Journal of proteome research

Article

pubs.acs.org/jpr

Reproducibility of Differential Proteomic Technologies in CPTAC Fractionated Xenografts

David L. Tabb,*[,†] Xia Wang,[‡] Steven A. Carr,[§] Karl R. Clauser,[§] Philipp Mertins,[§] Matthew C. Chambers,[†] Jerry D. Holman,[†] Jing Wang,[†] Bing Zhang,[†] Lisa J. Zimmerman,[∥] Xian Chen,[⊥] Harsha P. Gunawardena,[⊥] Sherri R. Davies,[#] Matthew J. C. Ellis,[#,□] Shunqiang Li,[#] R. Reid Townsend,[#] Emily S. Boja,[▽] Karen A. Ketchum,[○] Christopher R. Kinsinger,[▽] Mehdi Mesri,[▽] Henry Rodriguez,[▽] Tao Liu,[◆] Sangtae Kim,[◆] Jason E. McDermott,[◆] Samuel H. Payne,[◆] Vladislav A. Petyuk,[◆] Karin D. Rodland,[◆] Richard D. Smith,[◆] Feng Yang,[◆] Daniel W. Chan,[¶] Bai Zhang,[¶] Hui Zhang,[¶] Zhen Zhang,[¶] Jian-Ying Zhou,[¶] and Daniel C. Liebler[∥]

[†]Department of Biomedical Informatics, [∥]Department of Biochemistry, Vanderbilt University, Nashville, Tennessee 37232, United States

[‡]Department of Mathematical Sciences, University of Cincinnati, Cincinnati, Ohio 45221, United States

[§]Proteomics Platform, Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02142, United States

[⊥]Department of Biochemistry and Biophysics, University of North Carolina, Chapel Hill, North Carolina 27599, United States

[#]Department of Medicine, Washington University, St. Louis, Missouri 63110, United States

[▽]Office of Cancer Clinical Proteomics Research, National Cancer Institute, Bethesda, Maryland 20892, United States

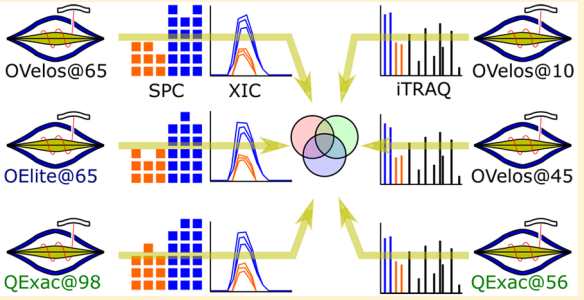
[○]Enterprise Science and Computing, Inc., Rockville, Maryland 20850, United States

[◆]Division of Biological Sciences, Pacific Northwest National Laboratory, Richland, Washington 99352, United States

[¶]JHMI and Division of Clinical Chemistry, Johns Hopkins University, Baltimore, Maryland 21231, United States

Supporting Information

ABSTRACT: The NCI Clinical Proteomic Tumor Analysis Consortium (CPTAC) employed a pair of reference xenograft proteomes for initial platform validation and ongoing quality control of its data collection for The Cancer Genome Atlas (TCGA) tumors. These two xenografts, representing basal and luminal-B human breast cancer, were fractionated and analyzed on six mass spectrometers in a total of 46 replicates divided between iTRAQ and label-free technologies, spanning a total of 1095 LC−MS/MS experiments. These data represent a unique opportunity to evaluate the stability of proteomic differentiation by mass spectrometry over many months of time for individual instruments or across instruments running dissimilar workflows. We evaluated iTRAQ reporter ions, label-free spectral counts, and label-free extracted ion chromatograms as strategies for data interpretation (source code is available from http://homepages.uc.edu/~wang2x7/Research.htm). From these assessments, we found that differential genes from a single replicate were confirmed by other replicates on the same instrument from 61 to 93% of the time. When comparing across different instruments and quantitative technologies, using multiple replicates, differential genes were reproduced by other data sets from 67 to 99% of the time. Projecting gene differences to biological pathways and networks increased the degree of similarity. These overlaps send an encouraging message about the maturity of technologies for proteomic differentiation.



KEYWORDS: *Differential proteomics, label-free, iTRAQ, quality control, xenografts, technology assessment, CPTAC*

INTRODUCTION

The Clinical Proteomic Tumor Analysis Consortium (CPTAC) was charged with establishing Proteome Characterization Centers to measure protein differences across large numbers of tumors from The Cancer Genome Atlas (TCGA).[1] To ensure that the differences they found were reproducible, each center conducted many replicate analyses for a pair of patient-derived xenograft tissues as a comparative reference material ("CompRef"), first as a preliminary validation of their

Special Issue: Large-Scale Computational Mass Spectrometry and Multi-Omics

Received: September 14, 2015
Published: December 14, 2015

DOI: 10.1021/acs.jproteome.5b00859
J. Proteome Res. 2016, 15, 691−706

workflows and later interspersed among the TCGA samples to ensure ongoing validation of their workflows. These data, spanning more than 1000 LC−MS/MS experiments, represent a unique opportunity to evaluate the variability of protein differentiation technologies across multiple platforms.

Variability can enter proteomic experiments through sample handling and proteolysis,[2] prefractionation,[3,4] liquid chromatography,[5] mass spectrometry configuration,[6] and bioinformatics.[7] At the start of the program in 2011, CPTAC evaluated ischemia time in the context of global proteomes[8] and phosphotyrosine signaling[9] to evaluate potential impacts for proteomic changes after blood supply to a tissue is lost but before the sample is frozen. At the same time, CPTAC conducted platform validation for each Proteome Characterization Center by distributing aliquots of the CompRef xenografts; data from each site could then be used to assess the variability contributed by analytical methodology. Most sites opted to modify their platforms in response to these data, most commonly by switching to bRPLC fractionation[10] from SCX or IEF fractionation[4] techniques.

Patient-derived tumor xenografts are a technology by which tumors can be induced in mice via cells taken from human tumors, affording better latitude for experimentation.[11] When tissue is harvested from a xenograft, it will contain both human and mouse proteins since the tumor cells are from a different species than the host environment. The xenografts employed in this study were WHIM2 (basal) and WHIM16 (luminal-B), drawn from a larger breast cancer study at Washington University in St. Louis.[12] Both xenografts produced large tumors; they were grown in a sufficient number of mice to generate a pool of protein that was large enough to serve all of the CPTAC Proteome Characterization Centers. While the differences between basal and luminal-B tumors have been studied at length, the proteomic differences between them cannot be scored against a comprehensive answer key.

These experiments with WHIM2 and WHIM16 represent the three chief techniques used for differentiating proteins in discovery proteomics experiments. iTRAQ is one of the most common isobaric labeling strategies in use at present;[13] after digestion with trypsin, peptides from each sample are labeled with different chemicals of nearly identical mass on their N-termini and lysine side chains, and then peptides from different samples can be mixed together for LC−MS/MS analysis. Ions of a particular peptide from multiple samples will be isolated and fragmented together, each producing a characteristic reporter ion at low m/z in the MS/MS scan. The intensities of these reporter ions reflect the quantity of that peptide in the source samples. Spectral counting is a label-free strategy by which the number of spectra attributed to a particular protein from an individual sample can be compared to the number of spectra observed for that protein in another sample of similar complexity.[14] Alternatively, data of this type may be analyzed by integrating the extracted ion chromatograms for intact peptide ions drawn from successive MS scans.[15] A protein that matches to intense ions in the first experiment but less intense ions in a second experiment may be judged to be higher in quantity for the first. Both analyses of label-free data, however, rely on the ability to control for higher overall identification rates in one sample than in another.

Measuring the reproducibility of differential proteomics has a noteworthy history. In 2004, a team from SurroMed evaluated the stability of peptide intensities measured from many replicates of a reference plasma sample in LC−MS on a Waters Micromass LCT.[16] This work was extended to encompass isotopic labeling strategies with an AB SCIEX QSTAR XL by Kim et al. in 2007.[17] Old et al. found that spectral counting methods detected differences better and that peptide intensities estimated enrichment ratios better from ThermoFinnigan LCQ Deca data using SCX or gel exclusion fractionation in 2005.[18] A similar 2005 investigation by Zybailov et al. found that spectral counting methods and isotope enrichment methods produced highly correlated results, but spectral counting was more reproducible across their replicates.[19] In 2006 and 2007, the Phillip C. Wright laboratory published two investigations into the reproducibility of iTRAQ quantitation, producing a coefficient of variation of 0.09 and probing sources of variations in iTRAQ reporter ion intensities and documenting the value of replicates in these studies.[20,21] Patel et al. applied iTRAQ labeling and label-free mass spectrometry-based proteomics approaches to the proteome of the bacterium *Methylocella silvestris*, and the results showed good agreement between the iTRAQ experiment and the label-free approach for relative quantification.[22] In 2012, Wang et al. reported a comparative study of iTRAQ and label-free LC-based quantitative proteomics approaches using two *Chlamydomonas reinhardtii* strains.[23] The comparison indicated that the label-free method provided better quantitation accuracy for high fold change ratios and identified 40% more proteins, but the iTRAQ method showed better quantitation accuracy and reproducibility. In a 2013 study of human adenovirus infections by Trinh et al.,[24] a label-free method showed higher levels of protein up- or downregulation in comparison to iTRAQ-labeled samples, and data suggested that the label-free method was more accurate than the iTRAQ method. Megger et al. compared label-free and label-based strategies for proteome analysis of hepatoma cell lines,[25] showing that the label-free approach outperformed TMT methods regarding proteome coverage, but the label-free method was found to be less accurate than TMT approaches. In addition to these comparative studies, the literature is also rich with papers on the advantages and disadvantages of different strategies of quantitative mass spectrometry in proteomics.[26−28]

In general, these technology assessments have been limited in the number of replicates and instruments they include, and the time frame from first to last replicate has been relatively short. The CPTAC CompRef experiments, on the other hand, span six different instruments and range up to 10 months in duration. Because the same pair of pooled samples is represented in each experiment, the data are informative for technical replicate variation rather than biological replicate variation.[29] The data supporting this study were generally drawn from interstitial experiments; the WHIM2 and WHIM16 samples were run as periodic quality controls in a longer series of TCGA samples. This study evaluates these data to answer the following key questions:

- Are differential proteomics experiments repeatable for a particular workflow?
- Are difference lists reproducible in comparing two different workflows?
- Do the various differential proteomics technologies see the same biology?

DOI: 10.1021/acs.jproteome.5b00859
J. Proteome Res. 2016, 15, 691−706

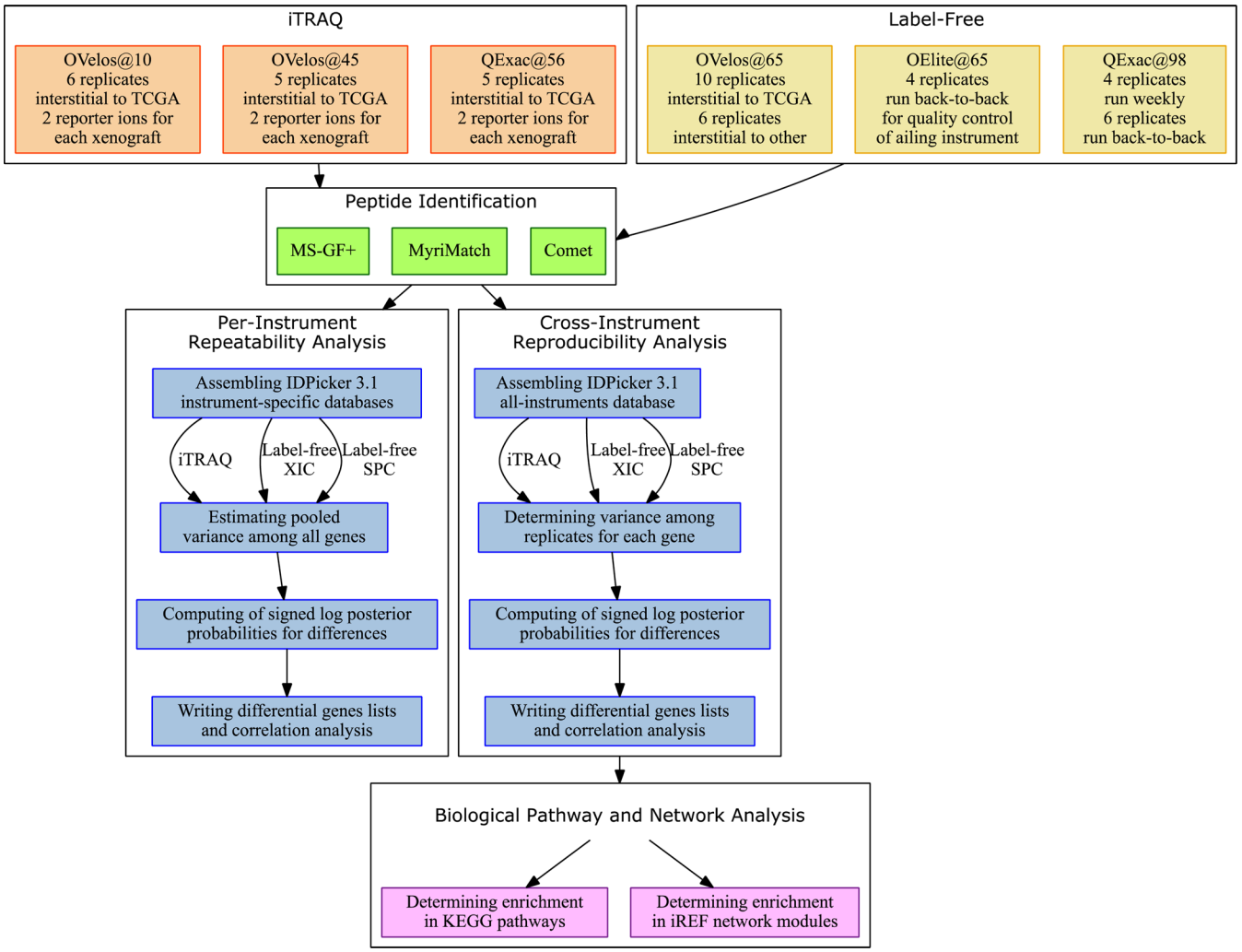


Figure 1. High-level view of the bioinformatics pipeline employed in this study. Six instruments analyzed the same xenograft pair. Label-free sets were processed once by a spectral counts method and once by extracted ion chromatograms. While instrument-specific assemblies of PSMs were used for repeatability analysis, an all-instrument assembly was analyzed for reproducibility and biological pathway and network enrichment.

EXPERIMENTAL PROCEDURES

Data Generation

Tissue Source. Patient-derived xenograft tumors from established basal (WHIM2) and luminal-B (WHIM16) breast cancer intrinsic subtypes[30,31] were raised subcutaneously in 8 week old NOD.Cg-Prkdc^{scid} Il2rg^{tm1Wjl}/SzJ mice (Jackson Laboratories, Bar Harbor, ME) as previously described.[12,32] These tumors have significantly different gene expression and proteomic signatures[12] that are related to their intrinsic biology and endocrine signaling. Tumors from each animal were harvested by surgical excision at approximately 1.5 cm^3 with minimal ischemia time by immediate immersion in a liquid nitrogen bath. The tumor tissues were then placed in precooled tubes on dry ice and stored at −80 °C. A tissue pool of cryopulverized tumors was prepared in order to generate sufficient material that could be reliably shared and analyzed among multiple laboratories.

Briefly, tumor pieces were transferred into precooled Covaris Tissue-Tube 1 Extra (TT01xt) bags (Covaris no. 520007) and processed in a Covaris cryoPREP CP02 device using different impact settings according to the total tumor tissue weight: <250 mg = 3; 250−350 mg = 4; 350−440 mg = 5; and 440−550 mg = 6. Tissue powder was transferred to an aluminum weighing dish (VWR no. 1131-436) on dry ice, and the tissue was thoroughly mixed with a metal spatula precooled in liquid nitrogen. The tissue powder was then partitioned (∼100 mg aliquots) into precooled cryovials (Corning no. 430487). (Note that cryopulverized tissue will melt if it is transferred to a plastic weighing boat.) All procedures were carried out on dry ice to maintain tissue in a powdered, frozen state. Each site processed the tissue powder by independent protocols; protein denaturation and digestion were not controlled by a CPTAC-wide Standard Operating Procedure.

Analytical Chemistry. Six instruments generated a total of 1095 LC−MS/MS experiments from the WHIM2 and WHIM16 samples; Thermo Fisher LTQ Orbitrap Velos, Orbitrap Elite, and Q-Exactive models were included. Three instruments (OVelos@10, OVelos@45, and QExac@56, where the number represents an anonymized CPTAC institution) produced iTRAQ 4plex experiments.[13] In each 4plex, two channels represented each of the two samples. Three other instruments (OVelos@65, OElite@65, and QExac@98) produced label-free experiments in which WHIM2 and WHIM16 were analyzed separately. All sites except site 65 employed HCD for data-dependent MS/MS production, with

DOI: 10.1021/acs.jproteome.5b00859
J. Proteome Res. 2016, 15, 691−706

Orbitrap measurement of fragment ions. Both OVelos@65 and OElite@65 employed CID instead, measuring fragments in the quadrupole ion trap. All raw data files are available from https://cptac-data-portal.georgetown.edu/cptacPublic/ under the name CompRef. All files were subjected to quality assessment through the QuaMeter IDFree mode.[33] The resulting tables are available as a Microsoft Excel spreadsheet in the Supporting Information. Instrument-specific details are provided in Supporting Information Method A.

Bioinformatics and Biostatistics

Proteomic Identification. To maximize information yield from these data, three different search engines were applied to each LC−MS/MS experiment (Figure 1). MyriMatch 2.1.138,[34] MS-GF+ versions 9630 and 9979,[35] and Comet version 2014.01[36] searched a FASTA sequence database containing NCBI RefSeq human (32 799 sequences, downloaded Sept 7, 2011), NCBI RefSeq Mouse (29 617 sequences, downloaded March 4, 2011), and the porcine trypsin sequence. Each sequence was reversed in silico by the search engines for target-decoy estimation of error rates; the decoy sequences from all three search engines were denoted by an XXX prefix. Fully tryptic and semitryptic peptides were allowable matches. In all cases, a precursor mass tolerance of 20 ppm was applied, allowing for an error of one neutron in monoisotope selection. Site 65 data were measured in the ion trap, so fragment ions were allowed to vary by up to 0.5 m/z in MyriMatch or within a one m/z bin by Comet and MS-GF+. For all other sites, fragments were required to fall within 20 ppm of expected m/z by MyriMatch, within a 0.01 m/z bin by Comet, or within the HCD model of MS-GF+. Post-translational modifications expected a mass shift of 57.021464 Da on all Cys, and all algorithms allowed Met to gain 15.994915 Da through oxidation as a dynamic modification. MS-GF+ also allowed for acetylation of N-termini (+42.010565 Da). Up to two missed cleavages were permitted, and up to three dynamic PTMs were allowed per peptide−spectrum match (PSM). In searches of iTRAQ data sets, a mass shift of 144.102063 Da was assumed to be found at both N-termini and Lys residues (and the iTRAQ protocol was applied in MS-GF+). Comet and MyriMatch results were exported to pepXML files,[37] whereas MS-GF+ results were written to mzIdentML format.[38]

Identification scores were translated to q-values[39] in IDPicker 3.1, build 599.[40] Data from each instrument were drawn into separate assemblies, at first. The PSM FDR (false discovery rate) was limited to 0.5%, with the FDR being estimated by doubling the number of decoy hits and dividing by the total passing a threshold. The number of spectra required per protein was then increased from two until the empirical protein FDR had fallen below 5%. For each site, the number of spectra required per protein was as follows: OVelos@10 = 7, OVelos@45 = 3, QExac@56 = 19, OVelos@65 = 4, OElite@65 = 2, and QExac@98 = 4. IDPicker assembled protein groups using parsimony rules that eliminated subset and subsumable proteins from the list and grouped indiscernible proteins.[41] IDPicker matched each protein name from RefSeq to an HGNC (HUGO Gene Nomenclature Committee) gene symbol for human or to a MGI (Mouse Gene Informatics) gene symbol for mouse. If neither was available for a protein name, then the software used the name of the protein preceded by Unmatched. The software allows for graphical or script-driven extraction of quantities in the form of spectral counts (SPC) or intensities, organized for this study by gene rather than protein groups. If spectra could be attributed to multiple transcripts of a particular gene, then reporting by gene would consolidate all of the PSMs to the single gene source, thus reducing the number of hypotheses to be tested in finding expression differences. Two kinds of intensities can be used: iTRAQ reporter ion values can be summed across all PSMs for each channel for each gene or extracted precursor ion chromatograms (XIC) in label-free experiments can be integrated and summed across all PSMs for each gene.[15,42] In iTRAQ, IDPicker computes sums rather than ratios, though users can subsequently compute ratios from the summed intensity for each gene. This places a greater weight on genes with intense PSMs than on those with weak signals. When inferring XIC intensities, the software was not directed to integrate precursor intensity in the absence of a confident PSM through retention time mapping due to the large numbers of LC−MS/MS experiments in this study. The software exported these quantitative tables by the command line idpQuery tool from the IDPicker SQLite databases.

For cross-site and biological pathway and network analysis, a conjoint assembly of data from all sites was created. Since three different identification sets were considered for each of 1095 LC−MS/MS experiments, the conjoint assembly spanned 3285 pepXML/mzIdentML files. As in the individual instrument assemblies, a PSM FDR of 0.5% was applied. Requiring 20 spectra per gene pushed the empirical protein FDR below 5%.

Statistical Differentiation. The raw abundance data, either spectral counts or iTRAQ and XIC intensities data, need to be normalized before any statistical analysis. Variation in sampling handling and sample loading across experiments can lead to variation in the gene abundance and thus introduce systematic biases into the differentiation study. To correct for the overall experiment-wise difference, the spectral counts were normalized by the overall abundance of each experiment using the total spectral counts in an experiment.[43] iTRAQ and XIC intensities data were normalized by the median intensities in an experiment.

A Bayesian hierarchical model was used to detect differential genes. The hierarchical construction allowed both single replicate and multireplicate comparisons. Though the QSpec method[43] was originally employed in label-free proteomics data, the modeling framework can be similarly applied to model the reporter ion intensities from iTRAQ data. To compare the log ratios between WHIM2 and WHIM16, a linear mixed effect regression model was constructed for XIC and iTRAQ intensities and a generalized linear mixed effect model for spectral counts. A detailed description of the method is presented in Supporting Information Method B. Fortran source code for this differentiation is available at http://homepages.uc.edu/~wang2x7/Research.htm.

The estimation of gene variance was a key factor differentiating the per-instrument and cross-instrument analyses (Figure 1). In one-versus-one replicate comparison, the normal distribution was assumed for log XIC and iTRAQ intensities. All of the nondifferential genes were assumed to have the same variance, and the constant variance was inferred from the pool of nondifferential genes detected by the model. This was a strong assumption that the variances are the same for all intensity levels, but a check on the fitted model confirmed that the equal variance assumption is approximately satisfied (see Supporting Information Figure S1). For spectral count data, the Poisson distribution was assumed, and thus the variance was set to equal the mean. A check on the data showed that

DOI: 10.1021/acs.jproteome.5b00859
J. Proteome Res. 2016, 15, 691−706

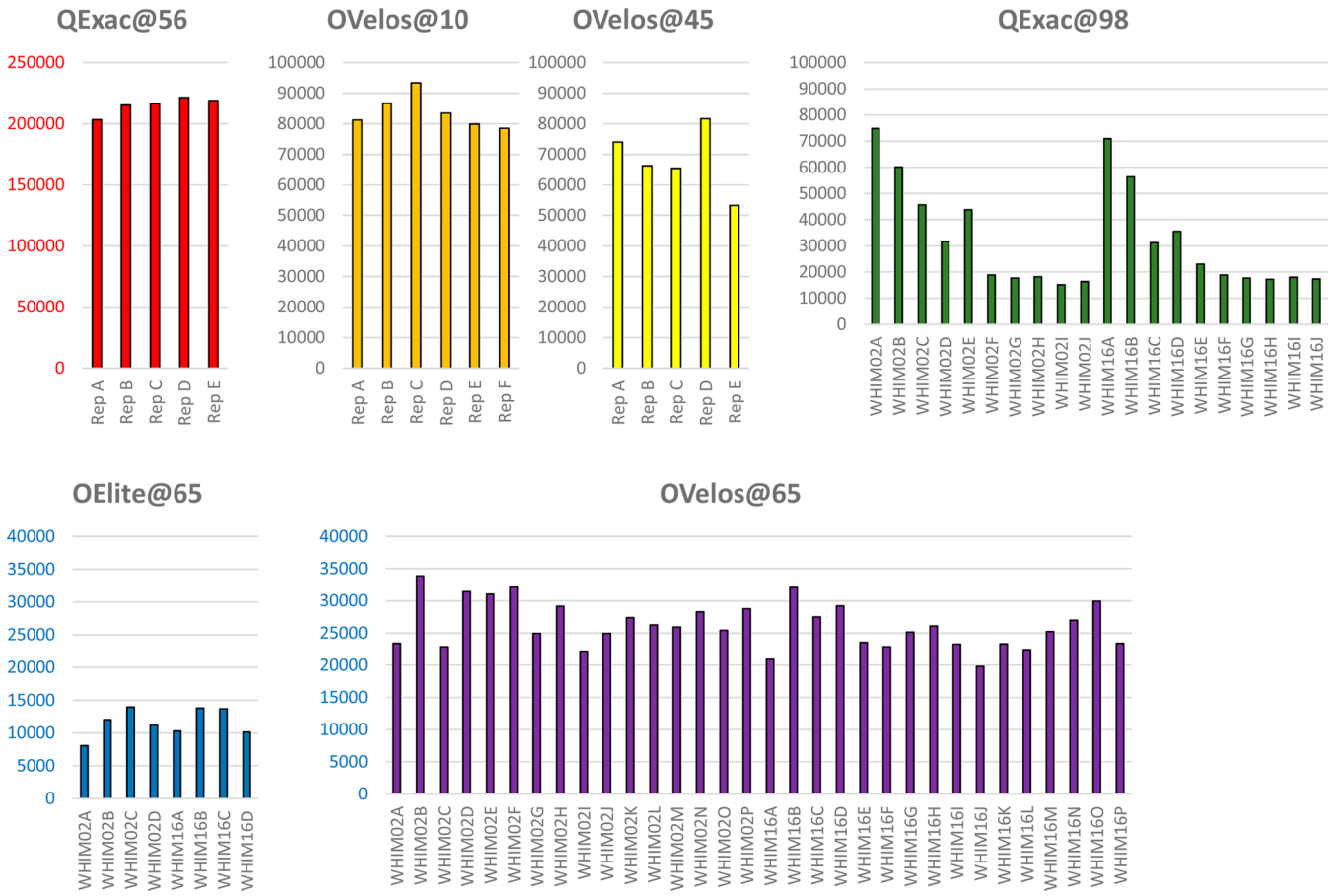


Figure 2. Number of identified distinct peptides per sample/replicate. When data from each instrument were assembled separately with 0.5% PSM FDR and <5% empirical protein FDR, identification sensitivity varied considerably by site. Ailing instrument OElite@65 yielded the lowest sensitivity by far, whereas QExac@56 produced a remarkable number of identifications. Label-free instruments ran WHIM2 and WHIM16 separately, whereas iTRAQ instruments combined these samples into a single 4plex.

overdispersion may have affected the data, leading to an underestimate of variance. A negative binomial model was fitted, and its results were very similar to those under a Poisson assumption. By contrast, in the multiple replicates comparison used for the cross-instrument reproducibility evaluation, gene variances were estimated individually for both the XIC/iTRAQ intensities and the spectral count cases.

For all genes observed from the xenograft samples, only genes for which the GeneID contained at least one human gene name were considered. Results on low-abundance genes were particularly unstable. To avoid random discovery, we set up a requirement of minimum evidence for a gene to be included in statistical differentiation analysis. Following Bing Zhang's minimal spectral count requirement,[1] we required a minimum of 1.4 spectra, on average, for multiple replicate comparison. For example, in OVelos@65 (A–J), 28 spectra were needed before a gene was evaluated for differentiation (1.4 × 10 replicates × 2 samples). In one-versus-one comparison, we set the threshold as 2 spectra per replicate.

Additionally, handling zero values for spectral counts, XICs, or iTRAQ intensities deserves special consideration as a zero may be interpreted as either a missing value (NAs) or a measured zero. Genes that had only zero spectral counts in either WHIM2 or WHIM16 were excluded (this may eliminate some number of genuinely infinite differences, though most will be low-information genes). These spectral count requirements were applied prior to iTRAQ or XIC differentiation as well. For genes with a mixture of zero and positive values in abundance, the handling of spectra counts data differs from that of iTRAQ or XIC data. In modeling of spectral counts, zeros are all treated as true zero counts in the Poisson distribution.[43−45] For the continuous iTRAQ and XIC intensity data, we noticed that zeroes mostly likely came from sources other than biological reasons, such as intensities below the detection limit or failure in peak detection. Treating zeros the same as positive values in the modeling may severely violate the log-normal distribution assumptions on the continuous data. We thus treated zeroes in iTRAQ and XIC as missing values; they were not used in differentiation analysis. The difference in modeling zero values may lead to contradictory differentiation results from using spectral count data as noted in the Results and Discussions; one potential solution is to model both zero and positive parts in differentiation analysis using zero-inflated models.[46]

Biological Pathway and Network Evaluation. Determining the intersection of differential genes in the context of biology emphasized a subset of the overall data. The analysis included these data sets: OVelos@10 (iTRAQ, 6 replicates), OVelos@45 (iTRAQ, 5 replicates), QExac@56 (iTRAQ, 5 replicates), OVelos@65 (SPC, first 10 replicates), OVelos@65 (XIC, first 10 replicates), QExac@98 (SPC, first 4 replicates), and QExac@98 (XIC first 4 replicates). The OElite@65 data were excluded because the instrument was known to be in deficient operating order at the time the 4 replicates were collected. The last 6 replicates of OVelos@65 were excluded

DOI: 10.1021/acs.jproteome.5b00859
J. Proteome Res. 2016, 15, 691−706

because they were interstitial with a different type of sample than the first 10 replicates. The final 6 replicates of QExac@98 were excluded because ion transmission through the quadrupole was considerably lower during data collection.

To perform pathway enrichment, we eliminated genes that were mapped only to the mouse genome. When both human and mouse gene names were associated with spectra, only the human name was retained. In cases where expression could be attributed to multiple isoforms, we chose to keep the first isoform listed, and when a natural read-through event was indicated (two gene symbols separated by a hyphen), only the first symbol was retained. Though these heuristic filters could conceivably introduce biases, we note that these rules were applied only to 20 multiple isoform proteins and 6 read-through proteins of the 8126 genes considered (0.25%).

We calculated enrichment in pathway sets using KEGG pathways.[47] Fisher's exact test[48] was used to examine the statistical enrichment of those genes called as more highly expressed in WHIM2 or in WHIM16 for each of the 188 pathways relative to all those genes not in the specified pathway. The genes upregulated for each of the two xenografts were handled separately to avoid losing statistical power.[49] Resulting p values were corrected using the Benjamini–Hochberg procedure,[50] and we considered multiple-test-corrected p values below 0.05 to be significant.

We also interpreted the genes expressed significantly highly in WHIM2 or in WHIM16 against the context of a global protein–protein interaction network using NetGestalt.[51] The network used in this article was downloaded from iRef.[52] We first identified the hierarchical modules from the iRef network based on the NetSAM algorithm.[51] Then, based on Fisher's exact test, we calculated the p values of enrichment for all modules in each differentially expressed gene set identified by the seven data sets. The p values were corrected by the Benjamini–Hochberg procedure. Applying an FDR limit of 5%, we finally identified WHIM2- or WHIM16-specific enriched modules for each of seven data sets. All of the above procedures were implemented in the NetGestalt web tool (http://www.netgestalt.org). We also used NetGestalt to covisualize the enriched modules of seven data sets for identification of the functional consistency.

RESULTS AND DISCUSSION

Identification Sensitivity and Differential Genes

The six instruments included in this study varied considerably in labeling and fractionation strategy, LC gradients and mass spectrometry instruments, and replicate schedule and data replacement policy. Correspondingly, large differences appeared in rates of MS/MS collection. The two Q-Exactive instruments in the study produced high maximum sustained rates of MS/MS collection: 9.10 Hz for QExac@56 and 9.28 Hz for QExac@98, exceeding the best rate of acquisition in Orbitraps for MS/MS scans (4.35 Hz for CID in OVelos@65 or 3.89 Hz for HCD in OVelos@45).

Translating these spectra to identified PSMs produced a less clear distinction based on instrument type (Figure 2). While OVelos@65 produced relatively low sensitivity in terms of distinct peptides, averaging 58 028 identified spectra (26 250 distinct peptide sequences) from its 15 fractions per sample, it showed an encouraging degree of consistency across 10 months of continuous operation. QExac@98 used essentially the same amount of time per sample, though split to only five fractions, and its first four experiments for the two samples identified an average of 99 916 spectra (53 045 distinct peptide sequences); unfortunately, this rate tapered off considerably, and by the time replicates six through ten were collected, identification rates were much lower. The instrument operator reported that the ion transmission through the quadrupole worsened significantly in later runs. Viewing these results instead in terms of discernible protein groups flattened the differences among instruments. The cumulative counts of distinct proteins at <5% empirical protein FDR were OVelos@10 = 11 428, OVelos@45 = 11 603, QExac@56 = 15 655, OVelos@65 = 8640, OElite@65 = 4199, and QExac@98 = 10 435. Note that these counts included both mouse and human proteins since a xenograft incorporates proteins from both species.

Rather than emphasize distinguishable protein groups, we projected PSMs to the genes from which those peptides derived. This projection ensured that spectra associated with peptides shared among multiple isoforms transcribed from the same gene would be counted only once.[53] In addition, reporting spectra per gene makes it simple to relate proteomics data to biological pathways and networks that are described by the use of HGNC identifiers. When multiple genes represented exactly the same sets of peptides, they were reported as an indiscernible gene group.

Repeatability of identification scaled with the numbers of spectra identified for each gene group.[6] The 16 replicates for OVelos@65 were evaluated for identification repeatability. First, the genes were separated into exponentially sized bins: 4–7, 8–15, 16-31, 32–63, 64–127, 128–255, 256–511, and 512–1023 spectra across all replicates of both WHIM2 and WHIM16. This excluded 585 proteins with disproportionately high spectral counts, such as the maximum for mouse albumin: 63 734. An average of 998 genes fell into each bin, with the highest spectral count bin including 610 genes. For each bin, we computed the average number of replicates in which the genes produced any identifiable spectra (out of 16 replicates). For WHIM2, the averages were 2.57, 4.96, 8.29, 11.71, 13.98, 15.18, 15.58, and 15.83. WHIM16 produced very similar averages: 2.12, 3.96, 7.00, 10.37, 13.23, 15.12, 15.72, and 15.93. Of the 8566 genes OVelos@65 identified, 3616 were identified universally in WHIM2 and 3306 were identified universally in WHIM16. A total of 2730 genes were identified universally across both WHIM2 and WHIM16. When a similar analysis was performed for the iTRAQ experiments of OVelos@10, the bins began at 7 spectra because that was the minimum number required to control the protein FDR. Even in the set of genes with only 7–13 spectra in evidence, the identification repeatability rate was quite high: 5.12 of 6 replicates. Once genes had produced 28 or more spectra in aggregate, the average identification repeatability was close to 6.00. Of the 11 284 genes that OVelos@10 identified, 9822 were identified in all six replicates.

Reports from IDPicker were used to find two sets of differential genes from each label-free replicate, one based on spectral counts (SPC) and another based on extracted ion chromatograms (XICs) and to find one set of differential genes from each iTRAQ replicate (see Experimental Procedures: Statistical Differentiation). This part of the study sought to determine the reproducibility of differentiation for individual experiments rather than across multiple replicates. For label-free experiments, this would mean comparing the spectral counts or precursor intensities from a set of fractions representing one analysis of WHIM2 to the comparable data

DOI: 10.1021/acs.jproteome.5b00859
J. Proteome Res. 2016, 15, 691–706

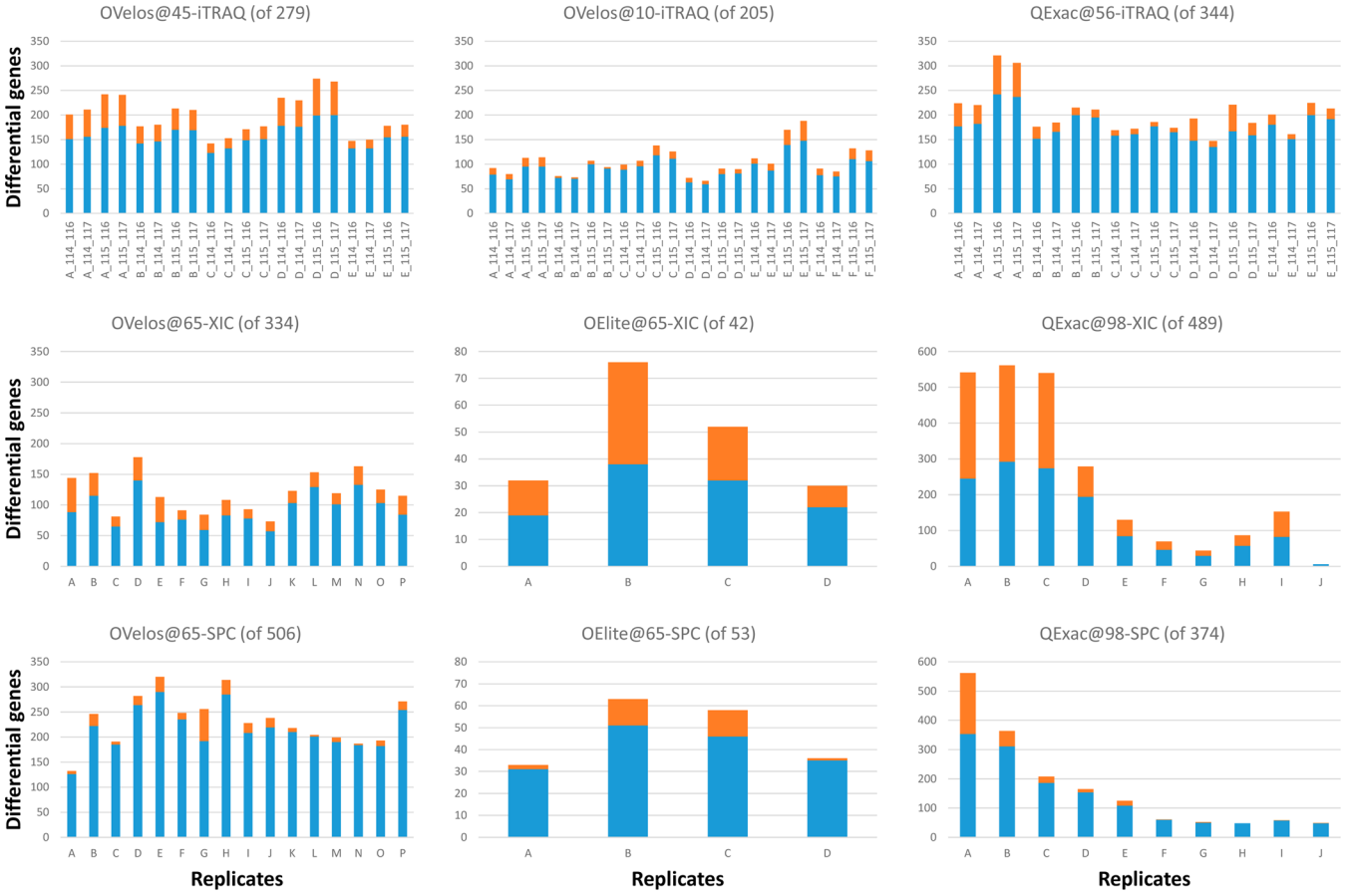


Figure 3. How many of the differential genes from each experiment are confirmed by at least one other replicate experiment? Blue differential genes are found in common by another experiment in this instrument, whereas orange ones are unique to a single replicate. For iTRAQ, the confirmation must come from a different set of LC−MS/MS data altogether. Note that instruments that produced more replicates were likely to have a higher proportion of common differential genes by random chance.

from a set of fractions representing one analysis of WHIM16 (analyses comparing multiple replicates are performed below). In these iTRAQ experiments, each experiment inherently produced duplicate information for WHIM2 and for WHIM16; this analysis, however, compares data from a single channel of WHIM2 to a single channel of WHIM16.

To determine differences, we chose strategies that have been evaluated in the peer-reviewed literature. We opted to employ a pooled variance model for the single-replicate comparisons; similar strategies have been implemented as local-pooled-error,[54] empirical Bayes,[55] and other tests for microarray experiments. The approach embodied in IDPicker for extracting precursor ion intensity has been published separately,[42] including an evaluation of the rate of false attribution of intensity to a peptide. Because the volume of data was so extensive, we opted not to take advantage of the ability to align retention time for fractions across replicates. As a result, a peptide that was present in both replicates contributed a spectrum count or precursor intensity when it was identified in a replicate but contributed no count or intensity when it was not identified in a replicate. In all analyses, evidence is summed at the level of the gene that gave rise to proteoforms[56] to which counts or intensity were identified. Working with ratios is certainly more common in iTRAQ analysis, but IDPicker 3.1 was capable of intensity output only, not ratios, summing the precursor intensities for all PSMs supporting the gene identification. Submitted work by Jian-Ying Zhou demonstrated that this method is equivalently resistant to false positives while producing high data set correlations. This determination was corroborated by Carrillo et al.,[57] who concluded that "Our results demonstrate that the sum of intensities and total least squares algorithms provided the most accurate estimates of protein abundance for a wide range of simulated and experimental conditions. The commonly used average of ratios algorithm consistently provided estimates with the highest errors."

The fraction of observed genes found to be differential at a 5% FDR varied strongly by instrument. The iTRAQ instruments yielded low average differential gene fractions: OVelos@10 = 1.6%, OVelos@45 = 3.4%, and QExac@56 = 2.5%; our investigation showed that the Bayesian model expected a high degree of variance in each gene's reporter ion intensities when no biological difference was present (the model assumed that at least half of genes would not differ between the two xenografts). Label-free data sets were analyzed twice: once solely on the basis of spectral counts (SPC) and once solely on the basis of extracted ion current (XIC) from precursor ions in MS scans. XIC analysis found somewhat higher fractions of genes to be differential: OVelos@65 = 5.0%, OElite@65 = 3.5%, and QExac@98 = 7.7%. QExac@98 produced a wide range of differential gene fractions, reflecting the drop in identification sensitivity (Figure 2); the first four replicates averaged 12.9% differential genes, whereas the last six averaged 4.3%. Shifting to SPC analysis instead (which shifts to a Poisson

DOI: 10.1021/acs.jproteome.5b00859
J. Proteome Res. 2016, 15, 691−706

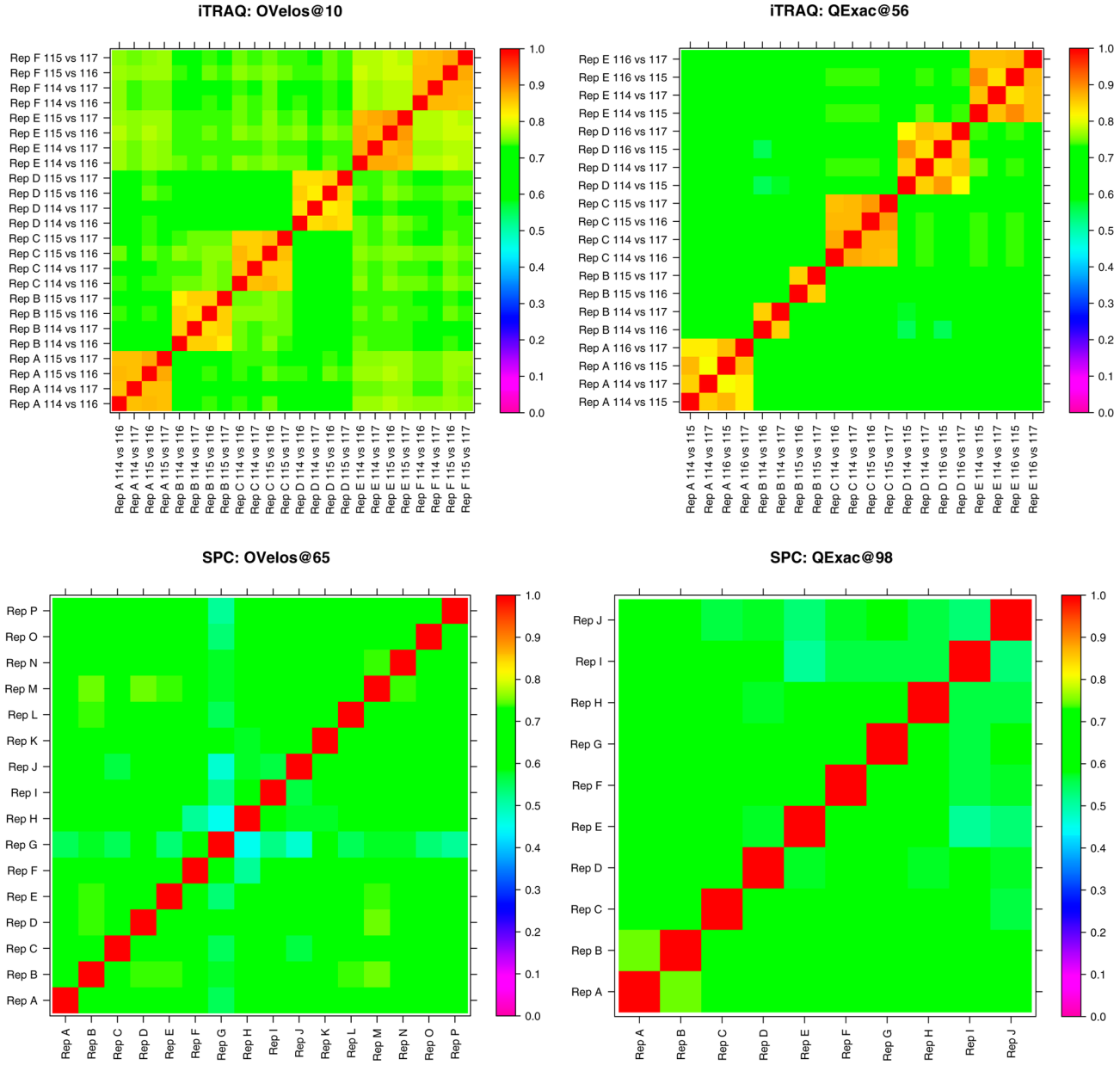


Figure 4. Rank correlations compare the ordering of genes by signed posterior probabilities. OVelos@10 illustrates the extent to which differential probabilities within an iTRAQ 4plex are more similar than those across multiple 4plexes. The QExac@56 shows an exception to this behavior in replicate B. Spectral count-based differentiation produces similar overall correlations to iTRAQ, though without the benefit of being able to compare within LC−MS/MS experiments. Replicate G from OVelos@65 correlated more poorly. The declining sensitivity of identification in QExac@98 gave low correlation values other than the one between the first two replicates.

expectation of variance in the statistical model) altered the differential gene fractions: OVelos@65 = 9.8%, OElite@65 = 3.5%, and QExac@98 = 5.3% (8.2% for the first four replicates, and 3.4% for the last six). We note that the fraction of genes found to be differential in the OVelos@65 set was larger under SPC analysis, whereas the fraction was higher in QExac@98 under XIC analysis; this may reflect the optimizations in methodology applied at each site.

Per-Instrument Repeatability

Two different strategies were applied to determine the repeatability of differential gene detection among experiments from a given instrument. The first simply asked which differences are consistent with those from at least one other replicate experiment on the same instrument; these overlaps were also evaluated through Cohen's kappa statistic. The second compared the rankings of genes on an axis from highly upregulated in WHIM2 to unchanging to highly upregulated in WHIM16. Both analyses compared data from a single iTRAQ channel to those from another single iTRAQ channel, despite the collection of duplicate data in iTRAQ 4plexes. This decision was intended to reflect the realities of large-scale iTRAQ experiments, where duplicate data collection is rare. Certainly, including biological replicates within experiments is preferred.

Figure 3 visualizes the set of differential genes from each replicate on each instrument. Differential genes were called common and colored blue if they were also found as differences in other experiments from the same instrument. They were labeled unique and colored orange if they were found in only

DOI: 10.1021/acs.jproteome.5b00859
J. Proteome Res. 2016, 15, 691−706

one replicate from this instrument. All four possible comparisons were generated from the duplicate iTRAQ experiments; both the low and high m/z channels for WHIM2 were compared to both the low and high m/z channels for WHIM16. Agreement in iTRAQ experiments was defined somewhat more stringently than in label-free experiments to reflect the use of this technology in experiments spanning many 4plexes. Consider the case where WHIM2 was presented in duplicate on channels 114 and 115, while WHIM16 was reported on the 116 and 117 channels. A comparison between channels 114 and 116 has mutual information with a comparison between channels 114 and 117. Even a comparison of 114/116 versus 115/117, however, will have shared information because they would be based on the same set of PSMs. As a result, common differential genes from iTRAQ experiments were required to be confirmed by altogether separate LC−MS/MS experiments from the same instrument.

The extent to which Figure 3 is dominated by common differential genes is an encouraging development for this field; the average percentage of differences in common for each instrument was as follows: OVelos@10 = 88%, OVelos@45 = 81%, QExac@56 = 87%, OVelos@65-XIC = 78%, OElite@65-XIC = 61%, QExac@98-XIC = 63%, OVelos@65-SPC = 93%, OElite@65-SPC = 88%, and QExac@98-SPC = 91%. The ailing OElite@65 generated a large proportion of unique differences, but the overall numbers of differential genes were quite low, and only three other replicates could confirm any particular difference. QExac@98 also presented a fairly large number of unique differential genes, but it seems clear that the falloff in sensitivity for the final six replicates was costly. The high degree of agreement seen in the latter replicates for this instrument (E−J) was reassuring in that while reduced sensitivity led to shorter lists of differences it did not compromise reproducibility.

A more restrictive analysis, based on Cohen's kappa, compared the genes declared to be differential in pairs of experiments drawn from each instrument. Essentially, this is built around a contingency table; how many genes are differential in A and differential in B, how many genes are differential in neither A nor B, and how many genes are found to be differential in one but not the other. In computing this metric, we included only those genes that appeared in both experiments. A value of 1.0 implies perfect agreement. The distribution of these kappa values appears in Supporting Information Figure S2. The iTRAQ instruments produced median kappas in the range of 0.49 to 0.57, with maximum values universally higher than 0.88 resulting from comparisons with a channel in common. For all three label-free instruments, the kappa values were substantially higher for SPC analysis than for XIC analysis (a difference of 0.18 to 0.34), with median kappas of 0.50 for QExac@98-SPC, 0.58 for OVelos@65-SPC, and 0.74 for OElite@65-SPC.

Signed log posterior probabilities were produced for each gene in each replicate, with the magnitude of the negative log posterior probabilities indicating the significance with which a gene may be called differential and the sign marking the orientation of that potential difference. The ranks of genes on these lists can then be compared in a Spearman rank correlation. Figure 4 shows four example plots of correlation coefficients for signed log posterior probabilities from each replicate against the values from each other replicate. The average number of genes included in each instrument for differentiation was as follows: OVelos@10 = 6571, OVelos@45 = 5877, QExac@56 = 8880, OVelos@65-SPC = 2381, OElite@65-SPC = 1344, and QExac@98-SPC = 2643 (XIC counts were only slightly different).

Data from OVelos@10 differentiate the correlation values within a 4plex from those between different 4plex replicates. Within a 4plex, the correlation coefficient did not fall below 0.8. Correlation coefficients between 4plexes, however, never rose as high as 0.8 but rarely fell below 0.7. This disparity may result in large part from the fact that intensities within a 4plex are drawn from exactly the same set of PSMs for a given gene, rendering them more comparable than a case where intensities are drawn from entirely different PSMs (and likely distinct sets of peptide sequences) as disparate sets of LC−MS/MS experiments are compared. QExac@56 generally follows the same pattern of behavior, with an even greater difference between internal-to-4plex and external comparisons. The behavior for replicate B, however, bears closer attention. The data from channels 114 and 115 were considerably different, and the external correlations for channel 114 were lower against all other replicates. The result is a reminder that a great many different factors must all go well for reproducible proteomic differentiation experiments. (Note, however, that the internal data analysis pipeline by site 56 did not show a reduced correlation for this channel. The internal pipeline used a different search engine, a peptide ratio-based assessment, and a Pearson correlation on log fold change.)

Label-free experiments may be simpler to conduct at the bench, but they are also subject to variation. The 16 replicates from OVelos@65 in Figure 4 reveal a general decrease in correlation for replicate G. Looking back to Figure 3 reveals an increased proportion of unique differential genes for this replicate; replicate G also produced the lowest kappa metric for this instrument in both SPC and XIC analysis. The experimental origin of this variation is unknown. QExac@98 was subject to considerable variation in identification sensitivity (Figure 2), and its correlation values were consistently lower than those for other instruments. Correlation within the earliest replicates is stronger than for later replicates. All of the correlation plots can be found in Supporting Information Figure S3.

Cross-Instrument Reproducibility across Instruments and Technologies

Comparing across instruments made it possible to take advantage of the replicates produced for each instrument. Instead of building differential gene lists from individual instrument assemblies, the cross-instrument evaluation was developed from an assembly that spanned all of the data (Figure 1 and Experimental Procedures: Bioinformatics and Biostatistics: Protein Identification). At an empirical protein FDR of 4.58%, the assembly included 12 741 protein groups supported by a total of 6 980 499 PSMs. The Bayesian model was able to estimate variance per gene through the use of replicates, improving its discrimination power.

The iTRAQ instruments benefited most substantially from the inclusion of replicates. The number of gene groups tested for differentiation in each instrument was as follows: OVelos@10 = 6570, OVelos@45 = 5744, and QExac@56 = 8081. Single-replicate analysis had shown a maximum of 3.4% differential genes for these three instruments, but those percentages rose significantly when the software was able to estimate variance using replicate measurements for each gene: OVelos@10 =

DOI: 10.1021/acs.jproteome.5b00859
J. Proteome Res. 2016, 15, 691−706

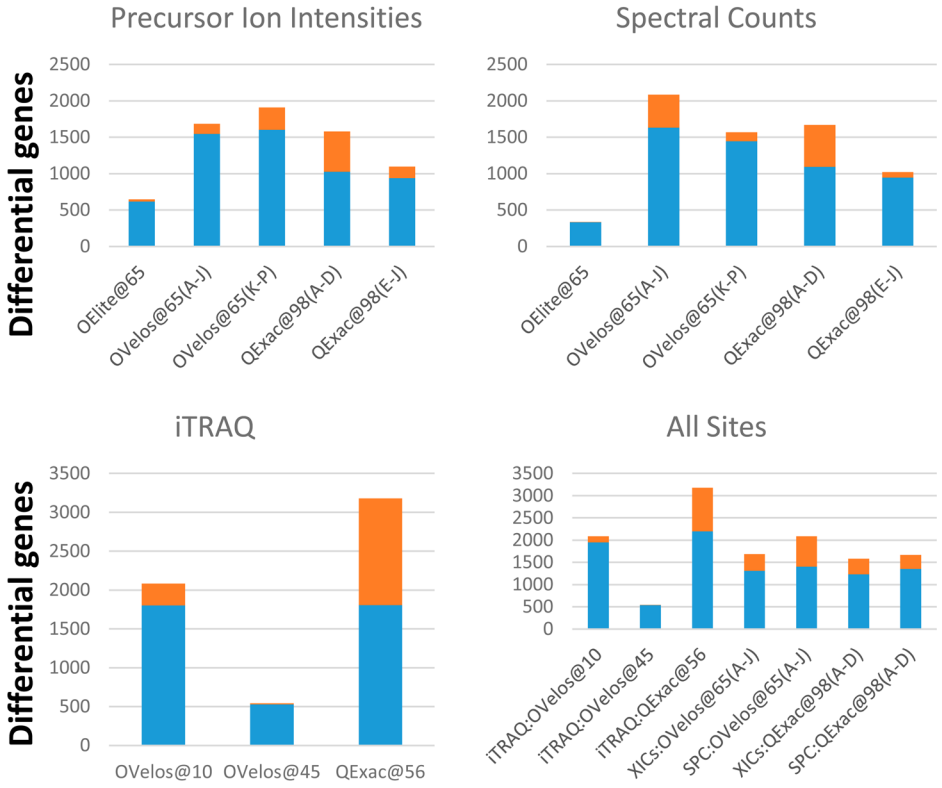


Figure 5. To what extent are the differential genes found for each data set confirmed by other data sets? Those represented by blue were found to be differential in common with another data set in that graph panel, as well. Orange genes, on the other hand, were unique to a particular data set. High identification sensitivity for QExac@56 led to many instrument-specific differences.

31.7%, OVelos@45 = 9.4%, and QExac@56 = 39.3%. When the first three and last three replicates for OVelos@10 were considered separately, the number of differential genes plummeted to 13.3 and 14.2% for the early and late sets, respectively; doubling from three to six replicates greatly improved the statistical power for detecting differential genes. OVelos@45, by contrast, identified the fewest overall genes and yielded the least differentiation among the iTRAQ instruments. Examining the underlying data revealed higher typical variance per gene than that for other sites, and the operators of this instrument noted that each replicate was a full process replicate (repeating digestion, labeling, and mixing) rather than a technical replicate that simply repeated LC−MS/MS on the same vials of peptides. An internal analysis by this site showed good consistency in the genes found to be differential for each replicate (see Supporting Information Figure S4).

The label-free data from OVelos@65 and QExac@98 provided enough replicates to allow for subset comparisons, whereas OElite@65 lagged enough in performance that it was not examined in detail. The first 10 and last 6 replicates from OVelos@65 were split to two sets, reflecting that the former were interstitial with TCGA samples and the latter were interstitial with normal tissue. Similarly, the data from QExac@98 for the first 4 replicates were separated from the last 6 replicates to reflect the different schedules for collecting these two sets. The number of genes observed in each set was as follows: OVelos@65A−J = 3248, OVelos@65K−P = 3219, OElite@65 = 1896, QExac@98A−D = 4893, and QExac@98E−J = 2758. The counts for label-free instruments were taken from the SPC analysis; the number of genes compared by XIC was slightly lower since some precursor ions were not matched. The differential gene fractions by SPC analysis were 64.2 and 48.7% for the early and late sets for OVelos@65, whereas the QExac@98 percentages were 34.1 and 37.0% (note, though, that the later set from this instrument included substantially fewer genes). Analyzing the same data by XIC, though, yielded 51.9 and 59.6% differential genes for OVelos@65, demonstrating that having more replicates is not a guarantee for a higher fraction of differential genes. The XIC analysis for QExac@98 produced 32.3% differential genes in the high-sensitivity early set but 39.8% in the later, less sensitive runs for this instrument.

Since SPC and XIC analyses are generated from the same underlying PSMs, one might expect that their lists of differences are highly similar. However, the sum of precursor ion intensity observed for a given gene may be stable while spectral counts vary between two cohorts, or vice versa. Difference lists from three data sets were interrogated for this purpose: OVelos@65A−J, OVelos@65K−P, and QExac@98A−D. Venn analysis by the Venny 2.02 web tool (J.C. Oliveros) compared the lists of up-in-WHIM16 and up-in-WHIM2 genes for both XIC and SPC analysis (see Supporting Information Figure S5). Of all genes named as up-in-WHIM2 in these three sets, from 47 to 64% were named in both analyses. Of all genes named as up-in-WHIM16 in the three sets, from 53 to 75% were named in both analyses. The two techniques contradicted each other in only one category: genes named as up-in-WHIM2 by SPC but up-in-WHIM16 by XIC. The first 10 replicates of OVelos@65 produced two genes in this conflict, while the last 6 replicates produced another 9. QExac@98 conflicted in six genes. Examining underlying data from these conflicts yielded examples of three phenomena. In the first case, the numbers of identified spectra and the sums of precursor intensity simply point in opposite directions. In the second case, zero counts

DOI: 10.1021/acs.jproteome.5b00859
J. Proteome Res. 2016, 15, 691−706

Table 1. Consistency of Enriched Pathways in Genes Expressed More Highly in WHIM16[a]

pathway	QExac@56	OVelos@10	OVelos@45	XICs: OVelos@65 (A–J)	SPC: OVelos@65 (A–J)	XICs: QExac@98 (A–D)	SPC: QExac@98 (A–D)
glycolysis/gluconeogenesis	×	×	×	×	×	+	×
arginine and proline metabolism	+	+	×	×	×	×	×
valine, leucine, and isoleucine degradation	×	+		×	×	×	×
ECM receptor interaction	×	×	×	×	×		
focal adhesion	×	×	×	×	×		
endocytosis	×	×		×	×	+	
antigen processing and presentation	+	+	×	+	+	×	×
glutathione metabolism	+	+	×	+	+	×	+
amino sugar and nucleotide sugar metabolism	+	×		×	+	+	+
fructose and mannose metabolism	+	+		+	+	×	×
propanoate metabolism	+	+		+	+	×	×
cell adhesion molecules	+	×	×	+	+		
hematopoietic cell lineage	+	×	×	+	+		
regulation of actin cytoskeleton	+	+	+	×	×		
starch and sucrose metabolism	+	+		×	×	+	
butanoate metabolism				+	+	×	×
citrate cycle/TCA_cycle				+	+	×	×
complement and coagulation cascades	×	+		+	×		
fatty acid metabolism				+		×	×
graft versus host disease	+	+	×	+	+	+	
allograft rejection	+	×	+	+	+		
tryptophan metabolism		+	+		+	+	×
lysosome	+	+		×	+		
pentose and glucuronate interconversions	+	+		+	×		
glycosaminoglycan degradation	×	+		+			

[a]× indicates that the pathway was in the top 10 most significant pathways for that data set. + indicates that the pathway was significant (corrected p value < 0.05) in that data set.

were treated as observations in SPC analysis but as missing data in XIC analysis, placing greater weight on replicates that reported nonzero intensities. In the third case, sorting genes by $\log_2$ ratios for FDR determination introduced noise that could be avoided by sorting on regression coefficients instead. Given that the number of differential genes in a list ranged from 928 to 1328, these conflicts never rose above 1%, which we judged to be tolerable in this study.

Venn analysis of differential genes from the three iTRAQ instruments shows that almost every difference found by OVelos@45 was also found by the other two instruments (see Supporting Information Figure S6 for these cross-instrument images). Among the 1403 up-in-WHIM2 genes, only 5% were observed by all three sites. The bulk of the gene differences was found by QExac@56 only (58%), by OVelos@10 only (18%), or by both instruments (18%). A similar pattern appears in the 2269 up-in-WHIM16 genes, with 33% found by QExac@56 only, 9% by OVelos@10 only, 38% found by both these instruments, and 19% found by all three instruments. The data for label-free instruments showed great consistency between early and late sets of OVelos@65 data along with a substantial number of differences found only by QExac@98 (presumably due to its high sensitivity). Whereas more SPC differences were observed by OVelos@65A–J, the XIC analysis favored the later OVelos@65K–P set instead.

Figure 5 visualizes the extent of agreement for differential genes found in one set of replicates versus those found in others. Whether analyzing the data by SPC or by XIC, many of the differences found in the first four replicates of QExac@98 were not reproduced by other label-free sets (perhaps because of the increased sensitivity achieved in these experiments). A similar phenomenon arose from QExac@56; this instrument produced by far the highest PSM sensitivity of the study, and the many additional differential genes inferred from its data could not be confirmed by other data sets. The low number of overall differences found in the OVelos@45 set reflects a relatively low overall gene sensitivity and a low degree of differential discrimination; by contrast, the differences found by this instrument were almost entirely consistent with the differences produced by others.

Reproducibility of Biological Appraisal

The fraction of differential genes confirmed by another instrument in the fourth panel of Figure 5 ranged from 67% for OVelos@65A–J by SPC to 99% for OVelos@45. However, biological insight is often derived by projecting data at the level of the functional pathway or subnetwork. To examine the

DOI: 10.1021/acs.jproteome.5b00859
J. Proteome Res. 2016, 15, 691–706

stability of biological insight that could be derived from data generated on each platform, we calculated the statistical enrichment of genes identified as more highly expressed in WHIM2 or in WHIM16 in each KEGG pathway and each network module generated from the iRef protein−protein interaction network (see Experimental Procedures) using Fisher's exact test with Benjamini−Hochberg multiple hypothesis correction.

Overall, many more proteins were identified as more highly expressed in WHIM16 than were identified as more highly expressed in WHIM2. Four KEGG pathways (antigen processing and presentation, arginine/proline metabolism, glutathione metabolism, and glycolysis/gluconeogenesis) were enriched in up-in-WHIM16 lists from all platforms. The top 10 most significant pathways from each platform showed consistency, though these sets varied (Table 1). Focal adhesion was ranked as first or second most significant by the three iTRAQ data sets, and it was in the top 10 most significant pathways for both partitions of the OVelos@65 data under both SPC and XIC methods, but it was not determined to be significant by QExac@98 by either analysis. Similarly, five network modules were identified as the top 10 most significant modules enriched with up-in-WHIM16 lists from at least four platforms (Table 2 and Figure 6). Only two modules were enriched for the QExac@98 data under the SPC method, and no module was enriched for the QExac@98 data under the XIC method (Figure 6). Specially, modules Level_2_Module_3 (response to wounding) and Level_2_Module_11 (immune system processes) were ranked as first or second most significant modules by four platforms. The up-in-WHIM2 genes were less consistent in both pathway and network analyses, but this was largely due to the fact that far fewer genes were observed to be significantly differential in WHIM2 by any platform, and pathway enrichment overall was much lower.

We also evaluated the possibility of permutation-based FDR estimation for pathway enrichment at a 10% threshold. Under these circumstances, QExac@98 resembled the other instruments to a greater degree, with significant hits occurring in an additional 10 pathways appearing in Table 1 by SPC analysis. The permutation analysis also found 5 additional pathways to be significant in Table 1 by XIC analysis. At this time, NetGestalt has not yet gained the ability to perform permutation-based FDR correction.

CONCLUSIONS

Differential proteomic technologies are complex, and that complexity opens the door to many potential sources of variability. Just as proteomic inventory experiments sample from a large potential pool of peptides, differential proteomics experiments sample from a large potential pool of differential proteins/genes. Given that the underlying inventories from these six instruments ranged more than an order of magnitude in the number of distinct peptides they identified (Figure 2), the degree of conformity in the differential gene lists across instruments was rather surprising (Figure 5). Projecting into KEGG pathways or iREF biological networks further reinforced this agreement among instruments.

Quality control methods and standard operating procedures are a necessary part of clinical proteomics, and sharing results for common samples between laboratories can be a powerful means to evaluate performance in both proteomic inventories and differentiation. The Association of Biomolecular Resource Facilities has been particularly influential in such cross-

Table 2. Consistency of Enriched Network Modules in Gene Expressed More Highly in WHIM16[a]

module	enriched function	QExac@56	OVelos@10	OVelos@45	XIC OVelos@65 (A–J)	SPC: OVelos@65 (A–J)	XICs: QExac@98 (A–D)	SPC: QExac@98 (A–D)
Level_2_Module_3	response to wounding	×	×	×	×	×		
Level_2_Module_11	immune system process	×	×	×	×	×		
Level_3_Module_37	cell-substrate adhesion	×	×	×	×	×		
Level_4_Module_32	microtubule-based process	×	×	×	×	×		
Level_2_Module_2	proteolysis	×	×	×		×		

[a]Enriched function is the most significant GO (gene ontology) biological process term enriched with genes in the module. × indicates that the pathway was in the top 10 most significant pathways for that data set. All modules can be found in NetGestalt (www.netgestalt.org).

DOI: 10.1021/acs.jproteome.5b00859
J. Proteome Res. 2016, 15, 691−706

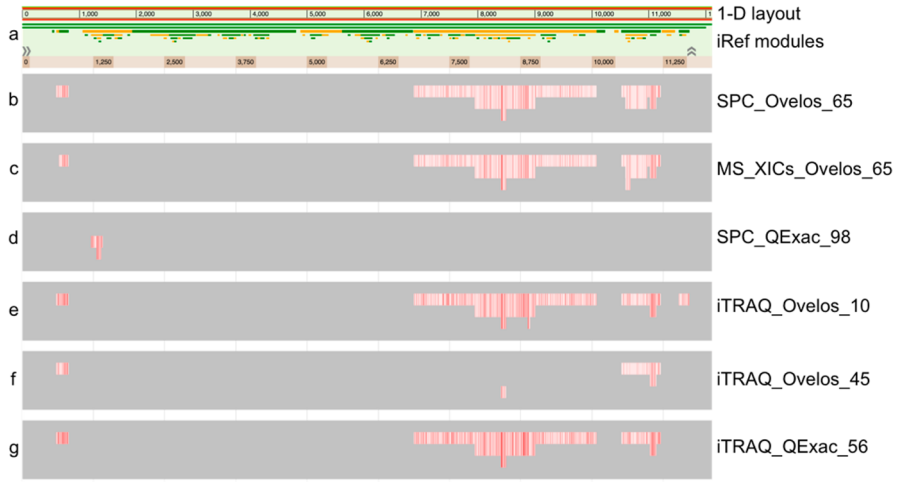


Figure 6. NetGestalt-based analysis of functional consistency for up-in-WHIM16 genes from seven data sets. (a) Proteins in the iRef network are placed in a linear order together with the hierarchical modular organization of the network. Alternating green and orange bar colors distinguish neighboring modules. (b–g) For each data set, the network modules found to be enriched are colored light red. Within these modules, up-in-WHIM16 genes are represented by a dark red stripe.

laboratory studies. When QC methods highlight an LC–MS/MS experiment from a fractionated set as deficient, however, the ideal response is to rerun the entire fractionated set rather than replacing a subset of the fractions. To replace only the problematic fractions can increase the variability of the set by adding in the drift of the instrument as a function of time.

Figure 4 illustrates an important principle for isobaric labeling studies: intensity values within an iTRAQ 4plex replicate were more comparable than values between different sets of replicates because they are drawn from the same set of PSMs. The speed and capacity of these methods clearly exceed those of label-free approaches as a result of multiplexing. Current isobaric reagents can support analysis of the greater than 100 sample comparisons carried out by the CPTAC data production centers through the use of a common pooled sample included in each 4plex or through a randomized block design. The degree of multiplexing possible through use of current isobaric labeling reagents is at 10plex through use of the tandem mass tag (TMT, Thermo) reagents, and further increases in the degree of multiplexing are anticipated as new chemistries and encoding strategies[58] are developed.

To return to the questions that motivated this study, it appears clear that repeatable performance in differential proteomics is feasible, whether isobaric labels or label-free methods are employed. The repeatability of OVelos@65 over 10 months of nearly continuous data acquisition was quite high, though the sixth of 16 replicates was somewhat questionable in the correlation analysis of Figure 4. The diminished performance in the final six replicates of QExac@98 is a clear reminder that instrument performance needs to be comparable for differentiation to be repeatable. The large increase in differential genes for the cross-instrument reproducibility analysis reminds us of the value of multiple replicates for differential proteomics. While this study did not take advantage of full biological replicates for its two samples, these are certainly valuable in most experiments. Using the information from technical replicates for this study greatly increased the number of differential genes by enabling per-gene variability estimates. Technical variation remains a fact of life for all proteomics laboratories; this study, like many others in the field, might have benefited from the use of randomization and blocking designs in addition to its use of replication.[59]

If laboratories deploy different methodologies to analyze the differences between the same two complex samples, then they will assuredly see differences in the gene or protein lists produced by the two technologies. The degree of conformity observed in this study, however, was encouraging. When label-free data were analyzed by spectral counting rather than precursor intensity, the differences yielded a high degree of overlap. When iTRAQ rather than label-free methods are deployed, the differential genes were again quite similar. These overlaps suggest a degree of maturity in proteomic methods that has grown through years of development along multiple tracks.

At base, biologists need to know that differential proteomics technologies can produce meaningful results. Our assessment showed that biological pathway and network analysis is highly consistent across instruments. Which subset of genes is observed to be differentially regulated may be sampled from a larger set associated with a pathway or network. The pattern of these differences yields a stronger signal in combination. With these questions addressed, we hope to see wider deployment of differential proteomics in support of clinical and biological studies.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jproteome.5b00859.

Correlation plots for repeatability, an alternative analysis for consistency in OVelos@45, Venn diagrams comparing differential genes from spectral count and extracted ion chromatogram analysis, and Venn diagrams comparing differential genes within platforms of the same type (PDF)

QuaMeter IDFree quality metrics for each LC–MS/MS experiment included in the study (XLSX)

DOI: 10.1021/acs.jproteome.5b00859
J. Proteome Res. 2016, 15, 691–706

AUTHOR INFORMATION

Corresponding Author

*E-mail: dtabb@sun.ac.za. Tel.: +27 21 938 9403. Fax: (615) 322-0502.

Present Address

□(M.J.C.E.) Smith Breast Center, Baylor College of Medicine, Houston, Texas 77030, United States.

Author Contributions

David L. Tabb designed the study, selected data sets, conducted proteomic identification, and wrote the manuscript. Xia Wang developed and applied the Bayesian hierarchical model with mixtures prior to determine sets of differential genes based on spectral counts, precursor intensity, and iTRAQ reporter ion intensity. At Broad Institute, Steven A. Carr provided guidance on aims in the study and proposed use of replicate analyses of the pair of patient-derived xenograft tissues as a comparative reference material ("CompRef") for initial validation of workflows and ongoing QC during patient sample analyses. Karl R. Clauser and Philipp Mertins optimized iTRAQ–LC–MS/MS experimental and bioinformatics methodology. At Vanderbilt, Matthew C. Chambers and Jerry D. Holman contributed source code to incorporate gene orientation and precursor intensity differentiation in the IDPicker protein assembler. Jing Wang and Bing Zhang conducted biological network enrichment analysis on the differential gene lists across multiple sites. Lisa J. Zimmerman ensured that the mass spectrometer continued in working order through a long queue of samples. At UNC, Xian Chen and Harsha P. Gunawardena collected replicates on their mass spectrometer explicitly in support of this article. At Washington University in St. Louis, Shunqiang Li and Matthew J. C. Ellis developed the xenograft tumors that were used for this study. Sherri R. Davies and Reid Townsend coordinated the special preparation and use of xenograft samples across CPTAC sites. At ESAC International, Karen A. Ketchum coordinated raw data collection and access through the CPTAC Public Portal. At NCI, Chris Kinsinger and Henry Rodriguez provided overall coordination, while Emily Boja tried to keep this project on schedule and Mehdi Mesri facilitated interactions with the CPTAC Steering Committee. At PNNL, Tao Liu and Feng Yang fielded methods development and data collection. Sangtae Kim assisted with the configuration of the MS-GF+ algorithm, and Samuel H. Payne coordinated discussion of this article with the bioinformatics working group. Jason E. McDermott conducted pathway enrichment analysis on the differential gene lists across multiple sites. Vladislav A. Petyuk contributed original code in the R statistical environment for the differentiation and display of iTRAQ reporter ion differences. Karin D. Rodland and Richard D. Smith helped to shape the scope of the differential expression quality control project. Bai Zhang and Zhen Zhang developed initial code for spectral count assessment, and Jianying Zhou optimized sample preparation and instrument methods with support from Hui Zhang. Daniel C. Liebler perceived the need for a study evaluating the reproducibility of proteomic differences and the opportunity posed by the many replicates available for these xenografts.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

CPTAC includes support from U24-CA-160035 (Washington University in St. Louis), U24-CA-159988 (Vanderbilt University), U24-CA-160034 (Broad Institute), U24-CA-160019 (Pacific Northwest National Lab), and U24-CA-160036 (Johns Hopkins University). The PDX models were developed through grants to Matthew J. Ellis by Susan G. Komen for the Cure (grant nos. BCTR0707808 and KG090422). The HAMLET Core that provided the xenograft tumors was supported by CTSA grant UL1 RR024992. Public dissemination of underlying raw data at the CPTAC Public Portal was made possible through contract HHSN261201100106C to ESAC, Inc. Biological network analysis was enabled through Leidos Biomedical Research contract 13XS029.

REFERENCES

(1) Zhang, B.; Wang, J.; Wang, X.; Zhu, J.; Liu, Q.; Shi, Z.; Chambers, M. C.; Zimmerman, L. J.; Shaddox, K. F.; Kim, S.; et al. Proteogenomic characterization of human colon and rectal cancer. *Nature* **2014**, *513* (7518), 382–387.

(2) Picotti, P.; Aebersold, R.; Domon, B. The implications of proteolytic background for shotgun proteomics. *Mol. Cell. Proteomics* **2007**, *6* (9), 1589–1598.

(3) Fang, Y.; Robinson, D. P.; Foster, L. J. Quantitative analysis of proteome coverage and recovery rates for upstream fractionation methods in proteomics. *J. Proteome Res.* **2010**, *9* (4), 1902–1912.

(4) Slebos, R. J. C.; Brock, J. W. C.; Winters, N. F.; Stuart, S. R.; Martinez, M. A.; Li, M.; Chambers, M. C.; Zimmerman, L. J.; Ham, A. J.; Tabb, D. L.; et al. Evaluation of strong cation exchange versus isoelectric focusing of peptides for multidimensional liquid chromatography-tandem mass spectrometry. *J. Proteome Res.* **2008**, *7* (12), 5286–5294.

(5) Prince, J. T.; Marcotte, E. M. Chromatographic alignment of ESI-LC-MS proteomics data sets by ordered bijective interpolated warping. *Anal. Chem.* **2006**, *78* (17), 6140–6152.

(6) Tabb, D. L.; Vega-Montoto, L.; Rudnick, P. A.; Variyath, A. M.; Ham, A.-J. L.; Bunk, D. M.; Kilpatrick, L. E.; Billheimer, D. D.; Blackman, R. K.; Cardasis, H. L.; et al. Repeatability and reproducibility in proteomic identifications by liquid chromatography-tandem mass spectrometry. *J. Proteome Res.* **2010**, *9* (2), 761–776.

(7) Bell, A. W.; Deutsch, E. W.; Au, C. E.; Kearney, R. E.; Beavis, R.; Sechi, S.; Nilsson, T.; Bergeron, J. J. M.; et al. A HUPO test sample study reveals common problems in mass spectrometry-based proteomics. *Nat. Methods* **2009**, *6* (6), 423–430.

(8) Mertins, P.; Yang, F.; Liu, T.; Mani, D. R.; Petyuk, V. A.; Gillette, M. A.; Clauser, K. R.; Qiao, J. W.; Gritsenko, M. A.; Moore, R. J.; et al. Ischemia in tumors induces early and sustained phosphorylation changes in stress kinase pathways but does not affect global protein levels. *Mol. Cell. Proteomics* **2014**, *13* (7), 1690–1704.

(9) Gajadhar, A. S.; Johnson, H.; Slebos, R. J. C.; Shaddox, K.; Wiles, K.; Washington, M. K.; Herline, A. J.; Levine, D. A.; Liebler, D. C.; White, F. M.; et al. Phosphotyrosine signaling analysis in human tumors is confounded by systemic ischemia-driven artifacts and intra-specimen heterogeneity. *Cancer Res.* **2015**, *75* (7), 1495–1503.

(10) Yang, F.; Shen, Y.; Camp, D. G.; Smith, R. D. High-pH reversed-phase chromatography with fraction concatenation for 2D proteomic analysis. *Expert Rev. Proteomics* **2012**, *9* (2), 129–134.

(11) Boone, J. D.; Dobbin, Z. C.; Straughn, J. M.; Buchsbaum, D. J. Ovarian and cervical cancer patient derived xenografts: The past, present, and future. *Gynecol. Oncol.* **2015**, *138* (2), 486–491.

(12) Li, S.; Shen, D.; Shao, J.; Crowder, R.; Liu, W.; Prat, A.; He, X.; Liu, S.; Hoog, J.; Lu, C.; et al. Endocrine-therapy-resistant ESR1 variants revealed by genomic characterization of breast-cancer-derived xenografts. *Cell Rep.* **2013**, *4* (6), 1116–1130.

(13) Luo, R.; Zhao, H. Protein quantitation using iTRAQ: Review on the sources of variations and analysis of nonrandom missingness. *Stat. Interface* **2012**, *5* (1), 99–107.

DOI: 10.1021/acs.jproteome.5b00859
J. Proteome Res. 2016, 15, 691–706

(14) Liu, H.; Sadygov, R. G.; Yates, J. R. A model for random sampling and estimation of relative protein abundance in shotgun proteomics. *Anal. Chem.* **2004**, *76* (14), 4193−4201.
(15) Cox, J.; Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **2008**, *26* (12), 1367−1372.
(16) Anderle, M.; Roy, S.; Lin, H.; Becker, C.; Joho, K. Quantifying reproducibility for differential proteomics: noise analysis for protein liquid chromatography-mass spectrometry of human serum. *Bioinformatics* **2004**, *20* (18), 3575−3582.
(17) Kim, Y. J.; Zhan, P.; Feild, B.; Ruben, S. M.; He, T. Reproducibility assessment of relative quantitation strategies for LC-MS based proteomics. *Anal. Chem.* **2007**, *79* (15), 5651−5658.
(18) Old, W. M.; Meyer-Arendt, K.; Aveline-Wolf, L.; Pierce, K. G.; Mendoza, A.; Sevinsky, J. R.; Resing, K. A.; Ahn, N. G. Comparison of label-free methods for quantifying human proteins by shotgun proteomics. *Mol. Cell. Proteomics* **2005**, *4* (10), 1487−1502.
(19) Zybailov, B.; Coleman, M. K.; Florens, L.; Washburn, M. P. Correlation of relative abundance ratios derived from peptide ion chromatograms and spectrum counting for quantitative proteomic analysis using stable isotope labeling. *Anal. Chem.* **2005**, *77* (19), 6218−6224.
(20) Chong, P. K.; Gan, C. S.; Pham, T. K.; Wright, P. C. Isobaric tags for relative and absolute quantitation (iTRAQ) reproducibility: Implication of multiple injections. *J. Proteome Res.* **2006**, *5* (5), 1232−1240.
(21) Gan, C. S.; Chong, P. K.; Pham, T. K.; Wright, P. C. Technical, experimental, and biological variations in isobaric tags for relative and absolute quantitation (iTRAQ). *J. Proteome Res.* **2007**, *6* (2), 821−827.
(22) Patel, V. J.; Thalassinos, K.; Slade, S. E.; Connolly, J. B.; Crombie, A.; Murrell, J. C.; Scrivens, J. H. A comparison of labeling and label-free mass spectrometry-based proteomics approaches. *J. Proteome Res.* **2009**, *8* (7), 3752−3759.
(23) Wang, H.; Alvarez, S.; Hicks, L. M. Comprehensive comparison of iTRAQ and label-free LC-based quantitative proteomics approaches using two Chlamydomonas reinhardtii strains of interest for biofuels engineering. *J. Proteome Res.* **2012**, *11* (1), 487−501.
(24) Trinh, H. V.; Grossmann, J.; Gehrig, P.; Roschitzki, B.; Schlapbach, R.; Greber, U. F.; Hemmi, S. iTRAQ-Based and Label-Free Proteomics Approaches for Studies of Human Adenovirus Infections. *Int. J. Proteomics* **2013**, *2013*, 581862.
(25) Megger, D. A.; Pott, L. L.; Ahrens, M.; Padden, J.; Bracht, T.; Kuhlmann, K.; Eisenacher, M.; Meyer, H. E.; Sitek, B. Comparison of label-free and label-based strategies for proteome analysis of hepatoma cell lines. *Biochim. Biophys. Acta, Proteins Proteomics* **2014**, *1844* (5), 967−976.
(26) Schulze, W. X.; Usadel, B. Quantitation in mass-spectrometry-based proteomics. *Annu. Rev. Plant Biol.* **2010**, *61*, 491−516.
(27) Meissner, F.; Mann, M. Quantitative shotgun proteomics: considerations for a high-quality workflow in immunology. *Nat. Immunol.* **2014**, *15* (2), 112−117.
(28) Bantscheff, M.; Lemeer, S.; Savitski, M. M.; Kuster, B. Quantitative mass spectrometry in proteomics: critical review update from 2007 to the present. *Anal. Bioanal. Chem.* **2012**, *404* (4), 939−965.
(29) Vaux, D. L.; Fidler, F.; Cumming, G. Replicates and repeats−what is the difference and is it significant? A brief discussion of statistics and experimental design. *EMBO Rep.* **2012**, *13* (4), 291−296.
(30) Parker, J. S.; Mullins, M.; Cheang, M. C. U.; Leung, S.; Voduc, D.; Vickery, T.; Davies, S.; Fauron, C.; He, X.; Hu, Z.; et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. *J. Clin. Oncol.* **2009**, *27* (8), 1160−1167.
(31) Perou, C. M.; Sørlie, T.; Eisen, M. B.; van de Rijn, M.; Jeffrey, S. S.; Rees, C. A.; Pollack, J. R.; Ross, D. T.; Johnsen, H.; Akslen, L. A.; et al. Molecular portraits of human breast tumours. *Nature* **2000**, *406* (6797), 747−752.
(32) Ding, L.; Ellis, M. J.; Li, S.; Larson, D. E.; Chen, K.; Wallis, J. W.; Harris, C. C.; McLellan, M. D.; Fulton, R. S.; Fulton, L. L.; et al. Genome remodelling in a basal-like breast cancer metastasis and xenograft. *Nature* **2010**, *464* (7291), 999−1005.
(33) Wang, X.; Chambers, M. C.; Vega-Montoto, L. J.; Bunk, D. M.; Stein, S. E.; Tabb, D. L. QC metrics from CPTAC raw LC-MS/MS data interpreted through multivariate statistics. *Anal. Chem.* **2014**, *86* (5), 2497−2509.
(34) Tabb, D. L.; Fernando, C. G.; Chambers, M. C. MyriMatch: highly accurate tandem mass spectral peptide identification by multivariate hypergeometric analysis. *J. Proteome Res.* **2007**, *6* (2), 654−661.
(35) Kim, S.; Pevzner, P. A. MS-GF+ makes progress towards a universal database search tool for proteomics. *Nat. Commun.* **2014**, *5*, 5277.
(36) Eng, J. K.; Jahan, T. A.; Hoopmann, M. R. Comet: an open-source MS/MS sequence database search tool. *Proteomics* **2013**, *13* (1), 22−24.
(37) Deutsch, E. W.; Mendoza, L.; Shteynberg, D.; Farrah, T.; Lam, H.; Tasman, N.; Sun, Z.; Nilsson, E.; Pratt, B.; Prazen, B.; et al. A guided tour of the Trans-Proteomic Pipeline. *Proteomics* **2010**, *10* (6), 1150−1159.
(38) Jones, A. R.; Eisenacher, M.; Mayer, G.; Kohlbacher, O.; Siepen, J.; Hubbard, S. J.; Selley, J. N.; Searle, B. C.; Shofstahl, J.; Seymour, S. L.; et al. The mzIdentML data standard for mass spectrometry-based proteomics results. *Mol. Cell. Proteomics* **2012**, *11* (7), M111.014381.
(39) Käll, L.; Storey, J. D.; MacCoss, M. J.; Noble, W. S. Posterior error probabilities and false discovery rates: two sides of the same coin. *J. Proteome Res.* **2008**, *7* (1), 40−44.
(40) Holman, J. D.; Ma, Z.-Q.; Tabb, D. L. Identifying proteomic LC-MS/MS data sets with Bumbershoot and IDPicker. *Curr. Protoc. Bioinf.* **2012**, DOI: 10.1002/0471250953.bi1317s37.
(41) Zhang, B.; Chambers, M. C.; Tabb, D. L. Proteomic parsimony through bipartite graph analysis improves accuracy and transparency. *J. Proteome Res.* **2007**, *6* (9), 3549−3557.
(42) Chen, Y.-Y.; Chambers, M. C.; Li, M.; Ham, A.-J. L.; Turner, J. L.; Zhang, B.; Tabb, D. L. IDPQuantify: combining precursor intensity with spectral counts for protein and peptide quantification. *J. Proteome Res.* **2013**, *12* (9), 4111−4121.
(43) Choi, H.; Fermin, D.; Nesvizhskii, A. I. Significance analysis of spectral count data in label-free shotgun proteomics. *Mol. Cell. Proteomics* **2008**, *7* (12), 2373−2385.
(44) Lundgren, D. H.; Hwang, S.-I.; Wu, L.; Han, D. K. Role of spectral counting in quantitative proteomics. *Expert Rev. Proteomics* **2010**, *7* (1), 39−53.
(45) Booth, J. G.; Eilertson, K. E.; Olinares, P. D. B.; Yu, H. A bayesian mixture model for comparative spectral count data in shotgun proteomics. *Mol. Cell. Proteomics* **2011**, *10* (8), M110.007203.
(46) Gleiss, A.; Dakna, M.; Mischak, H.; Heinze, G. Two-group comparisons of zero-inflated intensity values: the choice of test statistic matters. *Bioinformatics* **2015**, *31* (14), 2310−2317.
(47) Ogata, H.; Goto, S.; Sato, K.; Fujibuchi, W.; Bono, H.; Kanehisa, M. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **1999**, *27* (1), 29−34.
(48) Fury, W.; Batliwalla, F.; Gregersen, P. K.; Li, W. Overlapping probabilities of top ranking gene lists, hypergeometric distribution, and stringency of gene selection criterion. *Conf. Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. IEEE Eng. Med. Biol. Soc. Annu. Conf.* **2006**, *1*, 5531−5534.
(49) Hong, G.; Zhang, W.; Li, H.; Shen, X.; Guo, Z. Separate enrichment analysis of pathways for up- and downregulated genes. *J. R. Soc., Interface* **2014**, *11* (92), 20130950.
(50) Benjamini, Y.; Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B Methodol.* **1995**, *57* (1), 289−300.
(51) Shi, Z.; Wang, J.; Zhang, B. NetGestalt: integrating multidimensional omics data over biological networks. *Nat. Methods* **2013**, *10* (7), 597−598.
(52) Turinsky, A. L.; Razick, S.; Turner, B.; Donaldson, I. M.; Wodak, S. J. Interaction databases on the same page. *Nat. Biotechnol.* **2011**, *29* (5), 391−393.

DOI: 10.1021/acs.jproteome.5b00859
J. Proteome Res. 2016, 15, 691−706

(53) Wang, X.; Slebos, R. J. C.; Chambers, M. C.; Tabb, D. L.; Liebler, D. C.; Zhang, B. proBAMsuite, a bioinformatics framework for genome-based representation and analysis of proteomics data. *Mol. Cell. Proteomics* **2015**, *14*, M115.052860.
(54) Jain, N.; Thatte, J.; Braciale, T.; Ley, K.; O'Connell, M.; Lee, J. K. Local-pooled-error test for identifying differentially expressed genes with a small number of replicated microarrays. *Bioinformatics* **2003**, *19* (15), 1945−1951.
(55) Efron, B.; Tibshirani, R.; Storey, J. D.; Tusher, V. Empirical Bayes Analysis of a Microarray Experiment. *J. Am. Stat. Assoc.* **2001**, *96* (456), 1151−1160.
(56) Smith, L. M.; Kelleher, N. L. Consortium for Top Down Proteomics. Proteoform: a single term describing protein complexity. *Nat. Methods* **2013**, *10* (3), 186−187.
(57) Carrillo, B.; Yanofsky, C.; Laboissiere, S.; Nadon, R.; Kearney, R. E. Methods for combining peptide intensities to estimate relative protein abundance. *Bioinformatics* **2010**, *26* (1), 98−103.
(58) Hebert, A. S.; Merrill, A. E.; Bailey, D. J.; Still, A. J.; Westphall, M. S.; Strieter, E. R.; Pagliarini, D. J.; Coon, J. J. Neutron-encoded mass signatures for multiplexed proteome quantification. *Nat. Methods* **2013**, *10* (4), 332−334.
(59) Oberg, A. L.; Vitek, O. Statistical design of quantitative mass spectrometry-based proteomic experiments. *J. Proteome Res.* **2009**, *8* (5), 2144−2156.

DOI: 10.1021/acs.jproteome.5b00859
J. Proteome Res. 2016, 15, 691−706