Koziol et al. Proteome Science 2013, 11:5
http://www.proteomesci.com/content/11/1/5

PROTEOME SCIENCE

METHODOLOGY Open Access

On protein abundance distributions in complex mixtures

JA Koziol[1*], NM Griffin[2], F Long[2], Y Li[2], M Latterich[2] and JE Schnitzer[2]

Abstract

Mass spectrometry, an analytical technique that measures the mass-to-charge ratio of ionized atoms or molecules, dates back more than 100 years, and has both qualitative and quantitative uses for determining chemical and structural information. Quantitative proteomic mass spectrometry on biological samples focuses on identifying the proteins present in the samples, and establishing the relative abundances of those proteins. Such protein inventories create the opportunity to discover novel biomarkers and disease targets. We have previously introduced a normalized, label-free method for quantification of protein abundances under a shotgun proteomics platform (Griffin et al., 2010). The introduction of this method for quantifying and comparing protein levels leads naturally to the issue of modeling protein abundances in individual samples. We here report that protein abundance levels from two recent proteomics experiments conducted by the authors can be adequately represented by Sichel distributions. Mathematically, Sichel distributions are mixtures of Poisson distributions with a rather complex mixing distribution, and have been previously and successfully applied to linguistics and species abundance data. The Sichel model can provide a direct measure of the heterogeneity of protein abundances, and can reveal protein abundance differences that simpler models fail to show.

Introduction

Large-scale proteome analysis using mass spectrometry and subcellular fractionation techniques can provide inventories of proteins identified in organelles, cells and tissues (e.g., [1–3]). Such protein inventories create the opportunity to discover novel biomarkers and disease targets (e.g., [4–7]). But a more detailed description of cells, tissues and organisms in health and disease would benefit greatly from quantitative tools that can carefully and comprehensively quantify the individual building blocks, which comprise the living entity. The ability to quantify properly identified proteins in biological samples in a comprehensive fashion engenders an enhanced understanding of cellular behavior during development or in response to disease, and can lead to novel biomarker and target discoveries [4,8].

Much effort has gone into developing more accurate and cost effective technologies that can capture the dynamics of biomolecular diversity in more quantitative ways. While significant advances have been made to develop accurate genomic sequencing tools [9] and highly accurate gene expression analytical methods [10], reliable methods of quantifying protein expression and modification levels have been challenging [11].

This difficulty is in part due to the immense chemical complexity of proteins, which are made up from over twenty amino acid monomers with distinct chemical properties, as contrasted to biopolymers such as RNA that are constituted from four monomers with similar properties. Currently there are no feasible direct methods to establish protein sequences like that of nucleotide polymers; the only method to directly determine the identity and the quantity of proteins in a mixture in large scale is the mass spectrometer, which can determine peptide sequences based on fragmentation pattern analysis and expression levels via direct or indirect means of analysis.

Quantitative proteomic mass spectrometry is indispensable to providing valuable insights into protein content and activity in various cellular states. There are at present three principal methods of quantifying proteins via mass spectrometry: labeling approaches such as iTRAQ and SILAC, which aim to reduce experimental variance and allow relative comparison of peptides between samples [12,13]; absolute quantitative approaches such as MRM

* Correspondence: koziol@scripps.edu
[1]The Scripps Research Institute, 10550 N Torrey Pines Road, La Jolla, CA 92037, USA
Full list of author information is available at the end of the article

and SISCAPA [7,14], which are highly accurate but thus far at the expense of completeness; and, label free approaches that rely on counting spectra or peptide numbers as a proxy for expression level (reviewed in [15]), or on ion intensities [16], or that jointly consider peptide count, spectral count, and fragment-ion intensity [17]. The latter method is particularly well suited for comparing clinical specimens for biomarker identification where samples are collected over long time periods and may have to be compared across sites [6,18].

We have previously introduced a normalized, label-free method for quantification of protein abundances under a shotgun proteomics platform [17]. The introduction of this method for quantifying and comparing protein expression leads naturally to the issue of modeling protein abundances. In this note, we examine various models for patterns of relative protein abundance from typical 2 dimensional liquid chromatography mass spectrometry (2D-LC-MS/MS) experiments.

Characterization of the joint distribution of all protein abundances in a proteome is complicated by the fact that protein abundances typically differ over several orders of magnitude. As might be expected, this joint distribution can be rather complex, and we would not expect a Gaussian distribution would adequately characterize it [17,19]. Here, we make no Gaussian assumptions about any abundances. Rather, from a somewhat historical perspective, we have chosen distributions that have been proposed for modeling word counts and species abundances, as we are positing an analogous problem to these precedents. We formally compare different families of distributions for protein abundance, with goodness of fit criteria utilized to determine adequacy of the models for summarizing the underlying data. Our fitting criteria allow us to determine which models best capture the underlying data structure, and would be appropriate for characterizing protein abundance distributions.

The protein abundance distributions can be utilized to establish the success rate of the experiments as defined by Eriksson and Fenyo [19], or what we have referred to as coverage [20]. Our ultimate goal was to identify a distribution that would improve the quantitative accuracy of label-free stochastic mass spectrometry.

Methods

Sample preparation

Luminal vascular endothelial cell plasma membranes and their caveloae were directly isolated from rat lung as previously described [21,22]. Proteins were pre-fractionated on SDS-PAGE gels prior to 2 dimensional liquid chromatography mass spectrometry (2D-LC-MS/MS). Gel lanes were cut into slices, approximately 50 per lane, for in-gel proteolytic digestions. Digested peptides were extracted from each gel slice three times with 20% ACN and 10% formic acid solution. The peptides extracted from each gel slice were first pooled into 7 groups then lyophilized. Each sample, either plasma membrane (experiment 1) or caveolae (experiment 2) was separated into five different gel lanes, and each lane was subjected to a complete 2D-LC-MS/MS analyses resulting in five replicate MS analyses of each sample. Proteins were inferred from each replicate [with the implication, that some proteins were not observed in every replicate]. By convention, we dropped from consideration any proteins detected in one run only.

Mass spectrometry

2D-LC-MS/MS: Lyophilized peptides were resuspended with 15 μl of buffer A (0.1% formic acid, 5% Acetonitrile (ACN)), then loaded onto a two-dimensional microcapillary column (manually packed C_{18} reversed phase and strong cation exchange column). The loaded samples were directly introduced into the LTQ mass spectrometer equipped with ESI nanospray ion source by eluting the bound peptides with a 2D-LC-MS/MS scheme controlled by Agilent 1100 HPLC quaternary pump [3]. Briefly, 17 salt steps (ammonium acetate) were applied. Each salt step was followed by a 5 to 80% ACN gradient containing 0.1% formic acid to elute the peptides on the C_{18} column. The flow rate was maintained at 200 to 250 nl/min.

Data acquisition for the LTQ was carried out in data-dependent mode. Full MS scans were recorded on the eluting peptides over the 400–1400 m/z range with one MS scan followed by three MS/MS scans of the most abundant ions. The temperature of the ion transfer tube of both mass spectrometers was set at 180°C and the spray voltage was 2.0 kv. The normalized collision energy was set at 35%. A dynamic exclusion was applied for Repeat Count of 2, a Repeat Duration of 0.5 minute, and an Exclusion Duration of 10 min.

Database search for protein identification

The acquired MS/MS spectra were converted into mass lists using the Extract_msn program from Xcalibur and searched against a protein database containing rat sequences using the Sequest program in the Bioworks™ 3.1 for Linux (Thermo Fisher Scientific, Inc., Waltham, MA, USA). The searches were performed allowing for tryptic peptides only with peptide mass tolerance of 1.5 Da and a minimum of 21 fragmented ions in one MS/MS scan. Accepted peptide identification was based on a minimum Cn score of 0.1; minimum cross correlation score of 1.8(z=1), 2.5(z=2), 3.5(z=3). False positive identification rate was determined by the ratio of number of peptides found only in the reversed database to the total number of peptides found in both forward and reverse databases. The false positive identification rates were ≤ 1%. The positive protein identification results were extracted

from Sequest.out files, filtered and grouped with DTASelect software using above criteria. Proteins were identified based on 2 unique significantly identified peptides.

Statistical methods

We consider the following discrete probability distributions:

(1) The negative binomial (NB) distribution, with probability mass function

$$P_{nb}(k;\gamma,p) = \frac{\Gamma(\gamma+k)}{k!\Gamma(\gamma)}p^k(1-p)^\gamma, k = 0,1,\ldots,\gamma>0, 0<p<1.$$

(2) The discrete Weibull distribution, with probability mass function

$$Pw(k;\nu,p) = p^{k^\nu} - p^{(k+1)^\nu}, k = 0,1,\ldots,\nu>0, 0<p<1.$$

(3) The Zipf distribution, with probability mass function

$$P_z(k;p) = \frac{k^{-(1+\rho)}}{Zeta(1+\rho)}, k = 1,2,\ldots,\rho>0,$$

where Zeta(.) is Riemann's zeta function.

(4) The Zipf-Mandelbrot distribution, with probability mass function

$$P_{zm}(k;\rho,a) = \frac{(k+a)^{-(1+\rho)}}{Zeta(1+\rho,a)}, k = 1,2,\ldots,\rho>0, a>0.$$

where here Zeta(r,a) denotes the Hurwitz zeta function.

(5) The Sichel distribution, with probability mass function

$$P_s(k;\alpha,\theta,\gamma) = \frac{(1-\theta)^{\gamma/2}}{K_\gamma\left(\alpha\sqrt{1-\theta}\right)}\frac{(\alpha\theta/2)^k}{k!}K_{k+\gamma}(\alpha),$$
$$k = 0,1,\ldots,\alpha>0, 0<\theta<1, -\infty<\gamma<\infty$$

where $K_\gamma(z)$ denotes the modified Bessel function of the second kind of order γ and argument z.

(6) The Poisson inverse Gaussian (PIG) distribution. This is a special case of the Sichel distribution, obtained by setting γ = −1/2 in the probability mass function P_s. [Numerical evaluation of $K_\gamma(z)$ is enormously simplified if γ = −1/2 or differs from −1/2 by an integer, advantageous in an earlier era of less powerful computational capabilities].

Our choice of these distributions is based partly on historical considerations, as we now describe.

The Poisson distribution is a standard baseline model for discrete data, and is often used as a starting point for deriving more realistic models that meet the characteristics of an observed set of data. Mathematically, the Poisson is a one-parameter distribution, with the mean equal to the variance. If discrete data show overdispersion relative to the Poisson, generalizations might be introduced to accommodate this. Greenwood and Yule [23] suggested a model in which the mean in the Poisson distribution is itself random, following a gamma distribution. This leads to a two-parameter distribution, the negative binomial, for discrete data. In turn, the negative binomial has become a standard baseline model for discrete data overdispersed relative to the Poisson.

In a seminal article, Fisher and colleagues [24] introduced the notion of mathematically modeling species abundance data. Their motivation was to model butterfly abundance data from Malaya [25], and Fisher explored the truncated negative binomial distribution and extensions to this end. With species abundance data, as with our peptide setting, one must consider the zero-truncated forms of the underlying distributions, to accommodate the fact that certain species may not be observed in a finite sampling frame. This can lead to some added complexities relative to model fitting, as for example, described by Sampford [26] relative to the truncated negative binomial distribution. As with Greenwood and Yule, Fisher et al. [24] assumed that abundances could be modeled by a gamma distribution, which led to the negative binomial. A special case is Fisher's log-series model, where the shape parameter of the gamma distribution tends to zero. Engen [27] provides a comprehensive review of species abundance models in ecology.

The eponymous Zipf's law was introduced by Zipf [28] as a word frequency distribution: if one tabulates from an arbitrary text the number of words arranged in the order of their frequency of usage, the resulting word frequency distribution is generally reverse J-shaped, with a very long upper tail. Zipf's law is a mathematical power-law representation of this type of distribution. Zipf's frequency distribution was later generalized by Mandelbrot [29], again in a linguistics context.

The discrete Weibull [30] is another model for skewed, power-like discrete data. The incorporation of an additional parameter, as with Zipf-Mandelbrot, allows added flexibility, to accommodate situation in which the power-law relationship tends to decay in the tail. This is closely related to the stretched exponential distribution [31]. Newman [32] and Clauset et al. [33] give particularly lucid accounts of power-law distributions.

The Sichel distribution was introduced by Holla [34], and popularized in a series of papers by Sichel (e.g., [35–38]). Sichel and others have applied it both to linguistics and to species abundance data (e.g., [39]). The special case of an inverse Gaussian mixing distribution, leading to the Poisson

inverse Gaussian distribution, enjoys some computational advantages (e.g., [40]). The Sichel distribution is a mixed Poisson distribution, and can be generalized by using mixing distributions other than the inverse Gaussian (e.g., [41–44]).

From a theoretical perspective, the negative binomial and Sichel distributions are attractive models for protein abundance data. The frequencies of the different proteins in the sample can be taken as independent Poisson variables, where the Poisson parameters are heterogeneous; a mixing distribution should then be chosen to accommodate the overdispersion. In this regard, the Poisson inverse Gaussian distribution seems preferable to the negative binomial, but the Sichel distribution, with one additional free parameter relative to the Poisson inverse Gaussian distribution, is correspondingly even more flexible.

We used maximum likelihood techniques for fitting observed protein abundance data to all models: this typically provides more efficient and robust estimates than other methods, developed prior to the advent of inexpensive computing resources. Goldstein et al. [45] have cautioned against informal methods of parameter estimation with power-law based discrete distributions, and Clauset et al. [33] provide theoretical justification for maximum likelihood. We utilized Mathematica 8.0 (Wolfram Research, Inc., 2010) for numerical fitting using its default global optimization algorithm; in addition, the program also provides built-in numerical evaluation of the special functions incorporated in the probability mass functions above, which facilitates the optimization.

The method of maximum likelihood in our setting is straightforward. We describe the method generically, as follows. Let X denote a positive-integer valued random variable, with Prob(X=i)=P(i;θ) for some vector of parameters θ. We draw a finite random sample, and observe X=i with frequency f_i, for i=1,2,. . .,m. The method of maximum likelihood entails finding the vector $\hat{\theta}$ that maximizes the log of the likelihood function

$$LL = \sum_{i=1}^{m} f_i \log[P(i;\theta)].$$

[In practice it is generally more convenient to maximize the log of the likelihood function than the likelihood itself]. With our data, the X_i are the various protein abundances, and the P(i) are the probabilities determined from the models given above. Note, however, that the minimal observed protein abundance is 1, whereas the supports of the negative binomial, discrete Weibull, and Sichel distributions begin at 0. Hence for these distributions, we fit zero-truncated forms of the distributions: when maximizing the log likelihood for these distributions, the P(i) are replaced by P(i)/(1-P(0)) in the above formula for *LL*. The supports for the Zipf and Zipf-Mandelbrot distributions begin at 1, obviating the need to deal with truncated forms of these distributions.

Because the models are not always nested, we adopt the Akaike information criterion (AIC; [46]) as our general criterion for comparing models. [In the case of nested models, as with the Zipf nested within the Zipf-Mandelbrot, one might use a likelihood ratio test, to assess the relative improvement in fit with the more complex model relative to the simpler one.] The AIC value is defined as −2[log likelihood - # fitted parameters]. Given a set of potential models for the data, the minimum AIC value would be indicative of the preferred model. We remark that, there is one fitted parameter for the Zipf distribution, two fitted parameters for the negative binomial, discrete Weibull, Zipf-Mandelbrot, and Poisson inverse Gaussian distributions, and three fitted parameters for the Sichel distribution.

We display observed and fitted distributions with rank-frequency plots [47]. The rank-frequency plot of a frequency distribution is in log-log coordinates, with x denoting the ranks of the items in the distribution, and y the corresponding relative frequencies. [A Zipf distribution would be a straight line in a rank-frequency plot, and the plot can be utilized to estimate the parameter r characterizing the Zipf distribution]. Newman [32] describes these plots in greater detail, and astutely notes their equivalence to complementary cumulative distribution function plots, but with log-log and not linear coordinates. We utilize Newman's construction in the following. Specifically, we start with a listing of all the proteins, along with their frequency of occurrence (abundance), ranked in order of increasing abundance. The complementary cumulative distribution P(x) of the frequency x is defined as the fraction of proteins with abundance greater than or equal to x. Our plots depict both the observed and the fitted complementary cumulative distributions.

Table 1 Summary statistics for peptide counts

	Min	Max	Median	Mean	SD	Skewness	Kurtosis	Var/Mean
Expt 1	1	525	7	13.13	26.36	10.56	164.7	52.9
Expt 2	1	302	6	12.37	20.43	6.06	60.8	33.7

Caption. Experiments 1 and 2 refer to the membrane replicates and caveolae replicates respectively, as described in the methods. 2075 unique proteins were identified in experiment 1 and 1069 in experiment 2.

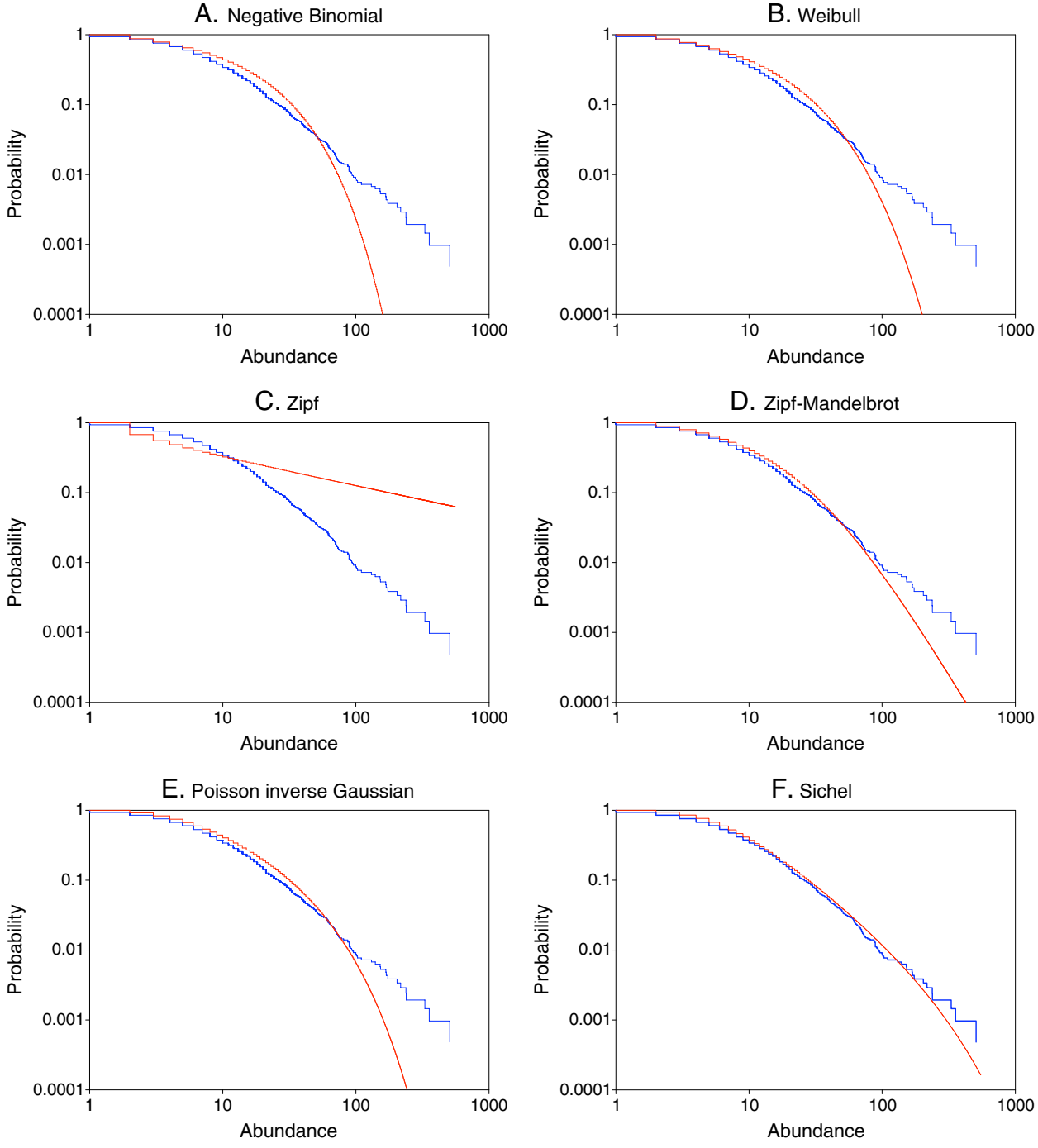

Figure 1 Rank-frequency plots of protein abundances from the first experiment, together with fitted distributions. A. Negative binomial. **B**. Discrete Weibull. **C**. Zipf. **D**. Zipf-Mandelbrot. **E**. Poisson inverse Gaussian. **F**. Sichel. Observed data are depicted in blue, and the fitted distributions are depicted in red. As described in the Methods, we start with a listing of all the proteins, along with their frequency of occurrence (abundance). The complementary cumulative distribution P(x) of the abundance x is defined as the fraction of proteins with abundance greater than or equal to x. Our plots depict both the observed and the fitted complementary cumulative distributions (ordinates) vs protein abundances (abscissas).

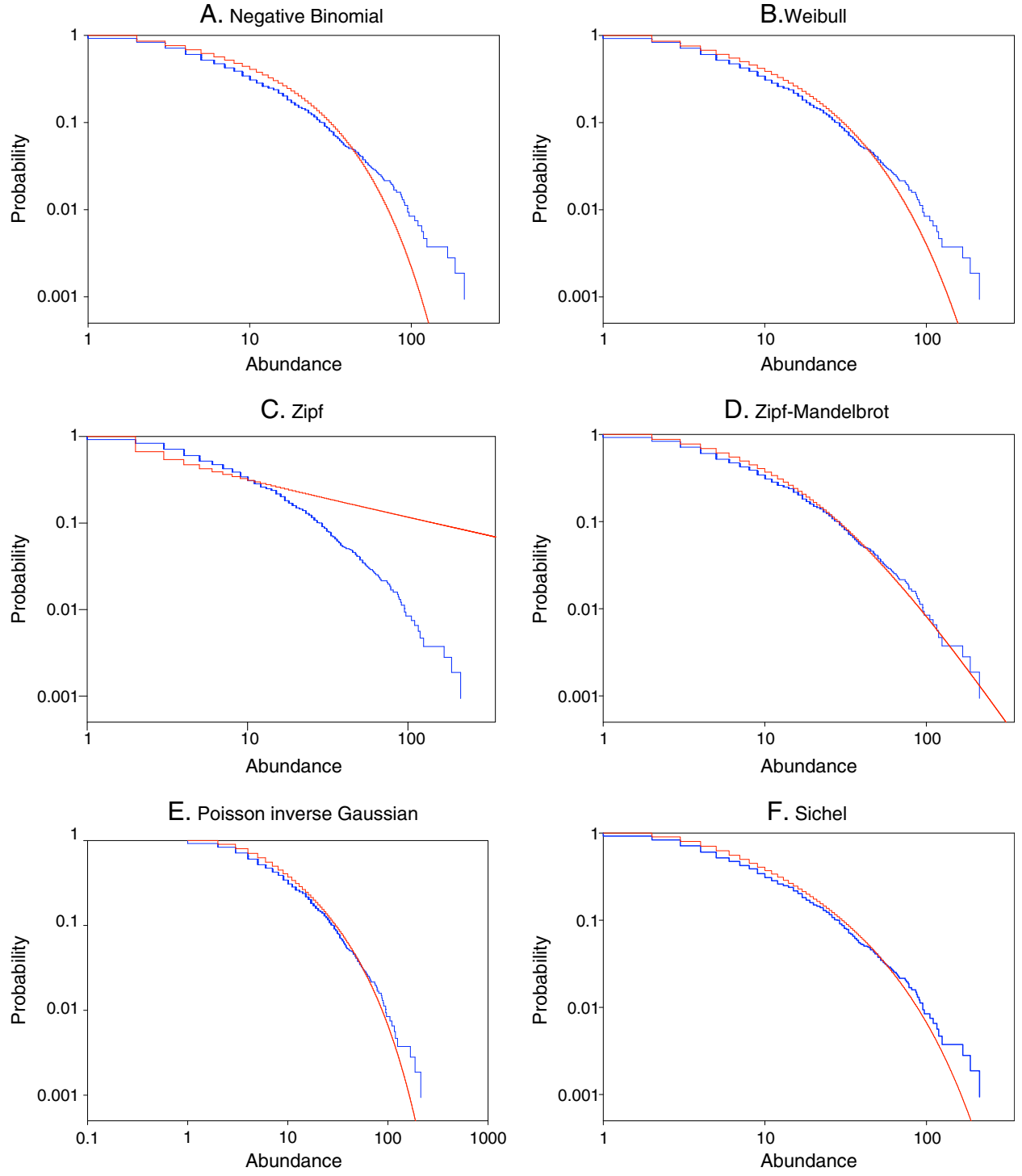

Figure 2 Rank-frequency plots of protein abundances from the second experiment, together with fitted distributions. A. Negative binomial. **B**. Discrete Weibull. **C**. Zipf. **D**. Zipf-Mandelbrot. **E**. Poisson inverse Gaussian. **F**. Sichel. Observed data are depicted in blue, and the fitted distributions are depicted in red. As described in the Methods, we start with a listing of all the proteins, along with their frequency of occurrence (abundance). The complementary cumulative distribution P(x) of the abundance x is defined as the fraction of proteins with frequency greater than or equal to x. Our plots depict both the observed and the fitted complementary cumulative distributions (ordinates) vs protein abundances (abscissas).

Table 2 Comparative statistics for six models

Model	AIC, Expt 1	AIC, Expt 2
negative binomial	14533.9	7326.9
discrete Weibull	14413.6	7280.8
Zipf	16146.9	8067.0
Zipf-Mandelbrot	14703.5	7482.7
Poisson inverse Gaussian	14238.4	7203.0
Sichel	14167.3	7189.8

Caption. Experiments 1 and 2 refer to the membrane replicates and caveolae replicates respectively, as described in the methods. AIC denotes Akaike's information criterion; smaller values connote better model fits.

Results

We are interested in quantitatively mapping the proteins expressed on the surface of vascular endothelial cells as they exist natively in tissue, and have developed subcellular tissue fractionation techniques to isolate luminal endothelial cell surface membranes directly from lung and other tissues. These endothelial plasma membranes (experiment 1) and their caveolae (experiment 2) were isolated from rat organs, and were subsequently analyzed by SDS-PAGE and mass spectrometry (see Methods). With the first experiment, a total of 27252 peptides were detected in 5 2D-LC-MS/MS cycles; these identified 2075 unique proteins, based on our model selection criteria outlined in the methods section. In the second experiment, a total of 13226 peptides were detected in 5 2D-LC-MS/MS cycles; these identified 1069 unique proteins. Summary statistics for the relative peptide counts are given in Table 1. If abundances were Poisson distributed, then the ratio of variance to mean would be about 1; the large variance/mean ratios are indicative of extra-Poisson variability. Within each experiment, the data are quite dispersed, and extremely right-skewed; heavy tails exist because of several extreme values of abundance counts.

In Figures 1 and 2 we display rank-frequency plots of the observed protein abundance distributions from the two experiments, along with the individual models fitted by maximum likelihood. The AIC values corresponding to the fits are given in Table 2. With both experiments, the ordering of the models would be

$$\begin{aligned}\text{Sichel} &< \text{PIG} < \text{discrete Weibull} < \text{NB} \\ &< \text{Zipf} - \text{Mandelbrot} < \text{Zipf},\end{aligned}$$

with the left to right ordering indicative of best to worst fitting. The added flexibility of the general Sichel distribution with arbitrary parameter γ provides an improvement in fit over the Poisson-inverse Gaussian distribution with $\gamma = -1/2$; in turn, both of these models are noticeably better than the other models, relative to AIC values. The rank-frequency plots for Set 1 show that only the Sichel model adequately represents the empirical frequency distribution in the right tail. With Set 2, the Poisson-inverse Gaussian model more closely resembles the Sichel model; lack of fit in the right tail is again noticeable for the other models. The Zipf distribution is particularly noteworthy for its lack of fit throughout the range of the empirical distribution of frequencies.

Discussion

It has become apparent that peptide and thus protein abundances, as measured by large scale high-throughput shotgun proteomics experiments, are not normally distributed [17,19]. This may be reflective of the complex nature of the proteome, especially when post-translational modifications are taken into account, or the inherent sampling limitations of the currently available MS technology as mentioned in the introduction. Nonetheless, we sought to characterize the protein abundance distributions in terms of their contributing peptides from two separate large-scale 2D-LC-MS/MS protein identification experiments. Our goal was to identify a distribution model that best fits or describes the protein abundance data, which can take into account the real world variation in protein abundances.

From the earliest reports of 2D-LC-MS/MS data [14,48,49], it has become clear that protein abundance differs over several orders of magnitude, with many proteins having a relatively small abundance, a few with relatively large abundances. This reflects the inherent dynamic range of any proteome, prior to identification by mass spectrometry. One must not forget that protein detection by traditional mass spectrometry methods is dependent on the inherent physical properties of the proteins and their resulting peptides. Peptide detection is highly dependent on the ease with which the peptide can be ionized. Ionization efficiency can be thought of as the tendency of the peptide to ionize and contribute to a mass spectrum thus facilitating the identification of the peptide and thus the protein. This is influenced mainly by the inherent structural properties of the peptide, such as length, mass, amino acid composition, and various biophysical properties, such as hydrophobicity, number of charges and potential modifications. Thus, one must be acutely aware that not every peptide in a given complex sample can and will be identified even though multiple methods have been developed in recent years to enhance peptide and protein coverage of a complex protein sample [3,50].

Let us next consider the issue of the external validity (generalizability) of our findings. To address this, we analyzed a smaller dataset reported by Ishihama et al. [51], Table 1. The relevant data consist of concentrations of 46 proteins that the authors had identified and quantified in mouse neuro2a cells [with a different quantitation method than that of Griffin et al.]. We proceeded to fit the 6 distributions described previously, and obtained the following ordering of the models:

Sichel < PIG < Zipf-Mandelbrot < discrete Weibull < NB < Zipf.

The respective AIC values were: 586.97, 592.45, 599.53, 603.54, 604.59, and 705.35. The pre-eminence of the Sichel distribution remains, as does the poor performance of the Zipf distribution. With this smaller dataset, Zipf-Mandelbrot outperforms the discrete Weibull and the negative binomial, although differences are at best modest. Nevertheless, we have insufficient evidence that a Sichel distribution would obtain with other quantification methods (e.g., spectral counting methods emPAI or RIBAR / xRIBAR); a cautious interpretation is, that we observed a Sichel distribution with the quantification method of Griffin et al. [17], but that the observed distribution may also depend on the mass spectrometer technology used.

From the analyses described in this study, one might infer that simple models of protein distribution do not adequately fit the experimental data, with empirical evidence pointing toward a more complicated mixing distribution. Indeed, the more complex Poisson inverse Gaussian or Sichel distributions work well to accommodate the heavy tail that is typically observed in proteomics experiments. These models accommodate the fact that protein abundances as reflected in the number of peptides detected per protein within a given sample and between identical samples can be different. This is not surprising giving the complex nature of the sample and the contribution of ion suppression effects which can mean that a peptide detected in one sample may not be detected in a subsequent MS analysis of the same sample. In fact, we previously found that each MS measurement of a shotgun proteomics analysis identifies only a subset of proteins and that second and third MS measurements of the same sample would reveal about 33% and 16% respectively of new proteins not detected in the previous analyses [1,20]. This means that multiple MS measurements should be performed to comprehensively define the full proteome to the degree possible with the technique used, hence why 5 replicate analysis of each sample were performed in the protein identification experiments analyzed in this paper. Furthermore, due to the intrinsic properties of some proteins, especially their large hydrophobicity peptides, or lack of accessible tryptic cleavage sites, some peptides may never be detected by the mass spectrometer. This suggests that, rather than total proteomic identification, the goal of these experiments should be adequate coverage of the entire proteome [20]. Thus, the ability to model protein abundance distributions from 2D-LC-MS/MS experiments or even fit the distributions to a specific model implies that one could theoretically exploit the properties of the model to improve protein coverage through optimizing experimental design [20].

Competing interests
The authors declare no competing financial interests.

Authors' contributions
The study was conceived by JAK, NMG, ML, and JES. NMG, FL, YL performed the sample preparation and mass spectrometry experiments. JAK and NMG undertook subsequent analyses. JAK, NMG, ML, and JES prepared the manuscript. All authors contributed to manuscript editing. All authors read and approved the final manuscript.

Author details
[1]The Scripps Research Institute, 10550 N Torrey Pines Road, La Jolla, CA 92037, USA. [2]Proteogenomics Research Institute for Systems Medicine, 11107 Roselle Street, San Diego, CA 92121, USA.

Received: 4 December 2011 Accepted: 15 May 2012
Published: 29 January 2013

References

1. Durr E, Yu J, Krasinska KM, *et al*: **Direct proteomic mapping of the lung microvascular endothelial cell surface *in vivo* and in cell culture.** *Nat Biotechnol* 2004, **22:**985–992.
2. Kislinger T, Gramolini AO, MacLennan DH, Emili A: **Multidimensional protein identification technology (MudPIT): technical overview of a profiling method optimized for the comprehensive proteomic investigation of normal and diseased heart tissue.** *J Am Soc Mass Spectrom* 2005, **16:**1207–20.
3. Li Y, Yu J, Wang Y, *et al:* **Enhancing identifications of lipid-embedded proteins in mass spectrometry for improved mapping of endothelial plasma membranes *in vivo*.** *Mol Cell Proteomics* 2009, **8:**1219–1235.
4. Addona TA, Shi X, Keshishian H, Mani DR, Burgess M, Gillette MA, Clauser KR, Shen D, Lewis GD, Farrell LA, Fifer MA, Sabatine MS, Gerszten RE, Carr SA: **A pipeline that integrates the discovery and verification of plasma protein biomarkers reveals candidate markers for cardiovascular disease.** *Nat Biotechnol* 2011, **29:**635–643.
5. Andersen JN, Sathyanarayanan S, Di Bacco A, Chi A, Zhang T, Chen AH, Dolinski B, Kraus M, Roberts B, Arthur W, Klinghoffer RA, Gargano D, Li L, Feldman I, Lynch B, Rush J, Hendrickson RC, Blume-Jensen P, Paweletz CP: **Pathway-based identification of biomarkers for targeted therapeutics: personalized oncology with PI3K pathway inhibitors.** *Sci Transl Med* 2010, **2:**45ra55.
6. Paweletz CP, Wiener MC, Bondarenko AY, Yates NA, Song Q, Liaw A, Lee AY, Hunt BT, Henle ES, Meng F, Sleph HF, Holahan M, Sankaranarayanan S, Simon AJ, Settlage RE, Sachs JR, Shearman M, Sachs AB, Cook JJ, Hendrickson RC: **Application of an end-to-end biomarker discovery platform to identify target engagement markers in cerebrospinal fluid by high resolution differential mass spectrometry.** *J Proteome Res* 2010, **9:**1392–1401.
7. Whiteaker JR, Zhao L, Anderson L, Paulovic AG: **An automated and multiplexed method for high throughput peptide immunoaffinity enrichment and multiple reaction monitoring mass spectrometry-based quantification of protein biomarkers.** *Mol Cell Proteomics* 2010, **9:**184–196.
8. Whiteaker JR, Lin C, Kennedy J, Hou L, Trute M, Sokal I, Yan P, Schoenherr RM, Zhao L, Voytovich UJ, Kelly-Spratt KS, Krasnoselsky A, Gafken PR, Hogan JM, Jones LA, Wang P, Amon L, Chodosh LA, Nelson PS, McIntosh MW, Kemp CJ, Paulovich AG: **A targeted proteomics-based pipeline for verification of biomarkers in plasma.** *Nat Biotechnol* 2011, **29:**625–634.
9. Lander ES: **Initial impact of the sequencing of the human genome.** *Nature* 2011, **470:**187–197.
10. Rosenberg S, Elashoff MR, Beineke P, Daniels SE, Wingrove JA, Tingley WG, Sager PT, Sehnert AJ, Yau M, Kraus WE, Newby LK, Schwartz RS, Voros S, Ellis SG, Tahirkheli N, Waksman R, McPherson J, Lansky A, Winn ME, Schork NJ, Topol EJ: **Multicenter validation of the diagnostic accuracy of a blood-based gene expression test for assessing obstructive coronary artery disease in nondiabetic patients.** *Ann Intern Med* 2010, **153:**425–434.
11. Wang K, Lee I, Carlson G, Hood L, Galas D: **Systems biology and the discovery of diagnostic biomarkers.** *Dis Markers* 2010, **28:**199–207.
12. de Godoy LM, Olsen JV, Cox J, Nielsen ML, Hubner NC, Frohlich F, Walther TC, Mann M: **Comprehensive mass-spectrometry-based proteome quantification of haploid versus diploid yeast.** *Nature* 2008, **455:**1251–1254.
13. Kuzyk MA, Ohlund LB, Elliott MH, Smith D, Qian H, Delaney A, Hunter CL, Borchers CH: **A comparison of MS/MS-based, stable-isotope-labeled,**

quantitation performance on ESI-quadrupole TOF and MALDI-TOF/TOF mass spectrometers. *Proteomics* 2009, **9:**3328–3340.
14. Anderson L, Hunter CL: **Quantitative mass spectrometric multiple reaction monitoring assays for major plasma proteins.** *Mol Cell Proteomics* 2006, **5:**573–588.
15. Yates JR 3rd, Gilchrist A, Howell KE, Bergeron JJ: **Proteomics of organelles and large cellular structures.** *Nat Rev Mol Cell Biol* 2005, **6:**702–714.
16. Wu Z, Fellenberg K, Lerner S, Kuster B: *Comparison of label-free protein quantification approaches for chemical proteomics*. Utah, USA: 58th ASMS Conference on Mass Spectrometry and Allied Topics; 2010.
17. Griffin NM, Yu J, Long F, *et al:* **Label-free, normalized quantification of complex mass spectrometry data for proteomic analysis.** *Nat Biotechnol* 2010, **28:**83–89.
18. Latterich M, Schnitzer JE: **Streamlining biomarker discovery.** *Nat Biotechnol* 2011, **29:**600–602.
19. Eriksson J, Fenyo D: **Improving the success rate of proteome analysis by modeling protein-abundance distributions and experimental designs.** *Nat Biotechnol* 2007, **25:**651–655.
20. Koziol JA, Feng AC, Schnitzer JE: **Application of capture-recapture models to estimation of protein count in MudPIT experiments.** *Anal Chem* 2006, **78:**3203–3207.
21. Schnitzer JE, McIntosh DP, Dvorak AM, *et al:* **Separation of caveolae from associated microdomains of GPI-anchored proteins.** *Science* 1995, **269:**1435–1439.
22. Oh P, Schnitzer JE: **Isolation and subfractionation of plasma membranes to purify calvaeolae separately from glycosylphosphatidylinositol-anchored protein microdomain**. In *Cell Biology: A Laboratory Handbook*. 2nd edition. Edited by Celis J. Orlando: Academic Press; 1998:34–36.
23. Greenwood M, Yule GU: **An inquiry into the nature of frequency distributions representative of multiple happenings with particular reference to the occurrence of multiple attacks of disease or of repeated accidents.** *J Roy Statist Soc* 1920, **83:**255–279.
24. Fisher RA, Corbet AS, Williams CB: **The relation between the number of species and the number of individuals in a random sample from an animal population.** *J Animal Ecology* 1943, **12:**42–58.
25. Corbet AS: **The distribution of butterflies in the Malay peninsula.** *Proc Roy Ent Soc Lond A* 1942, **16:**101–116.
26. Sampford MR: **The truncated negative binomial distribution.** *Biometrika* 1955, **42:**58–69.
27. Engen S: *Stochastic Abundance Models*. New York: John Wiley; 1978.
28. Zipf GK: *Selected Studies of the Principle of Relative Frequency in Language*. Cambridge, MA: Harvard University Press; 1932.
29. Mandelbrot B: **Information theory and psycholinguistics**. In *Language*. Edited by Oldfield RC, Marchall JC. London: Penguin Books; 1968.
30. Englehardt JD, Li R: **The discrete Weibull distribution: an alternative for correlated counts with confirmation for microbial counts in water.** *Risk Anal* 2011, **31:**370–381.
31. Guo L, Tan E, Chen S, Xiao Z, Zhang X: *The stretched exponential distribution of internet media access patterns. PODC '08, August 18–21*. Toronto, Ontario, Canada:; 2008.
32. Newman MEJ: **Power laws, Pareto distributions and Zipf's law.** *Contemp Phys* 2005, **46:**323–351.
33. Clauset A, Shalizi CR, Newman MJA: **Power-law distributions in empirical data.** *SIAM Review* 2009, **51:**661–703.
34. Holla M: **On a Poisson-inverse Gaussian distribution.** *Metrika* 1966, **11:**115–121.
35. Sichel HS: **On a family of discrete distributions particularly suited to represent long-tailed frequency data**. In *Proceedings of the Third Symposium on Mathematical Statistics*. Edited by Laubscher NF. South Africa: Council for Scientific and Industrial Research, Pretoria; 1971:51–97.
36. Sichel HS: **On a distribution representing sentence-length in written prose.** *J Roy Statist Soc Ser A* 1974, **137:**25–34.
37. Sichel HS: **On a distribution law for word frequencies.** *J Amer Statist Assoc* 1975, **70:**542–547.
38. Sichel HS: **Modelling species-abundance frequencies and species-individual functions with the generalized inverse Gaussian-Poisson distribution.** *South African Statist J* 1997, **31:**13–37.
39. Ord JK, Whitmore GA: **The Poisson-inverse Gaussian distribution as a model for species abundance.** *Commun Statist Theor Meth* 1986, **15:**853–871.
40. Atkinson AC, Yeh L: **Inference for Sichel's compound Poisson distribution.** *J Amer Statist Assoc* 1982, **77:**153–158.
41. Karlis D, Xekalaki E: **Mixed Poisson distributions.** *International Statistical Review* 2005, **73:**35–58.
42. Puig P, Valero J: **Count data distributions: some characterizations with applications.** *J Amer Statist Assoc* 2006, **101:**332–340.
43. Zhu R, Joe H: **Modelling heavy-tailed count data using a generalized Poisson-inverse Gaussian family.** *Statist Probab Letters* 2009, **79:**1695–1703.
44. El-Shaarawi AH, Zhu R, Joe H: **Modelling species abundance using the Poisson-Tweedie family.** *Environmetrics* 2011, **22:**152–164.
45. Goldstein ML, Morris SA, Yen GG: **Problems with fitting to the power-law distribution.** *Eur Phys J B* 2004, **41:**255–258.
46. Akaike H: **A new look at the statistical model identification.** *IEEE Transactions on Automatic Control* 1974, **19:**716–723.
47. Zipf GK: *Human Behaviour and the Principle of Least Effort*. Reading, MA: Addison-Wesley; 1949.
48. Wolters DA, Washburn MP, Yates JR III: **An automated multidimensional protein identification technology for shotgun proteomics.** *Anal Chem* 2001, **73:**5683–5690.
49. Liu H, Sadygov RG, Yates JR III: **A model for random sampling and estimation of relative protein abundance in shotgun proteomics.** *Anal Chem* 2004, **76:**4193–4201.
50. Fischer F, Poetsch A: **Protein cleavage strategies for an improved analysis of the membrane proteome.** *Proteome Science* 2006, **4:**2.
51. Ishihama Y, Oda Y, Tabata T, Sato T, Nagasu T, Rappsilber J, Mann M: **Exponentially modified protein abundance index (emPAI) for estimation of absolute protein amount in proteomics by the number of sequenced peptides per protein.** *Mol Cell Proteomics* 2005, **4:**1265–72.

doi:10.1186/1477-5956-11-5
Cite this article as: Koziol *et al.*: On protein abundance distributions in complex mixtures. *Proteome Science* 2013 **11**:5.