Multidimensional Protein Identification Technology (MudPIT): Technical Overview of a Profiling Method Optimized for the Comprehensive Proteomic Investigation of Normal and Diseased Heart Tissue

Thomas Kislinger,* Anthony O. Gramolini, David H. MacLennan, and Andrew Emili*

Banting and Best Department of Medical Research, University of Toronto, Toronto, Ontario, Canada

An optimized analytical expression profiling strategy based on gel-free multidimensional protein identification technology (MudPIT) is reported for the systematic investigation of biochemical (mal)-adaptations associated with healthy and diseased heart tissue. Enhanced shotgun proteomic detection coverage and improved biological inference is achieved by pre-fractionation of excised mouse cardiac muscle into subcellular components, with each organellar fraction investigated exhaustively using multiple repeat MudPIT analyses. Functional-enrichment, high-confidence identification, and relative quantification of hundreds of organelle- and tissue-specific proteins are achieved readily, including detection of low abundance transcriptional regulators, signaling factors, and proteins linked to cardiac disease. Important technical issues relating to data validation, including minimization of artifacts stemming from biased under-sampling and spurious false discovery, together with suggestions for further fine-tuning of sample preparation, are discussed. A framework for follow-up bioinformatic examination, pattern recognition, and data mining is also presented in the context of a stringent application of MudPIT for probing fundamental aspects of heart muscle physiology as well as the discovery of perturbations associated with heart failure. (J Am Soc Mass Spectrom 2005, 16, 1207–1220) © 2005 American Society for Mass Spectrometry

Cardiomyopathies are diseases of the heart which impair cardiac muscle function that can progress to heart dilatation and cardiac failure. Heart failure represents a leading cause of morbidity and death globally. Due to the poor prognostic outcome of late stage disease, innovative preventive and therapeutic measures are needed urgently for the early detection, categorization, and treatment of at-risk patients [1]. These developments will require a more complete molecular understanding of the molecular basis of normal heart function and the pathophysiological effects of impaired cardiac function associated with disease.

The trigger for cardiac contraction is the elevation of myoplasmic Ca^{2+} concentrations, mediated by Ca^{2+}-release channels (ryanodine receptors; RyRs) that tap the Ca^{2+} store in the lumen of the sarcoplasmic reticulum (SR) and plasma-membrane Ca^{2+} channels (dihydropyridine receptors; DHPRs) that tap the high concentrations of Ca^{2+} in the extracellular space [2]. The trigger for relaxation is the lowering of myoplasmic Ca^{2+} concentration by the combined activity of the sarco(endo)plasmic reticulum Ca^{2+}-ATPase (SERCA), the plasma membrane Ca^{2+}-ATPases (PMCAs), and Na^{+}/Ca^{2+} exchangers (NCXs). In humans, the activity of SERCA2 (the cardiac specific isoform) determines the rate of removal of >70% of cytosolic Ca^{2+} [2, 3], thereby determining the rate of relaxation of the heart, and influencing cardiac contractility by determining the size of the luminal Ca^{2+} store that is available for release in the next beat. Proper regulation of Ca^{2+} flux is central to normal heart function. This regulation is perturbed in most, if not all, cardiomyopathies.

Cardiac function is regulated on a beat-to-beat basis through the sympathetic nervous system [3]. When demand arises, the heart can respond to stress and increase blood flow to peripheral tissues within sec-

Published online June 23, 2005

Address reprint requests to Dr. A. Emili, CH Best Institute, Room 402, 112 College Street, Toronto, Ontario M5G 1L6, Canada. E-mail: andrew.emili@utoronto.ca

* Also with the Program in Proteomics and Bioinformatics, University of Toronto, Toronto, Ontario, Canada.

1044-0305/05/$30.00
doi:10.1016/j.jasms.2005.02.015

Received December 22, 2004
Revised December 22, 2004
Accepted February 23, 2005

onds. This is due to the large cardiac reserve in humans; the slow basal heart beat rate and submaximal contractility at rest are increased markedly after the release of adrenaline into the blood [3]. Adrenaline and other β-agonists initiate an important stimulatory signal-transduction pathway in the heart by binding to and activating β-adrenergic receptors present on the cell outer membrane. The signal proceeds through G_s proteins, leading to the formation and accumulation of cyclic AMP by adenylate cyclase. Elevations in cAMP concentration cause activation cAMP-dependent protein kinase (PKA), which then phosphorylates and alters the function of key cardiac proteins regulating overall cardiac function. Prominent among these proteins is phospholamban (PLN), a small, reversibly phosphorylated transmembrane protein that is located in the cardiac SR [4]. Depending on its phosphorylation state, PLN binds to and regulates the activity of SERCA2a (the cardiac specific isoform). The dephosphorylated form of PLN inhibits SERCA activity by reducing its affinity for Ca^{2+}. This inhibition is overcome by phosphorylation of PLN by either protein kinase A (PKA), calcium-calmodulin kinase (Cam-kinase), or protein kinase C (PKC) which, by relieving Ca^{2+}-pump inhibition, enhances relaxation rates and contractility [4].

Hypertrophic cardiac disease, caused by ischemic heart disease or tissue damage stemming from myocardial infarction, often progresses into dilated cardiomyopathy [5]. The cellular mechanisms that underlie this progression are poorly understood. Cardiac hypertrophy, hypertension and heart failure are linked to impaired cardiomyocyte function, and one form of impairment stems from perturbed Ca^{2+} regulation [2]. A better understanding of the full complement of proteins perturbed in cardiac tissue during progression to heart failure induced by impaired Ca^{2+} signaling and the role of these factors in disease pathogenesis is a major focus of our ongoing collaborative proteomic research program [6].

Several validated mouse genetic and transgenic models are available to investigate cardiomyopathy and heart failure at a detailed molecular level not currently feasible in humans. For instance, transgenic mice overexpressing a human disease point mutant variant in PLN (R9C) die early of severe dilated cardiomyopathy (within four to five months of age). It is possible to examine the phenotype of the affected cardiac tissue of such mice at various stages of pathology, even prior to overt presentation of clinical symptoms, and to examine the corresponding proteome at select time-points in the disease progression using shotgun proteomic approaches [7]. When this potential is combined with the use of high penetrance inbred strains to minimize the influence of genetic variance, a more refined investigation of the course of disease action can be carried out. Mouse models also provide a useful setting for investigation of the cellular responses and long-term effects of clinically relevant therapeutic interventions [8–10].

Heart myocytes are predicted to express several thousand distinct protein species [11–13], several hundred of which are likely to be tissue-specific and hence critical for proper cardiac performance and capacity. Although a number of gene products predisposing to cardiomyopathy have been reported to date (dystrophin, for example [14]) based on known or predicted heart-related functions, identification of the full set of proteins associated with this "complex trait" has proven to be a challenge [15]. A comprehensive, nonbiased description of the proteome or set of expressed proteins in healthy and diseased cardiac tissue could provide a breakthrough in understanding of the pathogenesis of heart disease by furthering knowledge of unknown critical disease pathways, leading to novel diagnostic and therapeutic targets.

Nevertheless, the complexity and markedly skewed composition of the cardiac muscle proteome represents a considerable experimental challenge and effective sample fractionation methods are required in order to detect low abundance proteins [16]. Historically, 2D-gel electrophoresis has provided a useful method for high-resolution separation of complex protein samples, including cardiac tissue [13]. However, gel-based proteomic techniques are generally biased towards detection of high abundance housekeeping enzymes, with reduced detection of low abundance proteins, membrane proteins, and proteins with extremes in isoelectric point and molecular weight [17, 18], limitations that are further compounded by the need to analyze many individual gel spots. To circumvent these problems, several groups have developed gel-free protein expression profiling strategies coupling high-efficiency liquid chromatography separation procedures with automated tandem mass spectrometry, allowing for large-scale 'shotgun' sequencing of complex mixtures [16]. The archetypal approach, termed MudPIT (for Multidimensional Protein Identification Technology) [19], pioneered in the laboratory of John Yates, III, has proven to be a remarkably effective and robust methodology for investigating global changes in protein expression as a function of development and disease [20–22].

Our group has been evaluating the utility of MudPIT as a method for investigating the molecular basis of normal heart physiology and disease perturbed cardiomyopathies in a controlled animal model setting [6]. Here, we outline some of the more challenging technical and analytical issues that have become apparent in our pilot studies, and offer helpful experimental, technical, and computational solutions that we have developed to date which allow for a more comprehensive and reliable analysis of healthy and diseased mammalian cardiac tissue. While centered on heart disease, the analytical approach described here is broadly applicable to a range of biomedical problems, and hence should be of general interest to investigators contemplating or currently applying MudPIT to generate a detailed molecular description of the proteomic patterns of cells, tissues and/or organelles of special focus.

Materials and Methods

Heart Homogenization and Organelle Isolation

Hearts were isolated from adult mice (24 weeks of age), atria were removed, and the ventricles carefully minced with a razor blade and rinsed extensively with ice-cold PBS (phosphate buffered saline) to remove excess blood. Tissue was homogenized for 30 s using a loose-fitting hand-held glass homogenizer in 10 ml lysis buffer (250 mM sucrose, 50 mM Tris-HCl pH 7.6, 1 mM $MgCl_2$, 1 mM DDT (dithiothreitol), 1 mM PMSF (phenylmethylsulphonyl fluoride). All subsequent steps were performed at 4 °C. The lysate was centrifuged in a benchtop centrifuge at 800 × *g* for 15 min; the supernatant served as a source for cytosol, mitochondria, and microsomal fractions. The pellet containing nuclei was diluted in 8 ml of lysis buffer and layered onto 4 ml of 0.9 M sucrose buffer (0.9 M sucrose, 50 mM Tris-HCl pH 7.6, 1 mM $MgCl_2$, 1 mM DDT, 1 mM PMSF) and centrifuged at 1000 × *g* for 20 min at 4 °C. The resulting pellet was resuspended in 8 ml of a 2 M sucrose buffer (2 M sucrose, 50 mM Tris-HCl pH 7.4, 5 mM $MgCl_2$, 1 mM DTT, and 1 mM PMSF), layered onto 4 ml of 2 M sucrose buffer and pelleted by ultracentrifugation at 150,000 × *g* for 1 h (Beckman SW40.1 rotor). The nuclei were recovered as a pellet. The mitochondria were isolated from the supernatant by re-centrifugation at 7500 × *g* for 20 min at 4 °C; the resulting pellet was washed twice in lysis buffer. Microsomes were pelleted by ultracentrifugation of the post-mitochondrial cytoplasm at 100,000 × g for 1 h in a Beckman SW41 rotor. The supernatant served as the cytosolic fraction.

Organelle Extraction

Nuclear proteins were extracted, followed by resuspension and incubation of the nuclei in 5 vol of high salt buffer (0.5 M NaCl, 20 mM HEPES (4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid) pH 7.9, 1.5 mM $MgCl_2$, 0.2 mM EDTA, 1 mM DTT, 1 mM PMSF, 20 % glycerol) on ice for 30 min with gentle shaking. The nuclei were then lysed by 10 passages through an 18-gauge needle, and debris was removed by centrifugation at 13,000 × *g* in a microfuge for 30 min. The supernatant served as the "NUC 1" fraction (see Figure 1), while the insoluble pellet was resuspended in 5 vol of high salt buffer containing 1% Triton-X-100 detergent and shaken gently for 30 min. The suspension was sheared by 10 passages through an 18-gauge needle and debris removed by centrifugation at 13,000 *g* for 30 min. The supernatant served as the "NUC 2" fraction.

Soluble mitochondrial proteins were extracted by incubating the mitochondria in hypotonic lysis buffer (10 mM HEPES, pH 7.9, 1 mM DTT, 1 mM PMSF), for 30 min on ice. The suspension was sonicated briefly and debris removed by centrifugation at 13,000 × *g* for 30 min. The supernatant served as the "MITO 1" fraction. The resulting insoluble pellet was resuspended in membrane detergent extraction buffer (20 mM Tris-HCl, pH 7.8, 0.4 M NaCl, 15% glycerol, 1 mM DTT, 1 mM PMSF, 1.5% Triton-X-100) and shaken gently for 30 min followed by centrifugation at 13,000 × *g* for 30 min; the supernatant served as "MITO 2" fraction.

Membrane-associated proteins were extracted by resuspending the microsomes in membrane detergent extraction buffer. The suspension was incubated with gentle shaking for 1 h and insoluble debris removed by centrifugation at 13,000 × *g* for 30 min. The supernatant served as the "MICRO" fraction.

Digestion of Organelle Extracts and MudPIT Analysis

An aliquot of ~100 μg total protein (as determined by Bradford assay) from each fraction was precipitated overnight with 5 vol of ice-cold acetone at −20 °C, followed by centrifugation at 13,000 × *g* for 15 min. The protein pellet was solubilized in a small volume of 8 M urea, 50 mM Tris-HCl, pH 8.5, 1 mM DTT, for 1 h at 37 °C, followed by carboxyamidomethylation with 5 mM iodoacetamide for 1 h at 37 °C in the dark. The samples were then diluted to 4 M urea with an equal vol of 100 mM ammonium bicarbonate, pH 8.5, and digested with a 1:150-fold ratio of endoproteinase Lys-C (Roche Diagnostics, Laval, Quebec, Canada) at 37 °C overnight. The next day, the samples were diluted to 2 M urea with an equal vol of 50 mM ammonium bicarbonate pH 8.5, supplemented with $CaCl_2$ to a final concentration of 1 mM, and incubated overnight with Poroszyme trypsin beads (Applied Biosystems, Streetsville, Ontario, Canada) at 30 °C with rotation. The resulting peptide mixtures were solid phase-extracted with SPEC-Plus PT C18 cartridges (Ansys Diagnostics, Lake Forest, CA) according to the instructions of the manufacturer and stored at −80 °C until further use.

A fully-automated 20 h long 12-step multi-cycle MudPIT procedure was set up as described previously [16]. Briefly, an HPLC quaternary pump was interfaced with an LCQ DECA XP ion trap mass spectrometer (Thermo Finnigan, San Jose, CA). A 100-μm i.d. fused silica capillary microcolumn (Polymicro Technologies, Phoenix, AZ) was pulled to a fine tip using a P-2000 laser puller (Sutter Instruments, Novato, CA) and packed with 8 cm of 5 μm Zorbax Eclipse XDB-C_{18} resin (Agilent Technologies, Mississauga, Ontario, Canada), followed by 6 cm of 5 μm Partisphere strong cation exchange resin (Whatman, Clifton, NJ). Individual samples were loaded manually onto separate columns using a pressure vessel. Chromatography solvent conditions were exactly as described earlier [16].

Protein Identification and Validation

The SEQUEST database search algorithm [23] was used to match peptide tandem mass spectra to peptide sequences in a locally-maintained minimally redundant FASTA formatted database populated with mouse and human protein sequences obtained from the Swiss-

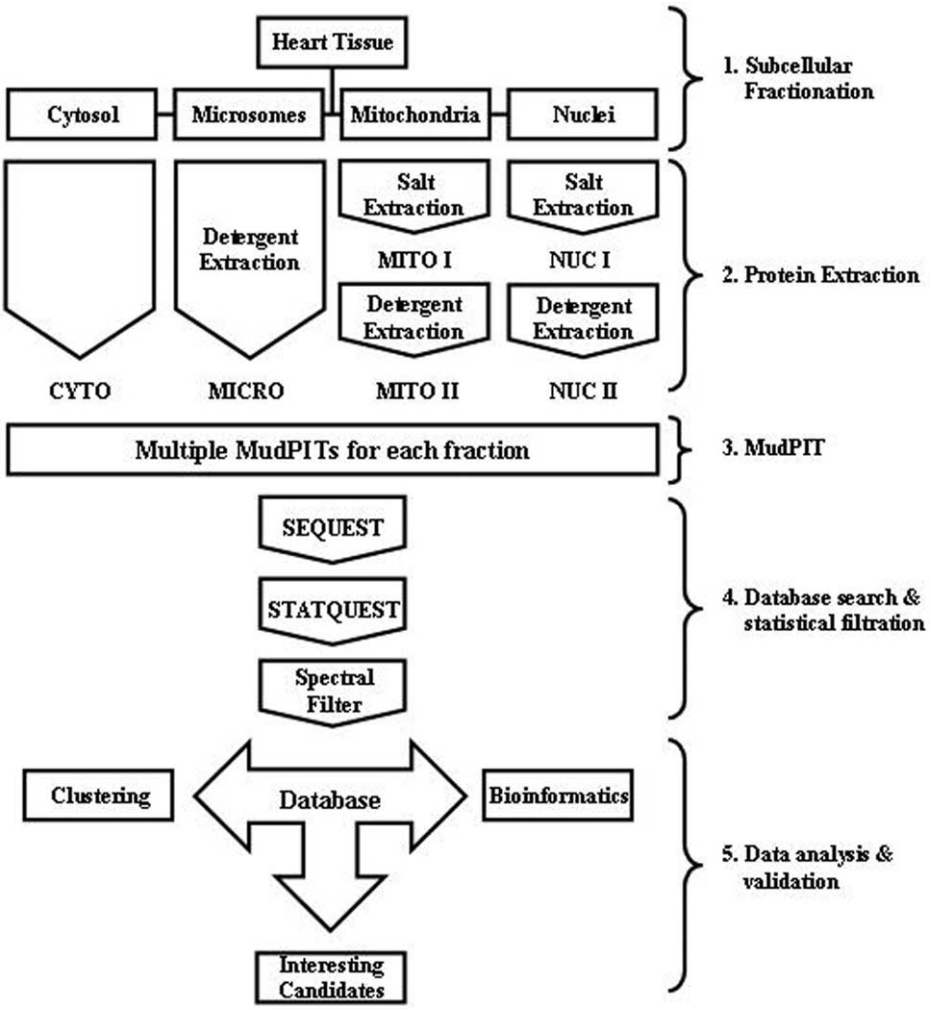

Figure 1. Overview of a systematic protein profiling methodology. Mouse heart tissue is homogenized and subcellular fractions isolated by differential ultracentrifugation using sucrose density gradients. Protein extracts generated from each organelle are digested, and the peptide mixtures analyzed by multiple independent MudPIT-based shotgun sequencing experiments. The generated tandem mass spectra are searched against a comprehensive non-redundant protein sequence database using the SEQUEST algorithm, and filtered statistically to minimize false positive identifications. High confidence putative protein identifications are parsed into a relational database and diverse data mining strategies used to find biologically interesting patterns for more extensive back-up analysis.

Prot/TrEMBL and IPI databases. To statistically assess the empirical False-Discovery Rate to control for, and hence, minimize false positive identifications [16], all of the spectra were searched against protein sequences in both the normal (Forward) and inverted (Reverse) amino acid orientations. The STATQUEST filtering algorithm was then applied to all putative search results to obtain a measure of the statistical reliability (confidence score) for each candidate identification (cutoff p-value $\leq$.15, corresponding to an 85% or greater likelihood of being a correct match).

Database

High-confidence matches were parsed into an in-house SQL-type database using a Perl-based script. The database was designed to accommodate database search results and spectral information (scan headers) for multiple peptides matching to a given protein, together with information regarding the sample name, experiment number, MudPIT step, organelle source, amino acid sequence, molecular mass, isoelectric point, charge, and confidence level. For this report, only those proteins with a predicted confidence p value of $\geq$95%, and for which at least two spectra were collectively detected, were retained for further analysis.

Immunoblots, Enzyme Activity, and DNA Levels

Protein samples were separated by denaturing sodium dodecyl sulphate-polyacrylamide gel electrophoresis (SDS-PAGE) using standard procedures. Commercial antibodies to calsequestrin (Affinity Bioreagents Inc;

ABI, Golden, CO), ryanodine receptor (RYR2; ABI), PKC-β (BD Transduction laboratories, Lexington, KY), α-actinin (Sigma-Aldrich, Oakville, Ontario, Canada) and myogenin (Developmental Studies Hybridoma Bank, Iowa City, IA) were used. Antibodies for the mitochondrial F1-ATP synthase α- and β-subunit were kindly provided by Dr. Peter Pedersen (Johns Hopkins Medical Institute, Boston, MA). Horseradish peroxidase-conjugated goat anti-mouse secondary antibody and enhanced chemiluminescence (Super Signal; Pierce, Rockford, IL) were used for visualization; signals were quantified using Image-J software (National Institutes of Health, Bethesda, MA).

Enzyme activities were determined using commercial assays (Sigma-Aldrich Lactate Dehydrogenase (LDH) assay kit, cat. no. DG 1340- K; creatine kinase assay kit (CK), cat. no. 47-20).

Microscopy of Isolated Nuclei

Aliquots of partially purified nuclei were applied to standard microscope slides and visualized using a Leica TCS SP laser scanning confocal system (Leica; Richmond Hill, Ontario, Canada).

Hierarchical Clustering, Data Visualization, and Cluster Evaluation

The cumulative spectral count was used as a semi-quantitative metric for estimating relative protein abundance, as described by Liu et al. [24]. Hierarchical clustering was performed using the Cluster 3.0 freeware software package [25] and the Spearman distance metric, with average linkage selected. To improve the consistency of data grouping, a nominal low non-zero (0.01) value was substituted for blank (missing) values in cases where a protein was not detected in a particular sample. The clustered profiles were visualized in heat map format using the TreeView software package [25].

Statistical enrichment of cluster membership to select functional annotation categories obtained from the Gene Ontology database (GO terms) was assessed using the hypergeometric distribution [26], which returns the probability (p-value) that the intersection of a given protein list with a given annotation class occurs by chance. To account for spurious significance due to multi-hypothesis testing (multiple GO-terms), a Bonferroni correction factor was applied; scores were amended by dividing the preliminary p-value by the number of tests conducted. A threshold cut-off p-value of 10^{-3} was used as a final selection criterion to highlight statistically significant and potentially biologically interesting clusters.

Results and Discussion

Heart disease is the leading cause of mortality and morbidity in the world [27]. Heart failure, in particular, is a major emerging epidemic, due to improved survival from acute cardiac syndromes (e.g., myocardial infarction) and the aging population with increasing risk factors such as hypertension and diabetes. Even though early detection usually permits successful therapeutic intervention, most early stage heart disease is not detected clinically until irreversible tissue damage accrues. Symptomatic heart failure occurs in ~1.5% of the population with an estimated population rate of asymptomatic ventricular dysfunction at ~5% in Western developed countries. Diagnosis only occurs when patients become overtly symptomatic (e.g., shortness of breath or peripheral edema). Once the symptoms occur, patients face recurrent hospitalization and a very high (~1/3) one year mortality [28]. Heart failure is presently the most costly health care diagnosis [28]. Although therapies for prevention are already available, screening of at-risk asymptomatic patients with early stages of heart failure is not routine due to a lack of effective tools. The discovery of biomarkers allowing for early diagnosis and therapeutic monitoring of at-risk patients is therefore needed urgently. Systematic analytical methods for determining the genetic, biochemical, and physiological basis of normal heart homeostasis and the deficiencies associated with progression of heart disease would be highly beneficial.

High-throughput high-resolution experimental technologies, such as DNA microarray gene chips [29] and gel-free mass spectrometry, [30, 31] have emerged over the past few years as powerful platforms for investigation of the molecular underpinnings of disease progression on a systems-wide level. Over the last two years, our group has been active in the development, optimization, and application of a proteomic expression profiling platform for the analysis of the global protein composition of heart and other organs using mouse as a primary model system [6, 16]. Our strategy is based on a combination of extensive tissue pre-fractionation, exhaustive MudPIT analysis of each isolated fraction, and back-end informatics analysis to provide a meaningful biological context. To this end, we have both adapted existing and developed novel biochemical, computational, and statistical tools for evaluating, validating, and mining large-scale protein expression datasets [16]. Below, we discuss the major conceptual, technical, and analytical difficulties that we have encountered along the way, as well as our attempts to resolve these problems. We also describe some important considerations that need to be taken into account when interpreting large-scale MudPIT-derived proteomic datasets.

Figure 1 provides a schematic overview of an analytical procedure optimized for routine global proteomic profiling of cardiac tissue. As a critical first step, simple, well-established, and highly reproducible biochemical procedures based on differential sucrose gradient ultracentrifugation are used to fractionate homogenized mouse heart prior to MudPIT analysis. This methodology yields four distinct subcellular fractions

[nuclei, mitochondria, microsomes (membranes), and cytosol]. Although these fractions are not pure and contain cross-contaminants [16], this easily implemented strategy offers several important analytical advantages for profiling studies. First, sample complexity is reduced significantly, thereby allowing for more comprehensive detection of the myriad of lower abundance proteins typically expressed in myocytes in addition to the high abundance components of the contractile apparatus. This is a particularly important consideration given the heterogeneous cell types found in tissues and organs, the sizeable overall dynamic range in absolute protein abundance (>five orders of magnitude; [32]), together with the substantive undersampling (detection bias) typically observed even with high-efficiency profiling methods such as MudPIT ([16, 20, 22, 33]; these critical issues are discussed further in detail below). Second, since cells are spatially organized, with significant physical clustering of protein modules linked to specific biochemical activities, organellar profiling provides a more relevant biological context for interpretation of the physiological status of the target tissue based on observed protein patterns. Moreover, subcellular fractionation can offer additional insights into the major biochemical pathways affected by disease processes, as well as provide hints as to the potential function(s) of previously uncharacterized (unannotated) proteins.

Once isolated organellar fractions are obtained, both soluble and membrane associated proteins are extracted. These are digested efficiently prior to MudPIT analysis using the proteases endoproteinase Lys-C and trypsin. The peptide mixtures are desalted and analyzed individually using a variant of the basic MudPIT procedure. As discussed further below, we have found that multiple repeat MudPIT analyses are generally needed to obtain comprehensive proteomic detection coverage, even when investigating simplified organellar fractions.

Database Searching and Statistical Validation

The large collections of acquired tandem mass spectra are searched against an extensive database of high quality (curated) protein sequences using the SEQUEST algorithm [23]. To estimate the false discovery rate due to incorrect spurious database matches, the database is populated with an equal number of inverted protein sequences corresponding to reverse amino acid sequence orientations. Putative matches to these bogus "dummy" decoy control sequences are interpreted as false positives. To reduce the rate of misidentification (false positives), individual candidate database matches are evaluated and statistically filtered using STATQUEST, a pattern-recognition software algorithm developed in-house trained to distinguish false positives based on weighted SEQUEST scoring parameters [16]. High-confidence peptide sequence matches are selected using a stringent first-step confidence filter (cut-off p-value $\leq .05$) based on the likelihood ratio of detecting bogus reverse matches to forward candidate sequences. Next, the candidate proteins are sorted based on the cumulative number of matching spectra obtained at the organelle level (Figure 2, Table 1 and Table 2), computational procedures are carried out on data that have been formatted and parsed into an SQL-style relational database. Although this database was developed for local use, the reader is directed to the many helpful stand-alone informatics software tools presently available for managing, assessing, and filtering large-scale MudPIT-type proteomic datasets that are freely available to academics from the Yates and Aebersold research laboratories (e.g., DTASelect [34]; http://fields.scripps.edu/ and ProteinProphet [35]; www.systemsbiology.org/).

As an objective measure of the effectiveness of these quality filters, we first calculate the empirical rate of false-discovery based on the observed ratio of forward-to-reverse protein sequence matches (with reverse sequence matches considered as false positives). Figure 2 shows a plot of the distribution of the ratio of forward-to-reverse sequence matches as a measure of database search accuracy versus the spectral count acquired for each candidate protein. Clearly, the vast majority of spurious identifications are predicted based on a single spectra only as evidence and, consequently, are readily discarded from further analysis by applying a minimum two-spectra cutoff (data summarized in Table 2). By applying this stringent two-step filter system to representative proteomic datasets generated by an exhaustive MudPIT analysis of three non-nuclear organellar fractions isolated from healthy adult mouse heart, >1200 high-confidence proteins were identified (Table 1). The estimated false positive rate <5%, with most candidate proteins having a predicted confidence p-value $<.002$).

Heart Homogenization and Organelle Fractionation

Subcellular fractionation of heart tissue is challenging due to a number of unique technical problems [6, 13], most stemming from the extreme fibrous structure and grossly elevated levels of sarcomeric and mitochondrial proteins found in cardiac muscle.

Traditional techniques, such as Western blotting and enzyme assays, were used initially to assess the purity of the isolated organelle fractions. As seen in Figure 3, these methodologies ostensibly suggested reasonable purity for each of the four main isolated cellular components. For example, the muscle-specific transcription factor myogeninwas detected exclusively in the nuclear fraction, the cytosolic enzyme protein kinase C β isoform (PKC-β) was detected mainly in the cytosol, while SR proteins such as ryanodine receptor RyR2 and calsequestrin were detected uniquely in the microsomal fraction. Moreover, standard enzyme assays of the

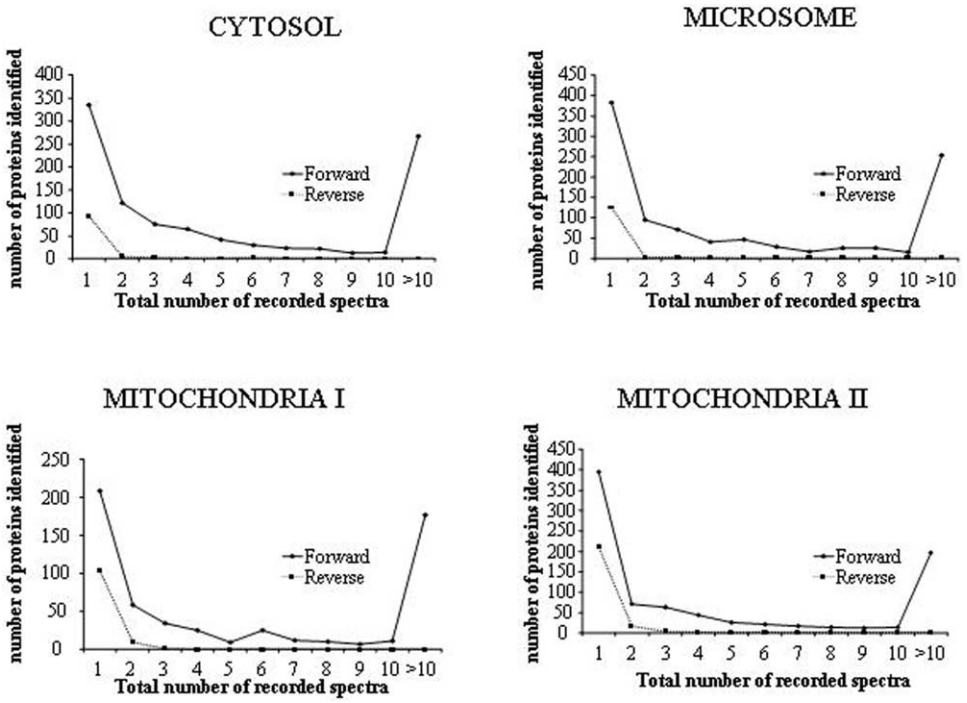

Figure 2. Two-step data filtration strategy. Proteins identified in each organelle fraction, passing a minimum 95%+ STATQUEST cutoff confidence filter, were sorted and placed into bins based on the observed spectral count (total number of spectra recorded for each protein identified). The plot shows the numbers of proteins found in each bin mapping to sequences in the forward SwissProt/TrEMBL database (Forward) versus those mapping to an equal sized control database composed of non-sense inverted decoy protein sequences (Reverse).

cytosolic enzymes creatine kinase and lactate dehydrogenase, clearly demonstrated enriched normalized activity in the cytosolic fraction.

Nonetheless, an early indication of substantive cross-contamination became apparent following proteomic analysis of these same fractions by MudPIT. For instance, the abundant muscle cytoskeletal protein α-actinin was detected in both the nuclear and mitochondrial fractions, a problem commonly seen with many other high abundance muscle factors (Figure 3a). The elevated load of mitochondria supporting heart function also leads to substantial cross-contamination of the nuclear compartment. Western blot analysis of mitochondrial markers, such as the β-subunit of the membrane-associated F_1-ATPase, generates very strong signals in the nuclear (and all other) fractions (Figure 3d; top panel). This cross-contamination was not entirely unexpected, given the highly specialized biomechanical adaptations of muscle, and has been reported previously by van Eyk and colleagues in 2D-gel proteomic studies of human cardiac tissue [36]. Nevertheless, it proved to be particularly debilitating for detection of proteins of biological interest. In fact, contaminating contractile apparatus (e.g., ventricular myosins) and even blood proteins (hemoglobins) were detected in nuclear extracts prepared from cardiac myocyte nuclei using a standard two-step centrifugation protocol following vigorous homogenization of heart tissue with an electrical blade (Polytron) [data not shown]. The degree to which this problem afflicts heart tissue is exemplified by a visual comparison of myocyte nuclear preparations to the highly pure nuclei isolated from mouse liver obtained using the same one-step protocol (Figure 3e; [16]).

Accordingly, we have optimized the heart nuclei isolation protocols, aiming to reduce the extent of sarcomeric and mitochondrial contaminants that result from aggressive disruption of tissue. We have found that gentle tissue homogenization using a handheld Dounce glass (B-type) pestle significantly minimizes gross cross-contamination (Figure 3d; middle panel). Nevertheless, given the vast excess of mitochondria relative to nuclei in differentiated cardiac myotubes, even more stringent fractionation procedures are still needed to eliminate cross-contamination of the nuclear compartment. We have determined that two sequential rounds of density gradient centrifugation using sucrose cushions can further alleviate this problem (Figure 3d; lower panel). The effectiveness of various isolation protocols is evident by visual inspection of light microscope images of the different nuclear preparations (Figure 3e). Most of the contractile contaminants seen in the single 0.9 M sucrose cushion [see the Materials and Methods section] were removed by passing the crude nuclei over a second 2 M sucrose cushion, albeit at the cost of obtaining a significantly reduced yield of nuclei

Table 1. Protein and spectral counts for Forward and Reverse database sequences for each repeat analysis

	Forward		Reverse	
Fraction (MudPIT experiment)	Proteins	Spectra	Proteins	Spectra
cyto1	452	4860	3	5
cyto2	538	3829	3	3
cyto3	501	3708	4	5
cyto4	472	3440	1	2
cyto5	456	3997	4	5
CYTO TOTAL	668	19834	7	20
micro1	413	4303	2	4
micro2	446	4178	1	1
micro3	390	4161	2	3
micro4	455	3714	6	11
micro5	434	3518	0	0
MICRO TOTAL	615	19874	7	19
mito I 1	279	2806	5	5
mito I 2	276	2687	4	4
mito I 3	210	1887	3	3
mito I 4	299	3095	1	1
mito I 5	313	5087	7	8
MITO I TOTAL	368	15562	10	21
mito II 1	349	3980	6	9
mito II 2	253	2960	9	11
mito II 3	346	3861	6	10
mito II 4	353	3827	9	11
mito II 5	397	4782	12	16
MITO II TOTAL	475	19410	23	57
TOTAL	1230	74680	41	117

in the most highly purified preparations (see Supplementary Figure 1 for a lower magnification view).

Protein Sampling

Shotgun sequencing of complex peptide mixtures involves a somewhat stochastic sampling process, whereby the mass spectrometer instrument sequentially selects individual peptide precursor ions for fragmentation. While database searching errors account for part of the variation limiting the overall reproducibility of profiling studies, the major deficiency is due to the finite duty cycle and limited overall dynamic range of current mass spectrometry instruments, such that not every precursor peptide ion species is sampled as peptides elute from the chromatography column. Even high-efficiency profiling methods, such as MudPIT, are generally biased towards preferential detection of higher abundance proteins (which typically produce higher intensity peptide signals). In contrast, lower abundance proteins (typically producing lower intensity peptide ion peaks) frequently go undetected or, at best, are prone to ion competition at the source, which ultimately leads to signal suppression and quantification artifacts. To overcome this "under-sampling" problem, repeated analysis of the same fraction has been suggested as a means of obtaining more complete proteomic coverage [24, 37].

Table 2. Effects of filler criteria on number of identified proteins and the false-positive rate

Fraction (Proteins identified)	1 spectra	2 spectra (sum)
CYTOSOL		
Forward	1002	668
Reverse	99	7
% (reverse/forward)	8.9	1
MICROSOMES		
Forward	998	615
Reverse	133	7
% (reverse/forward)	9.9	1.1
MITOCHONDRION I		
Forward	576	368
Reverse	113	10
% (reverse/forward)	19.6	2.7
MITOCHONDRION II		
Forward	870	475
Reverse	233	23
% (reverse/forward)	26.8	4.8

Figure 4 shows the proteomic coverage (detection efficiency) typically obtained by performing multiple repeat MudPIT analyses on the same heart cytosolic fraction. It can be seen from this plot that virtually complete sample saturation (plateau in the total number of proteins identified) is achieved after five individual MudPIT analyses. This pattern is typical of that seen with the other organellar fractions analyzed. Saturation of detection is an important consideration if one eventual aim is a comparison of the proteomic patterns of different samples (e.g., developmental time-points or healthy versus disease states).

To investigate further some of the reasons behind the

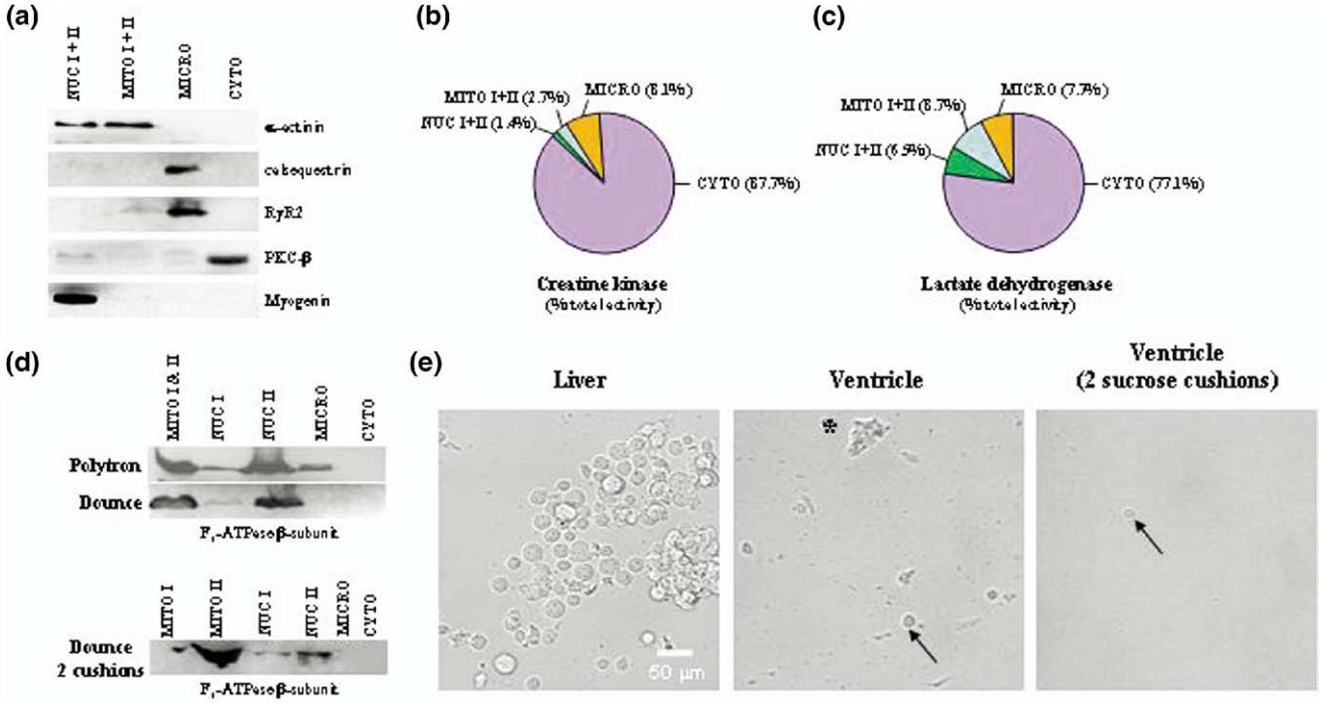

Figure 3. (**a**) Western blot analysis of subcellular fractions using antibodies directed against select marker proteins. (**b**) Normalized in vitro creatine kinase enzyme activity. (**c**) Normalized in vitro lactate dehydrogenase enzyme activity. (**d**) Western blot detection of the mitochondrial membrane protein F_1-ATPase β-subunit across various subcellular fractions. (**e**) Light microscopy images of nuclei isolated from liver using a basic single-step method or from heart using two different isolation protocols. Middle panel, in one protocol, crude ventricular nuclear preparations were passed through a single 2 M sucrose cushion. Far right panel, crude ventricular nuclei were subjected to two rounds of ultracentrifugation, first using a 0.9 M sucrose cushion, followed by a second 2 M sucrose cushion. Asterisk, muscle fiber; arrows, isolated nuclei.

random sample variation seen with profiling experiments, we examined more closely the behavior of an individual MudPIT injection of this representative fraction. As a first pass to assess the reproducibility of detection, each of the 668 high-confidence cytosolic proteins identified were binned according to the number of repeat MudPIT analyses in which they were detected. 40% of the proteins (266) were detected across all five experiments, while nearly two-thirds were found in four or fewer runs (Figure 5a). Moreover, only ~5% (32) of the proteins were detected in only a single dataset, demonstrating that proteomic coverage is saturated quite quickly by repeat analyses (Figure 5a). As might be expected, proteins found in every single MudPIT analysis were detected with a considerably higher average spectral count (67 spectra; range 5–1754 spectra) than proteins detected less reproducibly (a mean of 2.3 spectra was recorded for proteins detected in only one experiment) (Figure 5b), consistent with the notion that lower abundance proteins are more likely to remain undetected because of the effects of random sampling.

To exclude the possibility of poor run quality in a particular injection, each of the five MudPIT datasets were more thoroughly compared (Figure 5c, d, and e). Figure 5c shows the total number of proteins detected in each of the five individual MudPIT runs, and although not identical in number, the total number of proteins detected was within a close range (average number 483 ± 36 proteins). Similar results were observed for the total number of spectra (average number 3966 ± 539 total spectra) and the average number of spectra recorded for the proteins (8.2 ± 1.5 average spectra) (Figure 5d and e). These results indicate that individual MudPIT runs are highly similar in overall quality, consistent with comparable robust performance, and hence, experimental failure is likely not the major cause of biased sample detection. In summary, it appears that extreme sample complexity, combined with a wide range of absolute protein abundance, is the underlying factor resulting in under-sampling in MudPIT-based proteomics experiments.

Yates and colleagues have reported a comprehensive evaluation of the under-sampling problem associated with MudPIT profiling, using soluble yeast extract as a model system [24]. Based on their experimental findings (similar to those reported above), they were able to develop a statistical model for accurately predicting the number of analyses needed to obtain nearly complete sample saturation. Importantly, this same study concluded that under-sampling has a useful dividend in that it can be used to infer protein relative abundance

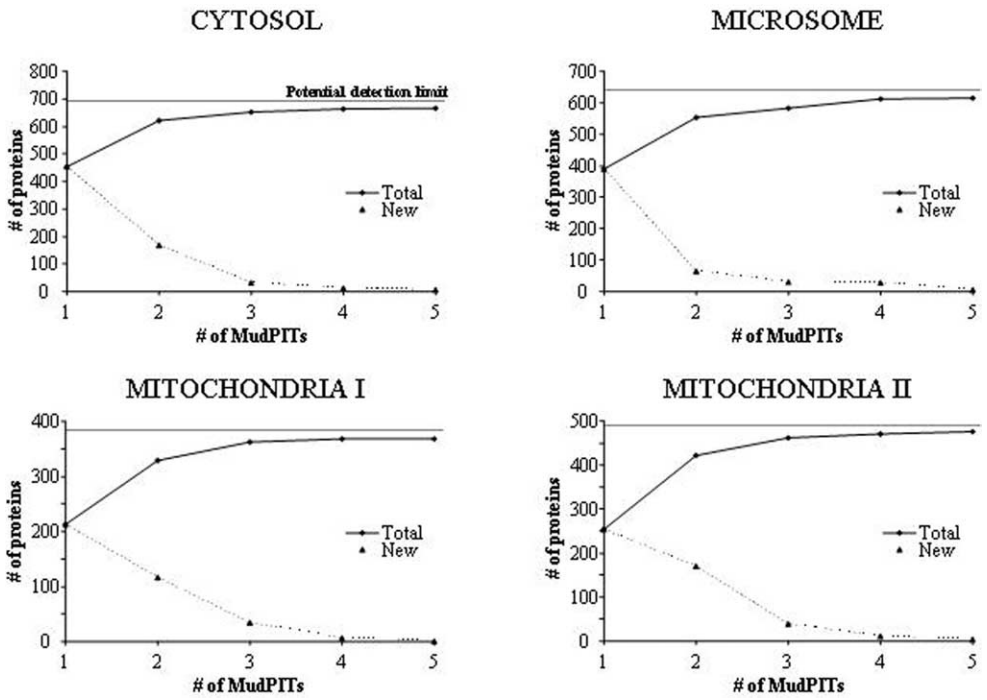

Figure 4. Saturation of protein identification coverage. Total number of putative high-confidence proteins detected in four subcellular fractions isolated from healthy adult heart tissue (after two-step filter system). Right panel, total number of proteins detected after a defined number of MudPIT experiments. Left panel, number of novel proteins detected after a defined number of MudPIT analyses.

simply and reliably, based on the ratio of the number of spectra recorded for each protein across different samples. This spectral count metric is far more straightforward to implement, and the results more readily interpreted, than other more sophisticated quantitative methods° [31,° 38,° 39].° Moreover,° by° looking° at° the relative distribution of spectral count recorded across different organellar fractions, one can readily deduce the "real" subcellular location of a given protein. For example, the highest spectral counts recorded for the α and β chains of the mitochondrial membrane protein ATP synthase were detected in the detergent solubi-

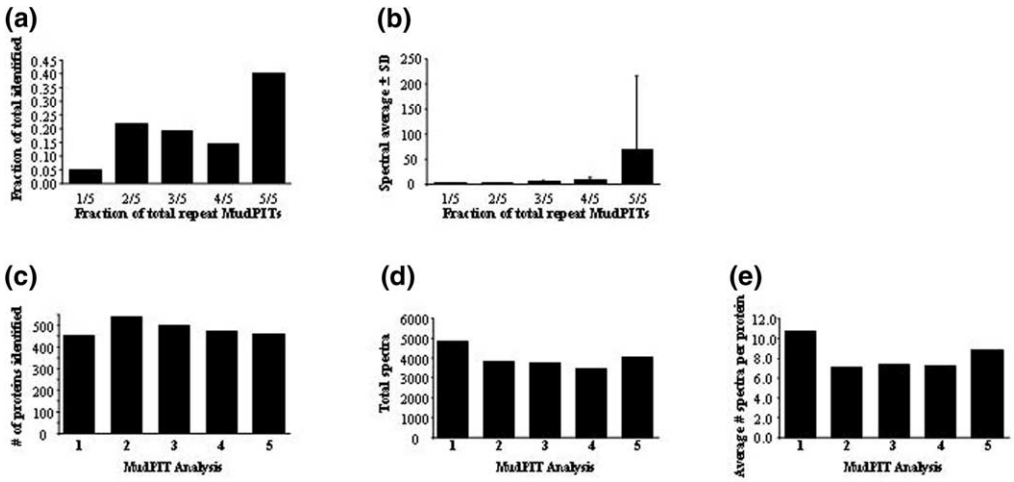

Figure 5. Assessment of MudPIT reproducibility. Statistical analysis of candidate protein identifications detected in heart cytosol using a two-step filtration process. (**a**) The fraction of proteins detected between 1 and 5 times in five repeat MudPIT analyses. (**b**) The average number of spectral counts (± standard deviation) versus the detection rate in the repeat analyses. (**c**) The total number of heart cytosol proteins detected for each of five individual MudPIT experiments. (**d**) Total spectral counts obtained for heart cytosol proteins detected for each of five individual MudPIT experiments. (**e**) Average spectral counts obtained for heart cytosol proteins detected in each of five individual MudPIT experiments.

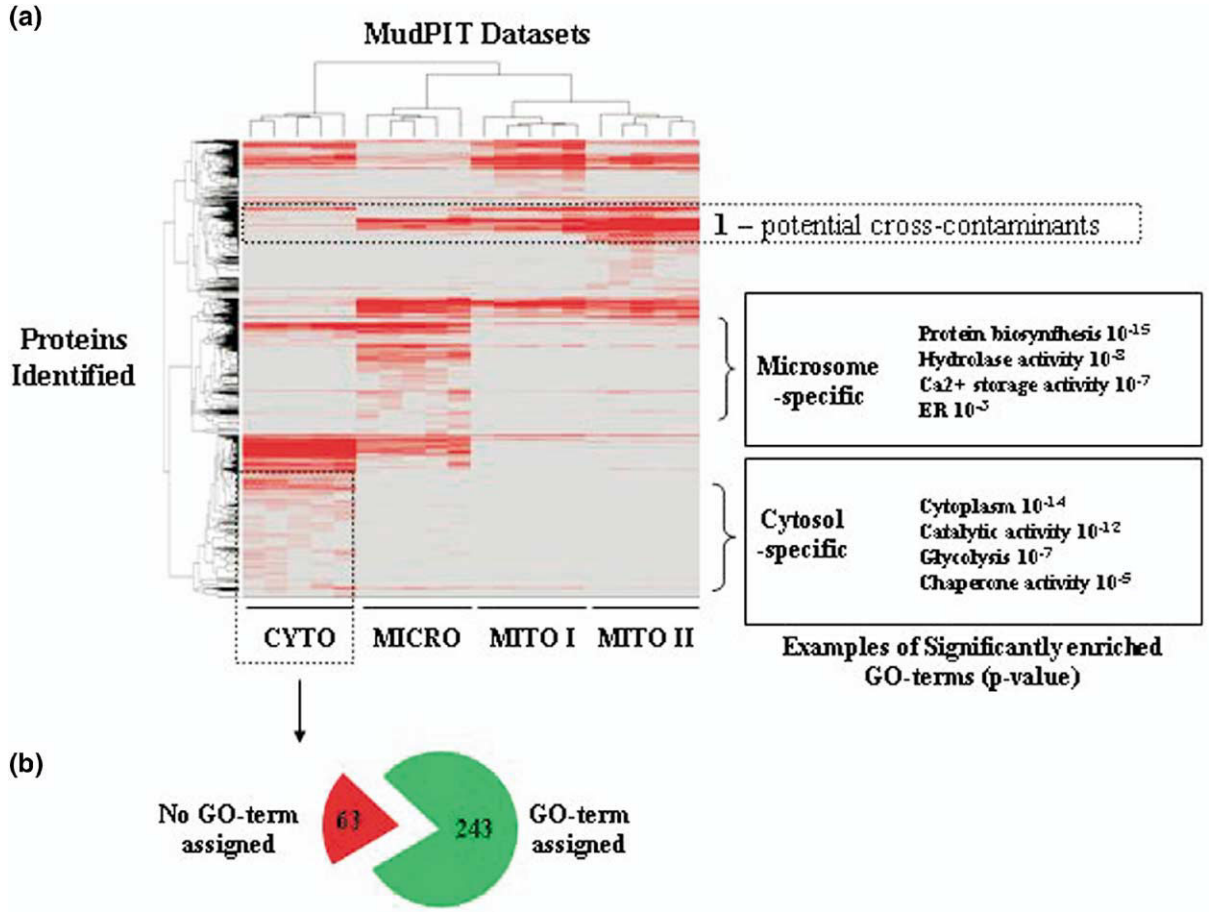

Figure 6. Data mining strategies. (**a**) The entire collection of MudPIT-generated datasets was clustered using the Spearman rank correlation and average linkage and a heat map displayed. Recorded protein spectral counts were used as a quantitative estimate (as indicated by the plotted intensity). Examples of significantly enriched GO-terms mapping to defined clusters are provided. (**b**) A cluster of cytosol unique proteins was mapped against all available GO-terms (circa November 2004). The proportion of proteins having one or more GO-terms and those without any available annotation are displayed.

lized mitochondrial extract (MITO II), whereas the majority of the soluble metabolic enzymes enolase β and fatty acid binding protein were detected in the cytosolic heart fraction [data not shown].

Data Mining

Expression profiling projects can generate enormous datasets quite rapidly, overwhelming the ability of a researcher to cope. For this reason, automated large-scale data mining tools are increasingly required to help organize and manage the data in order to find interesting patterns for biological follow-up or back-up studies. In Figure 6, we present an overview of some of the data mining procedures and software applications that we have found to be particularly valuable for global proteomic data mining efforts. In particular, we find data clustering as an effective starting point for finding hidden patterns among a large collection of proteomic profiles. Figure 6a shows a standard "heat map"-style visual display of a clustered collection of MudPIT datasets generated for healthy adult heart protein fractions. Several features can be readily noticed from this global perspective. First, the five individual repeat analyses recorded for each organellar fraction cluster tightly together, indicating that despite experimental variation, the multiple proteomic patterns recorded are highly similar. Second, good separation can be observed between the distinct subcellular fractions, indicating that each fraction has a distinct protein composition. Indeed, clustering provides a quick method for finding evidence of regulated patterns of expression or coexpression [25], and can be helpful for inferring the existence of a biological module. Lastly, evidence of significant cross-contamination by high abundance proteins is more readily observed (e.g., Box 1—potential cross-contaminations).

Clustering constitutes the logical first step in a multipronged informatics analysis tailored to address a specific biological-oriented research problem. Individual clusters of potential interest (e.g., tight groups of proteins displaying coherent features) can be exported

from the heat map display and the sets of proteins analyzed individually. As a first pass, one can examine the cluster membership looking for evidence of functional enrichment, referring to a publicly accessible annotation resource. The Gene Ontology (GO) database is particularly helpful in this regard as it reports the known or predicted molecular functions, subcellular locations, and biological roles of curated proteins in a computer-interpretable and biologist-friendly format. Manual inspection of cluster membership can also help highlight potential cross-contamination, such as the presence of groups of mitochondrial proteins, released presumably because of organellar damage and leakage during tissue homogenization.

To automate post-cluster analysis, several groups, including our own, have developed programs to calculate the statistical enrichment of a cluster of proteins to select functional categories (GO terms) using a hypergeometric distribution function [26]. For our heart project, we developed a stand-alone software application, MouseSpec, to map annotation terms to an input list of proteins and then calculate the probability (*p*-value) that the intersection of a given input list (cluster) of proteins with any given annotation term occurs by chance alone. To correct for so-called multi-hypothesis testing, the *p*-value threshold deemed significant for an individual test is determined by dividing the number of tests conducted, thereby accounting for spurious significance due to repeated testing over all of the categories in the GO database. We typically use a minimum cutoff value (e.g., 10^{-3}; i.e., association unlikely to happen by chance alone) as a final selection criterion to highlight promising, biologically interesting clusters.

As an illustrative example, a cluster of cytosol-unique heart proteins was exported and evaluated using MouseSpec. Instances of significantly enriched GO-terms are listed to the right of the cluster (hatched box) shown in Figure 6a. As expected, highly significant categories within this cluster included the GO-terms cytoplasm, catalytic activity, chaperone activity, glycolysis, and heat shock protein activity. Surprisingly, a significant fraction of the proteins detected in this same fraction (63 proteins; ~25%) could not be assigned any GO-term (Figure 6b). However, based on the observed subcellular localization, cluster neighborhood, and overall expression patterns, as well as on other sources of available information (such as domain structure or interaction partners), a potential function for many of these proteins can be predicted with reasonable confidence.

Future Perspectives

In the near future, we hope to incorporate the aforementioned MudPIT-based profiling methodology with microarray-based gene expression studies, the results obtained from phenotypic analysis, and select follow-up functional analyses in order to create an integrated systems-wide perspective of the main aspects of heart biology that become perturbed during the course of disease action leading to heart failure. To account for biological complexity, multiple validated mouse models of cardiomyopathy have been chosen for detailed study. We are currently completing an exhaustive evaluation of the proteomic patterns of heart tissue in transgenic mice carrying specific point mutations in a key regulatory protein, phospholamban, in comparison to those recorded with age-matched wild-type animals, tracking different stages as these animals progress to severe dilated cardiomyopathy, hypertrophy, and heart failure [7]. One key objective of the data analysis is to identify interesting candidates that appear to be mechanistically linked to disease progression arising from this study for follow up analysis using traditional molecular genetic methods.

Conclusions

In this report, we have endeavored to provide the reader with a basic framework for analyzing complex tissue expression patterns using systematic large-scale MudPIT-based protein expression profiling in a manner best suited to gaining biologically interpretable datasets. We have attempted to confront the reader with some of the less-appreciated problems associated with successful implementation of the MudPIT technology, while aiming to provide helpful guidelines and useful solutions to the more common problems encountered with these types of studies. Specifically, the issue of data validation and filtration to minimize the rate of false positive identifications, and the challenge of random sampling leading to incomplete detection, were discussed, together with rules for evaluating the optimal number of MudPIT experiments needed to achieve full coverage. We have highlighted some basic bioinformatics approaches that can help to facilitate data organization, mining, and interpretation. Although we have emphasized sample work-up problems that we have encountered that are specific to heart, particularly in isolating discrete organellar fractions from this fibrous tissue, we strongly encourage investigators to fine-tune these basic protocols in a project-specific manner. While the benefits of subcellular fractionation is a biologically meaningful approach for reducing sample complexity, other more extensive fractionation protocols, in particular those involving non-denaturing conventional chromatography, may prove to be even more helpful in revealing of biological modules. In this same vein, the recent introduction of a new generation of linear ion-trap mass spectrometers, with significantly faster scan speeds and more rapid duty cycles, will likely markedly improve overall detection limits, helping surmount the under-sampling problem.

Acknowledgments

Research in the authors' laboratories was supported by Heart and Stroke Foundation of Canada grant T5042, Canadian Institutes of Health Research grant MOP 49493, and the Neuromuscular Research Partnership to DHM, grant funds from the Natural Sciences and Engineering Research Council of Canada (NSERC), the Protein Engineering Network Center of Excellence (PENCE), and Genome Canada to AE, and by a developmental grant from the Muscular Dystrophy Association (USA) to AOG. AOG is a Research Fellow of the Heart and Stroke Foundation of Canada.

Supplementary Material

Supplementary data associated with this article can be found, in the online version, at 10.1016/j.jasms.2005.02.015

References

1. Lee, D. S.; Mamdani, M. M.; Austin, P. C.; Gong, Y.; Liu, P. P.; Rouleau, J. L.; Tu, J. V. Trends in heart failure outcomes and pharmacotherapy: 1992 to 2000. *Am. J. Med.* **2004,** *116,* 581–589.
2. MacLennan, D. H.; Kranias, E. G. Phospholamban: A crucial regulator of cardiac contractility. *Nat. Rev. Mol. Cell. Biol.* **2003,** *4,* 566–577.
3. Bers, D. M. *Excitation-contraction coupling and cardiac contractile force*; Kluwer Academic Publishers: Boston, MA, 2002.
4. Simmerman, H. K.; Jones, L. R. Phospholamban: Protein structure, mechanism of action, and role in cardiac function. *Physiol. Rev.* **1998,** *78,* 921–947.
5. Yamazaki, T.; Komuro, I.; Yazaki, Y. Signaling pathways for cardiac hypertrophy. *Cell. Signaling* **1998,** *10,* 693–698.
6. Pan, Y.; Kislinger, T.; Gramolini, A. O.; Zvaritch, E.; Kranias, E. G.; MacLennan, D. H.; Emili, A. Identification of biochemical adaptations in hyper- or hypocontractile hearts from phospholamban mutant mice by expression proteomics. *Proc. Natl. Acad. Sci. U.S.A.* **2004,** *101,* 2241–2246.
7. Schmitt, J. P.; Kamisago, M.; Asahi, M.; Li, G. H.; Ahmad, F.; Mende, U.; Kranias, E. G.; MacLennan, D. H.; Seidman, J. G.; Seidman, C. E. Dilated cardiomyopathy and heart failure caused by a mutation in phospholamban. *Science* **2003,** *299,* 1410–1413.
8. Dorn, G. W., II; Molkentin, J. D. Manipulating cardiac contractility in heart failure: Data from mice and men. *Circulation* **2004,** *109,* 150–158.
9. Hoshijima, M. Models of dilated cardiomyopathy in small animals and novel positive inotropic therapies. *Ann. N.Y. Acad. Sci.* **2004,** *1015,* 320–331.
10. Vichinsky, E. New therapies in sickle cell disease. *Lancet* **2002,** *360,* 629–631.
11. Zhang, W.; Morris, Q. D.; Chang, R.; Shai, O.; Bakowski, M. A.; Mitsakakis, N.; Mohammad, N.; Robinson, M. D.; Zirngibl, R.; Somogyi, E.; Laurin, N.; Eftekharpour, E.; Sat, E.; Grigull, J.; Pan, Q.; Peng, W. T.; Krogan, N.; Greenblatt, J.; Fehlings, M.; van der Kooy, D.; Aubin, J.; Bruneau, B. G.; Rossant, J.; Blencowe, B. J.; Frey, B. J.; Hughes, T. R. The functional landscape of mouse gene expression. *J. Biol.* **2004,** *3,* 1–22.
12. Su, A. I.; Wiltshire, T.; Batalov, S.; Lapp, H.; Ching, K. A.; Block, D.; Zhang, J.; Soden, R.; Hayakawa, M.; Kreiman, G.; Cooke, M. P.; Walker, J. R.; Hogenesch, J. B. A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl. Acad. Sci. U.S.A.* **2004,** *101,* 6062–6067.
13. Dos Remedios, C. G.; Liew, C. C.; Allen, P. D.; Winslow, R. L.; Van Eyk, J. E.; Dunn, M. J. Genomics, proteomics, and bioinformatics of human heart failure. *J. Muscle Res. Cell Motil.* **2003,** *24,* 251–260.
14. Ferlini, A.; Sewry, C.; Melis, M. A., et al. X-linked dilated cardiomyopathy and the dystrophen gene. *Neuromuscul. Disord.* **1999,** *9(5),* 339–346.
15. Liew, C. C.; Dzau, V. J. Molecular genetics and genomics of heart failure. *Nat. Rev. Genet.* **2004,** *5,* 811–825.
16. Kislinger, T.; Rahman, K.; Radulovic, D.; Cox, B.; Rossant, J.; Emili, A. PRISM, a generic large scale proteomic investigation strategy for mammals. *Mol. Cell Proteom.* **2003,** *2,* 96–106.
17. Gorg, A.; Weiss, W.; Dunn, M. J. Current two-dimensional electrophoresis technology for proteomics. *Proteomics* **2004,** *4(12),* 3665–3685.
18. Gygi, S. P.; Corthals, G. L.; Zhang, Y.; Rochon, Y.; Aebersold, R. Evaluation of two-dimensional gel electrophoresis-based proteome analysis technology. *Proc. Natl. Acad. Sci. U.S.A.* **2000,** *97,* 9390–9395.
19. 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.
20. Koller, A.; Washburn, M. P.; Lange, B. M.; Andon, N. L.; Deciu, C.; Haynes, P. A.; Hays, L.; Schieltz, D.; Ulaszek, R.; Wei, J.; Wolters, D.; Yates, J. R., III. Proteomic survey of metabolic pathways in rice. *Proc. Natl. Acad. Sci. U.S.A.* **2002,** *99,* 11969–11974.
21. Le Roch, K. G.; Johnson, J. R.; Florens, L.; Zhou, Y.; Santrosyan, A.; Grainger, M.; Yan, S. F.; Williamson, K. C.; Holder, A. A.; Carucci, D. J.; Yates, J. R., III; Winzeler, E. A. Global analysis of transcript and protein levels across the *Plasmodium falciparum* life cycle. *Genome Res.* **2004,** *14,* 2308–2318.
22. Schirmer, E. C.; Florens, L.; Guan, T.; Yates, J. R., III; Gerace, L. Nuclear membrane proteins with potential disease links found by subtractive proteomics. *Science* **2003,** *301,* 1380–1382.
23. Eng, J. K.; McCormack, A. L.; Yates, J. R., III. An approach to correlate tandem mass-spectral data of peptides with amino-acid-sequences in a protein database. *J. Am. Soc. Mass Spectrom.* **1994,** *11,* 976–989.
24. Liu, H.; Sadygov, R. G.; Yates, J. R., III. A model for random sampling and estimation of relative protein abundance in shotgun proteomics. *Anal. Chem.* **2004,** *76,* 4193–4201.
25. Eisen, M. B.; Spellman, P. T.; Brown, P. O.; Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. U.S.A.* **1998,** *95,* 14863–14868.
26. Robinson, M. D.; Grigull, J.; Mohammad, N.; Hughes, T. R. FunSpec: A web-based cluster interpreter for yeast. *BMC Bioinformatics* **2002,** *3,* 35.
27. Stanley, B. A.; Gundry, R. L.; Cotter, R. J.; Van Eyk, J. E. Heart disease, clinical proteomics, and mass spectrometry. *Dis. Markers* **2004,** *20,* 167–178.
28. American Heart Association: Heart Disease and Stroke Statistics: 2004 Update.
29. Mantripragada, K. K.; Buckley, P. G.; de Stahl, T. D.; Dumanski, J. P. Genomic microarrays in the spotlight. *Trends Genet.* **2004,** *20,* 87–94.
30. Yates, J. R., III. Mass spectral analysis in proteomics. *Annu. Rev. Biophys. Biomol. Struct.* **2004,** *33,* 297–316.
31. Aebersold, R.; Mann, M. Mass spectrometry-based proteomics. *Nature* **2003,** *422,* 198–207.
32. Adkins, J. N.; Varnum, S. M.; Auberry, K. J.; Moore, R. J.; Angell, N. H.; Smith, R. D.; Springer, D. L.; Pounds, J. G. Toward a human blood serum proteome: Analysis by multidimensional separation coupled with mass spectrometry. *Mol. Cell. Proteom.* **2002,** *1,* 947–955.
33. Skop, A. R.; Liu, H.; Yates, J., III; Meyer, B. J.; Heald, R. Dissection of the mammalian midbody proteome reveals conserved cytokinesis mechanisms. *Science* **2004,** *305,* 61–66.

34. Tabb, D. L.; McDonald, W. H.; Yates, J. R. III. DTASelect and Contrast: Tools for assembling and comparing protein identifications from shotgun proteomics. *J. Proteome Res.* **2002,** *1,* 21–26.
35. Pedrioli, P. G.; Eng, J. K.; Hubley, R.; Vogelzang, M.; Deutsch, E. W.; Raught, B.; Pratt, B.; Nilsson, E.; Angeletti, R. H.; Apweiler, R.; Cheung, K.; Costello, C. E.; Hermjakob, H.; Huang, S.; Julian, R. K.; Kapp, E.; McComb, M. E.; Oliver, S. G.; Omenn, G.; Paton, N. W.; Simpson, R.; Smith, R.; Taylor, C. F.; Zhu, W.; Aebersold, R. A common open representation of mass spectrometry data and its application to proteomics research. *Nat. Biotechnol.* **2004,** *22,* 1459–1466.
36. Labugger, R.; McDonough, J. L.; Neverova, I.; Van Eyk, J. E. Solubilization, two-dimensional separation, and detection of the cardiac myofilament protein troponin T. *Proteomics* **2002,** *2,* 673–678.
37. 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,* 985–992.
38. Cagney, G.; Emili, A. De novo peptide sequencing and quantitative profiling of complex protein mixtures using mass-coded abundance tagging. *Nat. Biotechnol.* **2002,** *20,* 163–170.
39. Gygi, S. P.; Rist, B.; Gerber, S. A.; Turecek, F.; Gelb, M. H.; Aebersold, R. Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat. Biotechnol.* **1999,** *17,* 994–999.