Evaluation of Strong Cation Exchange versus Isoelectric Focusing of Peptides for Multidimensional Liquid Chromatography-Tandem Mass Spectrometry

Robbert J. C. Slebos,[†,‡,#] Jonathan W. C. Brock,[†,§,#] Nancy F. Winters,[†] Sarah R. Stuart,[†]
Misti A. Martinez,[†] Ming Li,[¶] Mathew C. Chambers,[‖] Lisa J. Zimmerman,[§,⊥] Amy J. Ham,[†,§]
David L. Tabb,[‖] and Daniel C. Liebler*,[†,§]

*Jim Ayers Institute for Precancer Detection and Diagnosis, Vanderbilt-Ingram Cancer Center, Medical Research
Building III, 465 21st Avenue South, Nashville, Tennessee 37232, Department of Cancer Biology, Vanderbilt
University School of Medicine, U9207 Medical Research Building III, 465 21st Avenue South, Nashville,
Tennessee 37232, Department of Biochemistry, Vanderbilt University School of Medicine, U1213 Medical
Research Building III, 465 21st Avenue South, Nashville, Tennessee 37232, Department of Biostatistics,
Vanderbilt University School of Medicine, 571 Preston Research Building, 2220 Pierce Avenue South, Nashville,
Tennessee 37232, Department of Biomedical Informatics, Vanderbilt University School of Medicine,
U9211 Medical Research Building III, 465 21st Avenue South, Nashville, Tennessee 37232, and Mass
Spectrometry Research Center, Vanderbilt University School of Medicine, 9110 Medical Research Building III,
465 21st Avenue South, Nashville, Tennessee 37232*

Shotgun proteome analysis platforms based on multidimensional liquid chromatography-tandem mass spectrometry (LC-MS/MS) provide a powerful means to discover biomarker candidates in tissue specimens. Analysis platforms must balance sensitivity for peptide detection, reproducibility of detected peptide inventories and analytical throughput for protein amounts commonly present in tissue biospecimens (<100 μg), such that platform stability is sufficient to detect modest changes in complex proteomes. We compared shotgun proteomics platforms by analyzing tryptic digests of whole cell and tissue proteomes using strong cation exchange (SCX) and isoelectric focusing (IEF) separations of peptides prior to LC-MS/MS analysis on a LTQ-Orbitrap hybrid instrument. IEF separations provided superior reproducibility and resolution for peptide fractionation from samples corresponding to both large (100 μg) and small (10 μg) protein inputs. SCX generated more peptide and protein identifications than did IEF with small (10 μg) samples, whereas the two platforms yielded similar numbers of identifications with large (100 μg) samples. In nine replicate analyses of tryptic peptides from 50 μg colon adenocarcinoma protein, overlap in protein detection by the two platforms was 77% of all proteins detected by both methods combined. IEF more quickly approached maximal detection, with 90% of IEF-detectable medium abundance proteins (those detected with a total of 3–4 peptides) detected within three replicate analyses. In contrast, the SCX platform required six replicates to detect 90% of SCX-detectable medium abundance proteins. High reproducibility and efficient resolution of IEF peptide separations make the IEF platform superior to the SCX platform for biomarker discovery via shotgun proteomic analyses of tissue specimens.

Keywords: shotgun proteomics • isoelectric focusing • ion exchange • LTQ-Orbitrap • cancer

Introduction

An emerging paradigm for biomarker development begins with unbiased discovery of biomarker candidates in tissues, cell models and biofluids proximal to sites of disease.[1] Of the

* To whom correspondence should be addressed. E-mail: daniel.liebler@vanderbilt.edu.
† Vanderbilt-Ingram Cancer Center.
‡ Department of Cancer Biology, Vanderbilt University School of Medicine.
§ Department of Biochemistry, Vanderbilt University School of Medicine.
# These authors contributed equally to this work.
¶ Department of Biostatistics, Vanderbilt University School of Medicine.
‖ Department of Biomedical Informatics, Vanderbilt University School of Medicine.
⊥ Mass Spectrometry Research Center, Vanderbilt University School of Medicine.

existing proteomics technology platforms, none are better suited for unbiased biomarker discovery than shotgun proteomics, which has revolutionized cell biology and biochemistry by enabling identification of the protein components of multiprotein complexes, complex subcellular proteomes and even whole cell, tissue and biofluid proteomes.[2–10] In shotgun analyses, protein mixtures are digested to peptides, which then are analyzed by liquid chromatography-tandem mass spectrometry (LC-MS/MS) to identify peptide and protein sequences.[11] Shotgun proteomics platforms use multidimensional peptide separations to fractionate complex peptide mixtures prior to reverse phase LC-MS/MS.[12] Each fraction presents a simplified peptide mixture for LC-MS/MS and this enables acquisition of MS/MS spectra for lower abundance peptides.

The combined data set generated from all of the fractions represents peptides from dozens to hundreds or even thousands of proteins.

The most widely adopted strategy for shotgun proteomics uses strong cation exchange (SCX) to fractionate peptide digests, followed by reverse phase LC-MS/MS to acquire peptide MS/MS spectra.[13-15] The SCX fractionation step can be done either in-line with multiphasic columns[13-15] or by off-line SCX fractionation, followed by LC-MS/MS of the collected fractions.[16] An alternate approach to SCX for peptide fractionation is isoelectric focusing (IEF) using either capillary systems[17-20] or a polyacrylamide gel containing an immobilized pH gradient (IPG).[21-23] IPG strips are commonly employed for first-dimension separation of intact proteins in two-dimensional sodium dodecyl sulfate polyacrylamide electrophoresis (2D-SDS-PAGE). IPG strips for separations in multiple p*I* ranges are available from several commercial sources and can be conveniently employed for peptide fractionation.

Despite the power of this technology platform, current shotgun proteomics approaches are limited in both sample throughput and reproducibility of peptide detection and identification, particularly for lower abundance proteins. The degree of proteome coverage is proportional to the extent of peptide fractionation, which creates a tradeoff between numbers of confidently identified proteins and overall analytical throughput. Moreover, instrument control algorithms for automated selection of peptide ions for MS/MS ("data-dependent scanning"[24]) result in a semirandom sampling of lower abundance components of peptide mixtures. Multiple technical replicates thus are necessary for identification of the largest numbers of proteins. A less-appreciated source of variation is the performance of the SCX or IEF peptide fractionation methods themselves.

Use of a multidimensional LC-MS/MS shotgun proteomics platform for biomarker discovery is also constrained by additional considerations involving use of clinical specimens. Many tissue biopsy specimens are relatively small (<1 mg wet weight) and analyses should accommodate correspondingly small protein input (<100 μg of protein). In addition, appropriate experimental designs may require analysis of multiple independent samples, so a shotgun analysis platform must balance analytical throughput with depth of proteome coverage. Most importantly, the variation in performance of the platform and component analytical steps must be minimized and the degree of variation characterized. This is essential if detected characteristics of tissue proteomes are to be reliably attributed to biological characteristics, rather than variability in platform performance.

Here, we describe studies to compare SCX and IEF methods for peptide fractionation for multidimensional LC-MS/MS and to evaluate their implementation in a shotgun proteomics platform. We used cell lysate and tumor tissue samples corresponding to protein inputs of 10-100 μg, which is typical of protein amounts present in small, macrodissected tissue biospecimens. To compare the performance of these multidimensional separations, our analyses evaluated resolution of peptides by fractionation, numbers of peptide and protein identifications and cumulative identifications with replicate analyses. The data illustrate the advantages and limitations of SCX and IEF fractionation in shotgun proteomics and suggest that IEF-based platforms offer clear advantages in reproducibility for unbiased biomarker discovery in tissue samples.

Materials and Methods

Cell and Tissue Digest. A human colon adenocarcinoma cell line (RKO) was obtained from ATCC (Manassas, VA) and cultured in 100 mL flasks in McCoy's 5A media (Mediatech, Herndon, VA) supplemented with 10% fetal bovine serum (Atlas Biologicals, Fort Collins, CO) at 37 °C in 5% CO$_2$. RKO cells were grown to >90% confluence, then harvested in 5 mL of 0.25% trypsin-EDTA, washed with PBS, split into 1×10^7 cells per tube, and frozen at −80 °C.

A frozen rectal adenocarcinoma biopsy specimen was obtained from the National Cancer Institute Cooperative Human Tissue Network-Western Division (Vanderbilt University, Nashville, TN) under an IRB-approved protocol that included informed consent from the patient. The tissue was embedded in polyvinyl alcohol and three 60 μm slices were placed in separate centrifuge tubes. Polyvinyl alcohol was removed with two washes of 70% ethanol followed by a single wash with deionized water.

Solubilization and tryptic digestion of proteins in both the RKO cell pellet and adenocarcinoma slices was done by a modification of the method of Wang et al.,[25] in which trifluoroethanol (TFE) is used to solubilize cell and tissue proteins. RKO cell pellet or adenocarcinoma slices were resuspended in 200 μL of TFE/50 mM ammonium bicarbonate, pH 8.0 (1:1, v/v). Samples were sonicated for 20 s, followed by 30 s incubation on ice; this was done three times. The homogenate then was heated with shaking at 1000 rpm for 1 h at 60 °C followed by a second series of sonication steps, as described above. The homogenate was reduced with tris(carboxyethyl)phosphine (10 mM) and dithiothreitol (25 mM) at 60 °C for 30 min, followed by alkylation with iodoacetamide (50 mM) in the dark at ambient temperature for 20 min. The reduced and alkylated protein mixture was diluted to 1 mL with ammonium bicarbonate (50 mM, pH 8.0) followed by addition of trypsin (Promega, Cat#TB309, Madison, WI) at a trypsin/protein ratio of 1:50 (w/w). The digest mixture was incubated overnight at 37 °C, then frozen at −80 °C and lyophilized. Samples were resuspended in 1 mL of deionized water and applied to SEP-Pak vac 1 cc (100 mg) C-18 cartridges (Waters Corp., Milford, MA), which were prewashed with 1 mL of acetonitrile and equilibrated with 2 mL of deionized water. The flow-through was discarded, the cartridges were washed with 1 mL of deionized water, the bound peptides were eluted with 80% acetonitrile in deionized water, and the eluate was evaporated *in vacuo*.

IEF of Peptides. Tryptic peptides from 10 or 100 μg of protein (RKO cells) or 50 μg of protein (adenocarcinoma) were redissolved in 500 μL of 6 M urea and loaded in a IPGphor rehydration tray (GE Healthcare, Piscataway, NJ). For some analyses, carrier ampholyte (2% (v/v) was included in the loading solution (IPG buffer pH 3.5-5.0 (GE catalog #17-6002-02)). Immobiline IPG strips (24 cm, pH 3.5-4.5) (GE Healthcare) were placed over the samples and allowed to rehydrate overnight at ambient temperature. The loaded strips were focused at 20 °C on an Ettan IPGPhor-III IEF system (GE Healthcare) using an initial focusing step at 300 V for 900 Vh, then a gradient to 1000 V for 3900 Vh, then a gradient to 8000 V for 13 500 Vh, then a step to 8000 V for 93 700 Vh. The strips were then cut into either 10 (24 mm) or 15 (16 mm) pieces and placed in separate wells of a 96-well Falcon flat bottom polystyrene ELISA plate (Fisher Scientific). Peptides were eluted from the strips with 200 μL of 0.1% formic acid for 15 min,

followed by 200 μL of acetonitrile/0.1% formic acid (1:1, v/v) for 15 min, then with 200 μL of acetonitrile containing 0.1% formic acid for 15 min. The combined eluates for each IPG strip fraction were evaporated *in vacuo* and then redissolved in 0.1% trifluoroacetic acid (TFA) and applied to a 96 well C-18 Oasis HLB plate (30 μm particle size, 10 mg packing) (Waters Corp., Milford, MA) prewashed with 1 mL of acetonitrile and equilibrated with 2 mL of 0.1% TFA. The flow-through was discarded and the cartridges were washed with 1 mL of 0.1% TFA. The bound peptides were eluted with 0.3 mL each of 30% acetonitrile/0.1%TFA, 70% acetonitrile/0.1% TFA and 100% acetonitrile/0.1% TFA and the combined eluate was evaporated *in vacuo* and redissolved in 100 μL of 0.1% formic acid for LC-MS/MS analysis.

SCX of Peptides. Tryptic peptides from 10 or 100 μg of protein (RKO cells) or 50 μg of protein (adenocarcinoma) were resuspended in 10 μL of 0.1% formic acid and loaded onto a LUNA polysulfethyl SCX column (100 μm i.d. × 100 mm, 5 μm particles with 300-Å pore size; Phenomenex, Torrance, CA). SCX chromatography was performed with an Agilent 1100 series high-performance liquid chromatography system (Santa Clara, CA) at a flow rate of 0.45 mL min^{-1}. Peptides were eluted with a step gradient using acetonitrile/ammonium formate buffers (buffer A, 25% acetonitrile /75% 10 mM ammonium formate, pH 3.0; buffer B, 25% acetonitrile/75% 200 mM ammonium formate, pH 8.0; buffer C, 25% acetonitrile/75% 500 mM ammonium formate, pH 8.0). The elution program was 100% buffer A for 10 min (flow-through fraction), followed by a gradient from 0 to 25% buffer B over 15 min, then 7 steps of 5 min each increasing buffer B by 3%, then a gradient from 46% to 100% buffer B over 5 min, followed by a step to 100% buffer C and finally a 10 min wash. Fractions (10 total) were taken for 10 min flow-through elution, the 15 min 0−25% buffer B gradient, each 3% buffer B step, and the 15 min wash. Collected SCX fractions were evaporated *in vacuo* and resuspended in 100 μL of 0.1% formic acid for LC-MS/MS analysis.

Reverse Phase LC-MS/MS. LC-MS/MS analyses were performed on an LTQ-Orbitrap hybrid mass spectrometer (Thermo Electron, San Jose, CA) equipped with an Eksigent nanoLC and autosampler (Dublin, CA). Peptides were resolved on a 100 μm × 11 cm fused silica capillary column (Polymicro Technologies, LLC., Phoenix, AZ) packed with 5 μm, 300 Å Jupiter C18 (Phenomenex, Torrance, CA). Liquid chromatography was carried out at ambient temperature at a flow rate of 0.6 μL min^{-1} using a gradient mixture of 0.1% (v/v) formic acid in water (solvent A) and 0.1% (v/v) formic acid in acetonitrile (solvent B). Centroided MS/MS scans were acquired on the LTQ-Orbitrap using an isolation width of 2 *m/z*, an activation time of 30 ms, an activation *q* of 0.250 and 30% normalized collision energy using 1 microscan with a max ion time of 100 ms for each MS/MS scan and 1 microscan with a max ion time of 500 for each full MS scan. (One set of analyses in Figure 2D was generated as described above, but with an isolation width of 3 *m/z*, which did not affect the numbers of confident peptide identifications compared to the two other replicate analyses done at the 2 *m/z* setting.) The mass spectrometer was tuned prior to analysis using the synthetic peptide TpepK (AVAGKA-GAR), so that some parameters may have varied slightly from experiment to experiment, but typically the tune parameters were as follows: spray voltage of 2 kV, a capillary temperature of 150 °C, a capillary voltage of 50 V and tube lens of 120 V. The AGC target value was set at 100 000 for the full MS and 10 000 for the MS/MS spectra. For the IEF fractionation in

Figure 2, an AGC target value of 500 000 was used, but this setting did not significantly affect numbers of confident peptide identifications and did not affect our comparisons of peptide separation methods. A full scan obtained for eluting peptides in the range of 400−2000 amu was collected on the Orbitrap portion of the instrument at a resolution of 60 000, followed by five data-dependent MS/MS scans on the LTQ portion of the instrument with a minimum threshold of 1000 set to trigger the MS/MS spectra. MS/MS spectra were recorded using dynamic exclusion of previously analyzed precursors for 60 s with a repeat of 1 and a repeat duration of 1.

To assess instrument performance, a quality control (QC) standard consisting of a tryptic digest of bovine serum albumin (BSA) (2 μL of a 0.6 μg mL^{-1} solution) was analyzed using the instrument settings described above several times daily during analyses of sample sets. Acceptable instrument performance required a signal intensity of 1−2 E7 (base peak chromatogram), <5 ppm mass accuracy for BSA peptide ions, and that BSA peptide ions were the predominant signals observed in the summed full scan mass spectrum across the chromatographic region of peptide elution. In addition, database search of the MS/MS spectra acquired during the BSA QC runs yielded >60% coverage of the BSA protein sequence based on the assigned MS/MS spectra.

Data Analyses. Captured peak lists from the mass spectral .RAW files were transcoded to mzData v1.05 format by a version of the open-source ReAdW tool that had been modified to support mzData conversion (http://www.mc.vanderbilt.edu/msrc/bioinformatics/index.php). The software was configured to transcode only tandem mass spectra; MS scans were excluded. Tandem mass spectra were identified to peptides from the IPI Human database version 3.31 (67 764 sequences) by the MyriMatch algorithm, version 1.0.321.[26] The sequence database was doubled to contain each sequence in both forward and reversed orientations, enabling false discovery rate estimation. Myri-Match was configured to expect all cysteines to bear carboxamidomethyl modifications and to allow for the possibility of oxidation on methionines. Candidate peptides were required to feature trypsin cleavages or protein termini at both ends, though any number of missed cleavages was permitted. Precursor error was allowed to range up to 0.6 *m/z* in either direction, but fragment ions were required to match within 0.5 *m/z*. The IDPicker algorithm[27] filtered the identifications for each LC-MS/MS run to include the largest set for which a 5% identification false discovery rate could be maintained. These identifications were pooled for each sample. False discovery rates (FDR) rates were computed by the formula:[28]

$$FDR = (2 \times reverse)/(forward + reverse) \qquad (1)$$

Proteins were required to have at least two distinct peptide sequences observed within a sample set of SCX or IEF fractions. This requirement allows for the observation of different peptides between the SCX or IEF platforms or in different sample sets. Peptides are distinct if they are of unique sequence or if they harbor distinct dynamic modifications or miscleavages that are allowed within the limits of our search parameters. Any two distinct peptides from a protein constitute independent confirmation of that protein identification, which is not the case when the same two original peptides are identified. Implementation of the IDPicker algorithm also identifies protein clusters with shared peptides, to derive a minimal list of proteins, termed protein groups.[27] The algorithm reported the number of spectra and number of distinct sequences
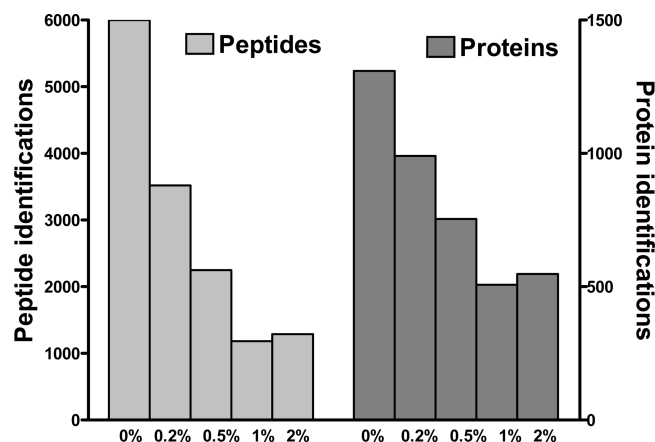
Figure 1. Effect of ampholyte concentration on peptide and protein identifications from tryptic peptides corresponding to 100 μg of RKO cell lysate. Peptides were fractionated with the indicated concentrations of ampholytes on a 24 cm IGPhor pI 3.5–4.5 strip and 10 fractions were collected and analyzed by LC-MS/MS on an LTQ-Orbitrap instrument. Identifications from all 10 fractions were combined and represented for each ampholyte concentration.

observed for each protein and protein group within in each replicate analysis.

Results

Comparison of Peptide Fractionation and Resolution between SCX and IEF. We began our studies of SCX and IEF peptide separations with tryptic peptide mixtures generated from the human colon carcinoma RKO cell line. Digestion of unfractionated RKO cells with the TFE-assisted method described by Wang et al.[25] generates tryptic peptides from a broad spectrum of cellular proteins, including hydrophobic and membrane-associated proteins. The first separation step in a multidimensional LC-MS/MS analysis resolves a complex mixture of peptides into multiple fractions prior to analysis by LC-MS/MS. The SCX and IEF methods we studied were based on previously published methods[13,16,21,23,29] as well as our own experience with both separation methods. IEF fractionation employed 24 cm IPG strips with a relatively narrow 3.5–4.5 pI range (GE Healthcare), which encompasses tryptic peptides from a majority of *Escherichia coli* proteins and provides broad proteome coverage.[21,23]

Initial studies with IEF examined the effect of supporting ampholyte on peptide identifications. Although exogenous ampholyte is required for optimal focusing of intact proteins, peptide identifications were actually suppressed by increasing concentrations of exogenous ampholyte (Figure 1). Indeed, peptides ionize efficiently throughout the pI range used for these analyses and essentially serve as ampholytes in these analyses. Thus, no additional exogenous ampholyte was used for subsequent IEF analyses.

Figure 2 shows the patterns of elution and numbers of peptide identifications from IEF (panels A and B) and SCX (panels C and D) of tryptic peptides from 10 and 100 μg of RKO cell lysate protein as assessed by LC-MS/MS analysis of the fractions. For IEF, peptide identifications were spread broadly across the 3.5–4.5 pI range with a bias toward the higher pI end, as reported previously.[23] In SCX, the majority of peptides eluted between fractions 2 and 5, with a "burst" of peptide elution occurring either in steps 2 or 3. There is significant

variability in peptides eluted in SCX fractions 2 and 3, suggesting that SCX is highly sensitive to subtle variations in pH and osmotic strength. The run-to-run variability in peptides eluted within any IEF fraction was less than 10%, indicating that IEF reproducibly focuses the peptides across the pI range.

Figure 3 depicts the resolution of tryptic peptides from 10 μg and 100 μg of RKO cell proteins by IEF (panels A and B) and SCX (panels C and D). IEF clearly produced much more efficient resolution of peptides, as 87–89% of peptides were detected in only a single fraction. In SCX separations, only 50–60% of peptides were found in a single fraction, and 29–39% and 7–9% found in two and three fractions, respectively. These results indicate that resolution is independent of peptide load in the concentration range studied. The data also are consistent with the previous observation that overall peptide identifications were modestly impacted by collecting only alternate fractions,[16] which suggests that many peptides are redundantly identified in adjacent SCX fractions.

Comparisons of Peptide and Protein Identifications between SCX and IEF. We fractionated tryptic peptides from 10 and 100 μg of RKO cell proteins by IEF or SCX and then analyzed each fraction by LC-MS/MS. As shown in Figure 4. The total number of unique peptide and protein identifications was higher in the 100 μg protein samples than in the 10 μg protein samples for both analytical methods. However, both peptide and protein identifications for the 10 μg sample fractionated by IEF separation were only about one-third of the identifications achieved in mixtures fractionated by SCX. For the 100 μg samples, this disparity was substantially decreased, but numbers of identifications were still greater with SCX fractionation. The difference in protein identifications between SCX and IEF was not as great as that observed for peptide identifications, which indicates that SCX identified additional peptides that mapped to the same proteins also identified by IEF. The higher reproducibility of IEF is also apparent in the number of peptide identifications across replicates; at both 10 μg and 100 μg about 45% of the peptide identifications are made in all 3 replicates, whereas for SCX only 38% and 29% of the identified peptides were found in all three replicates for the 10 μg and 100 μg samples, respectively.

SCX fractionation yielded more protein identifications than did IEF fractionation. IEF peptide fractionation identified 440 and 1756 proteins from the 10 and 100 μg samples, respectively. For SCX peptide fractionation, 1256 and 2188 proteins were identified for the 10 and 100 μg samples, respectively. The disparity between the methods was greatest with the 10 μg samples, where SCX outperformed IEF with roughly 3 times the number of protein identifications (Figure 4). These data suggest that some aspect of the IEF fractionation suppresses detection of peptides from small samples. This effect may be attributed either to poor recovery of peptides from the IPG strips or the coextraction of contaminants from the strips that suppress ionization of peptides in LC-MS/MS runs, or possibly some combination of these two effects. We have not detected LC-MS/MS signals indicative of contaminants from the strips, although nonionizable components could conceivably suppress peptide ionization. We had hypothesized that this effect may be attributable to the urea, which is introduced into the samples when the IPG strips are rehydrated in 6 M urea. However, analyses from strips rehydrated with and without urea indicated no difference in protein identifications (data not shown). These considerations collectively suggest that poor
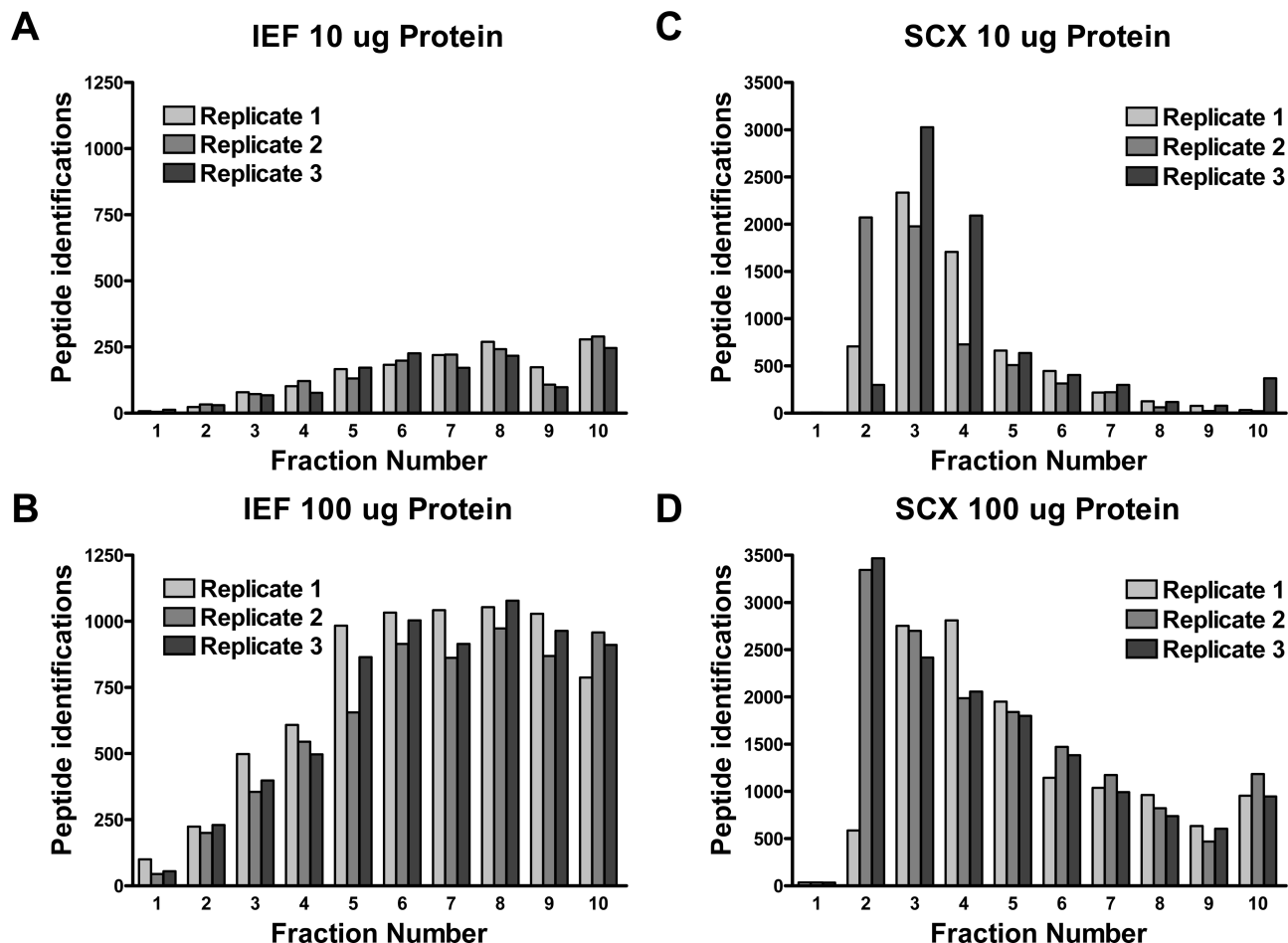
A

B

C

D

Figure 2. Fractionation of RKO cell tryptic peptides by IEF (A and B) and SCX (C and D). Digests corresponding to either 10 μg (A and C) or 100 μg (B and D) of protein were fractionated in triplicate on either a 24 cm IGPhor p*I* 3.5–4.5 strip or on a capillary SCX column with a step gradient as described under Materials and Methods. For both IEF and SCX, 10 fractions were collected and analyzed by LC-MS/MS on an LTQ-Orbitrap instrument.

peptide recovery is the most reasonable explanation for lowered sensitivity with samples <100 μg.

Effects of Replicate Analyses on Peptide and Protein Detection by Shotgun Platforms Using IEF and SCX. A major contributor to variability in peptide detection is the acquisition of MS/MS spectra by "data-dependent scanning", in which precursor ions are automatically selected by instrument control software for MS/MS.[24] This approach results in semirandom selection of lower abundance peptide precursor ions in full scan spectra. Accordingly, some peptide ions not sampled in one analysis are sampled on a subsequent run. A number of reports have demonstrated that optimum inventory coverage of complex proteomes requires multiple replicate analyses.[28,30–32] We asked whether this phenomenon was affected by the choice of SCX versus IEF peptide fractionation and by the relative abundance of the proteins, as assessed by spectral counting.

For these studies, we performed replicate analyses of tryptic peptides derived from 50 μg protein from a single colon adenocarcinoma sample. A total of 9 replicate analyses (each comprising 10 SCX or IEF fractions analyzed by LC-MS/MS) were performed with each separation method. The combination of all 180 LC-MS/MS runs with peptide identifications at 5% FDR yielded 128 013 confident peptide matches representing a total of 14 480 distinct peptides in the database. The most parsimonious summary of this large "universe" of peptides

yielded a total of 1766 protein groups, of which only 19 (1%) were identifications of reversed peptide sequences in the decoy component of the database. Of the 1766 protein groups, 1404 (78%) were identified by IEF and 1720 (97%) by SCX. The two platforms both identified 1358 protein groups (77%) in common. A total of 427 proteins were identified by 2 distinct peptides, 284 by 3 distinct peptides, 191 by 4 distinct peptides and the remaining 864 proteins by 5 or more distinct peptides. Of the total of 1766 proteins, 683 (39%) proteins were detected in all 9 SCX replicates, 434 (25%) proteins were detected in all 9 IEF replicates, whereas only 280 (16%) proteins were detected in all 18 analyses.

To allow an assessment of the proportion of proteins detected with different numbers of replicates, we further limited our data set to the 1355 proteins that were detected by 2 distinct peptides *within* any of the 18 analyses. We then employed a sampling technique whereby random combinations of 1–9 analyses were chosen to estimate the average number of proteins that would have be identified had this number of replicates been performed as separate experiments. The result of this simulation is shown in Figure 5. After 9 replicates, the SCX platform detects, on average, 83% of these 1355 proteins, whereas the IEF platform yields 55% on average. Not unexpectedly, our simulation demonstrates that the first 2 or 3 replicate
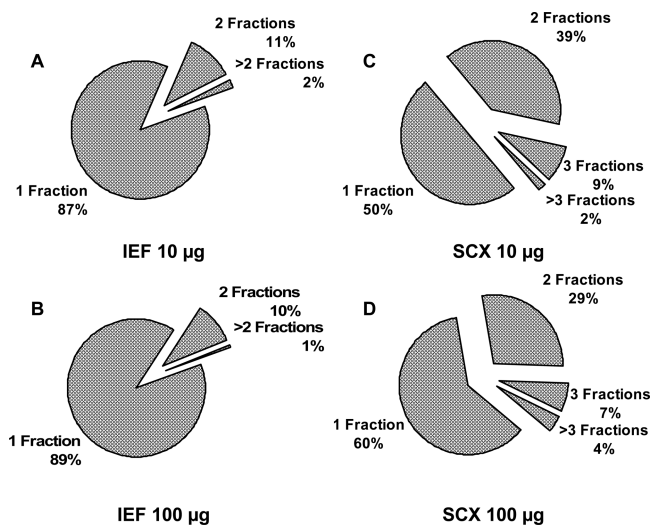
Figure 3. Efficiency of resolution of RKO cell tryptic peptides by IEF (A and B) and SCX (C and D) in analyses depicted in Figure 1. Pie charts depict percentages of peptide identifications associated with only one fraction, two fractions, three fractions or more than three fractions.
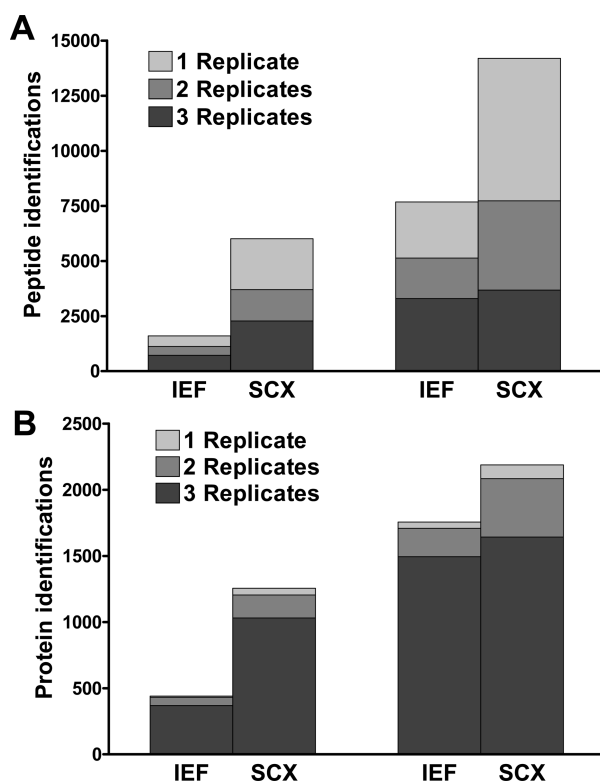

Figure 4. Accumulation of peptide (A) and protein (B) identifications in triplicate analyses of RKO cell tryptic peptides by IEF or SCX. Aliquots of peptide mixtures corresponding to 10 or 100 μg of protein were fractionated in triplicate on either a 24 cm IGPhor p*I* 3.5–4.5 strip or on a capillary SCX column with a step gradient as described under Materials and Methods. For both IEF and SCX, 10 fractions were collected and analyzed by LC-MS/MS on an LTQ-Orbitrap instrument. Bar graph shading indicates peptides and proteins found in only a single replicate analysis, in two of three replicates and in all three replicates.

runs add proportionally more protein identifications than subsequent replicate runs.
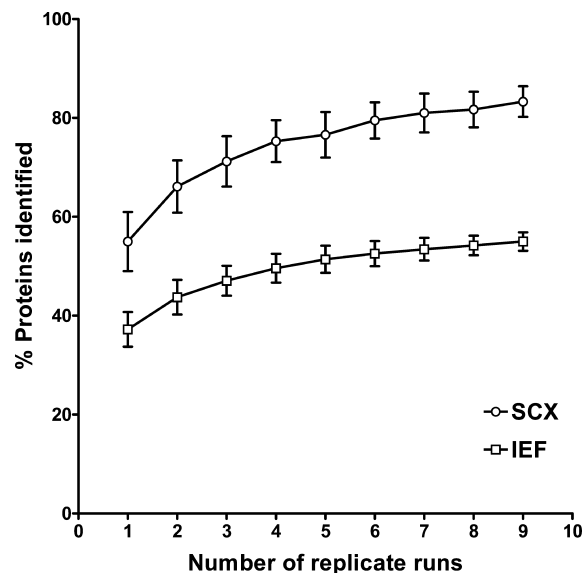

Figure 5. Accumulation of protein group identifications during nine successive replicate analyses of tryptic peptides from colon adenocarcinoma tissue. Each replicate analysis corresponding to 50 μg of tissue protein. Peptides were fractionated on either a 24 cm IGPhor p*I* 3.5–4.5 strip or on a capillary SCX column with a step gradient as described under Materials and Methods. For each replicate analysis by both IEF and SCX, 10 fractions were collected and analyzed by LC-MS/MS on an LTQ-Orbitrap instrument.

The simulation results summarized in Figure 5 incorporate average numbers obtained from proteins spanning a wide range of abundance. For instance, myosin-11 was identified by 302 distinct peptides, which matched to 5821 separate MS/MS scans, whereas a large number of proteins were detected by the minimum of 2 distinct peptides within a single analysis. Detection profiles for proteins of different abundance are likely to vary. To study this, we binned proteins by ranges based on the numbers of distinct peptides by which they were identified. The 165 proteins that were identified by more than 25 distinct peptides always met the selection criteria for detection by at least 2 distinct peptides *within* at least one of the 18 replicate analyses. To determine the sensitivity of detection, proteins that were detected by 20 or fewer distinct peptides were binned and detection profiles within each bin were determined by additional simulation tests (Figure 6). This simulation showed that 67% of the proteins that were SCX-detectable by 3 or 4 peptides were found in two SCX platform replicates, whereas 88% of the proteins that were IEF-detectable by 3 or 4 peptides could be detected by two IEF platform replicates. This same trend applied to proteins detected by only two peptides (Figure 6). Thus, the IEF platform achieves maximal numbers of protein identifications in fewer replicate analyses than does the SCX platform.

Discussion

We began this work with the goal of determining whether an SCX or IEF-based shotgun proteomics platform would be best suited to unbiased discovery of protein biomarkers through shotgun proteomic analysis of tissue samples. The key questions we posed were (1) Which platform identifies the most peptides and proteins? (2) Which platform generates the most reproducible results? and (3) Which platform is best suited to
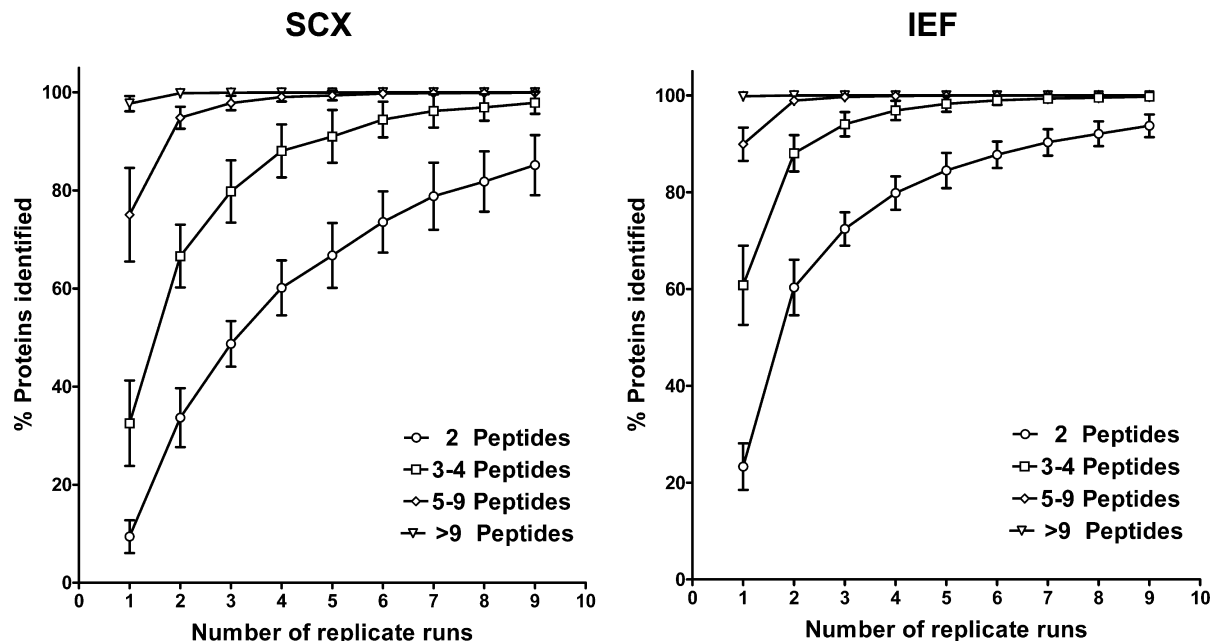
Figure 6. Accumulation of protein group identifications during nine successive replicate analyses of tryptic peptides from colon adenocarcinoma tissue, as classified by the number of distinct peptides that characterized each protein. The data are taken from the analyses described in Figure 5. Listed are proteins that were identified by 2 distinct peptides within a single replicate MS/MS run, those identified by 3 or 4 peptides, by 5 through 9 peptides and by 10 or more peptides. Proteins that are identified by higher average numbers of distinct peptides have higher probabilities of detection in a single replicate MS/MS run.

unbiased biomarker discovery in tissue specimens? In the following sections, we address these questions in the context of our results and the potential application of these platforms to candidate biomarker discovery in tissues.

Of the platforms tested, the SCX-based shotgun platform generates the greatest numbers of peptide and protein identifications. For small RKO cell protein samples (10 μg protein), the superiority of SCX is particularly evident, as SCX generated over twice as many identifications as the 24 cm IPGphor IEF strips (Figure 4). The advantage of SCX for peptide identification translated into a smaller advantage in protein identification, as the main effect of the additional peptide identifications was to increase sequence coverage for identified proteins. The difference between SCX and IEF decreased for larger RKO cell protein samples (100 μg). In analyses of 50 μg adenocarcinoma protein samples, SCX also generated the greatest number of protein identifications.

The SCX and IEF platforms differed not only in the mode of peptide separation, but also in the pI range of peptides sampled. The IEF system employed narrow pI range (pI 3.5−4.5) focusing, which excludes many peptides from LC-MS/MS analysis altogether, whereas SCX entails no such restriction. In principle, this would enable SCX fractionation to present a greater diversity of peptides for LC-MS/MS analysis. However, the logic behind a narrow range IEF approach is that most proteins are represented by peptides in the selected pI range and that both high- and low-abundance proteins are represented by a few peptides in the narrow range.[23] The exclusion of other peptides from highly abundant proteins is intended to reduce bias toward highly redundant identifications of abundant proteins. Whether this strategy actually succeeds is hard to determine from our results. Any advantage of the narrow pI range peptide fractionation may have been compromised either by limited recovery of peptides from IPG strips or by components from the IPG strips that interfere with LC-

MS/MS peptide detection, particularly when small protein samples are analyzed with the IEF platform. We also found that carrier ampholytes, which are commonly used in protein IEF separations, are unnecessary for peptide resolution by IEF and actually suppress peptide identifications.

The IEF platforms offered the greatest reproducibility in peptide and protein identifications. This is best illustrated by the incremental gains in peptide and protein identifications with replicate analyses depicted in Figure 4; the second and third replicate analyses generated fewer additional identifications than did the second and third SCX replicates. This same characteristic of IEF is also illustrated by the steepness of the curves for accumulation of protein identifications in Figures 5 and 6. The IEF platform approaches the plateau with fewer replicates than does the SCX platform. This indicates that fewer replicate IEF analyses are needed to achieve a given degree of sampling of the peptide mixture.

The advantage in reproducibility for IEF over SCX is likely due in large part to superior resolution and reproducibility of peptide fractionation, as illustrated in Figure 2, which indicates that approximately 90% of all peptide identifications are found in a single fraction at both low and high sample loads. In contrast, SCX is characterized by spread of peptides into adjacent fractions. Peptides at lower abundance or those generating lower signal intensity are more likely to be selected for MS/MS if they appear in multiple fractions. SCX thus affords peptides more chances for detection, which accounts for the greater sensitivity and the greater diversity of peptide identifications with SCX. However, this sensitivity advantage is accompanied by a much higher variability in peptide detection at each fractionation step between replicate analyses (note the run-to-run differences in detected peptides for SCX fractions 2−4 in Figure 2).

What is the best shotgun proteomics platform for unbiased biomarker discovery in tissue specimens? Our results suggest

that the major differences between platforms are (1) sensitivity for protein detection and identification and (2) reproducibility of the protein inventories. Moreover, there is a tradeoff between IEF and SCX platforms between these two characteristics, particularly with smaller protein samples. For small samples (10 μg), SCX detected more than twice as many peptides and proteins as did IEF. This disparity narrowed considerably for larger protein inputs (50 and 100 μg), where the advantage of SCX is modest. Our observations are consistent with those of Essader et al., who showed that the use of SCX and IEF yielded similar numbers of peptide and protein identifications from complex proteomes and relatively large (mg) amounts of protein.[33]

IEF provides superior reproducibility of peptide and protein inventories, regardless of protein input. This is due mainly to superior resolution and reproducibility of peptide separations by IEF. The reproducibility advantage has practical implications for biomarker discovery work, because it dictates the number of replicate analyses needed to achieve a given level of inventory of proteins in the sample. Figure 6 illustrates this point. For both SCX and IEF, the most abundant proteins (those detected by >9 peptides) are all detected in a single replicate analysis. For less abundant proteins (those detected by 3−4 peptides), 6 replicate SCX analyses are needed to detect 90% of these proteins, whereas IEF detects 90% of these in 3 replicates. An IEF-based platform thus achieves a more consistent sampling of proteins in fewer replicate analyses than does an SCX-based platform.

Another point of comparison not explicitly addressed in our studies is the advantage of IEF over SCX in ease of use and potential for standardized adoption in multiple laboratories. Most laboratories using SCX prepare their own capillary columns and methods of column preparation and performance of the columns varies considerably. Failed analyses due to clogged columns, inadequate sample desalting leading to poor retention of peptides and other mishaps are relatively common. Aside from numbers of high quality MS/MS spectra and peptide identifications in the SCX fractions, there are no accessible metrics of system performance for SCX separations. In contrast, IEF separations using commercially obtained IPG strips provide a much more robust system. The match of commercially produced IPG strips and focusing apparatus also makes it considerably easier to implement standardized protocols. In addition, focusing current versus time traces provide a quality control record and metric for system performance. In our experience, failed analyses with the IEF platform described here occurred far less frequently than did failed SCX analyses. One final point in favor of an IEF-based peptide fractionation approach is the potential value of using calculated peptide pI information and IEF fraction pI as a filter in database searching and protein identification. The utility of the approach has been demonstrated previously.[21,23,34] We did not employ pI filtering in our analyses because there is no equivalent approach applicable to SCX-fractionated peptides.

In a biomarker discovery project, there is a tradeoff between the need for depth of protein inventory (i.e., numbers of identified proteins) and sample throughput. Identification of candidate biomarkers requires statistical comparisons of data sets from replicate shotgun proteome analyses of multiple tissue specimens for each phenotype. Variability in platform performance between replicate analyses can obscure real differences between tissue proteomes or create the appearance of difference when none actually exists. Any advantage of SCX over IEF in numbers of protein identifications could be compromised by the greater variability of the SCX platform. To compensate for the higher variability of SCX, greater numbers of replicate analyses are needed for a meaningful statistical analysis, which compromises overall sample throughput. Thus, platform reproducibility becomes the paramount consideration favoring the IEF-based platform for biomarker discovery. On the basis of these results, we have adopted an IEF-based platform for shotgun proteomics-based biomarker discovery in tissue specimens.

References

(1) Rifai, N.; Gillette, M. A.; Carr, S. A. Protein biomarker discovery and validation: the long and uncertain path to clinical utility. *Nat. Biotechnol.* **2006**, *24* (8), 971–983.

(2) Cravatt, B. F.; Simon, G. M.; Yates, J. R. 3rd, The biological impact of mass-spectrometry-based proteomics. *Nature* **2007**, *450* (7172), 991–1000.

(3) Omenn, G. S.; States, D. J.; Adamski, M.; Blackwell, T. W.; Menon, R.; Hermjakob, H.; Apweiler, R.; Haab, B. B.; Simpson, R. J.; Eddes, J. S.; Kapp, E. A.; Moritz, R. L.; Chan, D. W.; Rai, A. J.; Admon, A.; Aebersold, R.; Eng, J.; Hancock, W. S.; Hefta, S. A.; Meyer, H.; Paik, Y. K.; Yoo, J. S.; Ping, P.; Pounds, J.; Adkins, J.; Qian, X.; Wang, R.; Wasinger, V.; Wu, C. Y.; Zhao, X.; Zeng, R.; Archakov, A.; Tsugita, A.; Beer, I.; Pandey, A.; Pisano, M.; rews, P.; Tammen, H.; Speicher, D. W.; Hanash, S. M. Overview of the HUPO Plasma Proteome Project: results from the pilot phase with 35 collaborating laboratories and multiple analytical groups, generating a core dataset of 3020 proteins and a publicly-available database. *Proteomics* **2005**, *5* (13), 3226–3245.

(4) Taylor, S. W.; Fahy, E.; Zhang, B.; Glenn, G. M.; Warnock, D. E.; Wiley, S.; Murphy, A. N.; Gaucher, S. P.; Capaldi, R. A.; Gibson, B. W.; Ghosh, S. S. Characterization of the human heart mitochondrial proteome. *Nat. Biotechnol.* **2003**, *21* (3), 281–286.

(5) Hall, N.; Karras, M.; Raine, J. D.; Carlton, J. M.; Kooij, T. W.; Berriman, M.; Florens, L.; Janssen, C. S.; Pain, A.; Christophides, G. K.; James, K.; Rutherford, K.; Harris, B.; Harris, D.; Churcher, C.; Quail, M. A.; Ormond, D.; Doggett, J.; Trueman, H. E.; Mendoza, J.; Bidwell, S. L.; Rajandream, M. A.; Carucci, D. J.; Yates, J. R., 3rd; Kafatos, F. C.; Janse, C. J.; Barrell, B.; Turner, C. M.; Waters, A. P.; Sinden, R. E. A comprehensive survey of the Plasmodium life cycle by genomic, transcriptomic, and proteomic analyses. *Science* **2005**, *307* (5706), 82–86.

(6) Yi, E. C.; Marelli, M.; Lee, H.; Purvine, S. O.; Aebersold, R.; Aitchison, J. D.; Goodlett, D. R. Approaching complete peroxisome characterization by gas-phase fractionation. *Electrophoresis* **2002**, *23* (18), 3205–3216.

(7) Adachi, J.; Kumar, C.; Zhang, Y.; Olsen, J. V.; Mann, M. The human urinary proteome contains more than 1500 proteins, including a large proportion of membrane proteins. *GenomeBiology* **2006**, *7* (9), R80.

(8) de Souza, G. A.; Godoy, L. M.; Mann, M. Identification of 491 proteins in the tear fluid proteome reveals a large number of proteases and protease inhibitors. *GenomeBiology* **2006**, *7* (8), R72.

(9) Adachi, J.; Kumar, C.; Zhang, Y.; Mann, M. In-depth analysis of the adipocyte proteome by mass spectrometry and bioinformatics. *Mol. Cell. Proteomics* **2007**, *6* (7), 1257–1273.

(10) States, D. J.; Omenn, G. S.; Blackwell, T. W.; Fermin, D.; Eng, J.; Speicher, D. W.; Hanash, S. M. Challenges in deriving high-confidence protein identifications from data gathered by a HUPO plasma proteome collaborative study. *Nat. Biotechnol.* **2006**, *24* (3), 333–338.

(11) Yates, J. R., III. Mass spectral analysis in proteomics. *Annu. Rev. Biophys. Biomol. Struct.* **2004**, *33*, 297–316.

(12) Liu, H.; Lin, D.; Yates, J. R., III. Multidimensional separations for protein/peptide analysis in the post-genomic era. *BioTechniques* **2002**, *32* (4), 898, 900, 902.

(13) Link, A. J.; Eng, J.; Schieltz, D. M.; Carmack, E.; Mize, G. J.; Morris, D. R.; Garvik, B. M.; Yates, J. R., III. Direct analysis of protein complexes using mass spectrometry. *Nat. Biotechnol.* **1999**, *17* (7), 676–682.

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

(15) Wolters, D. A.; Washburn, M. P.; Yates, J. R. An automated multidimensional protein identification technology for shotgun proteomics. *Anal. Chem.* **2001**, *73* (23), 5683–5690.

(16) Peng, J.; Elias, J. E.; Thoreen, C. C.; Licklider, L. J.; Gygi, S. P. Evaluation of multidimensional chromatography coupled with tandem mass spectrometry (LC/LC-MS/MS) for large scale protein analysis: the yeast proteome. *J. Proteome Res.* **2003**, *2*, 43–50.

(17) Chen, J.; Balgley, B. M.; DeVoe, D. L.; Lee, C. S. Capillary isoelectric focusing-based multidimensional concentration/separation platform for proteome analysis. *Anal. Chem.* **2003**, *75* (13), 3145–3152.

(18) Guo, T.; Wang, W.; Rudnick, P. A.; Song, T.; Li, J.; Zhuang, Z.; Weil, R. J.; DeVoe, D. L.; Lee, C. S.; Balgley, B. M. Proteome analysis of microdissected formalin-fixed and paraffin-embedded tissue specimens. *J. Histochem. Cytochem.* **2007**, *55* (7), 763–772.

(19) Wang, W.; Guo, T.; Rudnick, P. A.; Song, T.; Li, J.; Zhuang, Z.; Zheng, W.; Devoe, D. L.; Lee, C. S.; Balgley, B. M. Membrane proteome analysis of microdissected ovarian tumor tissues using capillary isoelectric focusing/reversed-phase liquid chromatography-tandem MS. *Anal. Chem.* **2007**, *79* (3), 1002–1009.

(20) Chen, J.; Lee, C. S.; Shen, Y.; Smith, R. D.; Baehrecke, E. H. Integration of capillary isoelectric focusing with capillary reversed-phase liquid chromatography for two-dimensional proteomics separation. *Electrophoresis* **2002**, *23* (18), 3143–3148.

(21) Cargile, B. J.; Bundy, J. L.; Freeman, T. W.; Stephenson, J. L., Jr. Gel based isoelectric focusing of peptides and the utility of isoelectric point in protein identification. *J. Proteome Res.* **2004**, *3* (1), 112–119.

(22) Cargile, B. J.; Talley, D. L.; Stephenson, J. L., Jr. Immobilized pH gradients as a first dimension in shotgun proteomics and analysis of the accuracy of pI predictability of peptides. *Electrophoresis* **2004**, *25* (6), 936–945.

(23) Cargile, B. J.; Sevinsky, J. R.; Essader, A. S.; Stephenson, J. L., Jr.; Bundy, J. L. Immobilized pH gradient isoelectric focusing as a first-dimension separation in shotgun proteomics. *J. Biomol. Tech.* **2005**, *16* (3), 181–189.

(24) Stahl, D. C.; Swiderek, K. M.; Davis, M. T.; Lee, T. D. Data-controlled automation of liquid chromatography/tandem mass spectrometry analysis of peptide mixtures. *J. Am. Soc. Mass Spectrom.* **1995**, *7*, 532–540.

(25) Wang, H.; Qian, W. J.; Chin, M. H.; Petyuk, V. A.; Barry, R. C.; Liu, T.; Gritsenko, M. A.; Mottaz, H. M.; Moore, R. J.; Camp Ii, D. G.; Khan, A. H.; Smith, D. J.; Smith, R. D. Characterization of the mouse brain proteome using global proteomic analysis complemented with cysteinyl-peptide enrichment. *J. Proteome Res.* **2006**, *5* (2), 361–369.

(26) 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.

(27) 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.

(28) Elias, J. E.; Haas, W.; Faherty, B. K.; Gygi, S. P. Comparative evaluation of mass spectrometry platforms used in large-scale proteomics investigations. *Nat. Methods* **2005**, *2* (9), 667–675.

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

(30) Durr, E.; Yu, J.; Krasinska, K. M.; Carver, L. A.; Yates, J. R.; Testa, J. E.; Oh, P.; Schnitzer, J. E. Direct proteomic mapping of the lung microvascular endothelial cell surface in vivo and in cell culture. *Nat. Biotechnol.* **2004**, *22* (8), 985–992.

(31) Resing, K. A.; Meyer-Arendt, K.; Mendoza, A. M.; Aveline-Wolf, L. D.; Jonscher, K. R.; Pierce, K. G.; Old, W. M.; Cheung, H. T.; Russell, S.; Wattawa, J. L.; Goehle, G. R.; Knight, R. D.; Ahn, N. G. Improving reproducibility and sensitivity in identifying human proteins by shotgun proteomics. *Anal. Chem.* **2004**, *76* (13), 3556–3568.

(32) Whiteaker, J. R.; Zhang, H.; Zhao, L.; Wang, P.; Kelly-Spratt, K. S.; Ivey, R. G.; Piening, B. D.; Feng, L. C.; Kasarda, E.; Gurley, K. E.; Eng, J. K.; Chodosh, L. A.; Kemp, C. J.; McIntosh, M. W.; Paulovich, A. G. Integrated pipeline for mass spectrometry-based discovery and confirmation of biomarkers demonstrated in a mouse model of breast cancer. *J. Proteome Res.* **2007**, *6* (10), 3962–3975.

(33) Essader, A. S.; Cargile, B. J.; Bundy, J. L.; Stephenson, J. L., Jr. A comparison of immobilized pH gradient isoelectric focusing and strong-cation-exchange chromatography as a first dimension in shotgun proteomics. *Proteomics* **2005**, *5* (1), 24–34.

(34) Cargile, B. J.; Bundy, J. L.; Stephenson, J. L., Jr. Potential for false positive identifications from large databases through tandem mass spectrometry. *J. Proteome Res.* **2004**, *3* (5), 1082–1085.