwell BJ, Haynes PA (2011) Differential proteomic response of rice (Oryza sativa) leaves exposed to high- and low-temperature stress. Proteomics 11: 2839–2850
22. Mirzaei M, Pascovici D, Keighley T et al (2011) Shotgun proteomic profiling of five species of New Zealand Pachycladon. Proteomics 11:166–171
23. Mirzaei M, Soltani N, Sarhadi E et al (2012) Shotgun proteomic analysis of long-distance drought signaling in rice roots. J Proteome Res 11:348–358
24. Neilson KA, Mariani M, Haynes PA (2011) Quantitative proteomic analysis of cold-responsive proteins in rice. Proteomics 11:1696–1706
25. Wisniewski JR, Zougman A, Nagaraj N, Mann M (2009) Universal sample preparation method for proteome analysis. Nat Methods 6:359–362
26. Craig R, Beavis RC (2003) A method for reducing the time required to match protein sequences with tandem mass spectra. Rapid Commun Mass Spectrom 17:2310–2316
27. Craig R, Beavis RC (2004) TANDEM: matching proteins with tandem mass spectra. Bioinformatics 20:1466–1467
28. Pavelka N, Pelizzola M, Vizzardelli C et al (2004) A power law global error model for the identification of differentially expressed genes in microarray data. BMC Bioinformatics 5:203
29. Pascovici D, Keighley T, Mirzaei M, Haynes PA, Cooke B (2012) PloGO: plotting gene ontology annotation and abundance in multi-condition proteomics experiments. Proteomics 12:406–410