Journal of
Proteomics & Bioinformatics

Jung et al., J Proteomics Bioinform 2015, 8:10
DOI: 10.4172/jpb.1000374

Research Article Open Access

Systematic Analysis of Yeast Proteome Reveals Peptide Detectability Factors for Mass Spectrometry

Sunhee Jung[1,2], Samuel A Danziger[2,3], Alexandre Panchaud[4], Priska von Haller[5], John D Aitchison[2,3] and David R Goodlett[6]*

[1]*Molecular and Cellular Biology Program, University of Washington, Seattle, WA, USA*
[2]*Institute for Systems Biology, Seattle, WA, USA*
[3]*Center for Infectious Disease Research, Seattle, WA, USA*
[4]*Nestle, Lausanne, Switzerland*
[5]*Proteomics Resource, University of Washington, Seattle, WA, USA*
[6]*Department of Pharmaceutical Sciences, University of Maryland, Baltimore, MD, USA*

Abstract

Here we used a data-independent acquisition (DIA) method, Precursor Acquisition Independent from Ion Count (PAcIFIC), to systematically profile the *S. cerevisiae* proteome. Direct PAcIFIC analysis of a yeast whole cell lysate (WCL) yielded 90% reproducibility between replicates and detected approximately 2000 proteins. When combined with sub-cellular fractionation, reproducibility was equally high and the number of detected yeast proteins approached 5000. As noted previously, this unbiased DIA approach identified so-called "orphan" peptides that could only be detected by tandem mass spectra because there was no detectable precursor ion. Using this unique dataset we examined features associated with peptide detectability and demonstrated that orphans were more likely to arise from low copy number proteins than proteins with median or high copy number. Finally, an investigation into why some orphans also arose from high copy number proteins found that, aside from protein copy number, there was a bias toward physiochemical factors associated with regions flanking the proteolytic cleavage sites of orphan peptides. This suggested that those orphan peptides originating from high abundance proteins were likely the result of inefficient protease release, which has implications for quantitative bottom-up proteomics.

Keywords: Data-independent acquisition; Precursor acquisition independent from ion count; Orphan peptides; Peptide detectability

Abbreviation: DIA: Data-Independent Acquisition; PAcIFIC: Precursor Acquisition Independent from Ion Count

Introduction

Traditionally, shotgun proteomic analysis has been performed in a data-dependent manner where ion selection for collision-induced dissociation (CID) relies on a preliminary precursor ion scan from which peptide ions are selected and then subjected to CID [1]. While this general approach has been extremely powerful for determining the protein content of complex mixtures, the ability to compare related, but different, samples is complicated by the semi-random sampling process of data-dependent acquisition (DDA). This stochastic ion selection process results in under-sampling of all the ions available in a given precursor ion scan [1,2], reducing overall detectable dynamic range. This bias in ion selection by DDA methods is generally against ions of low signal/noise resulting in many more ions detected during a precursor ion scan than can be selected for CID in the time available.

Currently, the proteomic community seeks to circumvent this DDA-based loss of detectable dynamic range through use of targeted proteomics approaches, such as multiple reaction monitoring (MRM) that are often used once the proteome has been defined [3,4]. The other advantage of the MRM-based targeted strategies is that peptides are detected in a more sensitive manner than DDA methods because the time to acquire tandem mass spectra is reduced by monitoring select *m/z* channels and a subset of all available fragment ions that provides a gain in sensitivity. However, the sensitivity and time advantages provided by detecting only a few fragment ions comes at the cost of selectivity available when all fragment ions are recorded [5,6]. Here we refer to the IUPAC definition of selectivity which is the extent to which a method can determine the presence of a given analyte in a mixture without interference from other components [7]. While MRM-based targeted strategies allow researchers to overcome some of the dynamic range limitations of peptides detected, there are a number of challenges. One of the first is selection of the peptides from a given protein to be monitored. Complicating this decision is due to the fact that not all possible tryptic peptides in a protein are equally amenable to MS detection and identification [6]. Recent computational tools that attempt to predict "MS detectable" peptides from proteins are based, in large part, on such peptide physicochemical properties gleaned empirically from MS data [8,9]. Nevertheless, success in accurate *a priori* prediction of MS detectable peptides appears to be limited by some of the following factors. First, the various experimental conditions used in the proteomics community to isolate proteins and peptides are not standardized, making comparisons of empirically derived MS detectable peptide lists between experiments difficult as they are often laboratory-specific. Second, the manner in which mass spectrometry is performed also varies across laboratories and thus influences which peptides are perceived to be MS detectable. Given these two sources of variability involving platforms and laboratories, it would be valuable to develop and test rules for MS detectable peptides on a single platform.

Previously, we developed a data-independent acquisition (DIA)

***Corresponding author:** David R Goodlett, 20 North Pine Street, Department of Pharmaceutical Sciences, University of Maryland School of Pharmacy, Baltimore, MD 21217, USA, Tel: 410-706-1490; Fax: 410-706-0886; E-mail: dgoodlett@rx.umaryland.edu

Received June 30, 2015; **Accepted** October 27, 2015; **Published** October 30, 2015

Citation: Jung S, Danziger SA, Panchaud A, von Haller P, Aitchison JD, et al. (2015) Systematic Analysis of Yeast Proteome Reveals Peptide Detectability Factors for Mass Spectrometry. J Proteomics Bioinform 8: 231-239. doi:10.4172/jpb.1000374

method which we refer to as Peptide Acquisition Independent From Ion Count (PAcIFIC), eliminates the need for precursor ion scans by acquiring CID on every available m/z channel over the mass spectrometer's available m/z range regardless of whether a precursor ion is detected to be present or not [10]. This PAcIFIC approach has been shown to provide a higher number of protein identifications at greater sequence coverage without extensive pre-fractionation. Additionally, detectable dynamic range was shown to be extended by three orders of magnitude (base 10) as was the reproducibility of peptide and protein detection compared to DDA methods. Similar to the targeted MRM approaches, the DIA PAcIFIC method systematically acquires CID at every 1.5 *m/z* channel, but at higher selectivity because all available fragment ions for each *m/z* channel (i.e., peptide) are acquired. Notably, we have previously shown that DIA PAcIFIC can achieve dynamic range similar to that for Picotti et al.'s study using MRM, which identified proteins down to ~ 50 copies per cell in yeast [11,12]. We have also previously shown that dynamic range is extended in part with DIA PAcIFIC by sampling of orphan peptides, a class of peptides with no detectable precursor ion that are not usually detected by stochastic DDA methods. In fact, 18% to 30% of PAcIFIC identifications could come from these orphan peptides [10,13]. While orphan peptides are not universal in nature, but rather dependent on experimental conditions, they do constitute a very interesting class of peptides because (1) given their lack of a precursor ion signal they are not typically detectable in a DDA shotgun proteomic experiment (2), and not surprisingly they can have precursor ion detectable siblings peptides identified from the same parent protein. Therefore, from this discrepancy in ability to detect orphans only by tandem mass spectra and their sibling peptides by both a precursor ion signal and a tandem mass spectrum, we hypothesized that a difference in physicochemical properties between the orphan peptides and their MS1 and MS2 detected siblings might provide novel insight into peptide detectability in shotgun proteomic experiments.

To pursue this we first used a simplified subcellular fractionation strategy to extend proteomic dynamic range in yeast combined with a modified DIA PAcIFIC analysis. This approach allowed a high quality catalog of *Saccharomyces cerevisiae* proteins to be generated from only three fractions: 1) a whole cell lysate (WCL), 2) a pellet produced by a 20 Kg centrifugation of the WCL (i.e., the 20 KgP) known to be enriched for peroxisomes and mitochondria, and 3) a pellet produced by a 200 Kg centrifugation of the 20 KgP supernatant (i.e., the 200 KgP) known to contain various so-called "high speed" pelletable organelles. The catalog of all three fractions by triplicate PAcIFIC analysis contained 706,196 PSMs and 44,341 unique, semi-tryptic peptide identifications with a known and low cumulative false discovery rate of 1% corresponding to 5,026 distinct proteins, or 83% of the yeast proteome. As we show the major advantage of such a strategy over typical shotgun proteomic strategies that focus on generating many fractions analyzed by DDA based MS methods is that the full dynamic range of yeast may be detected from only these three fractions in a highly reproducible manner. Perhaps not surprisingly we also show that parent protein copy number is the single most important factor in determining whether a peptide will be MS detectable. Moreover, we show that orphan peptides are more likely to come from low copy number proteins than proteins with medium or high copy number. However, as discussed later they can also arise from high copy number proteins as well, which provided some clues as to their origins. By comparing the properties of orphans to both their precursor ion detected siblings and their siblings for which neither a precursor ion nor fragment ion spectrum are detected, our data largely confirm that the amino acid sequences neighboring orphan peptides in the parent protein are at least as important for prediction of peptide detectability as the peptide's own physicochemical properties. Finally, a unique peptide library produced as a result of this study presents a valuable resource for developing a high-throughput, low-redundancy proteome screening approach based on targeted mass spectrometry.

Experimental Procedures

Materials

All reagents and solvents were of the highest available purity. All reagents were purchased from Sigma-Aldrich or Fisher Scientific unless otherwise stated.

Sample preparations

Strain BY4742 was grown to log phase were grown to log phase in YPD (1% (w/v) yeast extract, 2% (w/v) peptone, 2% (w/v) glucose). Three fraction samples were prepared in the following manner. Briefly, cells were quickly pelleted, washed, and immediately lysed in a buffer 0.74% (v/v) β-mercaptoethanol and 0.815 M of NaOH while incubated at 4°C for 10 minutes. The proteins were precipitated by addition of trichloroacetic acid to a final concentration of 10% (v/v) followed by centrifugation. The pellet was washed twice with ice-cold acetone. Protein concentration was estimated by BCA protein assay (ThermoFisher Scientific Inc., Rockford, IL). Proteins were denatured by 6 M urea in 50 mM ammonium bicarbonate and reduced for 1 h at 37°C with 5 mM tris (2-carboxyethyl) phosphine. Alkylation of cysteine residues was performed with 30 mM iodoacetamide, for 1 h in the dark, followed by the addition of dithiolthreitol, to a final concentration 30 mM, and incubated for 1 h. The volume was increased eightfold with 50 mM ammonium bicarbonate to dilute the urea and the sample was incubated overnight at 37°C with sequencing grade trypsin (50:1 protein:trypsin ratio). Samples were desalted using MacroSpin C18 columns (30-300 µg capacity, SMMSS18V, The Nest Group, Southborough, MA, USA) according to the manufacturer's protocol. Eluates were stored at -80°C. For 20KgP and 200KgP fractions, subcellular fractionation was performed as previously described [14,15]. Briefly, harvested cells were converted to spheroplasts with 1 mg Zymolase 100 T/g of cells for 1 h at 30°C. Spheroplasts were lysed by homogenization in MES buffer (0.65 M sorbitol, 5 mM MES, pH 5.5) containing 1 mM KCl, 1mM EDTA, 0.2 mM PMSF, 0.4 µg pepstatin A/ml, 1X SigmaFAST™ Protease inhibitor (Sigma). Cell debris and nuclei were pelleted from the homogenate by centrifugation for 10 min at 2,000 x g to generate a postnuclear supernatant (PNS), which was subjected to 20,000 x g_{max} for 30 min at 4°C in a JS13.1 rotor (Beckman Instr., Inc.) to yield a pellet (20 KgP) enriched for peroxisomes and mitochondria and a supernatant (20 KgS) enriched for cytosol and high-speed pelletable organelles. The 20 KgS fraction was further subfractionated by differential centrifugation at 200,000 x g for 1 h at 4°C in a TLA120.2 rotor (Beckman Instrs., Inc.) to yield a pellet (200KgP) enriched for high-speed pelletable organelles and a supernatant (200KgS) highly enriched for cytosol.

Mass spectrometric analysis

All MS experiments were performed on a nanoACQUITY system (Waters, Milford, MA) connected to a hybrid LTQ-Orbitrap XL (Thermo Fisher scientific, San Jose, CA). For each injection, ~ 0.5 µg of peptide mixture was loaded, trapped on a 100 µm i.d. x 25 mm long precolumn packed with 200 Å (5 µm) Magic C18 particles (C18AQ; Michrom BioResources Inc., Auburn, CA), and separated in a gravity-pulled 75 µm i.d. x 200 mm analytical column packed with 100 Å (5 µm) Magic C18 particles (C18AQ; Michrom BioResources Inc.,

Auburn, CA) with a linear gradient of 0-35% (v/v) acetonitrile (ACN) in 0.1% (v/v) formic acid over 60 minutes. Peptides were eluted using an acetonitrile gradient flowing at 100 nL/min using mobile phase consisting of the following: A, water, 0.1% formic acid; B, acetonitrile, 0.1% formic acid. The gradient program was as follows: 0 min, A (95%), B (5%); 60 min, A (65%), B (35%); 65 min, A (15%), B (85%); 70 min, A (85%), B (15%); 75-95 min, A (95%), B (5%). The eluted peptides from the HPLC were directly electrosprayed into the mass spectrometer and analyzed in positive ion mode.

For DIA analysis, all data were acquired in triplicate (using three different technical replicates of each fraction) on the LTQ Orbitrap XL (Thermo Fisher scientific, San Jose, CA) using a modified PAcIFIC method [12]. Briefly, each single LC-MS/MS experiment comprised a cycle of 15 data-independent CID spectra covering 22.5 m/z units with a precursor ion survey scan inserted every 5 tandem mass spectra. Each survey scan was acquired from 400-2000 m/z in the Orbitrap analyzer at 60'000 resolution (at 400 Th) using an optimal ion population of 5 x 10^5 controlled by automatic gain control. For CID spectrum, ion population was set to 1 x 10^4, precursor isolation width to 2.5 Th, activation Q to 0.250, activation time to 30 ms and collision energy to 35%. A total of 45 LC-MS/MS analyses were performed on identical fashion to achieve a 1000 m/z mass range (400-1400 m/z). As 3.5 days are required for 45 injections, the entire analysis required 31.5 days to complete (3.5 d x 3 fractions x 3 technical replicates).

Pre-processing of data

For data acquired using accurate PAcIFIC method, data were pre-processed using the workflow as described previously [12]. Briefly, a feature detection step was first performed on the high resolution survey scans using Hardklor [16]. This information was further used to correct the precursor mass of the data-independent spectrum using an in-house Perl program (aPAcIFIC.pl). The output consists of two separate mzXML files containing either modified tandem MS spectra (to be searched at high mass accuracy) or raw tandem MS spectra (to be searched at low mass accuracy).

Peptide and protein identification

Database searches were performed against the yeast ORF database (release 2009-05-08) on the Saccharomyces genome database (SGD) website (www.yeastgenome.org) using SEQUEST v.27 algorithm [17]. For high mass accuracy search, precursor ion tolerance was set to 10 ppm and other search parameters include: one enzyme specific terminus required by trypsin, one missed cleavage allowed, alkylated cystein (+57 Da) set as a fixed modification, and oxidized methionine (+16 Da) as variable. For low mass accuracy search, precursor ion tolerance was set to 3.75 Da and other parameters were the same as in the high mass accuracy search. SEQUEST results were converted to pepXML files and probability assessments of identified peptides were computed with PeptideProphet [18]. For all individual searches (45 pepXML files for a given sample), peptides with an estimated false discovery rate of less than 1% were accepted unless specified. Only peptides mapping to unique protein were considered for protein identification and further analysis and only proteins with multiple unique peptides were considered for further analysis.

Gene ontology slim term enrichment

Go slim terms for each gene in the yeast genome were downloaded from the Saccharomyces Genome Database (SGD) website (www.yeastgenome.org) on 11/06/2010. For each term, the observed frequencies in each dataset (WCL, 20 KgP, 200 KgP) were compared with those expected by chance; i.e., the frequency of annotation for the 6310 yeast genes. For enriched terms, the probability that the observed distribution would be found by chance was determined by calculating binomial distribution probability using Microsoft Excel and the probability mass function. This algorithm has been used by others to estimate the significance of term enrichments with similar population and sample sizes [19].

Physicochemical properties that distinguish three groups of peptides using random forest classifier

For each protein for which orphans and their MS1 and MS2 detected siblings had been identified, *in silico* digests were performed and peptides were grouped into: (i) MS1 and MS2 detected siblings, (ii) orphan peptides, and (iii) undetected siblings. The 550 features using ESPP predictor [9] and 9 features using Peptide Detectability [8] were calculated for the list of peptides describing a set of 43476 instances where ~ 13% are MS1 and MS2 detected siblings, 5% are orphan peptides, and ~ 82% are undetected siblings. The original published versions of both the ESPP predictor and Peptide Detectability programs were used for our calculations. This data set is divided into equal sized training and test sets.

A Random Forest classifier was used to predict whether peptides would be detectable or non-detectable category. The Random Forest had the added benefit that it could report which features were most relevant for making the prediction and thus point to the underlying biology. Random Forest generates many classification trees by sampling the training data with replacement. These trees are grown by selecting a random subset of candidate features for each node and splitting on the best feature. Thus a large forest has considered many subsets of both the training data and the features describing that data. The forest predicts new examples based on the unweighted vote from all trees. When a Random Forest is allowed to construct many trees (10,000 in our study), it tends to outperform other popular machine learning methods such as support vector machines [20]. The best results were achieved when we down sampled to equalize the number of orphans, non-orphans, and non-detectable in the training set. The down sampling was performed at random. Ten-fold cross-validating was repeated four times using a different seed each time and cross-validated *P*-values reported reflect forty runs with different subsamples of the undetected siblings taken each run.

Results

Improved proteome coverage by a simple subcellular fractionation process

Given that we had previously identified only one-third of the yeast proteome from the WCL [12], we determined that additional fractionation was warranted to reach a higher number of protein identifications. In order to minimize the number of fractions needed for proteome characterization, we chose classical differential centrifugation, which gave us two new fractions: a 20 Kg pellet (20 KgP) and a 200 Kg pellet (200 KgP), each of which is known to be enriched in different organelles (Supplementary Figure 1A). Specifically, it is known that the 20 KgP contains enriched mitochondria and peroxisomes, while the 200 KgP contains enriched membrane vesicles and light organelles. Each fraction was subjected to triplicate DIA PAcIFIC analysis.

After combining both high and low mass accuracy search of the tandem MS data generated from all three fractions (i.e., WCL, 20 KgP and 200 KgP), we were able to identify 5,026 proteins with 99% certainty from the *Saccharomyce cerevisiae* proteome (Figure 1A,

Supplemental Table S4) representing 83% of the sequenced genome. Our data demonstrates that proteins with the lowest abundance level of 100 copies per cell can be effectively detected by our DIA PAcIFIC method when coupled to a simple and fast sample fractionation step. Previously, proteins likely to be translated were detected by a fused tandem affinity tag (TAP) [21] or green fluorescence protein (GFP) tagging genome wide experiments [22]. Our data overlaps 89% and 90%, respectively (Figure 1B), with these tagging approaches. In addition, our DIA PAcIFIC data overlaps with 94% of a prior yeast study generated by DDA using extensive fractionation technologies that included all of the following: gas-phase fractionation, SDS-PAGE and isoelectronic focusing (IEF) [23] (Figure 1B). Furthermore, we identified 608 proteins, which were detected neither by a genetic-based tagging approach nor a DDA method even using extensive pre-fractionation [23]. Taken together, our strategy coupled a modified DIA PAcIFIC method to a simple, bulk fractionation process known to produce discrete sub-proteomes allowed the whole MS-observable yeast proteome to be mapped.

Note that although more recent work on DDA without fractionation using the most recent instrumentation (e.g., Q-Exactive) has shown similar results to ours in terms of the number of proteins identified and reproducibility [24,25], our approach provides a method for those without access to the latest mass spectrometers, which are becoming out of reach for many laboratories, to outperform DDA strategies on similar instruments as used here without need to fractionate.

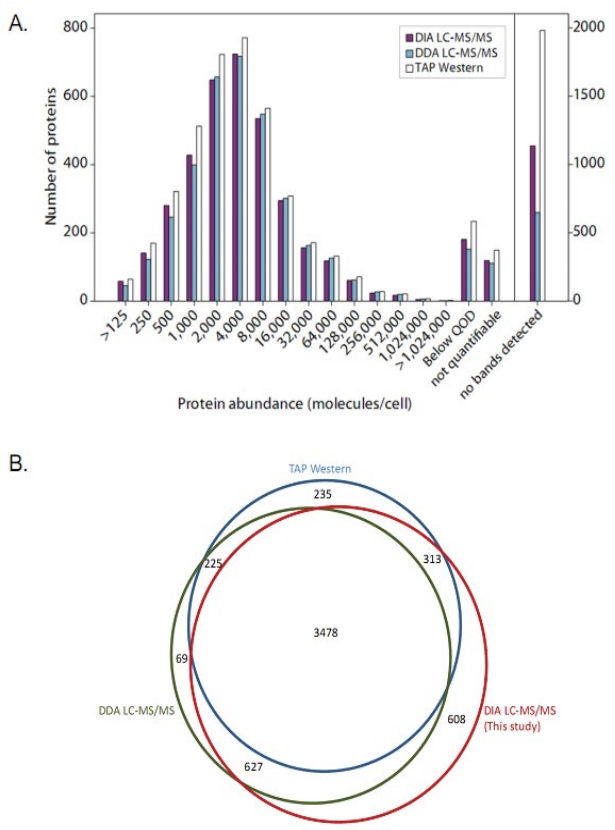

Figure 1: Proteome coverage from the all three combined fractions. *A*, Abundance distribution of the yeast proteome. Distribution of yeast proteins observed by TAP/Western-blot (white) [14], liquid chromatography/mass spectrometry (LC-MS/MS) combined with extensive pre-fractionation analysis (blue) [23], and DIA-MS analysis of the three combined fractions (magenta). *B*, Venn diagram comparing sets of proteins detected by Western blot of TAP-tagged strains (blue), DDA LC-MSMS (green), and DIA LC-MS/MS (red). Numbers are the identified proteins by each method.

DIA PAcIFIC metrics for a yeast WCL, 20 KgP, and 200 KgP analysis

Based on a threshold of two detected peptides in the 45 analyses carried out on single sample, an average of 2,280 ± 110 (mean ± SD) proteins were identified from triplicate DIA PAcIFIC analyses of the yeast WCL. Likewise, several DDA studies each analyzing many fractions have reported identification of 1500 to 2000 proteins from the same type of yeast WCL sample using Gel-based MS/MS method [26,27].

In order to determine the detectable dynamic range of DIA PAcIFIC, we compared our results to reported protein copy numbers determined by tagging with tandem affinity purification (TAP) followed by Western analysis [21]. Similarly to what was reported by Panchaud et al. [12], the 2,280 identified proteins ranged from very low (50 copies per cell) to very high abundance (10^6 copies per cell) which showed a remarkable overlap with the TAP dataset in a relative frequency distribution of the detected proteins across the entire dynamic range (Figure 2A).

Regarding the amount of sample needed for DIA PAcIFIC, we note that only 25 µg of total protein from the WCL sample was used to generate the data in a single DIA PAcIFIC replicate data set or 75 µg for triplicate analysis. In contrast, a typical DDA experiment employs 4-times more protein (100 µg) with extensive pre-fractionation [27], which results in unaccountable protein losses, simply to identify the same number of proteins at a similar dynamic range as the DIA PAcIFIC analysis. In agreement with the calculations of de Godoy et al. [27], it was estimated that the lowest abundance proteins (~ 50 to 100 molecules/cell) were present at a concentration of about 5 femtomoles in 25 µg of a WCL sample (i.e., 0.1 femtomoles on column per LC-MS/MS analysis; Figure 2A).

One measurement of a methods' effectiveness in analyzing a proteome is to correlate the number of proteins identified with each stage of analysis. In order to estimate the number of DIA PAcIFIC measurements required to reach 95% saturation in protein identification, using the triplicate DIA PAcIFIC measurements we statistically analyzed how many proteins are newly identified after each replicate. As shown in Figure 2B, after the second PAcIFIC replicate 253 more proteins (10%) were newly identified compared to the first replicate alone and after the third replicate 104 (4%) more new proteins were identified. An extrapolation of this data permits one to conclude that an additional fourth replicate would identify fewer than 4% more new proteins.

Reproducibility of analyses is a critical factor in use of quantitative proteomics where one wishes to distinguish true differences between different sample types. The typical DDA-based methods yield a low reproducibility between technical replicates due to the aforementioned semi-random sampling problem. In contrast to this low reproducibility of DDA approaches, the DIA PAcIFIC yielded high protein identification reproducibility (~ 90%) between triplicate data sets, indicating very low variance between replicates.

Performance metrics for both of 20 KgP and 200 KgP fractions were also assessed. For detailed results on method effectiveness, reproducibility, gene ontology and coverage of membrane proteins, please refer to the Supplementary Note.

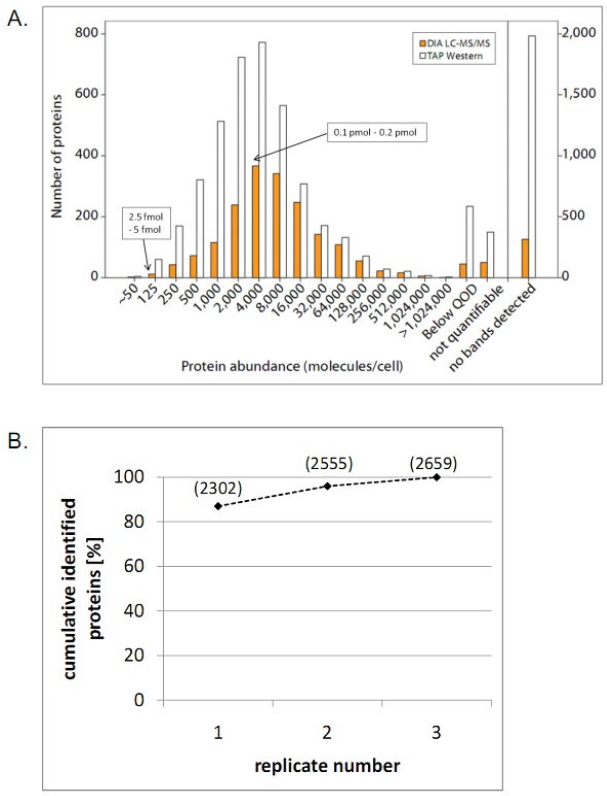

Figure 2: Analysis of identified proteins from a yeast whole cell lysate. *A*, Protein abundance in the yeast proteome and identified proteins from WCL by DIA MS analysis. White columns indicate the number of yeast proteins in copy number classes from TAP Western study [14]. Orange bars represent the proteins identified based on 2 peptide threshold in each copy number class in this study. *B*, Percentage of proteins identified as a function of the number of replicates for triplicate 45 HPLC–ESI-MS/MS analyses. Numbers in parenthesis indicate the total number of proteins.

Overall, these results demonstrate that the DIA PAcIFIC method surpasses typical DDA methods in terms of simplicity of use by circumventing pre-fractionation. In so doing results from the above PAcIFIC analysis indicate that a comparable number of proteins are identified from a WCL and using less protein than needed by a typical DDA method, the full detectable dynamic range of yeast proteins are detected in 45 LC-MS/MS injections, and finally that 90% reproducibility is achieved in duplicate analysis.

Significance of orphans for protein identification

By inserting an MS1 scan after every five DIA MS2 scans, we were able to re-investigate the rate of orphan peptide occurrence more accurately than prior attempts that used separately acquired MS1 and MS2 data [12]. Here the list of MS1 identified features was correlated with the list of MS2 identified peptides from the same HPLC-MS experiment. Analyzing the triplicate data sets from the WCL sample in this manner we identified approximately 16.6% (n=3816, SD ± 1.05%) of the total 22959 (SD ± 558) unique peptides as orphans. Fifty one percent of non-orphans and 29% of orphans were common to the three replicates.

We further investigated whether these orphan peptides detected by the DIA PAcIFIC approach had a significant contribution to the overall protein identifications. For this investigation only peptides identified (charge state +2 and FDR <0.005) in one of the WCL triplicate data sets were used. Non-orphan peptides accounted for a total of 1226 proteins identified, while orphan peptides added 493 more proteins to the list of 1719 proteins identified in this single replicate. The results were similar for the other replicates, which showed that the ability to detect orphan peptides significantly improves the number of protein identification.

Factors governing discrimination between non-orphans and orphans

Given the fact that orphans have no detectable precursor ions, the most obvious explanations for why a peptide is an orphan would be either due to their lower abundance relative to other peptides on the capillary column or some unique physicochemical property that creates an ionization efficiency bias or a bias in the binding or release from the C18 stationary phase used in these HPLC-MS/MS experiments, the latter two being difficult to examine here.

To investigate further, we first looked at whether there was a bias toward orphan peptides coming from proteins of low copy number, which in fact had already been suggested by prior analysis [10,13]. In order to investigate this hypothesis, we created a subset of detected peptides containing orphans and their MS1 and MS2 detected siblings with no charge state bias; i.e. all peptides in this data set were fully tryptic with two charges. Second, peptides that were homologous to more than one parent protein were discarded in order to avoid ambiguity. Third, the parental protein abundance for this culled peptide data set was derived from the known protein copy numbers from Ghaemmaghami et al. [21]. As shown in Figure 3 the results clearly indicate that orphan peptides were more likely to come from lower abundant proteins. This result strongly supports our hypothesis that the likelihood of a peptide being an orphan is due primarily to its lower abundance relative to other peptides on the capillary column.

In order to gain further insight into this outcome, we used two recently developed software tools [8,9] that attempt to predict which peptides from a genome are likely to be detected in an MS experiment to help us tease apart any hidden factors that might contribute to the conundrum of why one peptide is labeled an orphan and a different peptide from the same protein a non-orphan sibling. Specifically,

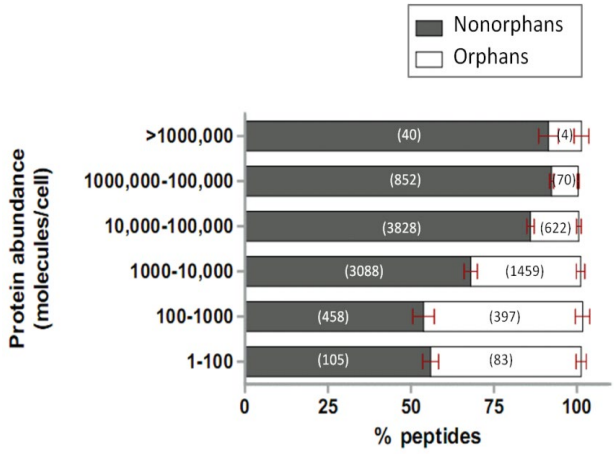

Figure 3: Abundance distribution of orphans and nonorphans derived from their parent protein copy numbers measured by TAP tagging approach [14]. Each bar represents the average of three technical replicates, and red error bars represent the standard deviation of the mean.

Physicochemical properties	*P*-value
Protein copy number	<<< 2.00E- 16
VL2S disorder	8.18E-11
Gas Phase basicity	4.82E-09
pI	2.58E-06
VL2 disorder	2.36E-05
VL2V disorder	0.000604
Hydrophobic moment c	0.013
Transfer free energy to surface	0.017
Partition energy	0.019
Transfer energy organic solvent water	0.028
Hydrophobic moment a	0.033
Bitterness	0.041

Table 1: Twelve best features estimated using LR test.

Peptide Detectability [8] and ESP Predictor [9] were used to calculate the top ranked physicochemical properties that most affect peptide detectability reported in these two studies. We chose these two tools because they had fundamentally different approaches to investigation of detectability. Specifically, Tang et al.'s [8] study focused on MS2 detectability, while Fusaro et al.'s [9] focused on MS1 detectability. Nine features from Tang et al.'s [8] study such as peptide flexibility, hydrophobic momentum, and intrinsic disorder and 35 features from Fusaro et al.'s [9] study were calculated for the set of peptides in this study (Supplemental Table S5). These features were chosen because they represented a set of forty-four unique features out of all of the features each group considered. Tang et al.'s nine features were based on protein primary sequence as well as regions neighboring the tryptic peptide of interest, while Fusaro et al.'s 35 features were based solely on primary sequence. A logistic regression test was performed to see how these features influenced the ability to discriminate non-orphans and orphans. As shown in Table 1, of all features examined, "abundance" was most strongly correlated with orphans, which had a *P*-value less than 2e-16. Thus, this result strongly supports our hypothesis that lack of a detectable precursor ion for orphan peptides is mainly due to the fact that these peptides come from low abundance proteins in a given sample. Interestingly, the second strongest correlation for whether a peptide might be an orphan peptide had to do with chemical properties of their parent proteins, not the peptide itself. Specifically, a property referred to as VL2S predictor that measures disorder of the parent protein [28] stood out as significant. The VL2S predictor is influenced mostly by entropy and hydrophobic amino acids (e.g., tyrosine). The third strongest correlation for discriminating between non-orphan peptides and orphan peptides was peptide gas phase basicity [29], which influences the quality of a tandem mass spectrum.

Detection of the critical features to predict peptide detectability

Recent studies have also observed that most proteins are identified by one to a few peptides observed at a high frequency [6,30]. In agreement with these observations, we also observed that surprisingly some orphan peptides come from proteins with high copy numbers where their MS1 and MS2 detected sibling peptides were detected as well, and of course, as in all proteomic experiments, where some sibling peptides were detected neither by an MS1 nor an MS2 signal (i.e., absent in the PAcIFIC proteomic data). In addition, our findings support the idea that peptide detectability is influenced by features (e.g., the VL2S) that are based on both its individual peptide sequence and the flanking peptide regions from the parent protein. Moreover, only ~ 1% of orphan peptides came from regions between siblings that were identified as non-orphan peptides. Thus, we further hypothesized that the difference in MS1 detectability between orphans and their siblings was due to the rate at which they are released from the parent protein, orphan peptides being released less efficiently due to inability of the protease to access the cut site.

In order to investigate this hypothesis for those proteins identified by both orphan peptides and their siblings we examined three groups of peptides: (1) sibling peptides detected by MS1 and MS2 signals, (2) sibling peptides detectable by neither MS1 nor MS2 signals which includes those peptides that trigger CID but do not generate a sequence match, and (3) orphan peptides detected only by MS2. In comparing differences between these three classes we extended the number of features examined to a total of 559 peptide physicochemical properties. Of the 559 features, 550 features were based only on peptide sequence and were analyzed using the ESP Predictor [9]. The additional nine features were based on peptide sequence as well as neighboring regions of the peptides in the parent protein and were calculated using Peptide Detectability (Supplemental Table S5). Next, a random forest algorithm, a nonlinear ensemble classifier composed of many decision trees, was used to rank the features with respect to their individual ability to distinguish the three classes of peptides. Table 2 lists the top 15 discriminatory properties. Among the top 15 discriminating properties, eight are related to protein structure properties. For example, Vihinen flexibility [31] and B factor [32], both of which are structural flexibility scales based on peptide sequence and juxtaposed neighboring amino acid regions, were selected as very informative. This result again strongly supports our hypothesis that peptide detectability is influenced not only by its amino acid sequence but also by the neighboring, linearly juxtaposed primary sequences from the parent protein. Thus, we can speculate that the existence of orphan peptides from high abundance proteins suggests inefficient proteolytic release from their parent proteins due to structural properties.

Having discovered that most of the discriminating physical properties were based on a peptide's sequence and its neighboring sequence regions, we lastly looked at whether these derived parameters could provide some better understanding of peptide detectability. We modeled two peptide response predictors using the same Random Forest algorithm used previously. One predictor was built using only 550 of Fusaro et al.'s features and the other predictor with both the 550 features of Fusaro et al. as well as the nine features of Tang et al.'s

Rank	Peptide properties
1	Mass
2	**Vihinen**
3	Length
4	**B-factor prediction**
5	Gas phase basicity
6	**Hydrophobic momentum c**
7	**Hydrophobic momentum a**
8	**Hydrophobic momentum b**
9	nBasic
10	Positive charge
11	**VLS disorder**
12	**VL2 disorder**
13	**VL2V disorder**
14	Activation Gibbs energy of unfolding
15	Isoelectric point

Features in bold are based on a peptide sequence and its neighboring regions, while the others are only based on a peptide sequence.

Table 2: List of top 15 ranked features.

report. The performance of the two predictors was evaluated within the data set (a 10-fold cross validation) as well as an independent data set. In order to avoid any obvious bias, four different measurements (F1, F2, accuracy, and weighted accuracy) were used to estimate accuracy of our predictions on the derived data set. The F1 and F2 metrics were developed to accurately assess classifier performance on unbalanced data sets [33,34]. The weighted accuracy is the average number of points if one assigns one point for a correct prediction, half a point if the classifier predicted "orphan" peptide was not an "orphan" (or vice versa), and zero points otherwise. *P*-values were calculated using a paired t-test on multiple predictions created by repeated down-sampling of the training set. The results (Table 3) showed that the Random Forest classifier performed significantly (P-value ≤ 0.05) better across all four accuracy measurements when trained with 559 physicochemical features based on peptide sequence and neighboring regions of the parent proteins than 550 features based only on peptide sequence. Important to the entire process of selecting MS detectable peptides, this result suggests that not only peptide sequence, but neighboring regions of the parent protein affect peptide detectability.

Discussion

We evaluated the performance of the DIA PAcIFIC method, which systematically acquires tandem mass spectra on all *m/z* channels, to characterize thoroughly the well-studied *Saccharomyces cerevisiae* proteome. The results clearly show that this relatively new proteomic method produces results from a single fraction that are comparable to DDA based shotgun proteomic technologies that analyze many fractions, e.g. MudPIT [26] and 2DGel [27]. While a wide variety of the popular DDA techniques require relatively higher amount of samples and labor intensive steps of sample pre-fractionation, the DIA PAcIFIC method requires no sample pre-fractionation, uses less sample and provides similar or better performance. Although some of the DDA methods such as MudPIT have been automated to circumvent the labor required to pre-fractionate [35], the use of multidimensional separations has been shown to introduce more variation with a relative decrease in reproducibility by up to 25% compared to a one-dimensional separation. With only one fraction, the DIA PAcIFIC method is much more automated making it much more like genomic sequencing than most labor intensive DDA proteomic methods that require exhaustive pre-fractionation. The addition of high mass accuracy MS1 scans employed here with PAcIFIC led to confident identification of more than 2,000 proteins in the yeast WCL without pre-fractionation and covered the whole dynamic range of the yeast proteome. Moreover, the addition of a single step of bulk, subcellular fractionation to reduce sample complexity increased the sensitivity of detection as witnessed by detection of ~ 83% of the yeast proteome (Supplementary Table S4). Furthermore, of the total proteins identified here, approximately 20% are proteins having known or predicted transmembrane domains that are usually considered as low abundant and rarely detected in a complex sample by MS [36,37], indicating that DIA PAcIFIC combined with basic subcellular fractionation is a suitable method to increase detection of membrane proteins (Supplemental Tables S1-S3). This new approach also allowed identification of a considerable number of proteins that previously have not been observed by high throughput tagging approaches nor the latest comprehensive DDA-based MS analysis (Figure 1).

One criticism of the DIA PAcIFIC approach has been the requirement for instrument time beyond what is perceived to be normal for popular DDA methods. While the PAcIFIC method requires multiple days of MS instrument time to complete one replicate (3.5 days on LTQ Orbitrap XL or 2.5 days on LTQ Velos Orbitrap), we have shown here that in only one replicate 90% saturation in protein identifications was achieved from a yeast WCL. Although one replicate of DIA PAcIFIC analysis reaches 90% saturation in protein identification, our data shows that two replicates of DIA PAcIFIC analysis are required to reach a statistically defined level of proteome saturation. Regardless of how the PAcIFIC experiments are conducted on current instrumentation, we expect that ultimately innovations in mass spectrometers and separations will provide an entire PAcIFIC data set from a single HPLC experiment [38].

Finally, the most fundamental new finding reported here relates to an investigation of the origin of orphan peptides that permitted the concept of peptide detectability to be distilled to a few basic parameters. These analyses revealed that orphans most often originate from proteins with low copy numbers, a result that indicates that protein abundance, a non-peptide physicochemical feature, primarily affects peptide detectability. Interestingly for the question of which peptides from a single protein are most likely to be detected, we found that some orphan peptides also came from proteins with high copy number. These orphan peptides often had MS1 and MS2 detected sibling peptides that were identified, which led to an interesting conundrum. Why would peptides from the same parent protein exhibit such different MS1 signals? One might expect that these sibling peptides would be present at equal stoichiometry, but their disparate MS1 signals suggested otherwise. This simple finding led us in turn to our investigation of peptide detectability using a novel set of three classes of peptides (1) orphans, (2) their MS1 and M2 detected siblings, and (3) their undetected siblings. All prior studies used a simplified two-class set of (1) detectable and (2) low/not-detectable for modeling detectability, but these studies lacked access to information on orphan peptides. Our results with three peptide classes showed that eight out of the top 15 ranked features to discriminate these three classes of peptides are based on peptide sequence as well as the detected peptide's neighboring amino acid sequence regions (Table 2). This result reinforces the concept that the abundance of peptides (not proteins) in a given enzymatically digested sample is influenced not only by their parent protein's abundance, but also by flanking amino acid sequences that can differ in protease accessibility which in turn can affect digestion

	Cross validation			Blind prediction	
	550 features[a]	550 features[a] and 9 features[b]	*P*-value	550 features[a]	550 features[a] and 9 features[b]
F1	0.4744	0.4865	5.58E-10	0.4874	0.4969
F2	0.557	0.5729	2.13E-09	0.5823	0.5962
Accuracy	0.6081	0.6197	1.02E-09	0.6138	0.6211
Weighted accuracy	0.6933	0.7007	7.16E-07	0.701	0.7045

[a] 550 features from ESP predictor [9]

[b] 9 features from Peptide Detectability [8] as shown in Table S3

Table 3: Performance comparison of two different predictors.

efficiency and thus peptide detectability.

In conclusion, unprecedented high proteome coverage of yeast was achieved from only three fractions by combining subcellular fractionation with DIA PAcIFIC analysis. This approach was robust and reproducible, affording a comprehensive simple analysis and yielding a large number of proteins including numerous membrane proteins. In addition, from the machine learning perspectives, we provide indications that in addition to properties only based on peptide sequence, peptide detectability is influenced by properties related to the flanking peptide regions in their parent protein; e.g., Vihinen, B-factor prediction, hydrophobic momentum, and VL disorders. This finding provides a basis for future research towards developing a better predictor for peptide detectability that incorporates protein structure. Moreover, due to the comprehensive nature of our empirical data set, it is an excellent resource for a rapid selection of MS detectable peptides for targeted MRM analyses. The simplicity, sensitivity and reproducibility of the approach in the analysis of complex samples make it an attractive tool for systems biology research and biomarker discovery.

Acknowledgement

This work was funded by NIH/NIGMS (R01-GM075152, P50-GM076547, U54-RR022220, 5R33CA099139-04, 3R33CA099139-04S1, and 1S10RR023044-01) and supported in part by the University of Washington's Proteomics Resource (UWPR95794). We also thank the Luxembourg Centre for Systems Biomedicine and the University of Luxembourg for support.

References

1. Aebersold R, Mann M (2003) Mass spectrometry-based proteomics. Nature 422: 198-207.
2. Liu H, Sadygov RG, Yates JR 3rd (2004) A model for random sampling and estimation of relative protein abundance in shotgun proteomics. Anal Chem 76: 4193-4201.
3. Anderson L, Hunter CL (2006) Quantitative mass spectrometric multiple reaction monitoring assays for major plasma proteins. Mol Cell Proteomics 5: 573-588.
4. Keshishian H, Addona T, Burgess M, Kuhn E, Carr SA (2007) Quantitative, multiplexed assays for low abundance proteins in plasma by targeted mass spectrometry and stable isotope dilution. Mol Cell Proteomics 6: 2212-2229.
5. Craig R, Cortens JP, Beavis RC (2005) The use of proteotypic peptide libraries for protein identification. Rapid Commun Mass Spectrom 19: 1844-1850.
6. Kuster B, Schirle M, Mallick P, Aebersold R (2005) Scoring proteomes with proteotypic peptide probes. Nat Rev Mol Cell Biol 6: 577-583.
7. Vessman J (1996) Selectivity or specificity? Validation of analytical methods from the perspective of an analytical chemist in the pharmaceutical industry. J Pharm Biomed Anal 14: 867-869.
8. Tang H, Arnold RJ, Alves P, Xun Z, Clemmer DE, et al. (2006) A computational approach toward label-free protein quantification using predicted peptide detectability. Bioinformatics 22: e481-e488.
9. Fusaro VA, Mani DR, Mesirov JP, Carr SA (2009) Prediction of high-responding peptides for targeted protein assays by mass spectrometry. Nat Biotechnol 27: 190-198.
10. Panchaud A, Scherl A, Shaffer SA, von Haller PD, Kulasekara HD, et al. (2009) Precursor acquisition independent from ion count: how to dive deeper into the Proteomics Ocean. Anal Chem 81: 6481-6488.
11. Picotti P, Bodenmiller B, Mueller LN, Domon B, Aebersold R (2009) Full dynamic range proteome analysis of *S. cerevisiae* by targeted proteomics. Cell 138: 795-806.
12. Panchaud A, Jung S, Shaffer SA, Aitchison JD, Goodlett DR (2011) Faster, quantitative and accurate precursor acquisition independent from ion count. Anal Chem 83: 2250-2257.
13. Scherl A, Shaffer SA, Taylor GK, Kulasekara HD, Miller SI, et al. (2008) Genome-specific gas-phase fractionation strategy for improved shotgun proteomic profiling of proteotypic peptides. Anal Chem 80: 1182-1191.
14. Titorenko VI, Smith JJ, Szilard RK, Rachubinski RA (1998) Pex20p of the yeast *Yarrowia lipolytica* is required for the oligomerization of thiolase in the cytosol and for its targeting to the peroxisome. J Cell Biol 142: 403-420.
15. Titorenko VI, Chan H, Rachubinski RA (2000) Fusion of small peroxisomal vesicles *in vitro* reconstructs an early step in the *in vivo* multistep peroxisome assembly pathway of *Yarrowia lipolytica*. J Cell Biol 148: 29-44.
16. Hoopmann MR, Finney GL, MacCoss MJ (2007) High-speed data reduction, feature detection, and MS/MS spectrum quality assessment of shotgun proteomics data sets using high-resolution mass spectrometry. Anal Chem 79: 5620-5632.
17. Eng JK, McCormack AL, Yates JR (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.
18. Deutsch EW, Shteynberg D, Lam H, Sun Z, Eng JK, et al. (2010) Trans-Proteomic Pipeline supports and improves analysis of electron transfer dissociation data sets. Proteomics 10: 1190-1195.
19. Begley TJ, Rosenbach AS, Ideker T, Samson LD (2004) Hot spots for modulating toxicity identified by genomic phenotyping and localization mapping. Mol Cell 16: 117-125.
20. Statnikov A, Wang L, Aliferis CF (2008) A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification. BMC Bioinformatics 9: 319.
21. Ghaemmaghami S, Huh WK, Bower K, Howson RW, Belle A, et al. (2003) Global analysis of protein expression in yeast. Nature 425: 737-741.
22. Huh WK, Falvo JV, Gerke LC, Carroll AS, Howson RW, et al. (2003) Global analysis of protein localization in budding yeast. Nature 425: 686-691.
23. de Godoy LM, Olsen JV, Cox J, Nielsen ML, Hubner NC, et al. (2008) Comprehensive mass-spectrometry-based proteome quantification of haploid versus diploid yeast. Nature 455: 1251-1254.
24. Nagaraj N, Kulak NA, Cox J, Neuhauser N, Mayr K, et al. (2012) System-wide perturbation analysis with nearly complete coverage of the yeast proteome by single-shot ultra HPLC runs on a bench top Orbitrap. Mol Cell Proteomics 11: M111.
25. Hebert AS, Richards AL, Bailey DJ, Ulbrich A, Coughlin EE, et al. (2014) The one hour yeast proteome. Mol Cell Proteomics 13: 339-347.
26. Washburn MP, Wolters D, Yates JR 3rd (2001) Large-scale analysis of the yeast proteome by multidimensional protein identification technology. Nat Biotechnol 19: 242-247.
27. de Godoy LM, Olsen JV, de Souza GA, Li G, Mortensen P, et al. (2006) Status of complete proteome analysis by mass spectrometry: SILAC labeled yeast as a model system. Genome Biol 7: R50.
28. Vucetic S, Brown CJ, Dunker AK, Obradovic Z (2003) Flavors of protein disorder. Proteins 52: 573-584.
29. Zhang Z (2004) Prediction of low-energy collision-induced dissociation spectra of peptides. Anal Chem 76: 3908-3922.
30. Mallick P, Schirle M, Chen SS, Flory MR, Lee H, et al. (2007) Computational prediction of proteotypic peptides for quantitative proteomics. Nat Biotechnol 25: 125-131.
31. Vihinen M, Torkkila E, Riikonen P (1994) Accuracy of protein flexibility predictions. Proteins 19: 141-149.
32. Radivojac P, Obradovic Z, Smith DK, Zhu G, Vucetic S, et al. (2004) Protein flexibility and intrinsic disorder. Protein Sci 13: 71-80.
33. Ko¨knar-Tezel S, Latecki LJ (2009) Improving SVM Classification on Imbalanced Data Sets in Distance Spaces. IEEE Int Conf on Data Mining.
34. van Rijsbergen CJ (1979) Information Retrieval. Butterworths (2nd Edn). Information Retrieval Group, University of Glasgow.
35. Motoyama A, Venable JD, Ruse CI, Yates JR (2006) Automated ultra-high-pressure multidimensional protein identification technology (UHP-MudPIT) for improved peptide identification of proteomic samples. Anal Chem 78: 5109-5118.

Citation: Jung S, Danziger SA, Panchaud A, von Haller P, Aitchison JD, et al. (2015) Systematic Analysis of Yeast Proteome Reveals Peptide Detectability Factors for Mass Spectrometry. J Proteomics Bioinform 8: 231-239. doi:10.4172/jpb.1000374

36. Yi EC, Marelli M, Lee H, Purvine SO, Aebersold R, et al. (2002) Approaching complete peroxisome characterization by gas-phase fractionation. Electrophoresis 23: 3205-3216.

37. Marelli M, Smith JJ, Jung S, Yi E, Nesvizhskii AI, et al. (2004) Quantitative mass spectrometry reveals a role for the GTPase Rho1p in actin organization on the peroxisome membrane. J Cell Biol 167: 1099-1112.

38. Wang H, Kennedy DS, Nugent KD, Taylor GK, Goodlett DR (2007) A Qit-q-Tof mass spectrometer for two-dimensional tandem mass spectrometry. Rapid Commun Mass Spectrom 21: 3223-3226.