MCP Papers in Press. Published on November 19, 2007 as Manuscript M700240-MCP200

Statistical Similarities Between Transcriptomics and Quantitative Shotgun Proteomics Data

Norman Pavelka(1), Marjorie L. Fournier(1), Selene K. Swanson(1), Mattia Pelizzola(2,§), Paola Ricciardi-Castagnoli(3), Laurence Florens(1), Michael P. Washburn(1,*)

1) Stowers Institute for Medical Research, 1000 E. 50th Street, Kansas City, Missouri 64110, U.S.A.

2) Department of Biotechnology and Bioscience, University of Milano-Bicocca, Piazza della Scienza 2, 20126 Milano, Italy

3) Singapore Immunology Network, 8A Biomedical Grove, Immunos Building, Singapore 138648

§) Present address:
Division of Biostatistics Epidemiology and Public Health, Yale University School of Medicine, New Haven, CT 06520, U.S.A.

*) To whom correspondence should be addressed:
Michael P. Washburn, Ph.D.
Stowers Institute for Medical Research
1000 E. 50th Street
Kansas City, MO 64110, U.S.A.
e-mail: mpw@stowers-institute.org
Phone: (816) 926-4457
Fax: (816) 926-4694

Running title: Similarities between transcriptomics and proteomics data

Abbreviations

MudPIT: Multidimensional Protein Identification Technology

NSAF: Normalized Spectral Abundance Factor

PLGEM: Power Law Global Error Model

FPR: False Positive Rate

FDR: False Discovery Rate

SD: Standard Deviation

CV: Coefficient of Variation

LP: Log Phase

SP: Stationary Phase

GO: Gene Ontology

Summary

If the large collection of microarray-specific statistical tools was applicable to the analysis of quantitative shotgun proteomics datasets, it would certainly foster an important advancement of proteomics research. Here, we analyze two large multi-dimensional protein identification technology (MudPIT) datasets – one containing 8 replicates of the soluble fraction of a yeast whole-cell lysate, one containing 9 replicates of a human immuno-precipitate – to test whether normalized spectral abundance factor (NSAF) values share substantially similar statistical properties with transcript abundance values from Affymetrix GeneChip data. First, we show similar dynamic range and distribution properties of these two types of numeric values. Next, we observe that the standard deviation (SD) of a protein's NSAF values is dependent on the average NSAF value of the protein itself, following a power law. This relationship can be modeled by a power law global error model (PLGEM), initially developed to describe the variance-versus-mean dependence that exists in GeneChip data. PLGEM parameters obtained from NSAF datasets prove to be surprisingly similar to the typical parameters observed in GeneChip datasets. The most important common feature identified by this approach is that, although in absolute terms the SD of replicated abundance values increases as a function of increasing average abundance, the coefficient of variation – a relative measure of variability – becomes progressively smaller under the same conditions. We next show that PLGEM parameters are reasonably stable to decreasing numbers of replicates. We finally illustrate one possible application of PLGEM in the identification of differentially abundant proteins, which might potentially outperform standard statistical tests. In summary, we believe that this body of work lays the foundation for the application of microarray-specific tools in the analysis of NSAF datasets.

Introduction

In recent years, the biomedical research community has recognized a need to shift its focus from the single-component level to the whole-system level, in order to understand complex physiological processes as well as elusive pathological conditions (1-3). Massive sequencing projects have provided comprehensive lists of the players in these games, and advancements in microarray technology (4, 5) and mass spectrometry (6, 7) allow today to measure abundances of all known mRNA and numerous protein species in a cell. The next challenge is to reverse engineer the ‘rules of the game’ by observing how players behave and how they interact with each other (8).

To this end, the first layer of complexity that can be addressed by such technologies is exemplified by the following question: Which transcripts or proteins change their abundance in a given cell as a result of a normal biological process, in response to a specific perturbation or as a consequence of disease? Although not conclusive, answering this type of question has already proven to help pinpoint the major players in several biological systems (9-12). In contrast to microarray-based transcriptomics, mass spectrometry-based proteomics (13) has unfortunately received less contributions from statistics and bioinformatics in terms of specific algorithms and software that are designed to answer the types of questions described above. Therefore, if the wealth of microarray-specific statistical tools could be directly applied to analyze proteomics data, this would most likely represent an enormous benefit for the rapid advancement of systems biology.

Conceptually, there are significant similarities between MS-based proteomics data and microarray-based gene expression data. Primarily, both technologies are believed to measure abundances of biological entities in a largely unbiased way (6, 14), which allows the use of a common mathematical representation of the data. Both types of datasets are typically represented as a matrix of numeric values, where rows represent different transcripts or proteins in a cell, columns represent distinct microarray hybridizations or MS runs and each entry represents the measured abundance level. Microarray data analysts have recognized long ago that standard statistical tools are not appropriate to analyze these data matrices, because of the 'many-genes-few-replicates' problem (15, 16). More precisely, all standard statistical methods rely on judging whether the difference in means between two series of values (here representing abundances of biological entities in two experimental conditions of interest) is significantly higher than the variation expected by chance. Classically, statistical tests estimate this random variation by measuring the variability between the replicated measures within each series of values. But when the number of available replicates is 100 or 1000 times smaller than the number of analyzed transcripts (or proteins), the chance of occasionally measuring artificially small or artificially large standard deviations becomes dominant, potentially leading to an increase in both false positive and false negative identifications. To address this issue, several microarray-specific tools have been developed (16-20). It would therefore be of particular interest to test whether these methods were applicable to the analysis of proteomics data as well.

One hindrance to the direct transfer of expertise between these two approaches has been the wide-spread belief that, due to the different chemistry of nucleic acids and polypeptides and the different technologies used to analyze them, transcriptome data and proteome data had to be analyzed with distinct sets of tools. Until very recently, for example, LC-MS/MS (also known as 'shotgun') proteomics was not even granted the definition of being a quantitative technique unless it was coupled with specific labeling methods that would make it suitable for relative quantification of proteins in an equimolar mixture of two samples of interest (21, 22). But sampling statistics, such as spectral counts, obtained by labeled or label-free shotgun proteomics have proven to allow quantification of proteins in single samples (23-25). For instance, we have recently used normalized spectral abundance factor (NSAF) values obtained by multi-dimensional protein identification technology (MudPIT) to determine the relative protein abundances inside the human Mediator complex (26) or for identifying abundance changes of yeast transmembrane proteins upon shift from a minimal to a rich culture medium (27). One feature of spectral counting based approaches, like NSAF, is that they provide measures of protein abundances between different proteins in datasets and are applicable to any sample type. In our view, these represent important steps forward that render shotgun proteomics data conceptually more similar to microarray gene expression data.

Besides conceptual similarities, applicability of microarray-specific statistical methods to the analysis of shotgun proteomics data will ultimately depend also on more substantial similarities. At the least, numeric values representing transcript or protein abundance levels should have similar statistical properties, such as dynamic range or overall shape of the distribution of values. Furthermore, it would be important if proteomics datasets and microarray datasets obeyed a similar global error model. Several authors, for example, have reported that variability of gene expression data is dependent on the average expression level of the gene itself and have termed this phenomenon ‘variance-versus-mean dependence’ (28, 29). Taking this relationship explicitly into account has shown to partially solve the ‘many-gene-few-replicates’ problem and to significantly improve the performance of the identification of differentially expressed genes (20, 30, 31). More specifically, we have previously reported that standard deviations from replicated Affymetrix GeneChip data could be modeled via a Power Law Global Error Model (PLGEM); and use of PLGEM-derived standard deviations allowed the detection of a higher number of truly differentially expressed genes without increasing the false positive rate (20). The PLGEM-based method was then implemented into a freely available Bioconductor (32) package, called ‘plgem’, as well as in an automated microarray data analysis pipeline, called ‘AMDA’ (33). These implementations have already been applied – both by us (34), as well as by other authors (35) – to successfully analyze real microarray data addressing real biological questions. Another study reported the successful application of a quadratic model to explain the dependence between noise variances and mean peak intensities in LC-MS proteomics datasets; and application of this error model resulted in a false positive rate (FPR) that was closer to the expectation value, compared to the FPR obtained by a standard Welch’s t-test (36). To the best of our knowledge, there is to date in the scientific literature no equally detailed error modeling study of shotgun proteomics data. If it was proven that NSAF data also obeyed a global error model, this could improve our ability to distinguish true protein abundance changes from random fluctuations.

The scope of the present work was therefore to compare general statistical properties of protein abundance values represented by NSAF values with those of transcript profiling data obtained by GeneChip experiments. Using two large MudPIT datasets (one containing 8 biological replicates of the soluble fraction of a yeast whole-cell lysate, one containing 9 technical replicates of a human protein complex preparation), we compared global distributions of major statistical parameters and tested whether NSAF datasets are characterized by a variance-versus-mean dependence similar to that governing GeneChip data. This work shows that there are indeed substantial similarities between the quantitative values obtained by these two apparently dissimilar technologies, and provides the basis for applying PLGEM-based methods – and possibly other microarray-specific tools – to NSAF datasets for the identification of differentially abundant proteins.

Experimental Procedures

Protein extraction for the Yeast proteome

For the control yeast dataset *Saccharomyces cerevisiae* strain BY4741 (37) was grown to middle log phase (OD at 600 nm of 1-1.5) in 2.5 l of rich media, consisting of: 100 ml of 10X concentrated BioExpress 1000 containing amino acids either labeled with ^{14}N or ^{15}N (Cambridge Isotope Laboratories, Andover, MA); 20 mg/l of uracil; 1.8 g/l of yeast nitrogen base without amino acids and ammonium sulfate; and 2% of dextrose. A total of 8 independent cultures were grown, four in ^{14}N and four in ^{15}N medium. Cells were collected and washed in cold ultrapure water by centrifugation for 20 min at 4,000 × g at 4 ºC. Cell pellets were resuspended in lysis buffer (310 mM of sodium fluoride, 3.45 mM of sodium orthovanadate, 12 mM of ethylenediamine tetraacetic acid, 250 mM of sodium chloride and 100 mM of sodium carbonate) and broken using silica glass beads by 10 cycles consisting of 1 min vortexing at 2,500 rpm followed by 30s incubation at 4 ºC. Unbroken cells were removed by centrifugation for 20 min at 4,000 × g at 4 ºC. The supernatant was transferred to a 50 ml centrifuge tube and soluble proteins were separated from the crude membrane fraction by centrifugation for 1 h at 22,000 × g at 4 ºC. The supernatant containing the soluble protein extract was collected, centrifuged, and transferred to a clean 50 ml tube and stored at -80 ºC. Protein concentration was determined by bicinchoninic acid (BCA) assay (Pierce, Rockford, IL). The eight independent samples were combined into four independent pools, each of which containing 500 µg total ^{14}N- and ^{15}N-labeled proteins mixed at a 1:1 ratio before TCA precipitation and MudPIT analysis.

For the comparative growth phase proteomic analysis, *Saccharomyces cerevisiae* strain BY4741 was grown in ^{14}N as described before. The logarithmic (LP) and stationary (SP) growth phase proteomic analyses were performed on cells collected respectively at an averaged OD at 600nm of 0.96 +/- 0.06 and 4.5 +/- 0.15 over four replicated experiments. Cells were collected and washed as described before and stored at -80 ºC before protein extraction. For protein extraction, cell pellets were resuspended in lysis buffer (310 mM of sodium fluoride, 3.45 mM of sodium orthovanadate, 12 mM of ethylenediamine tetraacetic acid, 250 mM of sodium chloride and 100 mM of sodium carbonate) and broken using silica glass beads by 12 cycles consisting of 30s bead beating, using a bead beater model 1107900 (BioSpec Products Inc.), followed by 1 min incubation at 4 ºC. The beads and cells debris were removed by centrifugation for 30min at 4,000 × g at 4 °C. The supernatant was collected and centrifuged for 1.5 hours at 45,000 × g at 4 °C. The supernatant containing the whole cells extract was collected and stored at – 80 °C. Protein concentration was determined by bicinchoninic acid (BCA) assay (Pierce, Rockford, IL). For each replicated experiments and growth condition, MudPIT analysis has been performed on 500µg of protein extract desalted by TCA precipitation.

Protein extraction for the Mediator complex

The mammalian Mediator of RNA polymerase II transcription is a multi-protein complex, composed of over 30 subunits. Stably transfected HeLa cell lines, each expressing a different FLAG-tagged Mediator subunit, i.e. human Med9, Med10, Med19, Med26, Med28, Med29 or the mouse orthologs of Med9 or Med19, were constructed. Nuclear proteins from these cell lines were extracted and purified by anti-FLAG agarose immunoaffinity chromatography (FLAG-IP) as described previously (38). The third elutions of all preparations involving a FLAG-tagged Mediator subunit were pooled, TCA precipitated and quantified by

BCA assay (Pierce). The pooled mixture was split into identical aliquots of 10 µg each, nine of which were independently analyzed in the present study.

MudPIT analysis

Protein mixtures were TCA precipitated, urea-denatured, reduced, alkylated and digested with endoproteinase Lys-C followed by modified trypsin digestion (both from Roche, Indianapolis, IN), as previously described (6). Peptide mixtures from the yeast proteins or the Mediator complex were respectively loaded onto split phase or 3-phase 100 µm fused silica microcapillary columns both packed with 5-µm C18 reverse phase (Aqua, Phenomenex), strong cation exchange particles (Partisphere SCX, Whatman), and reverse phase (39). Loaded microcapillary columns were placed in-line with a Quaternary Agilent 1100 series HPLC pump and a LTQ linear ion trap ion trap MS equipped with a nano-LC electrospray ionization source (ThermoFinnigan). Fully automated 7-step MudPIT runs were carried out on the electrosprayed peptides for the Mediator samples as described previously (40), while a 12-step MudPIT run was performed for the yeast proteome analyses as previously described (27). Each full MS scan (from 400 to 1600 m/z range) was followed by five MS/MS events using data-dependent acquisition, where the five most intense ions from a given MS scan were subjected to CID.

MS/MS data processing

Proteins were identified by database searching using SEQUEST software (41). The list of parameters used for the yeast and human datasets searches are available in Supplementary Tables 1A-D and 2A, respectively. Briefly, no enzyme specificity was imposed during searches, setting a mass tolerance of 3 amu for precursor ions and of 0 amu for fragment ions. In all searches, cysteine residues were considered to be fully carboxamidomethylated (+57 Da statically added). No variable modifications were searched. For the yeast proteome, tandem mass spectra were searched against a database containing 14176 protein sequences combining 6911 *S. cerevisiae* proteins (from the National Center of Biotechnology Information 2006-03-03 release), 177 common contaminants, such as keratin and immunoglobulins, and their corresponding 7088 randomized amino acid sequences. Each MS/MS dataset was searched four times following these criteria: 1) ^{14}N aminoacids; 2) ^{14}N aminoacids and +16 Da statically added to methionine (referred as methionine oxidation); 3) ^{15}N aminoacids for which the appropriate number of nitrogen atoms where statically added to their masses; 4) ^{15}N aminoacids and methionine oxidation (Supplementary Table 1A-D). The sqt files generated from the four independent searches were merged in the final dataset as described before (27). For the yeast log phase versus stationary phase comparative analyses, no ^{15}N was used so each dataset was searched using ^{14}N specific parameters found in Supplemental Table 1A-B. Each MS/MS dataset was searched two times following these criteria: 1) ^{14}N aminoacids and 2) ^{14}N aminoacids and +16 Da statically added to methionine. The sqt files generated from the two independent searches were merged in the final dataset as described before (27). For the Mediator samples, MS/MS spectra were searched against a database of 60234 amino acid sequences, consisting of 29890 human proteins (non-redundant entries from NCBI 2006-11-07 release), 160 usual contaminants (such as human keratins, IgGs and proteolytic enzymes), 67 epitope-tagged proteins (including mouse orthologs of Med9 and Med19) and 30117 randomized amino acid sequences derived from each non-redundant protein entry. Peptide/spectrum matches, including precursor ion m/z values and charge states, for the yeast control, human, and yeast log phase versus stationary phase datasets are respectively provided as Supplementary Tables 1E, 2B, and 3A and can be viewed under http://research.stowers-institute.org/washburnlab/Pavelka-MCP-2007/. The lists of detected peptides and proteins were sorted and selected using DTASelect (42) with the following criteria set: spectra/peptide matches were only retained if they had a DeltCn of at least 0.1, minimum XCorr of 1.5 for singly-, 2.5 for doubly-, and 3.0 for triply-charged spectra, and maximum Sp rank of 10. In addition, peptides had to be fully-tryptic and at least 7 amino acids long. Peptide hits from multiple runs were compared (Supplementary Tables 1F, 2C, and 3B) using CONTRAST (42) and contrast-report (43). Proteins that were subsets of others were removed using the parsimony option in DTASelect (42). The False Discovery Rate (FDR) was calculated as the number of spectra matching randomized peptides multiplied by 2 and divided by the total number of spectra, as described before (44), and ranged between 0 and 0.465% for all MudPIT runs (Supplementary Tables 1F, 2C, and 3B).

Protein abundances were estimated using Normalized Spectral Abundance Factor (NSAF) values, calculated from the spectral counts of each identified protein (27). Briefly, to account for the fact that larger proteins tend to contribute more peptide/spectra, spectral counts were divided by protein length to provide a Spectral Abundance Factor (SAF). SAF values were then normalized against the sum of all SAF values in the corresponding run, allowing the comparison of protein levels across different runs. No particular thresholds or outlier removal steps were applied prior to NSAF calculation. The NSAF values of each detected protein from the yeast, Mediator, and yeast log phase versus stationary phase MudPIT datasets are provided as Supplementary Tables 1F, 2C, and 3B, respectively. For subsequent statistical analysis, all datasets were further processed to retain only proteins that were identified at least in three replicated experiments. Finally, contaminant proteins were removed.

GeneChip datasets

The Mouse GeneChip dataset used in the present study is a subset of a previously published dataset (45). This subset contains 11 replicates of the transcriptome of untreated mouse dendritic cells, measured by MG-U74Av2 GeneChip arrays (Affymetrix, Santa Clara, CA) following standard procedures. All experimental details can be found in the original publication (45). All remaining microarray datasets were downloaded on 2007-02-22 from the Gene Expression Omnibus database (46), using the following search criteria: i) The microarray platform had to be an Affymetrix GeneChip; ii) Absolute signal intensities had to be obtained by standard image processing, background correction and summarization methods, as implemented either in the MicroArray Suite 5.0 or in the GeneChip Operating System software application (both from Affymetrix); iii) Datasets had to contain at least one experimental condition with a minimum of three replicates. The combination of these selection criteria yielded 26 distinct studies across 7 distinct platforms and 5 species (*Homo sapiens*: HG-U133Plus2.0 and HG-U133A; *Mus musculus*: MOE-430A and MG-U74Av2; *Rattus norvegicus*: RG-U34A; *Arabidopsis thaliana*: ATH1; *Saccharomyces cerevisiae*: YG-S98), with a total of 336 samples grouped in 101 sets of replicates. Each set of replicates represented either a unique experimental condition or a unique combination of experimental factors (in case that more than one experimental factor was annotated in the database for a particular dataset) and contained between three and five – either biological or technical – replicates. All accession numbers of the downloaded data can be found in Supplementary Table 4.

Statistical analysis

NSAF datasets and GeneChip datasets were imported into the R environment for statistical computing (47) and parsed into individual 'exprSet' objects, to allow recognition by specific Bioconductor packages (32). Missing values were replaced with zeros and data were normalized by dividing each value by the mean value of the corresponding column. The Bioconductor package 'plgem' (20) was used to fit a PLGEM to the individual datasets, evaluate the goodness-of-fit of the model to the data and detect differentially abundant transcripts or proteins. Relevant algorithmic details of the PLGEM method will be explained in the following section. All R scripts written to dynamically generate all the figures and tables in the present work are available from the authors upon request.

Results

Global statistical properties of NSAF datasets

In the present study, MudPIT was used to generate large-scale shotgun proteomics data and NSAF values were generated to obtain quantitative information from these datasets. We then compared the statistical properties of two previously unpublished NSAF datasets (Supplementary Tables 1 and 2) with those found in previously published GeneChip datasets. Before demonstrating the existence of significant similarities between these two types of numerical data, we first would like to acknowledge the presence of some important differences. One obvious difference among the datasets analyzed in the present work is related to the size of the corresponding data matrices (Table 1). By definition, a microarray experiment will provide abundance values for every transcript probed by the chip regardless of the actual presence of the corresponding transcripts in the analyzed sample. Instead, due to the sampling nature of shotgun proteomics approaches (23), MudPIT will detect only those proteins which are present in the sample with a concentration that is higher than the sensitivity threshold of the technology. In accordance to this view, the number of proteins present in the Yeast and in the Mediator NSAF dataset, were respectively ~15 and ~42 times smaller than the number of transcripts present in the Mouse GeneChip dataset (Table 1). For the same reason, an abundance value equal to zero (hereafter referred to as a 'zero value') was extremely unlikely in the Mouse GeneChip dataset (representing only ~0.02% of the total values), whereas it accounted for ~29% and ~35% of all values present in the Yeast and the Mediator NSAF datasets, respectively (Table 1). Interestingly, the percentage of transcripts associated with an 'Absent call' in the GeneChip dataset (~50%) was similar to the percentage of zero values in the two NSAF datasets, suggesting a possible semantic equivalence between these two types of information. Most probably as a third consequence of the phenomenon described above, the dynamic range of measured abundance values in the Yeast and the Mediator NSAF datasets ranged ~3.6-3.8 orders of magnitude, while the ones in the Mouse GeneChip dataset reached almost 4.7 orders of magnitude (Table 1). Nonetheless, these data confirm that despite important differences in the overall size and in the presence of zero values, microarray datasets and proteomics datasets are both able to measure abundances of biological entities over several orders of magnitude.

Such a wide dynamic range of values is unlikely to be produced by a Normal distribution. Instead, spot intensities from microarray data (48, 49) and NSAF values from shotgun proteomics datasets (27) have both been proposed to be approximately log-normally distributed. In a previous study, we have shown that the distribution of log-transformed NSAF values from a MudPIT dataset was not significantly different from a Normal distribution (27). In that study, in order to allow the log-transformation step, we analyzed only those proteins that were identified in a significant proportion of all performed MS runs and replaced the remaining zero values by a fraction of a spectral count before calculating the corresponding NSAF. Following the same approach, we observed a similar distribution of values also in the two NSAF datasets analyzed in the present work (data not shown). These results certainly support the hypothesis that the NSAF values of the most highly abundant proteins in a MudPIT dataset are log-normally distributed. Here, in order to provide a more general description of the distribution of values that would encompass also more lowly abundant proteins, and given the high percentage of zero values in the two NSAF datasets of the present study, we judged not to be appropriate to replace the zero values with a fractional value, to avoid introduction of a significant distortion in the data. Instead, we decided to focus our attention on the distribution of average (untransformed) NSAF values calculated for every protein in the dataset using all available replicates, which by definition have to be non-zero and will be referred hereafter as '*rowMean* values'. Interestingly, the overall distribution of *rowMean* values was more complex than a simple log-Normal distribution (Figure 1). In fact, it could be explained more realistically as a combination of multiple log-Normal distributions. In the case of the Mouse GeneChip dataset, the distribution of *rowMean* values could be clearly explained by two dominant log-Normal distributions, one representing transcripts flagged as 'Absent' across all 11 replicates, the other one representing transcripts without a single 'Absent call'. Only a minor proportion of transcripts had an intermediate number of 'Absent calls' (Figure 1A). Also the distribution of *rowMean* values in the NSAF datasets showed two dominant log-Normal components, one representing proteins with exactly three non-zero values, the other one representing proteins without zero values (Figure 1B-C). But in this case, the contribution of proteins with an intermediate number of zero values was more important, as compared to the contribution of transcripts with an intermediate number of 'Absent calls' in the GeneChip dataset. These results support a strategy of including in an NSAF data analysis also proteins identified only in a minor fraction of all performed MS runs, because these might simply represent more lowly abundant proteins that only occasionally pass the sensitivity threshold of the technology. Statistical methods capable of dealing with these rarely identified proteins will surely enhance our ability to fully interpret a shotgun proteomics dataset.

We next sought to provide a description of the distribution of the standard deviations measured for each transcript or protein across all available replicates, referred hereafter as *rowSD* values. The distribution of *rowSD* values was surprisingly similar to the distribution of *rowMean* values in the corresponding dataset (Figure 1D-F). This suggested the intriguing hypothesis that, as has been demonstrated in microarray datasets, also in proteomics datasets there is a relationship between the reproducibility of a protein's abundance values and the protein's average abundance level.

To identify the possible underlying relationship between data variability and average abundance levels, we drew two types of scatter-plots for each dataset (Figure 2). In the first case, we analyzed *rowSD* values – which can be seen as an absolute measure of data variability – as a function of the corresponding *rowMean* values in a log-log space (Figure 2A-C). These plots revealed a striking linear relationship over the whole dynamic range in all three analyzed datasets, with highly abundant transcripts or proteins showing a higher SD compared to lowly abundant ones. While the SD is considered an absolute measure of data variability, the coefficient of variation (CV) can be seen as a relative measure of data variability. The CV is defined as:

$$CV = \frac{SD}{mean}. \qquad \text{[Equation 1]}$$

In the second series of scatter-plots, we therefore analyzed the CV of the transcript or the protein, measured as the ratio between the corresponding *rowSD* and *rowMean*, hereafter referred to as *rowCV*. Also plots of the *rowCV* values as a function of the corresponding *rowMean* values in a log-log space revealed a striking linear relationship over the whole dynamic range in all three analyzed datasets (Figure 2D-F). But conversely to the behavior of the *rowSD* values, here, highly abundant transcripts or proteins had smaller *rowCV* values compared to lowly abundant ones.

Goodness-of-fit of PLGEM on NSAF datasets

The simplest model able to explain a linear relationship in a log-log space is a power law relationship in the linear-linear space. In mathematical terms, if

$$\ln(rowSD) = k \cdot \ln(rowMean) + c + \varepsilon, \qquad \text{[Equation 2]}$$

where *k*, *c* and ε respectively represent the slope, the intercept and a normally-distributed residual error of a linear regression, then

$$rowSD = rowMean^{k} \cdot \exp(c) \cdot \exp(\varepsilon). \qquad \text{[Equation 3]}$$

And since

$$rowCV = \frac{rowSD}{rowMean}, \qquad \text{[Equation 4]}$$

then

$$rowCV = rowMean^{(k-1)} \cdot \exp(c) \cdot \exp(\varepsilon). \qquad \text{[Equation 5]}$$

According to this model, if $k = 1$, then the *rowSD* would be directly proportional to the *rowMean*, while the *rowCV* would be constant over the whole dynamic range of *rowMean* values. Values of $k > 1$ would cause both the *rowSD* and the *rowCV* to increase as a function of the *rowMean*, while values of $k < 0$ would lead to a decrease of both the *rowSD* and the *rowCV*. Hence, there is a critical range $0 < k < 1$, in which the absolute variability increases with increasing average abundance (because of the positive power coefficient *k* in Equation 3), while the relative variability decreases (because of the negative power coefficient ($k-1$) in Equation 5). An error model with parameter *k* within this critical range, would therefore fully explain the observations made in Figure 2. In addition, such a model would also be consistent with the fact that the dynamic range of *rowSD* values was significantly smaller than the dynamic range of the *rowMean* values in the same dataset (Table 1).

We have previously described the above variance-versus-mean dependence to be at the basis of GeneChip data and we modeled this relationship via a Power Law Global Error Model (PLGEM) (20). Here, we tested whether and how PLGEM would be able to explain the variability present in a typical NSAF dataset as well. Using the Bioconductor package 'plgem' we fitted a PLGEM either to a simulated dataset (forced to obey a PLGEM) or to the GeneChip and the two NSAF datasets under investigation in the current study (Figure 3). Details about the robust PLGEM fitting method implemented in the 'plgem' package can be found in the original publication (20). Briefly, the dynamic range of *rowMean* values is partitioned into equally sized bins and a modeling point is determined in each partition, so that it captures the local median variation (20). Then, a linear regression is performed through the set of modeling points in the log-log space, to obtain the slope *k* and the intercept *c* of the PLGEM. As quality controls, a Pearson's correlation coefficient was calculated between all available ln(*rowSD*) values and the corresponding ln(*rowMean*) values and an adjusted r^2 value was calculated between the fitted PLGEM and the modeling points. In general, PLGEM fitted equally well on all analyzed datasets (Figure 3A-D), with correlation coefficients >0.96 and adjusted r^2 values >0.99. An additional evaluation of the goodness-of-fit of PLGEM was performed through an analysis of the residuals of the model. Residuals were calculated as differences between the modeled and the measured ln(*rowSD*). As expected from a good fit, in all analyzed datasets the residuals were relatively constant across the whole dynamic range (Figure 3E-H) and were approximately normally distributed (Figure 3I-P).

Once established that NSAF datasets could be modeled by a PLGEM similarly to GeneChip datasets, we next asked whether the model parameters obtained by fitting PLGEM on NSAF datasets were similar to the typical parameters observed in GeneChip data. To this end, we took advantage of the Gene Expression Omnibus database, a public repository of microarray experiments (46). We fitted PLGEM on 101 distinct GeneChip datasets downloaded from this database, which represented microarray experiments performed across 5 different species and 7 different platforms, and drew density distribution plots of the PLGEM slopes, the PLGEM intercepts, the correlation coefficients and of the adjusted r^2 values found in these datasets (Figure 4). The PLGEM slopes found in the 101 analyzed GeneChip datasets were all within the range $0.5 < k < 1$, which was well within the critical range described above (Figure 4C). Importantly, correlation coefficients and adjusted r^2 values found both in the Yeast and in the Mediator datasets were among the highest values observed for GeneChip datasets, suggesting that the fitting of PLGEM was particularly good in the analyzed NSAF datasets (Figure 4A-B). Notably, both the Yeast and the Mediator NSAF datasets had PLGEM slopes ~0.8, which was very close to the average PLGEM slope generally found in GeneChip datasets (Figure 4C).

The NSAF datasets analyzed in the present work contained an unusually high number of replicates, which was important for a solid investigation of the statistical properties of these types of datasets. However, in a realistic experimental setting, it would be unlikely to have 8 or 9 replicates. Therefore, if PLGEM was to be proposed as a novel tool in NSAF data analysis, we deemed important to test its behavior also when a significantly smaller number of replicates were available for a given experiment. We therefore simulated the effect of decreasing the number of available replicates by randomly removing 1 or more columns from the datasets analyzed above, until only 3 replicates were retained (Figure 5). As expected, a smaller number of replicates caused a less obvious linearity between the ln(*rowSD*) values and the ln(*rowMean*) values, as demonstrated by the progressive decay of the Pearson's correlation coefficient (Figure 5A), and a consequent decrease of the goodness-of-fit of PLGEM, as exemplified by the drop in the adjusted r^2 of the modeling points (Figure 5B). Nonetheless, even in the datasets with only 3 replicates, all measured correlation coefficients were >0.85 and the r^2 values were >0.96, demonstrating a reasonably good fit. In addition, PLGEM slopes and intercepts deviated only marginally from the parameters obtained from the full dataset (Figure 5C-D). However, there was a large benefit in both accuracy and precision in the determination of all parameters, when the number of available replicates was increased from three to four or, though to a lesser extent, from four to five. A further increase in the number of replicates mainly affected the precision but only marginally affected the accuracy by which PLGEM parameters were estimated (Figure 5C-D). Taken together, these data stress once more the importance of performing as many replicates as possible in these types of experiments. In addition, these results suggest that four or five replicates might represent a reasonable compromise between the cost of a MudPIT experiment and the accuracy and precision with which the underlying PLGEM parameters can be estimated from NSAF values.

Use of PLGEM to detect differentially abundant proteins

The main benefit of an error model relies in its ability to more accurately estimate data variability, compared to measuring it directly from the data alone (18). As a consequence, using model-derived rather than data-derived SD estimates has shown to significantly improve – in both GeneChip (20) and LC-MS proteomics data (36) – the performance of statistical methods designed to detect significant abundance changes between two experimental conditions of interest. We therefore asked whether PLGEM could improve the identification of differentially abundant proteins also in NSAF datasets.

In order to test the added value provided by the use of PLGEM in the analysis of NSAF-based proteomics datasets, we performed a MudPIT experiment designed to detect proteins that show differential abundance in different yeast growth phases. Whole-cell extracts from four biological replicates of a yeast cell culture grown in rich medium and harvested either in log-phase (LP) or in stationary phase (SP) were analyzed by a total of eight independent MudPIT runs and quantified using the NSAF approach, to search for proteins up- or down-regulated during the growth phase shift (Supplementary Table 3). A total of 783 proteins were consistently identified in at least 3 out of 4 replicates in either the LP or the SP samples. Out of these, 108 were identified only in the SP samples and 164 only in the LP samples. These two subsets respectively represent proteins induced or repressed in different growth phases and are consistent with prior knowledge on the biology of stationary phase in yeast (data not shown, (50)). Although these proteins provide insights into the global changes occurring in response to this physiological transition, they represent only a minor fraction of the total identified proteins. In addition, their behavior can be modeled as an ON/OFF response and are therefore less challenging to detect. The identification of differential abundance among the remaining majority of proteins (511/783, i.e. ~65%), which were consistently identified in most of the samples, represents instead a much more challenging task. It is in this type of analysis that a model-based statistical analysis might prove its benefits.

A standard procedure in quantitative proteomics data analysis makes use of the "fold change" (FC) as a measure of differential abundance of proteins across two groups of replicated samples. It is implicitly assumed that the higher the FC, the more the protein abundance level varies between the two experimental conditions of interest. A more rigorous procedure would take the within-group variability into account as well, in order to tell whether the signal we are interested in (the difference in abundance of the protein) is higher than the noise (the background variability caused by a combination of biological and technical variation). In such an analysis it becomes important to obtain accurate estimates of the standard deviation of NSAF measurements across different replicates of a same experimental condition, in order not to over- or under-estimate the background noise and thus under- or over-estimate the signal-to-noise (STN) ratio. We therefore tested the performance of PLGEM in providing more accurate estimates of standard deviation, by incorporating PLGEM-derived standard deviation into the following STN statistic:

$$rowSTN = \frac{rowMean_{SP} - rowMean_{LP}}{rowSD_{SP} + rowSD_{LP}}. \qquad \text{[Equation 6]}$$

Since they were independently analyzed, two distinct sets of PLGEM parameters were fit to the SP NSAF dataset and the LP NSAF dataset (Supplemental Figure 1). It has to be noted that although the above statistic has successfully proven to provide excellent results in the analysis of GeneChip data (9, 20, 34, 35, 45), it has not yet been used for the analysis of NSAF-based proteomics data.

We first compared the results obtained by analyzing the above-mentioned 511 yeast proteins either with the simple FC method or with the STN statistic incorporating classical data-derived estimates of standard deviation (Standard-STN). The FC statistic was implemented here as the log ratio of the average NSAF value in the SP samples over the average NSAF value in the LP samples. The 511 proteins were ranked based on the absolute value of either of the two statistics and the top 100 with the most extreme values selected as the most significantly changing (Supplementary Table 5). Whereas the FC method was biased towards detection of the most lowly abundant proteins, because these are the ones expected to vary most, the Standard-STN method selected several proteins with very little fold changes and missed other proteins with very high fold changes (Figure 6). Among the proteins with low FC values that were nonetheless selected by the Standard-STN method, some were identified with extremely small spectral counts like the transcriptional elongation protein Spt6, identified by 0, 1, 2 and 3 spectra in the four LP replicates and by 2, 3, 3 and 3 spectra in the SP samples (Figure 6). Proteins with very small spectral counts have been ranked among the 100 most differentially abundant ones by Standard-STN only because they happened to have reproducibly small NSAF values, but due to the variability of such low spectral counts they should likely be regarded as false positives. Among the proteins with large changes that were not ranked among the most significant ones using the Standard-STN method, many are well known to be down-regulated during a shift from LP to SP in yeast and should therefore be regarded as false negatives. An example of such a protein is the ribosomal protein Rpl8a, identified by 5, 11, 20 and 76 spectra in the LP samples and by 0, 2, 3 and 6 spectra in the SP samples (Figure 6). The most likely cause, for which these proteins were missed by the Standard-STN method, was their relatively high standard deviations.

We next analyzed the same dataset using STN ratios incorporating PLGEM-based estimates of standard deviation (PLGEM-STN). In contrast to the FC and Standard-STN methods, the PLGEM-STN statistic was more stringent in calling a significant hit among proteins with low average NSAF value and was less stringent for proteins with high abundance values (Figure 6). As a consequence, none of the proteins ranked by PLGEM-STN among the 100 most significantly changing proteins had reproducibly different but very low total spectral counts in both conditions, like Spt6p, which was instead selected by the Standard-STN method (Figure 6). On the other hand, none of the proteins that showed a large negative FC during the shift from LP to SP, like Rpl8a, were missed by PLGEM-STN, although many of them were missed by Standard-STN (Figure 6). These results demonstrate that incorporation of PLGEM into an STN-based ranking analysis of a NSAF-processed MudPIT dataset naturally selects for proteins with changes in abundance between samples that intuitively makes more sense compared to the use of FC or Standard-STN.

Ranking of proteomics hits based on some significance criterion is a common procedure to prioritize the follow-up of candidate proteins potentially involved in the biological phenomenon under investigation. We therefore tested the biological significance of the proteins identified with the FC, the Standard-STN or the PLGEM-STN method. To this end, significant enrichment of Gene Ontology (GO) annotation terms or Swissprot keywords among the top-ranking 100 proteins was evaluated. We submitted the three different lists of 100 Refseq IDs corresponding to the proteins selected by each method to the FatiGO+ website (http://babelomics.bioinfo.cipf.es/fatigoplus/cgi-bin/fatigoplus.cgi) (51), to test whether any functional annotation terms were significantly over-represented in the query list in comparison to the background list of 411 non-selected proteins. This website provides p-values from a Fisher's exact test, adjusted for multiple testing by an FDR-based method. Whereas no statistically significant hits were returned for the 100 proteins with the highest FC values or the highest Standard-STN values, FatiGO+ detected a significant enrichment of GO Biological Process annotation terms 'biosynthetic process' (FDR-adjusted p-value = 2.3×10^{-3}), 'cellular biosynthetic process' (FDR-adjusted p-value = 2.1×10^{-3}), 'macromolecule biosynthetic process' (FDR-adjusted p-value = 3.7×10^{-4}) and 'translation' (FDR-adjusted p-value = 5.7×10^{-4}) and for the SwissProt keyword 'ribosomal protein' (FDR-adjusted p-value = 2.6×10^{-6}) among the 100 proteins with the highest PLGEM-STN values. It has to be noted that from a biological perspective the shift from

LP to SP is well known in yeast to be accompanied by a progressive slow down of the whole biosynthetic machinery and especially of translation (52), and only by using PLGEM in this analysis did we capture this information.

Discussion

The major findings of the present study can be summarized in the following way: i) From a statistical point of view, NSAF datasets are more similar to GeneChip data than previously anticipated; ii) The variability of NSAF values can be accurately modeled by a PLGEM; iii) PLGEM-based methods can be used to identify differentially abundant proteins in NSAF datasets. The most important implications of these results are discussed below.

Similarities between NSAF and GeneChip data

Here, we have provided evidence that NSAF datasets share with GeneChip data substantial statistical similarities. Not only the dynamic range and the distribution of values were qualitatively very similar between the two technologies, but also – and perhaps more importantly – these two types of data have proven to obey the same global error model with surprisingly similar parameters (see next section for a more detailed discussion of the latter point). These similarities offer the exciting opportunity to take advantage of the multitude of statistical tools that have been designed to specifically deal with open issues in microarray data analysis and to test whether they perform as well in proteomics data analysis. There is for instance a wealth of literature, algorithms and software that has been devoted to solve microarray data analysis problems related to missing values (53-55), multiple testing (56, 57), variance-versus-mean dependence (20, 29, 30), etc. We foresee that most of these issues will be recapitulated also in shotgun proteomics data. Therefore, if these microarray-specific tools were directly applicable to the analysis of proteomics data, this would represent a significant shortcut in the advancement of proteomics research. Other authors already successfully applied specific microarray tools in the analysis of proteomics data (25), without providing a more general demonstration of the underlying assumption that proteomics data are substantially similar to transcriptomics data. The substantial similarities shown here between NSAF data and GeneChip data, suggest instead that most GeneChip-specific statistical tools should be applicable to the analysis of NSAF datasets as well.

PLGEM as an error model for shotgun proteomics

The most important similarity between NSAF and GeneChip datasets was that, not only both types of datasets obeyed a PLGEM, but the most critical parameter of the model, i.e. the power coefficient *k*, was surprisingly conserved. The fact that this parameter was always inside the critical range $0 < k < 1$ for more than 100 distinct GeneChip datasets from 5 distinct species as well as for four NSAF datasets, three from yeast and one from human samples, indicates that this global error model might really be a general model of GeneChip and NSAF data, regardless of the specific nature of the analyzed samples. The main consequence of such a model with such constraints would be that transcript or protein abundance levels of more highly expressed genes would be intrinsically more stable than those of more lowly expressed genes. This observation raises the question about the reason of this skew. A possible explanation for this is that cells might have skewed their gene expression control system, by concentrating their efforts in more precisely controlling the expression of genes with a potentially higher impact on cellular functions, rather than dissipating energy in controlling the expression of genes the products of which would be expressed at low levels anyway. What argues against this explanation is that it assumes a direct relationship between the expression level of a gene and the biological impact of the encoded protein, which might not always be the case. Investigation of the real reason behind this peculiar phenomenon goes well beyond the scope of the present work.

It has to be noted that there is nothing radically distinct about PLGEM as compared to previously proposed error models for these types of measurements. In fact, PLGEM could be seen as a generalization of these models. For instance, two-component error models have been previously proposed for atomic absorption spectroscopy (58), GC-MS (58), LC-MS (36) or microarray data (29). These models assume a constant *rowSD* for very low abundances and a constant *rowCV* for higher abundances. A constant *rowCV* model would in fact be able to explain an increase of the *rowSD* as the function of the *rowMean*, but would not account for the progressive decay of *rowCV* that we have observed in both GeneChip data (20) and NSAF data (Figure 2) for increasingly higher values of the *rowMean*. PLGEM, conversely, by not assuming any particular value of the power coefficient k, relies on more relaxed assumptions. Notably, a PLGEM with $k \approx 1$ would result in an approximately constant *rowCV* model. Thus, a PLGEM with $k \approx 1$, would be difficult to distinguish from a 'constant-CV' model, especially if the analyzed dynamic range was not sufficiently large. The wide dynamic range of abundance levels that can be measured with NSAF and the GeneChip technology, instead, allows a clear distinction between these two models. The fact that we have observed here the power coefficient k to be in the range 0.7-0.8 for most analyzed GeneChip and NSAF datasets (Figure 4), might therefore explain why in the past the 'constant-CV' assumption has been often taken for granted.

Identification of differentially abundant proteins

The unbiased sampling nature of shotgun proteomics approaches theoretically allows to detect virtually any protein in a sample regardless of its concentration, provided that the experiment is replicated a sufficiently large number of times (23). However, these extremely lowly abundant proteins pose numerous challenges in their statistical analysis, because of the presence of several zero values and the intrinsically low reproducibility described above. In order to increase the confidence of downstream statistical analyses, it is therefore common practice to discard proteins identified only in a minority of the analyzed replicates of a MudPIT experiment or transcripts flagged as 'Absent' in the majority of replicates in a GeneChip experiment. But in a comparative analysis, where significant differences between two experimental conditions are sought after, a transcript or a protein that passed the above criteria in one experimental condition but was virtually absent in the other condition, would represent a valuable candidate for follow-up studies. Statistical methods able to deal with these lowly abundant transcripts or proteins and to detect a significant difference between a virtual absence and a modest presence will certainly expand the coverage by which we can interpret the outcomes of these experiments.

We have shown here that PLGEM fits equally well over the whole dynamic range of average NSAF values, even to proteins identified in a minor fraction of all available replicates, i.e. 3/8 in case of the Yeast dataset and 3/9 in case of the Mediator dataset. In addition, we have observed that PLGEM fitted equally well also on NSAF datasets where ~50% of the proteins were identified in only one or two replicates (data not shown). This suggests that PLGEM has the potential to improve our ability to cope with these lowly abundant proteins, because it provides a reasonable estimate of the expected standard deviation in spite of the presence of only a small number of non-zero NSAF values.

The performance of a PLGEM-based method for the analysis of GeneChip experiments has already been thoroughly investigated and compared to the behavior of other commonly used statistical methods (20). In the present work, we have shown that the use of PLGEM-based standard deviations to calculate STN ratios in an NSAF dataset improves our ability to determine protein expression changes between yeast sampled at LP and SP (Figure 6 and Supplementary Table 3). While determining which proteins were present in one growth condition and absent in another is relatively straightforward, determining changes in abundance of proteins found in both LP and SP is challenging. The PLGEM-STN statistic outperformed both FC and Standard-STN by being more conservative with proteins of low abundance than proteins with high abundance. In conclusion, we envision a broad range of applications of PLGEM in the analysis of NSAF data. PLGEM might assist in prioritizing the follow-up analysis of candidate proteins that show significant abundance changes between any two samples of interest, i.e. in the comparison of a wild-type vs. a knock-out cell line, a diseased vs. a normal tissue, or a treated vs. an untreated patient.

References

1. Ideker, T., Thorsson, V., Ranish, J. A., Christmas, R., Buhler, J., Eng, J. K., Bumgarner, R., Goodlett, D. R., Aebersold, R., and Hood, L. (2001) Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science* 292, 929-934.

2. Kitano, H. (2002) Systems biology: a brief overview. *Science* 295, 1662-1664.

3. Albeck, J. G., MacBeath, G., White, F. M., Sorger, P. K., Lauffenburger, D. A., and Gaudet, S. (2006) Collecting and organizing systematic sets of protein data. *Nat Rev Mol Cell Biol* 7, 803-812.

4. Lipshutz, R. J., Fodor, S. P., Gingeras, T. R., and Lockhart, D. J. (1999) High density synthetic oligonucleotide arrays. *Nat Genet* 21, 20-24.

5. David, L., Huber, W., Granovskaia, M., Toedling, J., Palm, C. J., Bofkin, L., Jones, T., Davis, R. W., and Steinmetz, L. M. (2006) A high-resolution map of transcription in the yeast genome. *Proc Natl Acad Sci U S A* 103, 5320-5325.

6. Washburn, M. P., Wolters, D., and Yates, J. R., 3rd (2001) Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat Biotechnol* 19, 242-247.

7. Aebersold, R., and Mann, M. (2003) Mass spectrometry-based proteomics. *Nature* 422, 198-207.

8. Csete, M. E., and Doyle, J. C. (2002) Reverse engineering of biological complexity. *Science* 295, 1664-1669.

9. Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D., and Lander, E. S.

(1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 286, 531-537.

10. Hughes, T. R., Marton, M. J., Jones, A. R., Roberts, C. J., Stoughton, R., Armour, C. D., Bennett, H. A., Coffey, E., Dai, H., He, Y. D., Kidd, M. J., King, A. M., Meyer, M. R., Slade, D., Lum, P. Y., Stepaniants, S. B., Shoemaker, D. D., Gachotte, D., Chakraburtty, K., Simon, J., Bard, M., and Friend, S. H. (2000) Functional discovery via a compendium of expression profiles. *Cell* 102, 109-126.

11. Granucci, F., Vizzardelli, C., Pavelka, N., Feau, S., Persico, M., Virzi, E., Rescigno, M., Moro, G., and Ricciardi-Castagnoli, P. (2001) Inducible IL-2 production by dendritic cells revealed by global gene expression analysis. *Nat Immunol* 2, 882-888.

12. Florens, L., Washburn, M. P., Raine, J. D., Anthony, R. M., Grainger, M., Haynes, J. D., Moch, J. K., Muster, N., Sacci, J. B., Tabb, D. L., Witney, A. A., Wolters, D., Wu, Y., Gardner, M. J., Holder, A. A., Sinden, R. E., Yates, J. R., and Carucci, D. J. (2002) A proteomic view of the Plasmodium falciparum life cycle. *Nature* 419, 520-526.

13. Listgarten, J., and Emili, A. (2005) Statistical and computational methods for comparative proteomic profiling using liquid chromatography-tandem mass spectrometry. *Mol Cell Proteomics* 4, 419-434.

14. Wang, E. (2005) RNA amplification for successful gene profiling analysis. *J Transl Med* 3, 28.

15. Lonnstedt, I., and Speed, T. (2002) Replicated microarray data. *Statist Sinica* 12, 31-46.

16. Huang, X., and Pan, W. (2002) Comparing three methods for variance estimation with duplicated high density oligonucleotide arrays. *Funct Integr Genomics* 2, 126-133.

17. Ideker, T., Thorsson, V., Siegel, A. F., and Hood, L. E. (2000) Testing for differentially-expressed genes by maximum-likelihood analysis of microarray data. *J Comput Biol* 7, 805-817.

18. Baggerly, K. A., Coombes, K. R., Hess, K. R., Stivers, D. N., Abruzzo, L. V., and Zhang, W. (2001) Identifying differentially expressed genes in cDNA microarray experiments. *J Comput Biol* 8, 639-659.

19. Tusher, V. G., Tibshirani, R., and Chu, G. (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci U S A* 98, 5116-5121.

20. Pavelka, N., Pelizzola, M., Vizzardelli, C., Capozzoli, M., Splendiani, A., Granucci, F., and Ricciardi-Castagnoli, P. (2004) A power law global error model for the identification of differentially expressed genes in microarray data. *BMC Bioinformatics* 5, 203.

21. Gygi, S. P., Rist, B., Gerber, S. A., Turecek, F., Gelb, M. H., and Aebersold, R. (1999) Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat Biotechnol* 17, 994-999.

22. Tao, W. A., and Aebersold, R. (2003) Advances in quantitative proteomics via stable isotope tagging and mass spectrometry. *Curr Opin Biotechnol* 14, 110-118.

23. Liu, H., Sadygov, R. G., and Yates, J. R., 3rd (2004) A model for random sampling and estimation of relative protein abundance in shotgun proteomics. *Anal Chem* 76, 4193-4201.

24. Old, W. M., Meyer-Arendt, K., Aveline-Wolf, L., Pierce, K. G., Mendoza, A., Sevinsky, J. R., Resing, K. A., and Ahn, N. G. (2005) Comparison of label-free methods for quantifying human proteins by shotgun proteomics. *Mol Cell Proteomics* 4, 1487-1502.

25. Zhang, B., Verberkmoes, N. C., Langston, M. A., Uberbacher, E., Hettich, R. L., and Samatova, N. F. (2006) Detecting Differential and Correlated Protein Expression in Label-Free Shotgun Proteomics. *J Proteome Res* 5, 2909-2918.

26. Paoletti, A. C., Parmely, T. J., Tomomori-Sato, C., Sato, S., Zhu, D., Conaway, R. C., Conaway, J. W., Florens, L., and Washburn, M. P. (2006) Quantitative proteomic analysis of distinct mammalian Mediator complexes using normalized spectral abundance factors. *Proc Natl Acad Sci U S A* 103, 18928-18933.

27. Zybailov, B., Mosley, A. L., Sardiu, M. E., Coleman, M. K., Florens, L., and Washburn, M. P. (2006) Statistical analysis of membrane proteome expression changes in Saccharomyces cerevisiae. *J Proteome Res* 5, 2339-2347.

28. Chen, Y., Dougherty, E. R., and Bittner, M. L. (1997) Ratio-based decisions and the quantitative analysis of cDNA microarray images. *J Biomed Opt* 2, 364-374.

29. Rocke, D. M., and Durbin, B. (2001) A model for measurement error for gene expression arrays. *J Comput Biol* 8, 557-569.

30. Huber, W., von Heydebreck, A., Sultmann, H., Poustka, A., and Vingron, M. (2002) Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics* 18 Suppl 1, S96-104.

31. Wright, G. W., and Simon, R. M. (2003) A random variance model for detection of differential gene expression in small microarray experiments. *Bioinformatics* 19, 2448-2455.

32. Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., Hornik, K., Hothorn, T., Huber, W., Iacus, S., Irizarry, R., Leisch, F., Li, C., Maechler, M., Rossini, A. J., Sawitzki, G., Smith, C., Smyth, G., Tierney, L., Yang, J. Y., and Zhang, J. (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol* 5, R80.

33. Pelizzola, M., Pavelka, N., Foti, M., and Ricciardi-Castagnoli, P. (2006) AMDA: an R package for the automated microarray data analysis. *BMC Bioinformatics* 7, 335.

34. Vizzardelli, C., Pavelka, N., Luchini, A., Zanoni, I., Bendickson, L., Pelizzola, M., Beretta, O., Foti, M., Granucci, F., Nilsen-Hamilton, M., and Ricciardi-Castagnoli, P. (2006) Effects of dexamethazone on LPS-induced activation and migration of mouse dendritic cells revealed by a genome-wide transcriptional analysis. *Eur J Immunol* 36, 1504-1515.

35. Iovino, F., Lentini, L., Amato, A., and Di Leonardo, A. (2006) RB acute loss induces centrosome amplification and aneuploidy in murine primary fibroblasts. *Mol Cancer* 5, 38.

36. Anderle, M., Roy, S., Lin, H., Becker, C., and Joho, K. (2004) Quantifying reproducibility for differential proteomics: noise analysis for protein liquid chromatography-mass spectrometry of human serum. *Bioinformatics* 20, 3575-3582.

37. Brachmann, C. B., Davies, A., Cost, G. J., Caputo, E., Li, J., Hieter, P., and Boeke, J. D. (1998) Designer deletion strains derived from Saccharomyces cerevisiae S288C: a useful set of strains and plasmids for PCR-mediated gene disruption and other applications. *Yeast* 14, 115-132.

38. Sato, S., Tomomori-Sato, C., Parmely, T. J., Florens, L., Zybailov, B., Swanson, S. K., Banks, C. A., Jin, J., Cai, Y., Washburn, M. P., Conaway, J. W., and Conaway, R. C. (2004) A set of consensus mammalian mediator subunits identified by multidimensional protein identification technology. *Mol Cell* 14, 685-691.

39. McDonald, W. H., Ohi, R., Miyamoto, D. T., Mitchison, T. J., and Yates III, J. R. (2002) Comparison of three directly coupled HPLC MS/MS strategies for identification of proteins from complex mixtures: Single-dimension LC-MS/MS, 2-phase MudPIT, and 3-phase MudPIT. *Int J Mass Spectrom* 219, 245-251.

40. Florens, L., and Washburn, M. P. (2006) Proteomic analysis by multidimensional protein identification technology. *Methods Mol Biol* 328, 159-175.

41. Eng, J. K., McCormack, A. L., and Yates, J. R. (1994) An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J Am Soc Mass Spectrom* 5, 976-989.

42. Tabb, D. L., McDonald, W. H., and Yates, J. R., 3rd (2002) DTASelect and Contrast: tools for assembling and comparing protein identifications from shotgun proteomics. *J Proteome Res* 1, 21-26.

43. Florens, L., Carozza, M. J., Swanson, S. K., Fournier, M., Coleman, M. K., Workman, J. L., and Washburn, M. P. (2006) Analyzing chromatin remodeling complexes using shotgun proteomics and normalized spectral abundance factors. *Methods* 40, 303-311.

44. Elias, J. E., Haas, W., Faherty, B. K., and Gygi, S. P. (2005) Comparative evaluation of mass spectrometry platforms used in large-scale proteomics investigations. *Nat Methods* 2, 667-675.

45. Trottein, F., Pavelka, N., Vizzardelli, C., Angeli, V., Zouain, C. S., Pelizzola, M., Capozzoli, M., Urbano, M., Capron, M., Belardelli, F., Granucci, F., and Ricciardi-Castagnoli, P. (2004) A type I IFN-dependent pathway induced by Schistosoma mansoni eggs in mouse myeloid dendritic cells generates an inflammatory signature. *J Immunol* 172, 3011-3017.

46. Edgar, R., Domrachev, M., and Lash, A. E. (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res* 30, 207-210.

47. Ihaka, R., and Gentleman, R. (1996) R: A Language for Data Analysis and Graphics. *J Comp Graph Stats* 5, 299-314.

48. Baldi, P., and Long, A. D. (2001) A Bayesian framework for the analysis of microarray expression data: regularized t -test and statistical inferences of gene changes. *Bioinformatics* 17, 509-519.

49. Hoyle, D. C., Rattray, M., Jupp, R., and Brass, A. (2002) Making sense of microarray data distributions. *Bioinformatics* 18, 576-584.

50. Werner-Washburne, M., Braun, E., Johnston, G. C., and Singer, R. A. (1993) Stationary phase in the yeast Saccharomyces cerevisiae. *Microbio Rev* 57, 383-401.

51. Al-Shahrour, F., Minguez, P., Tarraga, J., Montaner, D., Alloza, E., Vaquerizas, J. M., Conde, L., Blaschke, C., Vera, J., and Dopazo, J. (2006) BABELOMICS: a systems biology perspective in the functional annotation of genome-scale experiments. *Nucleic Acids Res* 34, W472-476.

52. Fuge, E. K., Braun, E. L., and Werner-Washburne, M. (1994) Protein synthesis in long-term stationary-phase cultures of Saccharomyces cerevisiae. *J Bacteriol* 176, 5802-5813.

53. Sehgal, M. S., Gondal, I., and Dooley, L. S. (2005) Collateral missing value imputation: a new robust missing value estimation algorithm for microarray data. *Bioinformatics* 21, 2417-2423.

54. Kim, H., Golub, G. H., and Park, H. (2005) Missing value estimation for DNA microarray gene expression data: local least squares imputation. *Bioinformatics* 21, 187-198.

55. Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., and Altman, R. B. (2001) Missing value estimation methods for DNA microarrays. *Bioinformatics* 17, 520-525.

56. Dudoit, S., Yang, Y. H., Callow, M. J., and Speed, T. P. (2002) Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statist Sinica* 12, 111-139.

57. Storey, J. D., and Tibshirani, R. (2003) Statistical significance for genomewide studies. *Proc Natl Acad Sci U S A* 100, 9440-9445.

58. Rocke, D. M., and Lorenzato, S. (1995) A Two-Component Model for Measurement Error in Analytical Chemistry. *Technometrics* 37, 176-184.

Footnotes

The authors would like to acknowledge Ron and Joan Conaway and members of the Conaway laboratory for providing the Mediator samples. Michael Coleman and members of the Information Technology department at the Stowers Institute for Medical Research are to be acknowledged for providing and maintaining computational resources. Norman Pavelka is especially thankful to Rong Li for support. This work was supported by the Stowers Institute for Medical Research.

Figure Legends

Figure 1. NSAF and GeneChip data have similar distribution properties.

(A-C) The *rowMean* value was calculated for each transcript or protein in the indicated dataset and subsequently transformed to its base-10 logarithm. The black line in each plot represents the density distribution of all $\log_{10}$(*rowMean*) values in the corresponding dataset. The blue lines in each plot represent the density distribution of $\log_{10}$(*rowMean*) values of transcripts or proteins that were detected with a specific number of 'Absent calls' (in case of the GeneChip dataset) or zero values (in case of the NSAF datasets). The color intensity of the blue lines was chosen from a gradual color palette to reflect the actual number of 'Absent' or zero values, according the color bar depicted at the bottom of the figure. (D-F) The *rowSD* value for each transcript or protein was measured across all available replicates in the indicated dataset and subsequently transformed to its base-10 logarithm. The density distributions of the $\log_{10}$(*rowSD*) values were plotted according the same color-coding scheme described for the upper panels. (A) and (D) represent mouse GeneChip data, (B) and (E) represent yeast NSAF data, and (C) and (F) represent Mediator NSAF data.

Figure 2. NSAF and GeneChip datasets have a similar variance-versus-mean dependence.

(A-C) The *rowMean* and the *rowSD* of the abundance values for each transcript or protein were measured across all available replicates in the indicated dataset and subsequently transformed to their corresponding base-10 logarithms. Scatter-plots of $\log_{10}$(*rowSD*) vs. $\log_{10}$(*rowMean*) were color-coded according to the same scheme described in the legend of Figure 1. (D-F) The *rowCV* of each transcript or protein was measured as the ratio between the *rowSD* and the *rowMean* in the indicated dataset, and subsequently transformed to its corresponding base-10 logarithm. Scatter-plots of $\log_{10}(rowCV)$ vs. $\log_{10}(rowMean)$ were color-coded according to the same scheme described in the legend of Figure 1. Note that a linear relationship in a log-log space is mathematically equivalent to a power law relationship in a linear-linear space. (A) and (D) represent mouse GeneChip data, (B) and (E) represent yeast NSAF data, and (C) and (F) represent Mediator NSAF data.

Figure 3. PLGEM fits equally well on NSAF and GeneChip datasets.

(A-D) Contour-plots of $\ln(rowCV)$ vs. $\ln(rowMean)$ scatter-plots of the indicated datasets were drawn, to visualize regions with a higher (orange contours), a medium (green contours) or a lower density of points (light blue contours). The modeling points used to fit a PLGEM were superimposed on the corresponding contour-plots as black circles. Red lines represent the PLGEM fitted to the indicated dataset. (E-H) For each transcript or protein in the indicated dataset a residual was calculated as the difference between the measured $\ln(rowSD)$ value and the $\ln(rowSD)$ value predicted by PLGEM. Residuals were then plotted as a function of the rank of the *rowMean* value and visualized as contour-plots, following the same color-code described for the upper panels. (I-L) The distribution of the residuals in the indicated dataset was plotted as a histogram of counts in equally-sized bins. (M-P) The similarity between the distribution of residuals and a standard Normal distribution was visualized as a quantile-quantile (Q-Q) plot. (A), (E), (I), and (M) represent a simulated dataset, (B), (F), (J), and (N) represent mouse GeneChip data, (C), (G), (K), and (O) represent yeast NSAF data, and (D), (H), (L), and (P) represent Mediator NSAF data. The simulated dataset contained 10 columns and 1000 rows. The 1000 *rowMean* values of the simulated dataset were randomly drawn from a log-Normal distribution with $\ln(\mu) = 0$ and $\ln(\sigma) = 0.25$. The *rowSD* values of each row were then forced to obey a PLGEM with $k = 0.75$, $c = -1$ and ε randomly drawn from a Normal distribution with $\mu =$

0 and σ = 0.25. The 10 values in each row were finally randomly generated from a Normal distribution with $\mu = rowMean$ and $\sigma = rowSD$.

Figure 4. NSAF and GeneChip datasets have similar PLGEM parameters.

PLGEM was fitted on 101 publicly available GeneChip datasets, four relevant fitting parameters were recorded and density distributions were plotted for each of these parameters. (A) Pearson's correlation coefficients were calculated between all available ln(*rowSD*) values and the corresponding ln(*rowMean*) values. (B) Adjusted r^2 values were calculated between the fitted PLGEM and the modeling points. Also shown are slopes (C) and intercepts (D) of the fitted model. Superimposed on the density plots are the actual values of the same four parameters, as obtained from the Mouse GeneChip (blue circles), the Yeast NSAF (red squares) and the Mediator NSAF datasets (green diamonds).

Figure 5. PLGEM parameters are reasonably stable to decreasing number of replicates.

A series of simulations was performed to test the effect of randomly removing one or more replicates from the indicated dataset. A total of 100 random deletions were performed for each indicated number of retained replicates (x-axis label). Matrix rows associated only with zero values after the column removal step were discarded before fitting a PLGEM. For each generated dataset a Pearson's correlation coefficient (A), an adjusted r^2 value (B), a PLGEM slope (C) and a PLGEM intercept (D) were recorded. Circles and error bars respectively represent means and standard deviations of the indicated PLGEM parameters obtained from the corresponding 100 simulated datasets.

Figure 6. Identification of differentially abundant proteins in the Yeast Growth Phase NSAF dataset. The 511 proteins consistently identified in both the log-phase (LP) and the stationary

phase (SP) samples in at least 3 out of 4 biological replicates of the Yeast Growth Phase NSAF dataset (grey dots) were plotted in the space defined by the base-2 logarithm of the ratio of the average NSAF value of the protein in the SP samples over the average NSAF value in the LP samples (y-axis) and the base-10 logarithm of the average NSAF value of the protein in the LP samples (x-axis). Highlighted in the same plot are the 100 proteins with the most extreme fold changes (small red circles), with the 100 most extreme STN ratios (medium-sized goldenrod circles) or with the 100 most extreme PLGEM-STN ratios (large blue circles). The red dashed lines delineate the boundaries separating the 100 proteins with the largest fold changes from the other 411 proteins, while the blue dashed lines separate the 100 proteins with the highest PLGEM-STN ratios from the remaining 411 proteins. The data points for Spt6 and Rpl8a are highlighted and described in the Results section.

Tables

Table 1. Basic descriptive statistics of the datasets analyzed in the present study.

Descriptive statistic	Mouse GeneChip	Yeast NSAF	Mediator NSAF
Number of rows	12488	845	295
Number of columns	11	8	9
Total number of data-points	137368	6760	2655
Zero values (%)	0.02	29.1	34.61
Absent calls (%)	50.13	NA	NA
Dynamic range of values (OOM)	4.68	3.82	3.6
Dynamic range of *rowMean* values (OOM)	4.16	3.69	3.32
Dynamic range of *rowSD* values (OOM)	3.72	3.34	3.07

A summary of basic statistical properties is reported for the GeneChip and the NSAF datasets analyzed in the present study. In NSAF datasets, zero values were introduced in place of missing values. In GeneChip datasets, 'Absent calls' are reported by the microarray scanning software for those transcripts that are considered as not reliably detected. Dynamic ranges were calculated as the base-10 logarithm of the ratio between the 99.95-th percentile and the 0.05-th percentile, after removing the zero values. NA = not applicable. OOM = orders of magnitude.

Figures

Figure 1

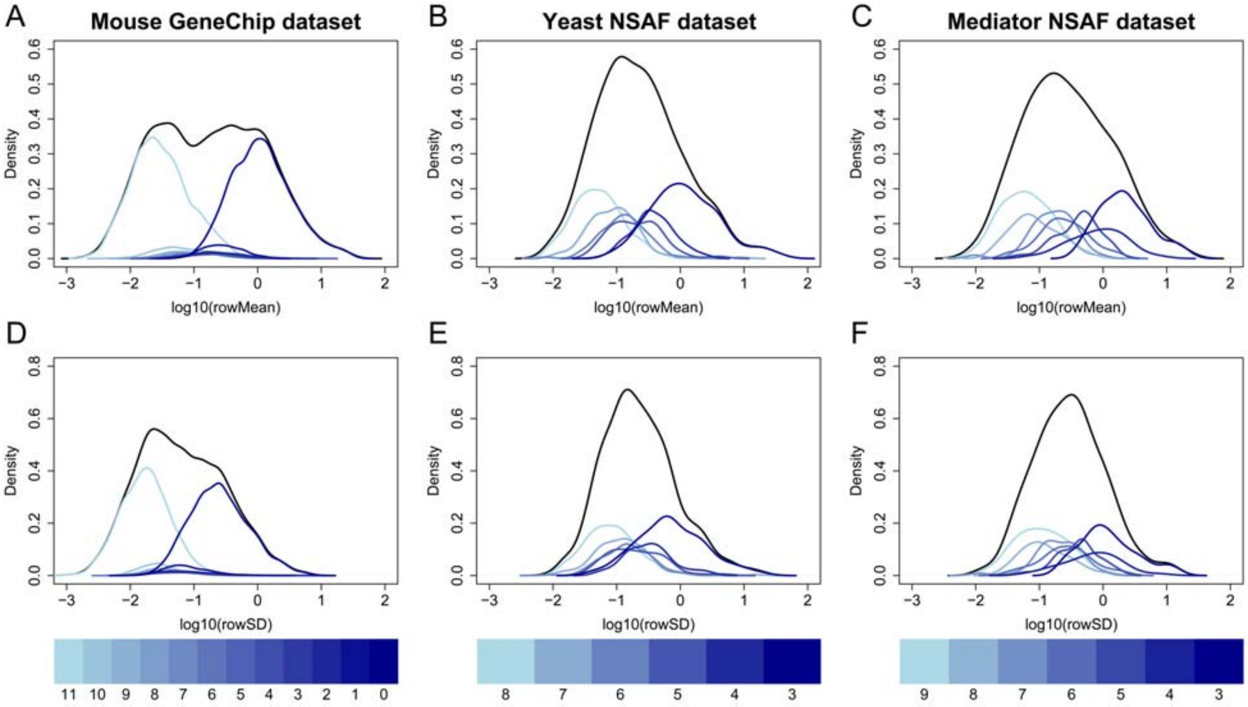


Figure 2

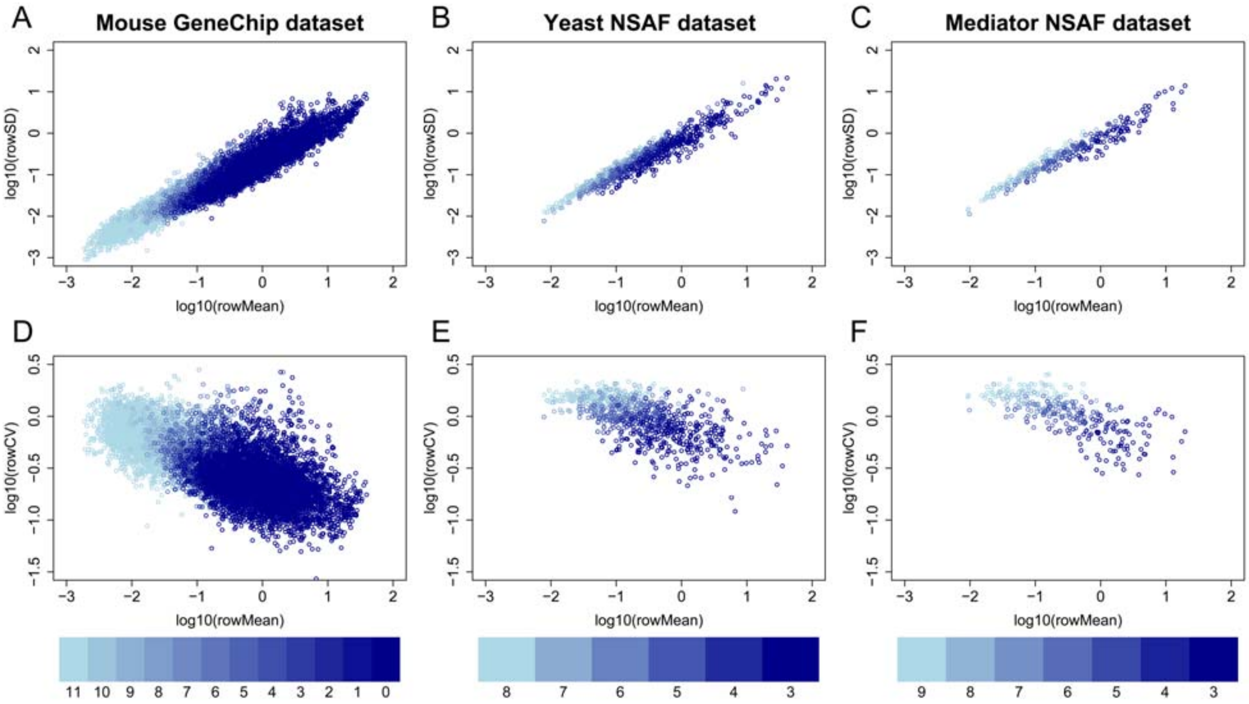

A
Mouse GeneChip dataset
B
Yeast NSAF dataset
C
Mediator NSAF dataset
log10(rowSD)
log10(rowMean)
D
E
F
log10(rowCV)
log10(rowMean)
11 10 9 8 7 6 5 4 3 2 1 0
8 7 6 5 4 3
9 8 7 6 5 4 3

Figure 3

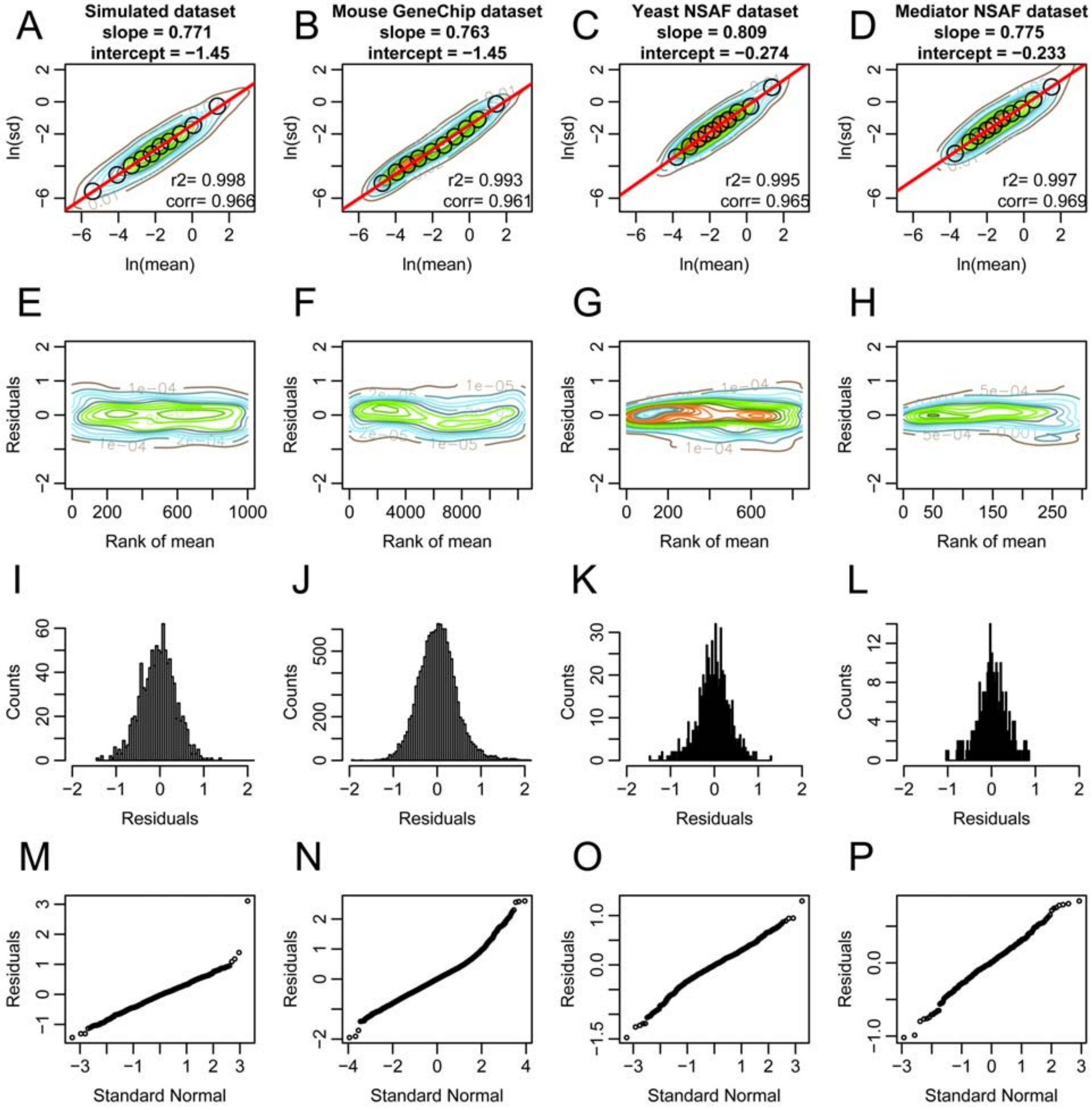

A
Simulated dataset
slope = 0.771
intercept = −1.45
r2= 0.998
corr= 0.966
B
Mouse GeneChip dataset
slope = 0.763
intercept = −1.45
r2= 0.993
corr= 0.961
C
Yeast NSAF dataset
slope = 0.809
intercept = −0.274
r2= 0.995
corr= 0.965
D
Mediator NSAF dataset
slope = 0.775
intercept = −0.233
r2= 0.997
corr= 0.969
ln(sd)
ln(mean)
E
F
G
H
Residuals
Rank of mean
I
J
K
L
Counts
Residuals
M
N
O
P
Residuals
Standard Normal

Figure 4

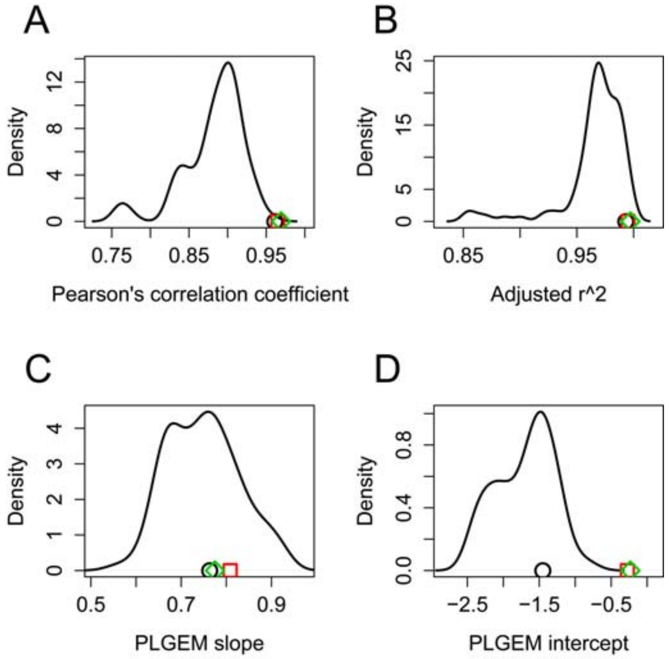

A
Density
0
4
8
12
0.75
0.85
0.95
Pearson's correlation coefficient
B
Density
0
5
15
25
0.85
0.95
Adjusted r^2
C
Density
0
1
2
3
4
0.5
0.7
0.9
PLGEM slope
D
Density
0.0
0.4
0.8
−2.5
−1.5
−0.5
PLGEM intercept

Figure 5

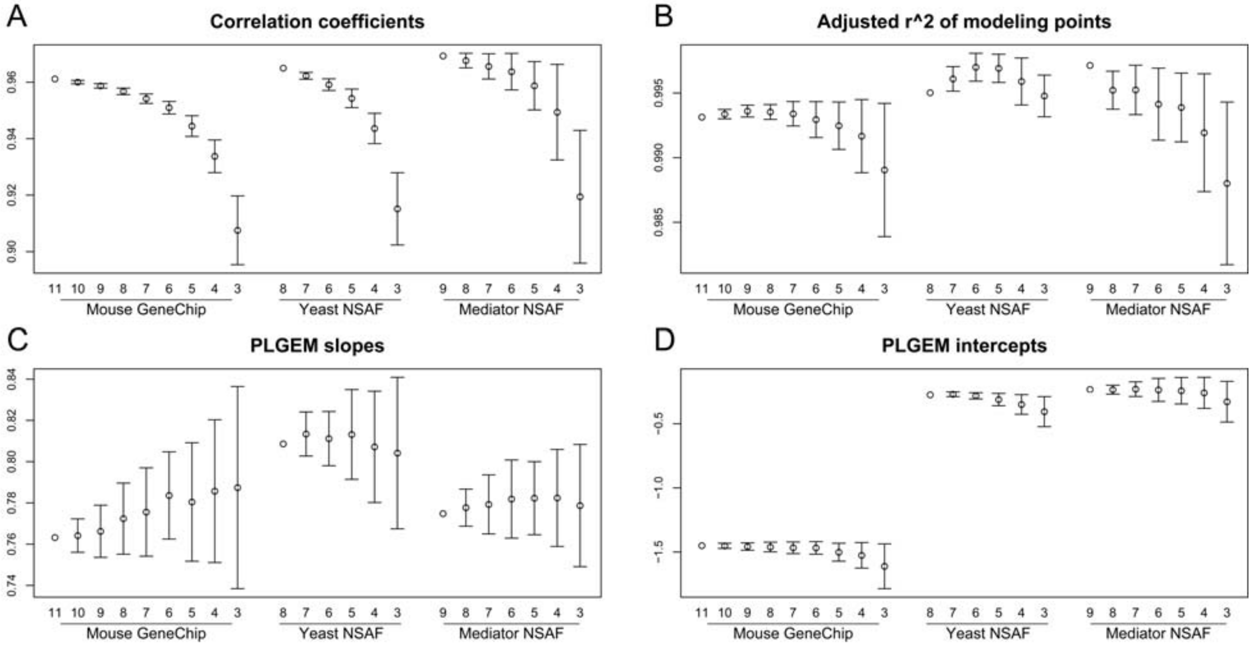
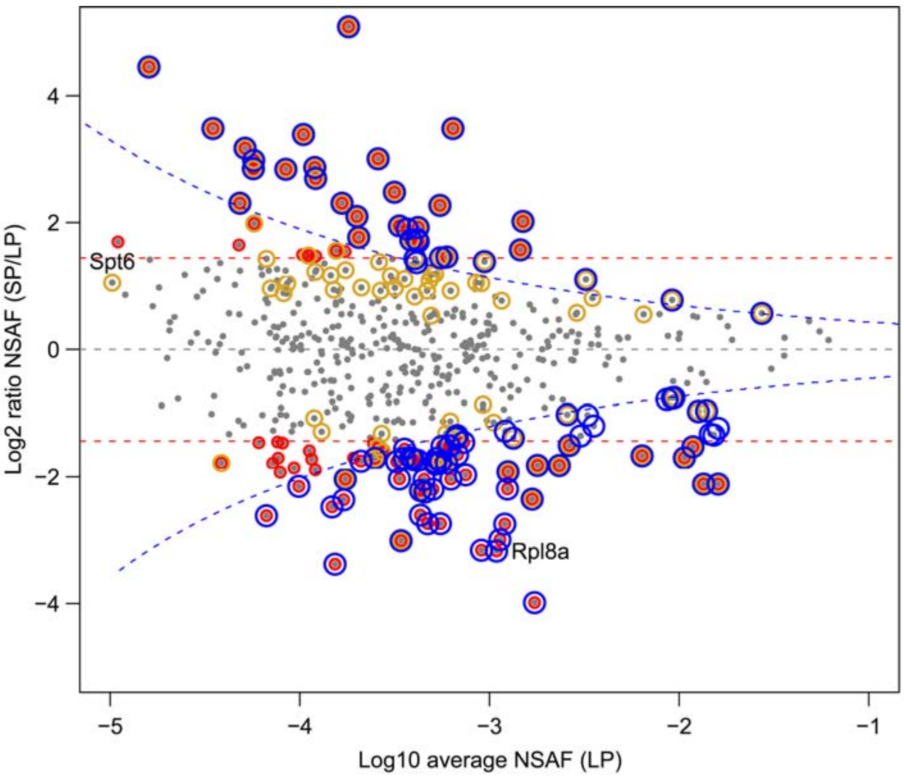

A
Correlation coefficients
0.90
0.92
0.94
0.96
B
Adjusted r^2 of modeling points
0.985
0.990
0.995
11 10 9 8 7 6 5 4 3
Mouse GeneChip
8 7 6 5 4 3
Yeast NSAF
9 8 7 6 5 4 3
Mediator NSAF
C
PLGEM slopes
0.74
0.76
0.78
0.80
0.82
0.84
D
PLGEM intercepts
-1.5
-1.0
-0.5

Figure 6


Log2 ratio NSAF (SP/LP)
4
2
0
−2
−4
Log10 average NSAF (LP)
−5
−4
−3
−2
−1
Spt6
Rpl8a