 DOI 10.1002/pmic.201000650 *Proteomics* 2011, *11*, 1340–1345

TECHNICAL BRIEF

Abacus: A computational tool for extracting and pre-processing spectral count data for label-free quantitative proteomic analysis

Damian Fermin[1], Venkatesha Basrur[1], Anastasia K. Yocum[1] and Alexey I. Nesvizhskii[1,2]

[1] Department of Pathology, University of Michigan, Ann Arbor, MI, USA
[2] Center for Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI, USA

We describe Abacus, a computational tool for extracting spectral counts from MS/MS data sets. The program aggregates data from multiple experiments, adjusts spectral counts to accurately account for peptides shared across multiple proteins, and performs common normalization steps. It can also output the spectral count data at the gene level, thus simplifying the integration and comparison between gene and protein expression data. Abacus is compatible with the widely used Trans-Proteomic Pipeline suite of tools and comes with a graphical user interface making it easy to interact with the program. The main aim of Abacus is to streamline the analysis of spectral count data by providing an automated, easy to use solution for extracting this information from proteomic data sets for subsequent, more sophisticated statistical analysis.

Received: October 13, 2010
Revised: December 15, 2010
Accepted: December 29, 2010

Keywords:
Bioinformatics / Protein inference / Shared peptides / Spectral counts / MS/MS

In the last several years, label-free MS based protein quantification methods have received significant attention and have become commonly used [1–3]. Label-free approaches offer several practical advantages over generally more accurate labeling-based methods. They often offer savings in terms of costs of the analysis, are easier to implement, and allow for complex experimental designs unlike labeling-based methods where comparisons can be made only for a limited number of samples.

One of the most commonly used label-free quantitation methods is spectral counting. In this approach, the number of tandem mass (MS/MS) spectra assigned to the peptides of a protein, after a proper normalization, is used to measure the protein's abundance in the sample [4, 5], for a recent review see, e.g. [3]. A number of statistical approaches and software tools have been described for assessing the significance of differential protein expression based on spectral count data [6–13], including our previously described program Qspec [7]. These programs take as input a spectral counts matrix (i.e. a table listing all proteins identified with high confidence and the corresponding spectral counts in each of the experiments) extracted from the MS data. While conceptually simple, accurate extraction of spectral counts and their use as a measure of the protein abundance nevertheless requires addressing several challenges. First, the analysis naturally involves processing of multi-sample data sets (biological and technical replicates, multiple cell lines or patient samples, etc.). The accession numbers of proteins identified in different samples need to be aligned across the experiments to create a single protein summary list. This task is complicated by the ambiguities in inferring protein identifiers from shotgun proteomic data – the protein inference problem [14]. Directly related to this issue is the uncertainty in the contribution from shared peptides, i.e. peptides present in multiple different proteins, to the spectral count of each their corresponding proteins [7, 15, 16]. Furthermore, protein-level data often needs to be matched to genomic data, which require mapping of protein to gene accession numbers.

Correspondence: Dr. Alexey I. Nesvizhskii, Department of Pathology, University of Michigan, 4237 Medical Science I, Ann Arbor, MI 48109, USA
E-mail: nesvi@umich.edu
Fax: +1-734-615-2500

Abbreviations: BAF, Brg/Brahma-associated factors; **ES,** embryonic stem

In this work we describe ''Abacus'' – a software tool for extracting spectral counts that is compatible with the Trans-Proteomic Pipeline (TPP) suite of computational tools [17]. The overview of Abacus is shown in Fig. 1. Abacus takes as input PeptideProphet [18] and ProteinProphet [19] output files (in pepXML and protXML format, respectively). The program aggregates data from multiple experiments, adjusts spectral counts based upon how the peptides are shared among the proteins reported in the experimental results, and performs common normalization steps. In order to use Abacus it is necessary to generate a list of all the proteins that were identified from all of the individual MS experiments. This is achieved by using ProteinProphet (by the user and prior to running Abacus) to create a combined (''merged'') protXML file from the peptide-level results of each independent experiment. The resulting combined file more accurately represents the protein-level identifications across all experiments. After a combined protXML file is obtained, Abacus parses the individual pepXML and protXML result files storing their respective data into an internal database. Abacus then performs the following steps to arrive at the final results it reports:

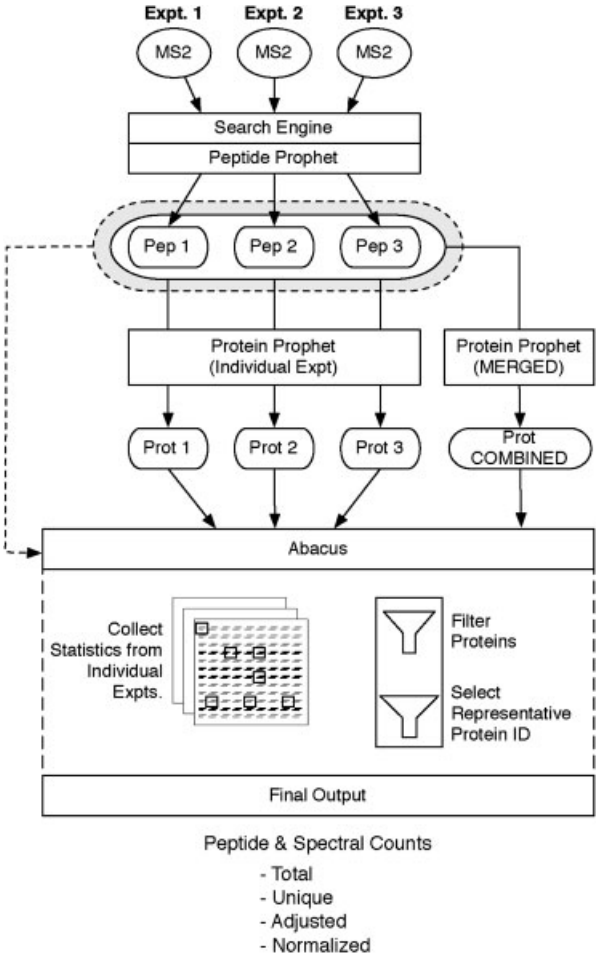

Figure 1. Overview of Abacus. Schematic representation of the workflow including processing of the data through Abacus. In this illustration, three individual MS experimental data sets were searched with a standard protein search engine (i.e. SEQUEST, MASCOT, etc.) and then post processed using PeptideProphet. The resulting pepXML files (Pep1, Pep2, and Pep3) were then processed through ProteinProphet to create three individual protXML files (Prot1, Prot2, and Prot3). The pepXML files were also processed together through ProteinProphet to merge all of their results into a single, combined protXML file (Prot COMBINED). All of the pepXML and protXML files were then parsed into Abacus and filters were applied to the merged results to filter out false positives and select a representative protein for each protein group. Statistics for the representative protein were then extracted from each of the individual protXML files and used to generate the final output that is produced by Abacus.

(i) Filter out low scoring proteins from the combined file.
(ii) Select representative protein identifiers from the combined file.
(iii) Collect summary information about the representative proteins from among independent experiments.
(iv) Calculate peptide/spectral counts for each protein.

The first step in the Abacus algorithm is to remove low scoring protein identifications recorded in the combined file. Abacus allows filtering of protein-level identifications based on several features: the ProteinProphet posterior protein probability from the combined file, the maximum protein probability observed across the individual experiments, as well as the maximum peptide probability observed for the protein. These parameters are adjustable allowing for precise control over what proteins are retained for subsequent analysis. It should be noted that in the context of this work, merging of multiple MS runs implies that all of the individual results from each MS run are combined together regardless of how often a particular protein is identified across all the replicates. This merging of multiple MS run results tends to increase the number of false positives [20]. Since the probability-based estimates provided by ProteinProphet may not be accurate in the case of very large multi-replicate experiments, filtering can be performed in such a way as to achieve a desired false discovery rate (FDR) based on the target-decoy strategy [20, 21].

In the second step of the algorithm, Abacus chooses a single protein identifier to represent each of the remaining entries in the combined file. ProteinProphet collects proteins that share peptide evidence into protein groups [14]. When there is ambiguity, Abacus selects a single representative identifier from within each of the protein groups of the combined file to report in the final output. This representative protein is chosen based on the following hierarchy. The first two heuristics are applied across the independent experimental results. The last four are executed on the data within the protein group of the combined file. These heuristics are followed sequentially until any ties are broken or the last rule is reached: (i) The protein identified the most often among independent experimental results;

(ii) The protein with the highest protein probability among the independent experiments; (iii) The protein with the highest scoring peptide assigned to it; (iv) The protein with the most number of distinct peptide sequences matched to it; (v) The protein with the highest spectral count; (vi) The top protein identifier after alphanumeric sorting of all remaining identifiers.

Having selected a representative protein accession number, the third step in the algorithm is to collect basic statistics about this protein from the results of the independent experiments. In addition to computing the total (i.e. regardless of whether the peptides are shared across multiple proteins) peptide and spectral counts for each protein, adjusted spectral counts are computed for the representative protein across each of the individual experiments. ProteinProphet already addresses the problem of shared peptides in the context of protein identification [14, 19] (and other approaches, e.g. [22–24]). In ProteinProphet, peptides shared across multiple proteins provide varying contributions (i.e. weight) to the protein's final score. Peptides that are highly redundant contribute less, whereas peptides unique to a single protein are given more importance. In Abacus, the same framework is applied to the calculation of spectral counts, leading to a more realistic quantitative measurement of a protein's abundance. A similar spectral count adjustment approach was investigated in [15], which showed that adjusting spectral counts based on how shared peptides are distributed gave the best agreement between computationally and experimentally derived measurements of protein abundance.

The adjustment procedure in Abacus is performed as follows. First, the number of unique spectra, s, assigned to each protein, i, is calculated. For each peptide p present in multiple proteins $j = 1, \ldots N$, its contribution to the spectral count of protein i is weighted by an adjustment factor, $\alpha_{p,i}$:

$$\alpha_{p,i} = \frac{s_i}{\sum_{j=1}^{N} s_j}$$

Given this definition, α weights range from zero to one. In essence, α determines what proportion of the spectral counts from peptide p should be ascribed to protein i. The sum of the peptides' adjusted counts constitutes the each protein's adjusted spectral count. The calculation of α and how it is applied to a single peptide case is illustrated in Fig. 2 using a subset of the prostatic secretion (EPS) data set from nine prostate cancer patients [25] (this data set is also provided as a sample data along with the Abacus software). After analyzing the X! Tandem search results with PeptideProphet and ProteinProphet, Abacus was used to extract adjusted spectral counts for one of the nine patient samples. Figure 2 focuses on four related immunoglobulin proteins (IGHG1 through IGHG4) that were identified with high confidence in the sample. These homologous proteins share numerous peptides in common. Figure 2 shows how the spectral count of a single one of these common peptides is distributed among these proteins. Following spectral count adjustment, IGHG1 protein is assigned to five of the peptide's eight spectral counts. The remaining three spectra are assigned to IGHG2 and IGHG3, respectively. IGHG4 is not assigned any of the spectra from this peptide. These adjusted counts are based on the unique spectral evidence ascribed to each of the proteins independently as described above. Without spectral adjustment, all eight of the peptide's spectra would have contributed equally to each protein's final spectral count.

In some cases, the use of adjusted counts is helpful for achieving accurate biological interpretation. This is illustrated in Fig. 3 using data from a recent study of an embryonic stem (ES) cell chromatin remodeling complex, esBAF [26]. In [26], the analysis was performed using a semi-manual spectral count adjustment procedure, for which Abacus now provides an automated software solution. The mammalian BAF (Brg/Brahma-associated factors) chromatin remodeling complexes play a key role in establishing and maintaining pluripotency. These complexes contain 11 core subunits, several of which are encoded by gene families. The diversity of BAF complexes is derived from the combinatorial assembly of alternative family members, some of which have a high degree of sequence homology to each other. As a result, accurate characterization of the composition of this protein complex requires appropriate adjustment of the spectral counts to account for the high number of shared peptides. Figure 3A shows the difference between the total and adjusted spectral counts for proteins identified in the analysis of the BAF complex in the mouse ES cells. A number of core complex components, notably BAF170 (protein Smarcc2), have substantially reduced protein abundance after adjustment. This is further illustrated in Fig. 3B, which shows the sequences of BAF170 and BAF155 (Smarcc1, 61.7% sequence homology with Smarcc2), and the peptides mapping to these two proteins identified from MS data. The comparison of the normalized abundances of these two proteins (NSAF factors) within the complex at two different stages of differentiations (ES and mouse embryonic fibroblast, MEF), computed using the spectral adjustment procedure is shown in Fig. 3D. Of key biological importance is the significant reduction of BAF170 in the BAF complex purified from ES cells, in agreement with quantitative immunoblotting data and other evidence [26]. At the same time, the abundance of BAF170 in ES cells based on the total (unadjusted) counts is overestimated (Fig. 3C). For example, without the adjustment the ratio of estimated BAF 155 versus BAF 170 protein abundances in ES cells is ~3:1, compared with ~12:1 after the adjustment for shared peptides (the latter shows a much better agreement with quantitative immunoblotting data shown in Fig. 3C in [26]).

Adjusted spectral counts allowed more accurate quantification in the study described above, and more accurate reconstruction of protein complexes from affinity purification-mass spectrometry (AP-MS) protein interaction data in

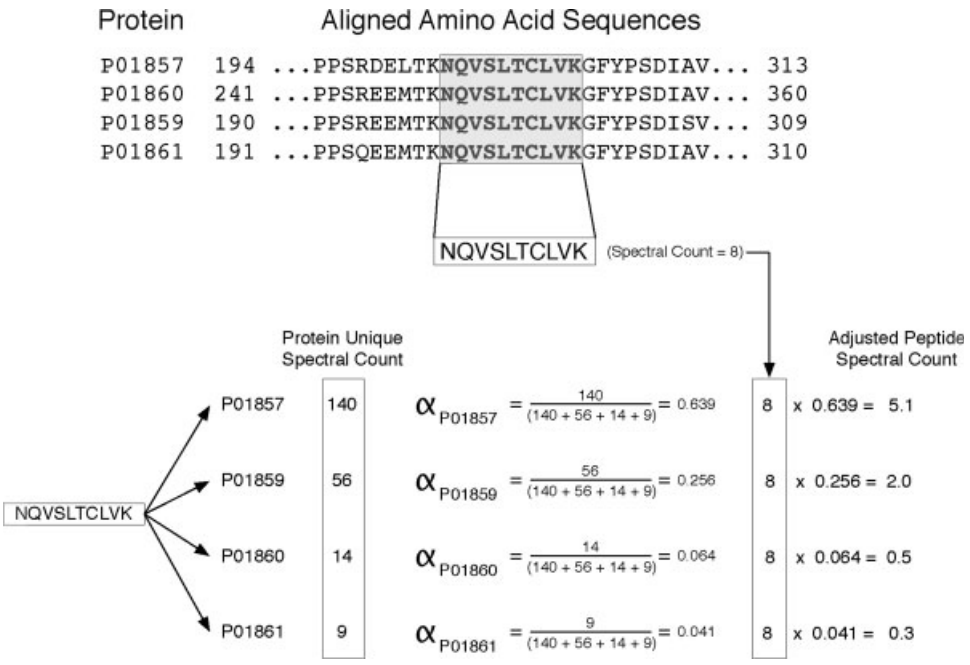

Figure 2. Illustration of the spectral count adjustment procedure. Example of how the α factor is calculated for a single peptide shared among four immunoglobulins (IGHG1–IGHG4) identified in the EPS data set. The upper panel shows the protein sequence alignment between the four immunoglobulins and highlights a peptide common to all of them. This specific peptide has eight unique spectra assigned to it making its spectral count 8. The lower panel shows how α is computed for each of the proteins that share this peptide. The α factor is based upon the unique spectral count of each protein. In this example, P01857, P01859, P01860, and P01861, each have unique spectral counts of 140, 56, 14, and 9 respectively. These unique spectral counts are derived from the peptides that are exclusive to each of the proteins in the example. The peptide's total spectral count of 8 is multiplied by the α factor of each protein. The resulting values indicate how the peptide's 8 spectral counts are to be distributed among each of the four proteins.

another of our recent works [27]. At the same time, our experience with label-free quantification data suggests that different research questions warrant the use of different counts (or simultaneous use of multiple measures). While adjusted counts can be more informative than total counts for relative quantification of highly homologous proteins, they may underestimate the absolute protein abundances. For example, when using spectral count-based quantification measures as a basis for separating between true and false protein interactions using the SAINT statistical model [28, 29], we routinely utilize total peptide or spectral counts (instead of adjusted counts) in order to perform more conservative assessments and to eliminate non-specific binders. Thus, Abacus reports a number of different abundance metrics, including adjusted, total, and unique counts, for spectra and peptides, as well as normalized spectral abundance factors, NSAF [30]. It is worth noting that similar challenges of dealing with shared sequence counts are present in other data, most notably RNA-Seq data (''multi-read'' counts), for examples see [31, 32].

Abacus can also provide counts (both spectral and peptide) at the gene level. The members of a protein group are often related isoforms derived from the same gene locus. For gene-centric consolidation, proteins are mapped to their parent gene. Once proteins have been mapped, their peptides are then associated with the parent gene. Spectral counts are still adjusted as described above except that α factors are computed based upon genes not proteins. The genes are assigned the maximum protein probability reported from among their observed protein products in the combined file. This gene-centric output is often useful for performing quantitative comparisons between protein and gene expression data. As next-generation sequencing methods become more established, such comparisons will become more popular [33–35]. It must be emphasized that mapping proteins back to their parent gene loci is not a trivial task and is an on-going challenge. For this reason, Abacus requires the user to provide a gene-to-protein mapping file. Such files can be easily generated for the public databases and we provide example programs in the Supporting Information.

A key aim in the development of Abacus was making it user friendly. Abacus comes with a graphical user interface making it easy to interact with the program (see Supporting Information for a detailed description of the software). This interface allows users to easily apply filters and choose what information is reported. The flexibility provided by Abacus is one of its strongest features. The user is given a great deal of control in deciding what to report as well as how the data are filtered. Currently there are few other open source tools that provide a platform independent method to extract spectral counts from proteomics data sets in a user-friendly manner.

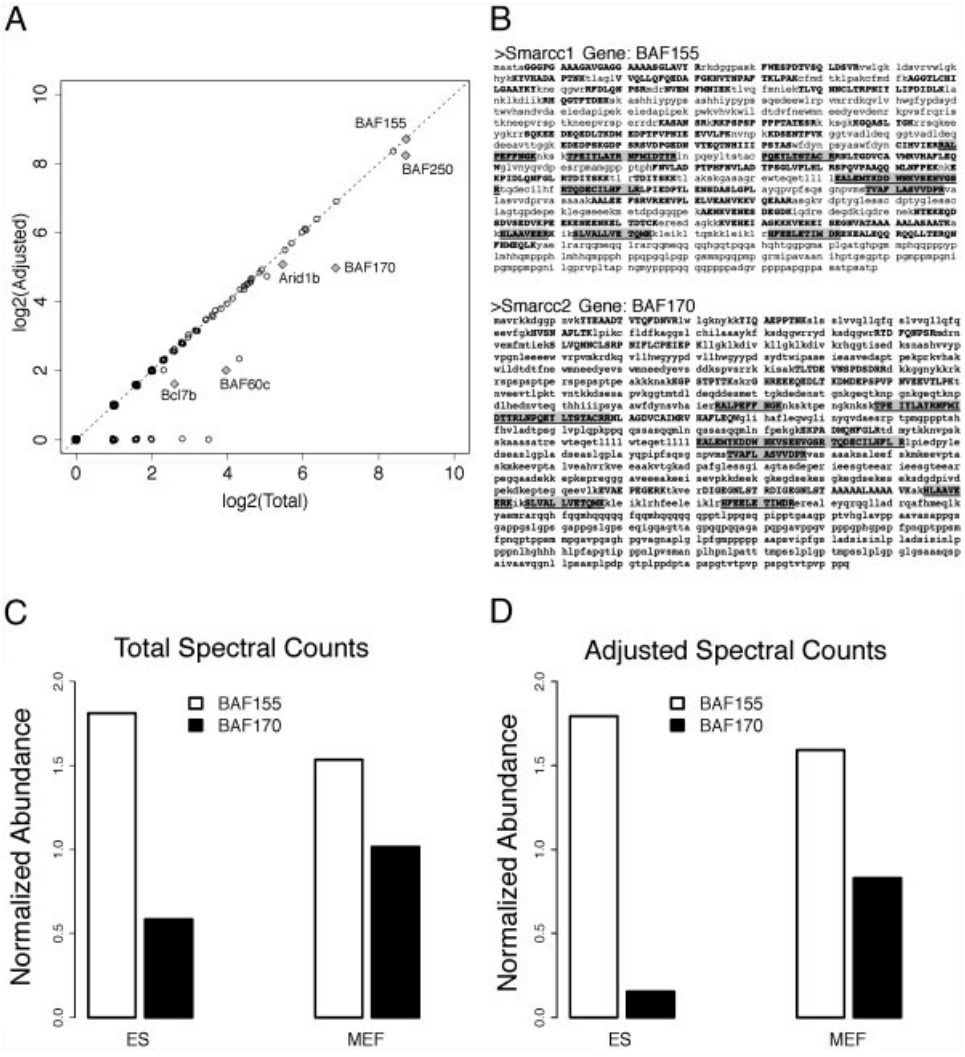

Figure 3. Comparison of total and adjusted spectral counts: application to BAF chromatin remodeling complex [26]. (A) The difference between the total and adjusted spectral counts for proteins identified in the analysis of the BAF complex in the mouse ES cells. Selected core components of the complex most affected by the adjustment procedure are marked. (B) The sequences of two homologous proteins, BAF170 and BAF155, and the peptides mapping to these two proteins identified from MS data in the ES cells (identified peptide sequences are in bold, and those that are shared between the proteins are in gray boxes). (C) The normalized abundances of these two proteins (NSAF factors, additionally normalized to levels of Brg protein [26] in each cell line) within the complex at two different stages of differentiations (ES and mouse embryonic fibroblast). (D) Same as (C), but using adjusted spectral counts.

In its simplest usage, Abacus can provide summary statistics for a large collection of experiments that would otherwise be too complicated to manage. An option for creating a QSpec-compatible [7] output format is available, which simplifies the subsequent statistical analysis of differential expression. Should the existing options be insufficient, users can directly query the database that holds all the information for their data. Abacus uses the HyperSQL database as a back end to store and query the information it extracts from the pepXML and protXML files (HyperSQL Database Engine, 2010, http://hsqldb.org/). Having this database distributed with Abacus allows users to directly access their data in a robust relational database should the default output of Abacus not fulfill their needs. Abacus is written in JAVA and has been tested to verify reproducible results on Windows, Linux, and MacOS X platforms. The software is open-source and distributed under the Apache License 2.0. The software, source code, documentation and sample data are available at http://abacustpp.sourceforge.net.

This work was supported in part by NIH grants R01-CA-126239 and R01-GM-094231. We would like to thank Lena Ho, Jeff Ranish, Hyungwon Choi, and Dattatreya Mellacheruvu for helpful discussions.

The authors have declared no conflict of interest.

References

[1] Elliott, M. H., Smith, D. S., Parker, C. E., Borchers, C., Current trends in quantitative proteomics. *J. Mass Spectrom.* 2009, *44*, 1637–1660.

[2] Zhu, W., Smith, J. W., Huang, C. M., Mass spectrometry-based label-free quantitative proteomics. *J. Biomed. Biotechnol.* 2010, 840518.

[3] Lundgren, D. H., Hwang, S. I., Wu, L. F., Han, D. K., Role of spectral counting in quantitative proteomics. *Expert Rev. Proteomics* 2010, *7*, 39–53.

[4] Liu, H. B., Sadygov, R. G., Yates, J. R., A model for random sampling and estimation of relative protein abundance in shotgun proteomics. *Anal. Chem.* 2004, *76*, 4193–4201.

[5] Old, W. M., Meyer-Arendt, K., Aveline-Wolf, L., Pierce, K. G. et al., Comparison of label-free methods for quantifying human proteins by shotgun proteomics. *Mol. Cell. Proteomics* 2005, *4*, 1487–1502.

[6] Carvalho, P. C., Fischer, J. S., Chen, E. I., Yates, J. R., 3rd., Barbosa, V. C., PatternLab for proteomics: a tool for differential shotgun proteomics. *BMC Bioinformatics* 2008, *9*, 316.

[7] Choi, H., Fermin, D., Nesvizhskii, A. I., Significance analysis of spectral count data in label-free shotgun proteomics. *Mol. Cell. Proteomics* 2008, *7*, 2373–2385.

[8] Heinecke, N. L., Pratt, B. S., Vaisar, T., Becker, L., PepC: proteomics software for identifying differentially expressed proteins based on spectral counting. *Bioinformatics* 2010.

[9] Li, M., Gray, W., Zhang, H. X., Chung, C. H. et al., Comparative shotgun proteomics using spectral count data and quasi-likelihood modeling. *J. Proteome Res.* 2010, *9*, 4295–4305.

[10] Little, K. M., Lee, J. K., Ley, K., ReSASC: a resampling-based algorithm to determine differential protein expression from spectral count data. *Proteomics* 2010, *10*, 1212–1222.

[11] Pham, T. V., Piersma, S. R., Warmoes, M., Jimenez, C. R., On the beta-binomial model for analysis of spectral count data in label-free tandem mass spectrometry-based proteomics. *Bioinformatics* 2010, *26*, 363–369.

[12] Searle, B. C., Scaffold: a bioinformatic tool for validating MS/MS-based proteomic studies. *Proteomics* 2010, *10*, 1265–1269.

[13] Pavelka, N., Fournier, M. L., Swanson, S. K., Pelizzola, M. et al., Statistical similarities between transcriptomics and quantitative shotgun proteomics data. *Mol. Cell. Proteomics* 2008, *7*, 631–644.

[14] Nesvizhskii, A. I., Aebersold, R., Interpretation of shotgun proteomic data: the protein inference problem. *Mol. Cell. Proteomics* 2005, *4*, 1419–1440.

[15] Zhang, Y., Wen, Z., Washburn, M. P., Florens, L., Refinements to label free proteome quantitation: how to deal with peptides shared by multiple proteins. *Anal. Chem.* 2010, *82*, 2272–2281.

[16] Zybailov, B., Rutschow, H., Friso, G., Rudella, A. et al., Sorting signals, N-terminal modifications and abundance of the chloroplast proteome. *PLoS One* 2008, *3*, 19.

[17] Deutsch, E. W., Mendoza, L., Shteynberg, D., Farrah, T. et al., A guided tour of the Trans-Proteomic Pipeline. *Proteomics* 2010, *10*, 1150–1159.

[18] Keller, A., Nesvizhskii, A. I., Kolker, E., Aebersold, R., Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal. Chem.* 2002, *74*, 5383–5392.

[19] Nesvizhskii, A. I., Keller, A., Kolker, E., Aebersold, R., A statistical model for identifying proteins by tandem mass spectrometry. *Anal. Chem.* 2003, *75*, 4646–4658.

[20] Nesvizhskii, A. I., A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. *J. Proteomics* 2010, *73*, 2092–2123.

[21] Elias, J. E., Gygi, S. P., Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat. Methods* 2007, *4*, 207–214.

[22] Li, Y. F., Arnold, R. J., Li, Y., Radivojac, P., Sheng, Q., Tang, H., A bayesian approach to protein inference problem in shotgun proteomics. *J. Comput. Biol.* 2009, *16*, 1183–1193.

[23] Li, J., Zimmerman, L. J., Park, B. H., Tabb, D. L., Liebler, D. C., Zhang, B., Network-assisted protein identification and data interpretation in shotgun proteomics. *Mol. Syst. Biol.* 2009, *5*, 303.

[24] Gerster, S., Qeli, E., Ahrens, C. H., Buhlmann, P., Protein and gene model inference based on statistical modeling in k-partite graphs. *Proc. Natl. Acad. Sci. USA* 2010, *107*, 12101–12106.

[25] Drake, R. R., Elschenbroich, S., Lopez-Perez, O., Kim, Y. et al., In-depth proteomic analyses of direct expressed prostatic secretions. *J. Proteome Res.* 2010, *9*, 2109–2116.

[26] Ho, L., Ronan, J. L., Wu, J., Staahl, B. T. et al., An embryonic stem cell chromatin remodeling complex, esBAF, is essential for embryonic stem cell self-renewal and pluripotency. *Proc. Natl. Acad. Sci. USA* 2009, *106*, 5181–5186.

[27] Choi, H., Kim, S., Gingras, A. C., Nesvizhskii, A. I., Analysis of protein complexes through model-based biclustering of label-free quantitative AP-MS data. *Mol. Syst. Biol.* 2010, *6*, 11.

[28] Choi, H., Larsen, B., Lin, Z.-Y., Breitkreutz, A. et al., SAINT: probabilistic scoring of affinity purification-mass spectrometry data. *Nat. Methods* 2011, *8*, 70–73.

[29] Breitkreutz, A., Choi, H., Sharom, J. R., Boucher, L. et al., A global protein kinase and phosphatase interaction network in yeast. *Science* 2010, *328*, 1043–1046.

[30] Zybailov, B., Mosley, A. L., Sardiu, M. E., Coleman, M. K., Florens, L., Washburn, M. P., Statistical analysis of membrane proteome expression changes in *Saccharomyces cerevisiae*. *J. Proteome Res.* 2006, *5*, 2339–2347.

[31] Li, B., Ruotti, V., Stewart, R. M., Thomson, J. A., Dewey, C. N., RNA-Seq gene expression estimation with read mapping uncertainty. *Bioinformatics* 2010, *26*, 493–500.

[32] Jiang, H., Wong, W. H., Statistical inferences for isoform expression in RNA-Seq. *Bioinformatics* 2009, *25*, 1026–1032.

[33] Greenbaum, D., Jansen, R., Gerstein, M., Analysis of mRNA expression and protein abundance data: an approach for the comparison of the enrichment of features in the cellular population of proteins and transcripts. *Bioinformatics* 2002, *18*, 585–596.

[34] Washburn, M. P., Koller, A., Oshiro, G., Ulaszek, R. R. et al., Protein pathway and complex clustering of correlated mRNA and protein expression analyses in *Saccharomyces cerevisiae*. *Proc. Natl. Acad. Sci. USA* 2003, *100*, 3107–3112.

[35] Fu, X., Fu, N., Guo, S., Yan, Z. et al., Estimating accuracy of RNA-Seq and microarrays with proteomics. *BMC Genomics* 2009, *10*, 161.

