

# Integrating Global Proteomic and Genomic Expression Profiles Generated from Islet $\alpha$ Cells

OPPORTUNITIES AND CHALLENGES TO DERIVING RELIABLE BIOLOGICAL INFERENCES\*<sup>§</sup>

Marlena Maziarz<sup>‡§¶</sup>, Clement Chung<sup>¶¶</sup>, Daniel J. Drucker<sup>‡§</sup>, and Andrew Emili<sup>¶¶††</sup>

Systematic profiling of expressed gene products represents a promising research strategy for elucidating the molecular phenotypes of islet cells. To this end, we have combined complementary genomic and proteomic methods to better assess the molecular composition of murine pancreatic islet glucagon-producing  $\alpha$ TC-1 cells as a model system, with the expectation of bypassing limitations inherent to either technology alone. Gene expression was measured with an Affymetrix MG\_U74Av2 oligonucleotide array, while protein expression was examined by performing high-resolution gel-free shotgun MS/MS on a nuclear-enriched cell extract. Both analyses were carried out in triplicate to control for experimental variability. Using a stringent detection  $p$  value cutoff of 0.04, 48% of all potential mRNA transcripts were predicted to be expressed (probes classified as present in at least two of three replicates), while 1,651 proteins were identified with high-confidence using rigorous database searching. Although 762 of 888 cross-referenced cognate mRNA-protein pairs were jointly detected by both platforms, a sizeable number (126) of gene products was detected exclusively by MS alone. Conversely, marginal protein identifications often had convincing microarray support. Based on these findings, we present an operational framework for both interpreting and integrating dual genomic and proteomic datasets so as to obtain a more reliable perspective into islet  $\alpha$  cell function. *Molecular & Cellular Proteomics* 4: 458–474, 2005.

The proglucagon gene is expressed in selected brainstem neurons, gut endocrine L cells, and  $\alpha$  (alpha) cells within the endocrine pancreas. Because glucagon excess contributes to the metabolic derangements of diabetes, while acquired defects in glucagon secretion are associated with an increased risk of hypoglycemia in diabetic subjects, understanding the

molecular features of the normal and dysregulated  $\alpha$  cell is of considerable relevance for strategies design to optimize  $\alpha$  cell function in diabetic patients.

Although glucagon-producing  $\alpha$  cells are the second most abundant islet cell type within the pancreatic islets of Langerhans, a majority of studies aimed at identifying islet gene products have used either a mixture of islet cells, or purified  $\beta$  cells or  $\beta$  cell lines (1–5). The current paucity of data on gene and protein expression in islet  $\alpha$  cells is due in part to the almost insurmountable technical challenges in isolating sufficient numbers of pure primary  $\alpha$  cells for exhaustive molecular studies. Our group (6–8) and others (9, 10) have therefore studied several transformed rodent  $\alpha$  cell-derived cell lines as model systems suitable for systematic experimental analyses aimed at determining the genetic, biochemical, and physiological mechanisms that underlie islet  $\alpha$  cell function.

While several laboratories have reported investigations of whole islet or islet  $\beta$  cell gene expression profiles (1, 2), only limited information is currently available concerning gene expression and proteomic profiles linked to the specification of the differentiated phenotypes associated with  $\alpha$  cells. To this end, we have initiated extensive molecular analyses of gene expression patterns in mature  $\alpha$  cells using an islet  $\alpha$  cell line as a model system, with the goal of deciphering key biochemical, regulatory, and cell biological properties. The  $\alpha$ TC-1 cell line, in particular, was chosen as a focus of this study because it is well differentiated, preserving normal endocrine characteristics such as correct hormone processing and secretion of glucagon in response to stimulation (11).

Several types of high-density DNA microarrays apt for global analysis of gene expression profiles are readily available to researchers today. These vary in terms of the technology used in their manufacture and use, as well as in terms of price, ease of use and analysis, and overall sensitivity and reproducibility. The main types of high-throughput gene expression arrays are spotted cDNA arrays (12–14), short oligonucleotide arrays (15), long oligonucleotide arrays (Agilent: [www.agilent.com](http://www.agilent.com)) (16), and fiber optic arrays (Illumina: [www.illumina.com](http://www.illumina.com)) (17). An informative web resource outlining the various platforms, as well as sources of commercial arrays, is available online at [ihome.cuhk.edu.hk/~b400559/array.html](http://ihome.cuhk.edu.hk/~b400559/array.html).

A particularly popular commercial research platform is the

From the <sup>‡</sup>Banting and Best Diabetes Centre, University of Toronto, Toronto, Ontario, Canada; <sup>§</sup>Department of Medicine, Toronto General Hospital, Toronto, Ontario, Canada; and <sup>¶</sup>Program in Proteomics and Bioinformatics, Banting and Best Department of Medical Research, University of Toronto, Toronto, Ontario M5G 1L6, Canada

Received, February 25, 2005, and in revised form, March 1, 2005  
Published, MCP Papers in Press, March 1, 2005, DOI 10.1074/mcp.R500011-MCP200

Affymetrix GeneChip™ short oligonucleotide arrays platform, which consist of panels of ~25-nucleotide-long probes synthesized *in situ* on the surface of a microscope slide. Each probe is built up consecutively on a predefined area (square element) on the array (typically called a cell or feature). The probe sequence is complementary to a unique target sequence present in a putative mRNA transcript, allowing specific hybridization (15). Affymetrix chip densities have been growing quickly in recent years, with newer arrays commonly capable of probing tens of thousands of gene transcripts—often representing the entire genome complement of an organism—simultaneously (18). Between 1998 and 2004, for instance, the average number of probes packed onto a single Affymetrix array shot up from ~2,000 to >50,000, due to significant reductions in probe feature dimensions, which have now decreased from >50 to ~11  $\mu\text{m}$  on average (18).

To interrogate the expression level of a single gene, a typical “probe set”—generally consisting of 11–20 distinct 25-mer oligonucleotide probes—is designed to hybridize to discrete sequences in a particular target gene transcript. These so-called “perfect match probes” (PM)<sup>1</sup> are aimed at detecting a specific mRNA presumed to be present in a biological sample and are usually optimized to prevent fortuitous similarity to off-target sequences resulting in confounding cross-hybridization to other gene products (19). In addition to the PM probes, Affymetrix chips also contain adjacent “mismatch probes” (MM), which are virtually identical to PM probes except for a single-nucleotide substitution at the 13th nucleotide of the sequence so as to perturb hybridization to the intended target. The PM and MM probe cells from the same probe set are usually evenly distributed over the entire chip area to minimize the effects of systematic variability—such as experimental artifacts due to localized hybridization defects—which often lead to inconsistencies in the inferences that can be drawn from a study.

The MM features were initially designed to measure non-specific binding so as to control for spurious background signal, but their use for estimating and removing noise has been criticized in recent years (20, 21). Indeed, it has been shown that more accurate estimates of relative transcript abundance between different biological samples can be obtained from interrogation of respective mRNA profiles using data derived from PM probes alone (22). Nevertheless, the MM probe data are used by the native Affymetrix software to determine if a target transcript is in fact present in a given sample. It has been shown that for target at low concentrations (<8 pM) the combination of PM-MM probe pairs can offer significantly higher sensitivity as compared with PM probes alone (19).

Because of steady improvements in technology, chip design, industrial standardization, and analysis methods, DNA microarrays have become a widespread experimental medium in the endocrine research community. This popularity, combined with the increasing density of microarray readouts, has resulted in a corresponding exponential increase in data production (23), a trend that is putting increasing strain on effective data analysis and biological interpretation (24). This is compounded by the many sources of irrelevant obscuring variance that also need to be carefully considered when dealing with microarray data, including irrelevant fluctuations due to scanner perturbations, changes in hybridization conditions such as variable humidity, different array batches, or even different personnel performing the experiments. Because these artifacts can substantially affect the results (and conclusions) obtained from a study (25–27), they need to be systematically identified, corrected for, or removed to ensure the reliability of an experiment. In certain platforms (e.g. cDNA arrays), the use of internal reference mRNA populations and reciprocal chromo/fluorophore label swaps can be used to minimize spurious measurements (24). Alternatively, with the Affymetrix array platform, multiple repeat analyses can be performed to increase confidence in distinguishing genuine signal from spurious deviances and baseline noise (28).

Ongoing challenges in evaluating the integrity of microarray data are spurring the development of novel computational methods for methodical data interpretation (29–31). Thanks to a concerted effort by the entire bioinformatics community, both academic and commercial algorithms for improving the reliability of quantification of chip signal and global comparative analyses have been steadily improving. For instance, seemingly effective methods for data quantification and normalization have been introduced in recent years (20, 24, 32–34). Indeed, there is now a bewildering array of open-source programs and proprietary software tools available for automated data preprocessing and analysis (35) (see [ihome.cuhk.edu.hk/~b400559/array.html](http://ihome.cuhk.edu.hk/~b400559/array.html) for a comprehensive list).

Many of the more commonly used tools are tailored specifically to the Affymetrix data format. Most of these employ some of the now standard statistical and machine learning methods (such as various types of clustering algorithms (e.g. SOM, k-means, hierarchical) and/or dimension reduction methods (e.g. PCA and many others) to extract meaningful information from typically complex, feature-rich gene expression datasets (36), but the list of interesting new approaches to the data analysis challenge is steadily growing.

Despite this progress, preliminary studies of gene expression patterns in endocrine cell lines have revealed a number of unforeseen obstacles that confound effective interpretation of microarray data,<sup>2</sup> several of which are the focus of this report. The most noteworthy of these concerns the problem of deciding on optimal transcript detection *p* values threshold cut-

<sup>1</sup> The abbreviations used are: PM, perfect match; MM, mismatch; MudPIT, multidimensional protein identification technology; RMA, Robust Multichip Average; GO, Gene Ontology; IM, idealized mismatch; ROC, receiver-operating characteristic.

<sup>2</sup> M. Maziarz and D. J. Drucker, unpublished observations.

offs to faithfully determine whether an mRNA target is indeed present or absent in a sample of interest. It is now broadly accepted that all microarray results should be treated as preliminary, and therefore putative mRNA expression patterns of special interest must necessarily be validated using an independent screening method, such as Northern blot analysis, RT-PCR, or real-time RT-PCR. Proper detection calls are an especially important factor for selecting optimal data points for use as features in pattern recognition and machine learning analyses aimed at defining robust molecular signatures (such as biomarkers). Detection  $p$  values can also be useful for assessing the reliability of extracted gene expression values for cross-sample comparative quantification analyses.<sup>3</sup>

Considering that proteins represent the ultimate effectors of gene function, and given that previous proteomic studies have suggested a relatively weak concordance between mRNA and protein levels (37), it may, in fact, be preferable to measure protein expression in islet  $\alpha$  cells directly. Effective methods for large-scale protein identification and quantitation have recently been developed (38, 39). Of these, high-throughput protein shotgun sequencing procedures coupling high-resolution multidimensional liquid chromatographic separation of proteolytic peptide digests to ultra-sensitive MS/MS (LC-MS) appear to be among the more powerful new emerging methods for examining complex protein mixtures like cell extracts (38, 39). By eliminating the need to first separate proteins on polyacrylamide gels, these gel-free profiling methods circumvent most limitations associated with gel-based procedures.

Although rapidly increasing in popularity, current LC-MS-based shotgun profiling methods are by no means nearly as robust, simple to execute, or as widely accessible as microarray technology. Moreover, gel-free proteomic screening methodologies still suffer from significant detection bias, leading to preferential detection of higher-abundance housekeeping proteins that are rarely of particular biological interest (38, 39) and generally do not achieve the same extent of global coverage as compared with that typically obtainable by microarray-based analysis (37). Despite these caveats, shotgun protein profiling methods can potentially be used to both guide the optimal interpretation of microarray data and to confirm the results of a gene expression study, and vice-versa.

To evaluate the advantages and challenges of combining proteomic and functional genomic screening platforms for deducing functional adaptations associated with enteroendocrine cells, we have performed pilot parallel large-scale analyses of gene and protein expression in mitotically active asynchronous cultures of murine  $\alpha$ TC-1 cells (11). We used an Affymetrix MG\_U74Av2 GeneChip, containing 12,488 distinct murine gene probe sets, to measure global mRNA levels, and

the multidimensional protein identification technology (MudPIT) procedure developed by Yates and colleagues (40, 41) to examine the global protein profile of nuclear-enriched cell extracts. Here, we summarize our key findings to date, outlining both the advantages and challenges of comparative cross-platform analyses so as to obtain a more complete, and rigorous, insight into the biochemical makeup of enteroendocrine cells. We compare and contrast the relative merits and limitations associated with each of the two platforms and list factors that influence the sensitivity and specificity of analysis with a particular emphasis on the effects of  $p$  value thresholds on the rate of false discovery and overall detection coverage. We also describe empirically derived criteria regarding optimal experimental design, together with simple guidelines that we believe allow for more reliable (and comprehensive) interpretation of complex global molecular profiling datasets. While centered on an experimental setting directly relevant to  $\alpha$  islet cell biology, the analytical approach reported here are broadly applicable to a range of analogous endocrine-related research problems and hence should be relevant to any researcher currently using (or contemplating) large-scale expression profiling at the mRNA and/or protein levels as a means of investigating the molecular makeup of endocrine cells and tissues of interest.

### EXPERIMENTAL PROCEDURES

**Cell Culture**—The immortalized rodent islet  $\alpha$ -TC-1 cell line (11) was passaged in Dulbecco's modified essential medium containing 25 mM glucose (DMEM, high glucose; Hyclone, Logan, UT) supplemented with 15% heat-inactivated FCS (Invitrogen, San Diego, CA) and Pen/Strep (Sigma-Aldrich, St. Louis, MO) as described previously (6, 11). Cell cultures were grown in triplicate to 80% confluence, with the tissue culture media replaced daily. After harvesting, the cells were snap-frozen and stored at  $-80^{\circ}\text{C}$  prior to protein and RNA extraction.

**Microarray/Statistical Analysis**—Total RNA was prepared using Trizol extraction (Invitrogen) and RNeasy kit (Qiagen, Valencia, CA) according to the manufacturers' instructions. Profiling was performed in triplicate for three distinct preparations of total RNA, using the Affymetrix GeneChip MG\_U74Av2 microarray. Each experiment was performed separately on different days using chips derived from different fabrication batches. Hybridization and washing were performed using standard conditions. Image processing was performed using an Affymetrix GeneArray 2500 scanner, with 570-nm, 3- $\mu\text{m}$  laser parameter settings. The data file was processed using Affymetrix Microarray Suite 5.0 (MAS5.0) using default parameter settings, and the average intensity of all probes on the array was scaled to target intensity of 150. Detection  $p$  value alone, or together with signal intensity, was used to make a prediction (present, marginal, or absent) as to gene expression. A freeware version of the Robust Multichip Average (RMA) algorithm was obtained from Bioconductor ([www.bioconductor.org](http://www.bioconductor.org)). Freely available statistical software "R" ([www.r-project.org](http://www.r-project.org)) and proprietary Matlab 7.0, Statistical and Bioinformatics Toolboxes from MathWorks™ ([www.mathworks.com](http://www.mathworks.com)) were also used for the analysis.

**Validation RT-PCR**—First-strand cDNA synthesis was generated from total RNA using the SuperScript Pre-amplification System (Fermentas, Hanover, MD) following the manufacturer's recommended protocol. Target cDNAs were amplified by PCR using gene-specific

<sup>3</sup> P. Hu, personal communication.

primer pairs. PCR products were loaded onto a 1% agarose gel, electrophoresed in Tris-acetate-EDTA buffer, transferred onto nylon membranes, and hybridized using internal oligonucleotide probes labeled by T4 kinase reaction with [ $\gamma$ - $^{32}$ P]ATP (Amersham Bioscience, Piscataway, NJ). Visualization was performed by standard film-based autoradiography. Primer sequences and detailed conditions used for RT-PCR are available upon request.

**Extracts**—Nuclear-enriched soluble protein extracts were prepared using a commercial protocol (Nu-CLEAR protein extraction kit; Sigma-Aldrich). Briefly, three separate cell pellets were thawed, resuspended in 5 volumes of hypotonic lysis buffer (10 mM HEPES, pH 7.9, 1.5 mM MgCl<sub>2</sub>, 10 mM KCl), and incubated on ice for 15 min. After repelling by centrifugation for 5 min at 420 × *g*, the nuclei were rinsed twice in 400  $\mu$ l of lysis buffer, and resuspended in high-salt extraction buffer (1 M HEPES, pH 7.9, 1 M MgCl<sub>2</sub>, 5 M NaCl, 0.5 M EDTA, pH 8.0, 25% (v/v) glycerol). Soluble proteins were salt-extracted from the isolated nuclei by incubation on ice for 15 min with occasional vigorous vortexing. After addition of Nonidet P-40 to a final concentration of 0.04% (v/v), the nuclei were disrupted with a glass homogenizer. Insoluble debris was removed by centrifugation at 13,000 rpm for 30 min, and the supernatants were retained for proteomic analysis (final protein concentration ~2 mg/ml).

**Proteolytic Digestion and Sample Preparation**—Equivalent amounts of total protein (100  $\mu$ g) from each processed cell batch were precipitated with 5 volumes of ice-cold acetone, followed by centrifugation at 13,000 × *g* for 15 min. The pellets were resolubilized in 40  $\mu$ l of 8 M urea, 50 mM Tris-HCl, pH 8.5, 1 mM DTT, for 1 h at 37 °C, followed by dilution to 4 M [urea] using an equal volume of 100 mM ammonium bicarbonate, pH 8.5 (AmBic Sigma-Aldrich). The denatured proteins were digested with a 1:150-fold ratio of endoproteinase Lys-C (Roche, Basel, Switzerland) at 37 °C overnight. The next day, the samples were further diluted to 2 M urea with an equal volume of AmBic supplemented with 1 mM CaCl<sub>2</sub>. Digestion was continued by adding poroszyme trypsin beads (Applied Biosystems) followed by incubation at 30 °C with rotation. The resulting peptide mixtures were solid phase-extracted with a SPEC-Plus PT C18 cartridge (Ansys Diagnostics, Lake Forest, CA). The eluates were concentrated by Speedvac to near dryness and stored at –80 °C until analysis.

**MudPIT Analysis**—A fully automated 12-cycle, 20-h long MudPIT procedure was set up as described previously (42). Briefly, an HPLC quaternary pump was interfaced with an LCQ DECA XP quadrupole ion trap mass spectrometer (Thermo Finnigan, Woburn, MA). A P-2000 laser puller (Sutter Instruments, Novato, CA) was used to pull a fine tip on one end of a 100- $\mu$ m inner diameter fused silica capillary microcolumn (Polymicro Technologies, Phoenix, AZ). The column was packed first with 6 cm of 5- $\mu$ m Zorbax Eclipse XDB-C<sub>18</sub> resin (Agilent Technologies, Palo Alto, CA), followed by 6 cm of 5- $\mu$ m Partisphere strong cation exchange resin (Whatman, Middlesex, United Kingdom). The extracts were loaded separately on to the column using a pressure vessel and analyzed sequentially using a fully automated 12-step, four-solvent, 24-h chromatographic cycle. A detailed description of the exact chromatographic conditions is provided in the supplementary information. Data-dependent MS/MS acquisition was performed in real time, with the ion trap instrument operated using dynamic-exclusion lists.

**Protein Identification and Statistical Validation**—The SEQUEST database search algorithm (43) was used to match the acquired tandem mass spectra to a minimally redundant database of mouse and human Swiss-Prot/TrEMBL protein sequences downloaded from the European Bioinformatics Institute (www.ebi.ac.uk). To further evaluate, and thereby minimize, the number of incorrect (false-positive) identifications, the spectra were searched against protein sequences in both the normal (forward) and fully inverted (reverse) amino acid

orientations. The STATQUEST algorithm was used to compute a confidence score (indicating the percentage likelihood or probability of being correct) for every candidate match. The Contrast and DTA-Select software tools were used to manage, organize, and filter the resulting dataset (44) (fields.scripps.edu). Spectral counts were summed across the replicates and used as a semiquantitative measure of protein abundance, as described by Liu *et al.* (45).

**Dataset Cross-referencing**—Matches between sequences identified by the proteomic and the complete Affymetrix probe sets were cross-mapped via annotation. Accession numbers corresponding to the MG\_U74Av2 GeneChip were extracted from the annotated gene table “MG\_U74Av2\_annot.csv” (Oct. 12, 2004) from the Affymetrix website: www.affymetrix.com/support/technical/. The probe IDs were cross-referenced to UniGene IDs, which were then mapped to the corresponding Swiss-Prot/TrEMBL accessions obtained for the identified proteins. Proteins matching multiple genes on the Affymetrix array were removed, as were matches where multiple MutPIT-derived proteins matched to one or more probe sets.

**Functional Classification**—Gene Ontology (GO) categorical annotations (www.ebi.ac.uk/GOA) were used as indicators of biological function and other properties, where annotations were available (~70% of the proteins examined in this study). Statistical testing for enrichment of functional categories among the set of identified proteins and the mRNA probe sets was based on a hypergeometric distribution model using the method of Hughes and co-workers (46). This method returns the probability (*p* value) that the intersection of a given protein list with any given annotation class occurs by chance. A Bonferroni scaling factor was applied to correct for spurious significance due to repeat testing (multihypotheses) over the many GO terms; scores were adjusted by dividing the preliminary *p* value by the number of tests conducted. A threshold cutoff *p* value of 10<sup>–3</sup> was used as a final selection criterion to highlight statistically significant, potentially biologically interesting clusters.

## RESULTS

**Microarray Analysis**—Mouse-derived islet  $\alpha$ TC-1 cells were cultured under standard conditions and analyzed for both global gene and protein expression (see “Experimental Procedures”). Because it has been persuasively argued that biological noise is far more of a concern than technical artifacts (24, 36), we performed both the proteomic and the microarray screening in triplicate, using three independently processed cell batches grown under identical tissue culture conditions. In practice, experimental reproducibility, as assessed by relative quantification (see further below), proved to very good at the mRNA level, and respectable at the protein level, indicating modest and acceptable levels of biological and experimental variability.

For the gene expression analysis, the total RNA sample was extracted and separately hybridized to Affymetrix Murine Genome U74Av2 GeneChip microarrays (MG\_U74Av2). The chip represents ~6,000 oligonucleotide sequences obtained from the Mouse UniGene database (build 74) that have been functionally characterized. The remaining ~6,000 probes represent ESTs. The chip has 16 probe pairs per probe set, with a 20- $\mu$ m feature size and a putative sensitivity rated at 1:100,000. Extensive gene annotation files are available (www.affymetrix.com) (47), providing detailed information for each probe, such as sequence source, transcript ID, target

description, UniGene ID, alignments, gene title and symbol, chromosomal location, identifiers for Ensembl, LocusLink, MGI, and Prot/TrEMBL databanks, as well as GO functional annotations, just to name a few.

**Data Processing**—The result of the microarray profiling experiments was a series of image files, corresponding to the intensities recorded for each feature, which needed to be processed and analyzed. Probe set quantification and data normalization was performed based on Absolute Analysis, using the Statistical Algorithm in the Affymetrix-supplied analytical software (Microarray Suite 5.0, MAS5.0). Signal was scaled to a target intensity of 150 across all probes (Fig. 1). At this step, the information extracted represented the absolute signal intensity for each probe set, together with a calculated detection  $p$  value, and an overall detection call (“present,” “marginal,” or “absent”), which is in turn based on a pre-defined default  $p$  value threshold. For this study, the default Affymetrix default  $p$  value threshold of 0.04 was used (that is, a gene was considered present if its calculated expression detection  $p$  value was equal to or below 0.04 in two of three experiments). The suitability of this detection  $p$  value threshold is examined below.

Additionally, to better account for systematic artifacts, the data files were next quantified using the RMA algorithm (20). This open-source algorithm corrects for spurious variations in background by compensating for nonspecific binding as determined based on the distribution of PM values, as opposed to PM-MM ratios as in the case of the Affymetrix-supplied algorithm. The algorithm performs probe-level quantile normalization (48–50) across all chips to unify their PM distributions, and lastly it summarizes the probe-set signals by median polishing. The output of RMA is log-transformed signal intensity for each probe (21). Supplemental Fig. 1 provides a comparison of the distributions of probe signal intensities obtained in two repeat experiments (chip A versus B), both before and after data normalization using either the Affymetrix Statistical Algorithm (MAS5.0) or RMA. The effects of normalization were also apparent by comparing a scatter plot of the average probe intensity versus the log-transformed ratio of signals on respective chips (so-called M versus A plot) before (Supplemental Fig. 1, left insets) and after RMA normalization (right insets). If the data was reproduced perfectly, the expected ratio ( $r$ ) of signals produced by scanning the two chips would be exactly 1 for all probes, resulting in a tight distribution around  $y = r = 1$ . Any deviation from this indicates spurious differences in signal readings between experiments. Clearly, there is a lot of variability seen in the unprocessed data, particularly in the low-intensity end. The quantile normalization procedure used by the RMA algorithm corrected this type of variability very well, with the RMA-processed data being much more closely reproduced across the entire intensity range.

**Data Analysis and Detection of Gene Expression**—Despite the effectiveness of normalization, a key, and frequently un-

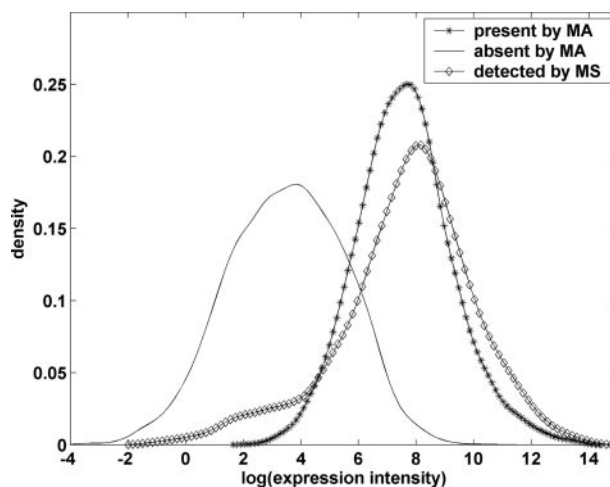


FIG. 1. **Distribution of microarray probe signal intensities.** The plot shows a comparison of distributions of the  $\log_2$  signal intensities of probes deemed present by MAS5.0, absent by MAS5.0, and those detected by MudPIT. The signal intensity of probes found to be present by either MAS5.0 or proteomics is on average 16 times higher than that recorded for undetected transcripts (probes called absent by MAS5.0).

derestimated, step in Affymetrix microarray signal processing is the determination of a probe detection call, wherein transcript signal of a probe set (and by inference, the corresponding gene product) is deemed to be present or absent. There are two pieces of information available per probe set on an Affymetrix chip that can be used to determine whether a particular transcript is indeed present (or absent) in a sample, signal intensity and the detection  $p$  value. The aim is to derive a rigorous, statistically based, and reliable measure of genuine signal over background (baseline), with a sufficiently high signal-to-noise ratio to unambiguously confirm the existence of a specific target mRNA species in the sample of interest.

The simplest method for deciding whether a transcript is expressed is to look at the “detection call” made by the Affymetrix algorithm, which is based on the calculated detection  $p$  value. The detection  $p$  value is a statistically derived confidence measure or indicator of whether the signal generated from a probe set is sufficiently above background (or not) to be deemed present. Using the default software settings, any transcript detected with a detection  $p$  value threshold cutoff defined as  $\alpha_1$  (or lower) will be defined as “present,” while those between  $\alpha_1$  and a second threshold  $\alpha_2$  will be called “marginal” and those with  $p$  values greater than  $\alpha_2$  deemed “absent.” Default values for  $\alpha_1$  and  $\alpha_2$  can differ between chip types and are defined by Affymetrix based on extensive in-house analysis of test arrays (51).

The detection  $p$  value is determined in two steps. First, a discrimination score,  $R$ , is calculated based on the signal intensities of the group of PM probes relative to the corresponding set of MM probes per probe set. Then, the scores from each of the probe pairs are tested against a user-defined threshold  $\tau$  (tau), which specifies the minimum  $R$  that “votes”

for presence, to calculate a detection  $p$  value for the entire probe set. These steps are described in more detail below.

From a set of intensity values for each group of  $PM$  and  $MM$  probes defining a probe set,  $R$  is calculated for each  $PM$  and  $MM$  pair according to the following formula:

$$R = \frac{PM - MM}{PM + MM} \quad (\text{Eq. 1})$$

This can be rewritten as:

$$R = 1 - \frac{2}{\frac{PM}{MM} + 1} \quad (\text{Eq. 2})$$

In other words, one can see that  $R$  depends strictly on the ratio of  $PM/MM$ , which varies between  $-1$  and  $1$ . The bigger the calculated ratio, the closer  $R$  will be to  $1$ . This would happen in cases of high signal-to-noise. Conversely, if the ratio is close to  $1$  (signal marginally above background),  $R$  will be close to  $-1$ .

The  $R$  scores are rank-sorted and evaluated using the one-sided Wilcoxon's signed rank test (see Refs. 48 and 49 for details). For every probe set, the algorithm ranks the probe pairs based on how far their discrimination score is from  $\tau$  and uses the information to calculate the detection  $p$  value. The Wilcoxon's signed rank test is a commonly used, simple nonparametric statistical method that has several desirable properties and certain limitations. Two notable advantages are that it is insensitive to spurious outliers and does not assume a normal distribution of the data (49). It is not necessarily the optimal test for class discrimination, however, and more advanced statistical measures (including methods better suited to the principles of machine learning) are likely to be far more effective for determining true probabilities, resulting in fewer incorrect or marginal classification calls.<sup>3</sup>

The calculated  $R$  scores are compared with  $\tau$ , and if the  $R$  score is higher than  $\tau$ , a "present" call is made for a given probe pair, while those with  $R$  scores falling below this value result in an "absent" call. The votes across all probe pairs comprising a complete probe set are tabulated and summarized in the detection  $p$  value, which is a confidence measure reflecting the detection call. The higher the  $R$  score is above  $\tau$ , the closer the detection  $p$  value is to  $0$ , and hence the greater confidence that a transcript is indeed expressed. Conversely, the lower the  $R$  score is below  $\tau$ , the closer the  $p$  value is to  $1$  and hence a lower confidence is placed on the detection call and, by inference, on the possible existence of the corresponding mRNA in the sample.

Increasing  $\tau$  increases the stringency of the test by decreasing the rate of incorrect predictions or false-positives. Although this effectively increases the specificity and reliability of the assay, in doing so the overall sensitivity of an experiment (namely, the detection of *bona fide* mRNA transcripts) is reduced. For example, increasing  $\tau$  from  $0.015$  to  $0.15$  de-

creases the proportion of present calls on a Rat Genome U34A array by  $25\%$  (4). Conversely, lowering  $\tau$  decreases the specificity of an analysis, thereby increasing the sensitivity. There is a default value for  $\tau$  for each particular type of chip, which is based on extensive controlled experimentation. In general, Affymetrix recommends that this default not be changed (52).

Redefining the  $\tau$  threshold has a somewhat different effect on the detection call than modifying the detection  $p$  value threshold. The former influences the votes of individual probe pairs for presence or absence, which in turn may drastically affect the final detection  $p$  value of a probe set—in other words, the  $\tau$  controls the number of probe pairs that vote for presence. The detection  $p$  value, on the other hand, considers the probe set as a whole. Ultimately, however, the effect of changing either threshold leads to a similar result—both modify the percentage of present calls made per dataset. Modifying the detection call by changing the detection  $p$  value threshold is much easier to do in practice, however, because it does not require reanalysis of the data outputted by Affymetrix analysis software, which can be very time consuming for large datasets. For our purposes, we decided not to modify  $\tau$  (or the statistical measure for calculating  $R$ ), but rather work with detection  $p$  value thresholds in order to optimize the detection call for our  $\alpha$ TC-1 dataset.

We also considered the differences in detection  $p$  values obtained in the three replicate arrays and found them to be very precisely reproduced between replicates, especially for probes with very low detection  $p$  values ( $<0.1$ ). In other words, we found that most probes with a detection  $p$  value below the default detection call threshold ( $0.04$  in the case of the MG\_U74Av2 array) in a single experiment were likewise similarly detected across the other replicates. Indeed,  $78\%$  of probes deemed present in any one of the three replicates were likewise present in all three, while  $12\%$  were present in exactly two of three replicates, with the remaining  $10\%$  detected in just one dataset. However, because roughly  $50\%$  of all probes on the  $12,000$ -gene array were called present using the default  $p$  value cutoff (Supplementary Table I), the decision as to whether some  $600$  or so transcripts are indeed expressed remains unclear. One may need to consider additional information when deciding detection in these cases, such as absolute signal intensity.

A detection  $p$  value close to  $0.5$  indicates that there is no significant difference in the relative abundance of the  $PM$  probes with respect to the  $MM$  probes. There are several possible causes for this: one is that the transcript is indeed not detected, so the signal will be very low (close to background) for both the  $PM$  and  $MM$  probes. In this case, the absent call is correct. Another reason may be nonspecific binding to the  $MM$  probe and possibly also the  $PM$  probe. In this case, the signal recorded for both may be very high, but the difference in signal between  $PM$  and  $MM$  may not be significant, resulting in a high detection  $p$  value and an absent

detection call. This is problematic, because it can be difficult to decide whether the probe is actually absent and what is detected is truly nonspecific binding, or that the mRNA is very abundant and so binds significantly even to the MM probes. Indeed, we have observed that highly abundant mRNAs often have detection  $p$  values as high as 0.8 (and hence are deemed “absent” by the Affymetrix analysis software) only to confirm these as truly highly expressed gene products using RT-PCR (as discussed below).

**Signal Intensity**—Signal is a quantitative metric summarizing the intensity of the probe set and representing the relative abundance of expression of transcript. Signal calculation takes into consideration the absolute and relative intensities of PM and MM probes in each probe pair and uses them to calculate an “idealized mismatch” (IM) signal which is then used instead of MM. If  $MM < PM$ , then  $IM = MM$ , otherwise IM is calculated from PM and depends on various other factors such as background intensity around the probe cell. The Affymetrix-supplied software uses a One-Step Tukey’s Biweight Estimate (a standard statistical-weighted average method that is relatively unaffected by outlier probe values) to first integrate the individual probe signals to obtain a robust mean of signal across the entire probe set. The algorithm starts by finding the median, and then the distance between each data point from this median is calculated. Based on these distances, the algorithm assigns a weight such that data points far away from the median will have low or even zero influence on the final value. The median is then recalculated once per probe set by incorporating these weights (see Ref. 49 for details).

In this manner, each probe pair contributes to the final estimate of signal intensity. The significance of this contribution is greater if PM signal is higher than MM and if the probe pair signal value is close to the mean of all probe signals. If the PM signal in a probe pair is higher than MM, the MM signal is considered to be informative and is often used as an estimate of stray nonspecific (background) signal. If MM signal for a given probe pair is higher than PM (which may occur in cases of significant cross-hybridization or when transcript levels are below detection limits), the MM signal is considered to be uninformative and is generally not used as a measure of noise. In this alternate scenario, an imputed value, IM, is used instead of MM signal. This IM value is usually deemed to be slightly smaller than PM so as to prevent negative signal values being generated for a given probe pair. If there are many such probe pairs present in a probe set, the detection algorithm will usually call the probe absent, even if its absolute signal intensity is high (53).

In order to increase the sensitivity of the transcript detection procedure, we now routinely consider probe signal intensity as well as the detection  $p$  value for probes with detection  $p$  values greater than the default cutoff.<sup>2</sup> But first, we performed independent measurements of transcript levels using the RT-PCR method (and, as discussed further below, proteomics) to guide our analyses.

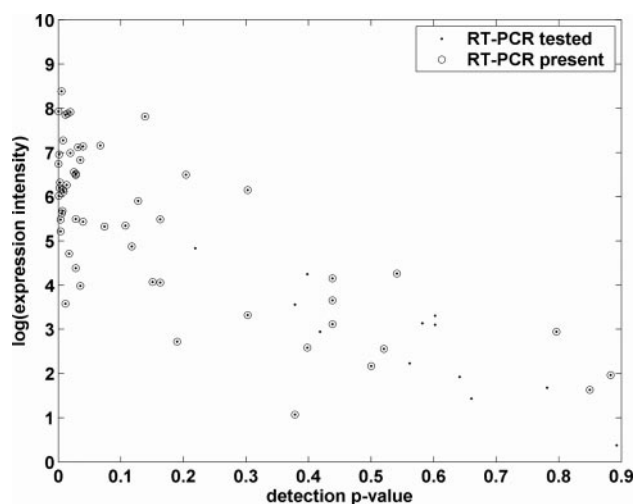
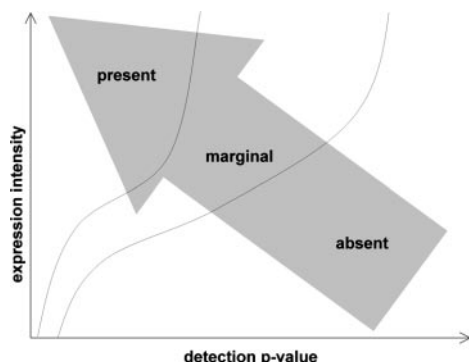


FIG. 2. **RT-PCR validation.** Scatter plot of microarray detection  $p$  values versus  $\log_2$  probe signal intensity obtained for 67 select gene products that were independently assessed for expression by RT-PCR. Probes with detection  $p$  value  $>0.04$  (dotted line) were deemed “absent” using the default Affymetrix detection call. While the expression of all probes predicted to be “present” by microarray was confirmed, many transcripts with predicted probe detection  $p$  values significantly above the default threshold were also identified as expressed by RT-PCR.

**RT-PCR Validation**—RT-PCR validation experiments are considered to be the “gold-standard” technique for transcript detection, both in terms of sensitivity and especially specificity. For this reason, we designed primers to amplify a selected set of 67 disparate transcription factors that were deemed either absent or present in all three replicate experiments to interrogate using RT-PCR. Based on our results, we found the detection  $p$  value threshold to be overly conservative—many gene products predicted to be absent (some with  $p$  values as high as 0.8) were in fact readily detectable by RT-PCR (Fig. 2). Although the specificity of the Affymetrix detection call (*i.e.*  $p$  value cutoff 0.04) proved to be good, with few false-positives, the sensitivity was unsatisfactory (excessive false-negatives). This is especially true for low-abundance transcripts (Fig. 2). These data question the suitability of the default  $p$  value thresholds (and the algorithms currently used to calculate these) and have forced us to reconsider the entire issue, such as increasing the default  $p$  value cutoffs for present detection call or evaluating signal intensity together with the detection  $p$  value when deciding on the presence of a probe.

Based on our interpretation of the RT-PCR experiments, we hypothesize that if a probe has a sufficiently elevated signal, the transcript is likely to be present even if the detection  $p$  value is deemed higher than the default cutoff. Fig. 3 shows a schematic cartoon representation of more flexible, heuristic detection “cutoffs” (lines) that incorporate knowledge of both typical distribution scatter plot of the  $p$  values and signal intensities obtained for endocrine cell lines like  $\alpha$ TC-1. The assumption here is that if a probe is detected with a reason-



**FIG. 3. Schematic guide for interpreting microarray-based gene expression profiles.** This plot provides a graphical representation of a suggested “rule” for evaluating transcript detection patterns. The gray arrow represents confidence in detection. If probe signal is high and the detection  $p$  value is low, a gene transcript is highly likely to be expressed and hence detected, while confidence decreases with decreasing signal and increasing detection  $p$  value. Low-intensity probes with elevated detection  $p$  values indicate an mRNA species is unlikely to be expressed. However, the status of gene products with intermediate probe signals and detection  $p$  values is more difficult to assign with confidence. Moreover, optimal thresholds for probe signal and  $p$  value cutoff values will depend on the actual experimental data and preprocessing, on how many replicates are available, and so on. The general assumption that 30–40% of mammalian genes will be expressed in any one tissue (70) can also be used as a second guide when deciding on suitable thresholds.

able signal (practically defined as >35th percentile) and has a relatively modest detection  $p$  value (<0.1 being considered reasonable, but the exact value will likely depend on the specific dataset (54)), then the gene product is most likely expressed. Visual inspection of hundreds of probe images indicated that mRNAs with a low signal (<25th percentile) and high detection  $p$  value are indeed most likely absent,<sup>4</sup> but one has to be especially careful with probes exhibiting intermediate signal intensity, regardless of the detection  $p$  value.

**Global-scale Proteomic Screening Using Multidimensional LC-MS**—Given that one cannot perform confirmation RT-PCR on a genome-wide scale, and considering the fact that RT-PCR may also be subject to artifacts when applied outside its useful dynamic linear range, we opted to measure protein levels directly. While classical biochemical methods for examining protein expression, such as Western blotting or ELISA, are quite effective, they are tedious and generally also limited in scope. Hence, we decided to perform large-scale proteomic profiling studies using high-throughput protein MS. The main objective was to identify as many proteins in  $\alpha$ TC-1 cells as possible and to compare this pattern to the microarray profile. To this end, we used an ultra-sensitive gel-free LC-MS-based shotgun microsequencing procedure to systematically identify proteins exhibiting detectable expression (see “Experimental Procedures” for details).

The MudPIT technique employed in this study (40–42, 55)

couples high-resolution fractionation of protein enzymatic (*i.e.* tryptic) digests using multidimensional capillary-scale HPLC to real-time high-efficiency data-dependent MS/MS via ESI (reviewed in Ref. 39). The principle advantage of this approach stems from the very good sample separation achieved by jointly performing ion exchange and reverse-phase chromatography at the peptide level combined with automated procedures for peptide selection and fragmentation. This technique yields very large collections of richly informative spectra, allowing for extensive sequence determination (55, 56).

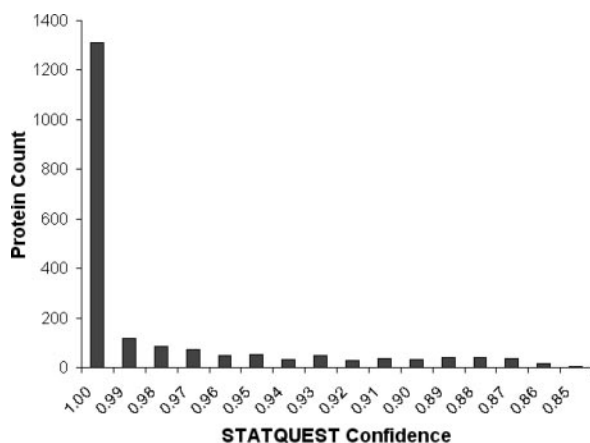
To improve the odds of detecting lower-abundance proteins of special biological interest, such as transcription factors, we first performed a simple subcellular fractionation procedure to enrich for lower-abundance nuclear factors prior to analysis (42, 57). This is a particularly important consideration given the sizeable overall dynamic range in protein abundances (commonly predicted to be over five orders of magnitude), which contrasts with the much more limited effective detection range of current instrumentation (*i.e.* two to three orders of magnitude). Although the resulting protein fraction was not pure and contained sizeable levels of cytoplasmic cross-contaminants (42, 57), this easily implemented procedure substantively reduced the levels of higher-abundance cytosolic proteins (*e.g.* housekeeping enzymes), which usually tend to dominate a proteomic analysis. In practice, this step considerably improves the overall comprehensiveness of proteomic detection (Refs. 42 and 57; data not shown).

For the shotgun analysis, the proteins were concentrated by precipitation, denatured using urea, and digested extensively and sequentially using two proteases; first with endoproteinase Lys-C, which functions under denaturing conditions, then with trypsin (after dilution of the urea), which is a more processive enzyme. The resulting peptide mixture was de-salted by solid-phase extraction and analyzed using the basic MudPIT procedure (see “Experimental Procedures”). Although this approach proved to be very effective, resulting in the identification of many proteins (see below), we have found that multiple repeat MudPIT analyses are generally needed to achieve blanket coverage (that is, a saturating level of detection), even when investigating simplified organellar fractions. Hence, for this study, we performed three analyses on independently isolated cell extracts.

**Database Searching and Statistical Validation**—To identify the proteins, the entire collection of ~100,000 acquired tandem mass spectra was searched against a minimally redundant database of curated human and mouse Swiss-Prot/TrEMBL protein sequences using the SEQUEST search algorithm (43). Because the candidate spectral matches and search scores are open to interpretation and are often inconclusive (56), it was critical to estimate both the accuracy of an individual putative identification and the overall rate of false discovery (proportion of incorrect identifications or false-pos-

<sup>4</sup> M. Maziarz, unpublished observations.





**FIG. 4. Distribution of preliminary proteomic confidence scores.** The graph shows the distribution of initial protein identification (database search) confidence scores, as determined by the STAQUEST probability algorithm, for all candidate proteins identified by MudPIT with a minimum predicted likelihood of 85% or greater. As can be seen from the markedly skewed distribution, the majority of the preliminary predictions are of very high confidence (>95% probability of being correct), with ~93% of all putative matches predicted with 90% or greater confidence and >65% of all proteins predicted with >99% confidence.

itives) when performing this analysis. Several heuristics guidelines have been developed to address these important (and related) concerns (56). More rigorous statistical measures have also been introduced recently to facilitate data interpretation (reviewed in Ref. 58).

For this study, each candidate database match was first evaluated using a statistical algorithm, STATQUEST (57), in order to calculate a probability score (likelihood a prediction is accurate) based on the preliminary SEQUEST results. An initial subset of higher-confidence proteins (with a minimum of 85% or greater predicted probability of being correctly identified in at least one of the three replicates) were selected for further consideration. It is important to note that this initial confidence filter threshold applies to individual sequence matches. In practice, due to the nonlinear distribution of database confidence scores (57), the majority of the filtered predictions actually have a very high likelihood of being correct. Indeed, as seen in Fig. 4, most of these putative identifications were predicted with a +99% likelihood of being correctly identified ( $p$  values <0.01; that is, far greater than the STATQUEST  $p$  value threshold of 0.15).

Because spurious database matches usually have limited supporting evidence, often corresponding to only a single spectral match (58), as an additional measure of stringency, we accepted only those candidate proteins for which at least two or more spectra were detected across the three datasets for further analysis. This additional filtering measure resulted in a final set of 1,651 high-confidence proteins (Supplemental Table II, A and B). Although significant variability in spectral counts was seen between the three repeat MudPIT runs, especially for proteins exhibiting lower cumulative spectral

counts, the overall reproducibility in terms of the overlap or fraction of shared proteins after filtering was about ~45%, as illustrated by the Venn diagram shown in Supplemental Fig. 2.

As an independent measure to assess the effectiveness of this filtering, we empirically calculated the preponderance of incorrect identifications in our set of putative high-confidence proteins. This is especially important with repeat experimental datasets, which can accumulate false-positives due to the sheer number of spectra examined. This was done by first populating the reference protein sequence database with an equal number of mock proteins, created by inverting the amino acid orientation of the normal Swiss-Prot/TrEMBL protein sequences. Because these “reverse” sequences are not expected to occur naturally, any matches to these decoy proteins represent spurious false-positives. The final proportion of identifications mapping to reverse relative to normal (or “forward”) proteins therefore provided an objective criterion for estimating the false-discovery rate. As mentioned, spurious database matches are often supported with minimal supporting spectral evidence. The plot provided in Supplementary Fig. 3 shows the decreasing ratio of reverse-to-forward sequence database matches typically observed with increasing number of observed spectral counts. Because the vast majority of incorrect matches were preferentially detected with single spectra alone, they can be readily discarded from further analysis by establishing a minimal two-spectral minimum as a final measure of accuracy. Indeed, after filtering the data using this and the initial criteria outlined above (e.g. STATQUEST), fewer than 5% false-positives (reverse proteins) were detected in the final dataset (data not shown).

Regardless of which method of data validation is employed, the filtering process should aim to balance the trade-off between specificity (precision), which reflects the proportion of incorrect identifications or false-positives, and sensitivity (recall), which indicates the proportion of missed identifications or false-negatives. Receiver-operating characteristic (ROC) plots are often used to assess the effects of varying filter stringency on precision and recall (59). Although we do not have knowledge of the correct class labels (because, in most proteomic studies, one usually does not know *a priori* which proteins are in fact present in a sample), in practice, one can estimate this trade-off empirically based on the fraction of database matches to forward and reverse sequences after various filters are applied. The ROC-like plot shown in Fig. 5 shows the effects of applying different confidence filters to the  $\alpha$ TC-1 dataset, based either on the preliminary STATQUEST probability scores or the reproducibility of detection (detected in one or more repeat analyses) or the cumulative spectral counts. The relationship between the number of credible candidates (namely, matches to normal forward protein sequences) and false-positives (represented by matches to reverse sequences) passing each filter is complex. Of course, as one applies a more-stringent filter, the overall false-positive rate decreases (as evidenced by the reduced proportion of

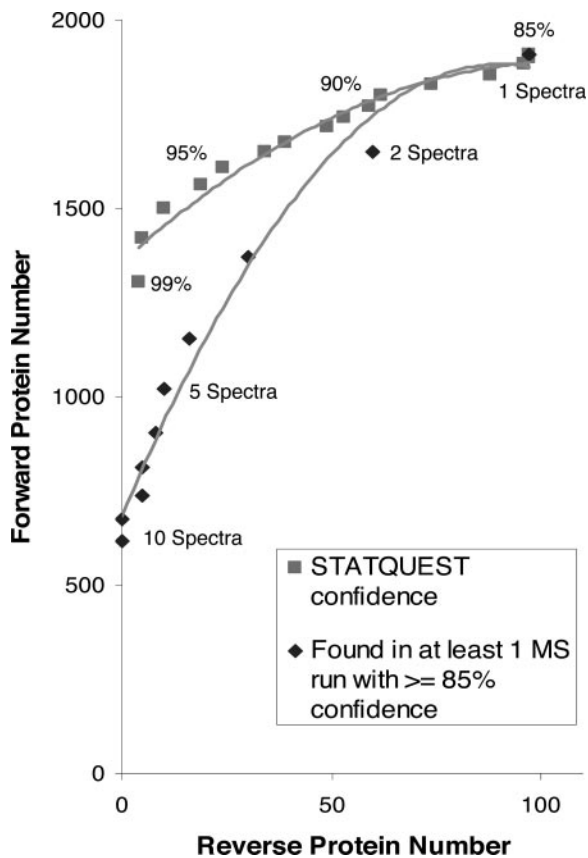


FIG. 5. **The interplay between precision and recall in database searches.** ROC-like plot of the predicted specificity (precision) and sensitivity (coverage or recall) obtained by applying various quality filters to the preliminary proteomic (database search) results. The presumed true-positive fraction is calculated based on the proportion of normal “forward” proteins passing the filter criteria, while the false-positive fraction is calculated based on the proportion of “reverse” proteins passing the same criteria. The data points represent the number of proteins identified with either a given minimum recorded spectral count (with the data also subdivided to highlight run-to-run detection reproducibility) or based on the preliminary STATQUEST database confidence cutoff score. Note that while the false-positive (reverse protein) fraction decreases as one increases the stringency of the STATQUEST quality filter (*i.e.* moving from right to left), the decrease is marginal as compared with the much sharper decrease in the positive fraction (detection sensitivity or coverage). A filter based on a STATQUEST cutoff of +85% confidence and a minimum of two spectra was chosen as it provided reasonable overall detection coverage for a single experimental dataset without excessive predicted false-positives.

reverse proteins), whereas the false-negative rate increases (as represented by the decrease in the number of forward proteins detected). But this trade-off is clearly nonlinear. For the most stringent filter, such as only considering proteins identified with high confidence in all three repeat samples or else protein detected with a minimum of 10 or more spectra, the false-positive rate can be reduced to virtually zero, but at the expense of a severely limited sensitivity (*i.e.* very high false-negative rate).

Depending on the desired balance of specificity/sensitivity, one may select a mixture of different filter criteria. In practice, the two-stage filtering chosen here ensures reliable identifications (limiting false-positives to <5%) without overly penalizing overall proteomic detection coverage. However, as discussed below, this issue can and should be revisited as other supporting data become available.

**Reproducibility and Variance**—Despite the effectiveness of statistical filtering, MudPIT-based profiling is still affected by many other sources of error that should be considered. One of the most troubling is the problem of under-sampling. Shotgun MS/MS fragmentation involves a somewhat stochastic peptide selection step, which is generally biased toward higher-intensity peptides even when data-driven exclusion lists are used (45, 56). Given that the duty cycle (scan rate) of current instrumentation is limited, only a fraction of all eluting peptides will be selected for fragmentation, while many others will be missed. This trend results in preferential detection of higher-abundance proteins (because these often give rise to higher-intensity peptide ions), while lower-abundance proteins frequently go undetected (45). This limitation is compounded by significant variability in the peptide spectral quality run-to-run (60).<sup>5</sup> Due to this, MudPIT experiments must generally be repeated several times to achieve adequate overall detection coverage (*i.e.* saturation) (61). The number of runs needed to reach a useful saturation point will depend on the complexity of the sample under study. In our experience, three repeat analyses are usually sufficient to detect >80% of all detectable proteins present in a sample.<sup>5</sup>

Saturation of detection is a particularly important consideration if one wishes to compare proteomic profiles between different samples (*e.g.* different cell lines). Relative protein abundance between samples can be estimated based on the ratio of median cumulative spectral count detected for each protein in the respective samples (45). This method is far simpler to implement and interpret than ones relying on sophisticated chemistries or based on internal standards labeled with stable isotopes (62), but requires a sufficient number of repeat analyses to be robust to outlier effects (45).<sup>5</sup> Spurious variance (experimental noise) can be assessed by calculating the standard deviation and overall reproducibility of protein detection run to run (data not shown).

Additional subcellular fractionation and proteomic enrichment procedures aimed at sample simplification can also help surmount the under-sampling problem (42). Alternatively, further advances in instrumentation resulting in higher scan rates, more-efficient modes of peptide sampling, or better spectral analysis, may effectively bypass this concern.

**Data Cross-referencing**—In addition to the RT-PCR validation, we sought to combine the results from the MudPIT profiling experiments with the microarray datasets to validate our hypothesis that if a probe has high enough signal (raw

<sup>5</sup> C. Chung and A. Emili, unpublished observations.

intensity), then it is likely to be present even if the detection  $p$  value is higher than the default cutoff. This first required cross-referencing the MudPIT and microarray datasets using their respective gene product identifiers (IDs or accession numbers).

When working with commercial microarrays, as we have done here, it is often straightforward to obtain the relevant cross-reference mappings between probe IDs and Swiss-Prot/TrEMBL protein accessions. Affymetrix, for example, provides current mappings between its probe IDs to various knowledge databases for all its chip designs, which can be readily downloaded through their website ([www.affymetrix.com](http://www.affymetrix.com)). As for the proteomic dataset, we needed to obtain the latest Swiss-Prot accession numbers in order to most efficiently map these to the corresponding gene probes. The ExPASy website ([us.expasy.org/sprot/](http://us.expasy.org/sprot/)) provides a convenient tool, Swiss-Prot ID tracker ([us.expasy.org/cgi-bin/idtracker](http://us.expasy.org/cgi-bin/idtracker)), to retrieve up-to-date Swiss-Prot IDs, including primary accession numbers for any deleted gene product IDs. Also, within ExPASy, is a Swiss-Prot/TrEMBL entry retrieval list tool ([us.expasy.org/sprot/sprot-retrieve-list.html](http://us.expasy.org/sprot/sprot-retrieve-list.html)), which allows a quick retrieval of the primary accession number (latest unique alphanumeric protein identifier) and secondary accession number (older accession numbers retained after sequence merger) for a given input list of proteins.

To ensure completeness for this study, we obtained all the primary and secondary accession numbers for each of the protein IDs before mapping these to the microarray probes. We were able to map 72% of the reliably identified proteins to probe IDs on the microarray (Supplemental Table III). Presumably, we were not able to map the remaining proteins either because there were no corresponding gene probes were present on the microarray or because a human ortholog was preferentially identified (a possibility, because both human and mouse protein sequences were used in the database search).

One challenge working with disparate proteomic and functional genomic datasets is the constant updates to the relevant annotation (knowledge) databases, such as Swiss-Prot/Trembl in the case of proteins and UniGene in the case of mRNA, from which one derives the respective IDs. To be comprehensive, these tables need to be up-to-date, but without creating legacy issues. Hence, an alternative, and perhaps more rigorous, approach to linking the proteins and microarray probes is to identify matches using a sequence alignment tool like BLAST ([www.ncbi.nlm.nih.gov/BLAST/](http://www.ncbi.nlm.nih.gov/BLAST/)). One can utilize a bulk sequence retrieval tool, like the ExPASy Swiss-Prot/Trembl entry retrieval list tool, to obtain both protein and gene sequence information using accession IDs in the correct format for alignment.

When interpreting a BLAST result, one needs to consider more than just the raw score or expectation ( $e$ ) value as match criteria. This stems from the fact that BLAST is a locally weighted context algorithm, and even highly significant  $e$ -values approaching 0 do not always indicate a perfect alignment.

Therefore, additional criteria, such as fraction percent identity, should be used to define an acceptable threshold cutoff. In our study, we opted for an  $e$ -value  $<10^{-20}$  and  $>95\%$  fraction identity as thresholds. We do note, though, that a reciprocal validation search, in which BLAST is repeated after swapping the queries and subjects, can help to identify spurious matches between paralogous gene products. But again, one needs to be cautious selecting appropriate cutoff scores in order to maintain a balance between sensitivity and specificity. Finally, it is worth considering that the reference source of the gene and protein sequence information, and whether the sequences were retrieved from the same database, can markedly affect the stringency of a test.

*Comparison of the mRNA and Protein Profiles*—After cross-referencing, we could evaluate the overlap and relative sensitivity, specificity, and biases of the proteomic and genomic approaches. Using the same detection call procedure as for the RT-PCR analysis, which is based on a predefined default threshold detection  $p$  value of 0.04, we analyzed the protein distribution relative to the microarray-derived detection  $p$  values and raw probe signal intensities.

A total of 5,945 gene products were detected by the gene chips alone (in at least two of the three repeat experiments, using a  $p$  value cutoff of 0.04), reflecting the high coverage attainable by microarray. A subset of the cognate proteins encoded by these transcripts may have been missed by MudPIT because we selectively analyzed a nuclear fraction, or because the corresponding sequences were not represented in the reference database used for the spectral search. Of the 888 gene products that could be unambiguously cross-mapped between the platforms, 762 were deemed to be expressed by microarray in addition to being identified by our proteomic method. Conversely, 126 gene products were detected (*i.e.* identified with high confidence) exclusively by MudPIT, being somehow missed by the gene chip platform using the default  $p$  value filter ( $p$  value  $>0.04$ ; therefore deemed “absent”). Taken at face value, the fact that the genomic and proteomic data are in agreement for the majority of cross-mapped gene product pairs (in that both the gene and corresponding protein were detected) validates the reliability of the two platforms. As seen in Fig. 6A, there was a marked tendency for the MudPIT screening to detect the putative translation end-products of transcripts detected with low  $p$  values (that is, at or below the default cutoff). These data further confirm the stringency of the default threshold.

Although no clear correlation existed between the predicted detection  $p$  value and the corresponding proteomic evidence, as seen in the scatter plot shown in Supplemental Fig. 4, many of the orphan proteins were identified with substantial spectral counts (and, by inference, with very high confidence). Although we cannot exclude experimental error, these data could be taken as evidence that the detection  $p$  value filter cutoffs used for the microarray (and even the proteomic analyses) are too stringent. Alternatively, these data may

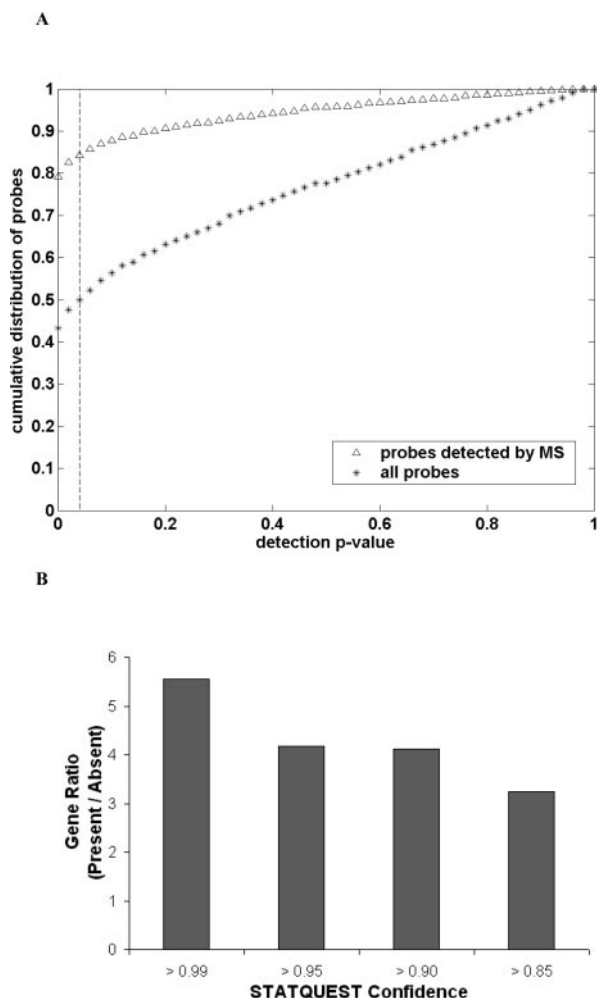


FIG. 6. **Comparison of the proteomic and microarray data.** A shows the cumulative fraction of high-confidence proteins detected by MudPIT profiling *versus* the corresponding microarray probe detection  $p$  values. For comparison, the complete cumulative profile of all probe  $p$  values is shown. The default detection call cutoff value is indicated with a dotted line. B shows a bar graph of the ratio of microarray probe detection “present” to “absent” calls, based on the default detection  $p$  value ( $\leq 0.04$ ), made for proteins identified by MS based on only a single medium- to high-confidence spectra (singleton peptide) but excluded from further consideration due to application of a stringent proteomic filter. As can be seen, considerable supporting evidence of gene expression is evident for these marginal protein identifications. This suggests one can use parallel gene expression data to validate inconclusive results from a shotgun proteomic profiling study.

point to an unexpected biological uncoupling between the corresponding levels of the respective mRNA and protein species.

Considering that probe signal and the detection  $p$  value are intrinsically inversely related due to the nature of the Affymetrix statistical algorithms used to calculate the two, it was not too surprising to see that microarray probes with low detection  $p$  values ( $\leq 0.04$ , or “present”) tended to have higher signal intensity than those corresponding to transcripts

deemed “absent” (Fig. 1). Likewise, and reassuringly, the probes for transcripts encoding the proteins identified by MS also tended to have higher average signal intensities (Fig. 1). In other words, the average mRNA signal for gene products confirmed to be expressed by MudPIT is much closer to probes detected by microarray analysis than those that were deemed absent (by microarray). These results are in alignment with the heuristic rule summarized above (Fig. 3).

*Improving Proteomic Coverage by Data Integration*—Using the relatively stringent two-stage quality criteria, false-negatives are an inevitable consequence of proteomic dataset filtering (*i.e.* Fig. 5). On the other hand, one generally prefers to avoid compromising the integrity of the data by loosening the filters too much. Hypothetically, one could seek to incorporate alternate supporting evidence, such as gene expression data, both as a guide to assess the effectiveness of the proteomic filtering criteria (specificity *versus* sensitivity) and to validate marginal protein identifications. To address this possibility, we examined the suitability of using microarray results to eliminate false-positives in the subset of protein identifications with lower-confidence database scores, considering only those proteins with solid preliminary probability scores ( $>85\%$  confidence according to STATQUEST) but that were nonetheless removed because the identifications were based on only a single spectrum (a commonly accepted quality criteria) (56).

Fig. 6B shows the ratio of “present” and “absent” calls made for the microarray probe pairs corresponding to transcripts predicted to encode these marginal proteins using a default detection  $p$  value  $\leq 0.04$ . Because half of all genes are predicted to be expressed, a false-positive database match has an equal chance of mapping to an expressed gene (hence, the ratio of forward-to-reverse sequences should be  $\sim 1$ ). However, the probes corresponding to even the most marginal protein identifications exhibited a far greater chance of being detected as “present” by microarray (even using the overly stringent  $p$  value threshold) than one would expect by chance alone. This implies that one could apply the results of a parallel gene expression study to reduce the false-negative rate resulting from stringent proteomic data filtering, without increasing the number of contaminating false-positives.

*Annotation and Functional Inference*—To obtain a holistic sense of the global similarities and differences between the two datasets, functional annotation was used to compare the transcriptome and proteomic datasets. Obtaining simple gene product descriptions from a reference annotation resource like ExPASy is often a good starting point for interpreting the biological significance of expression profiles and a logical first place for deducing the functions of proteins of special note (*e.g.* those with interesting expression patterns), which can then be individually followed up. Alternatively, one can look for general patterns of functional enrichment using a more-generalized annotation resource such as the GO database ([www.geneontology.org](http://www.geneontology.org)), which reports on the functional properties

of gene products using a more-standardized, computer-friendly schema. The idea is to look for underlying trends in an input list of gene products, by calculating the relative membership enrichment (*versus* chance alone) in various select functional categories (e.g. GO terms) using a suitable statistical measure. While the GO curation is far from complete, and potentially error prone, it is currently the most comprehensive resource of its type. Moreover, one can use a dedicated computer program or web application, such as GOMiner (discover.nci.nih.gov/gominer/), to automate the analysis.

As might be expected, gene transcripts expressed in  $\alpha$ TC-1 were enriched for cell communication, regulatory, and signaling proteins known to act as key determinants of differentiation, tissue-specificity, and endocrine cell function (data not shown). As shown in Supplemental Table IV, a disproportionate fraction of the gene products also identified by LC-MS were found to be involved in nucleic acid binding, transcription, mRNA splicing, and genome maintenance (e.g. DNA replication, cell-cycle control, etc), consistent with the subcellular fractionation and enrichment procedure used in this study. Hence, based on the observed subcellular localization and concordant expression patterns, the potential function of at least some of these proteins in  $\alpha$  cell physiology can be reasonably predicted.

### DISCUSSION

The introduction of high-throughput, high-resolution experimental technologies over the past few years, especially DNA microarray gene chips and protein MS, has raised optimism that the molecular underpinnings of endocrine cell function can now be elucidated on a systems-wide level. Our expectation is that systematic molecular profiling studies, when performed rigorously, will also help address fundamental biomedical research questions relating to the basis for common metabolic disorders, such as insulin resistance and overt diabetes. Nevertheless, there is growing recognition of the serious limitations and biases associated with most expression profiling technology in their current forms, particularly with regards to the reliability of the biological inferences that can be made (24, 36).

In this pilot study, we have attempted to combine large-scale gene and protein expression technology platforms together with the aim of making more reliable, and more comprehensive, inferences regarding endocrine cell function. We have outlined some of the less well-appreciated, but nevertheless important, technical and analytical issues associated with practical implementation of global profiling methods, while offering potential solutions to the most pressing problems. Although the overlap between the gene chip and proteomic patterns reported here represents only a small portion of the total predicted genetic complement of mouse, our data strongly suggest that integration of multiple types of experimental techniques truly does allow one to gain a more complete, statistically sound, and biologically meaningful picture of the sample under study.

It is important to stringently filter genomic datasets using proven, rigorous statistical measures of reliability (24, 36, 56, 58) and to regard preliminary results with caution. Detection calls are an important first step for determining the reliability of probe values. Accurate predictions, and their associated  $p$  values, are proving to be important metrics as microarray datasets are increasingly being used as the basis for disease or sample classification systems (e.g. see Ref. 63). The robustness of probe features should be carefully investigated prior to their use as in machine-learning and classification studies so as to avoid spurious biomarker discovery. In this study, we have found the default microarray  $p$  value threshold, as suggested by Affymetrix, to be overly conservative with respect to detection specificity at the expense of sensitivity. Using RT-PCR, often considered to be the “gold-standard” technique for transcript detection, we have validated the expression of a number of mRNAs with high-predicted probe  $p$  values ( $>0.04$ ; therefore, classified as marginal or absent by the Affymetrix detection call). Moreover, we have confirmed these initial findings using high-stringency global proteomic profiling. Viewed together, these results indicate that a significant number of genuine transcripts are likely commonly missed in most microarray screens, presumably as false-negatives due to overly restrictive filtering criteria.

Based on this study, we propose, as a heuristic guideline, that researchers incorporate other information, like probe signal intensity as well as other orthogonal experimental data, such as RT-PCR and (ideally) protein MS, to define suitable detection filters for microarray data. Although the exact nature of these criteria would differ from experiment to experiment, this approach is generalizable. By considering additional biologically pertinent factors, such as detected related properties in gene function across a global dataset, one might better discern between the significance of a (noisy) expression profile. This is particularly the case when taking into account biological variation on top of experimental noise.

The problem of making a reliable detection call seems like a far simpler problem as compared with determining differential gene expression for example. In the latter, one is usually dealing with data from several different samples—either a time series or treatment-control-response experiments, or both—that first need to be normalized so that the signals are comparable. Then, thresholds for deciding upon “significant change” in transcript levels must be decided on. A “yes or no” answer for deciding whether an mRNA was detected requires only a single chip (although, the confidence and power of a result can be increased by repeat analysis, if required), and normalization is not necessary because there is no need for comparing signals. However, detection of genuine gene expression is a very important and interesting research problem in and of itself, though not nearly as alluring as the issue of differential expression, which is possibly a reason why it has received relatively little attention to date.

For starters, it is important to know if a gene is in fact

expressed (or not) before one attempts to establish if transcript levels change under different experimental conditions. That is because low-abundance or marginal transcripts are frequently misidentified as “differentially expressed” by standard analysis software, because the absolute signal is often not taken into account if merely looking at fold change. Small fluctuations in background signal, due to technical variability, can sometimes override the signal intensity of an otherwise blank probe, resulting in classification as “differentially expressed” when in fact the transcript is absent altogether. Normalization tries to account for this, but not all normalization algorithms work equally well at dampening or removing variability in low-intensity signal. Hence, one must carefully consider both the absolute signal and detection call to reliably judge the significance and amplitude of log ratios when searching for differentially expressed probes.

This consideration has also been recognized by Affymetrix, as their new proprietary expression analysis algorithm, PLIER (Probe Logarithmic Intensity Error estimation), takes probe hybridization affinity information into account and uses a sophisticated error model to calculate background and nonspecific hybridization, accepting >10% more probes compared with the Statistical Algorithm (based on Rat 230 2.0 array) (18). Other related algorithms are continuously being developed to improve the overall sensitivity and specificity of mRNA detection, especially for lower-abundance transcripts that are often presumed to be of particular biological relevance.<sup>2</sup>

For quantitative multisample comparisons, microarray data is usually normalized to minimize technical variability, such as variation in manufacture and processing of the arrays, scanning (and scanner calibration), differences in detection efficiency between the fluorescent dyes, and systematic spatial biases in measured expression levels. Normalization also aims to minimize unwanted variability due to unequal quantities of starting RNA, variable sample preparation, differences in labeling and hybridization efficiencies, the time of day the experiment is performed, and even technician performing the experiment (20, 21). Ultimately, however, biological variation remains the major issue of concern.

The techniques used to normalize Affymetrix data differ in two major aspects: whether they use probe-level data or summarized expression intensity, and whether they perform normalization across the entire dataset or normalize each chip individually. For example Li and Wong’s model-based expression index (MBEI) measure uses probe-level data of the entire dataset, excluding outliers, to fit their statistical model, which is then used to calculate gene expression (64). This is a nonlinear normalization algorithm, which takes into account information specific to each probe.

Quantile normalization, used as the normalization method in the widely accepted RMA expression quantitation method (20, 21), uses probe-level data and does not require a baseline array. An RMA expression measure that uses sequence information (GCRMA) is a related, though more sophisticated,

expression quantitation method (34). The difference is that in GCRMA the background adjustment algorithm takes sequence information into account (such as GC content) and uses a stochastic model for binding affinities. For normalization the quantile normalization is used, and median-polish is used as a summary method, as is the case with RMA. This probe-specific background adjustment can also be used to improve the accuracy of other methods that summarize or normalize probe-level data (65, 66).

Because most microarray studies are ultimately aimed at predicting corresponding protein levels, in most scenarios it would probably be advantageous to measure global protein expression patterns directly. The proteome (or, possibly more correctly, the translome or population of expressed proteins), is conceptually analogous to the transcriptome (population of mRNA transcripts) and is defined as the entire set of proteins produced by a cell or tissue at any given time point (67). MS-based protein expression profiling represents an increasingly attractive, albeit still very challenging, approach for investigating the global biochemical properties of cells and tissues (38, 62). When combined with biochemical prefractionation and enrichment procedures, proteomic screening also offers the potential for extensive functional discovery, such as by determining both the subcellular location and possible interacting partners of proteins of special interest. Nevertheless, much remains to be accomplished in this regard with respect to endocrine cell biology.

Despite recent advances, large-scale proteomic measurements of endocrine cells and tissues represents a daunting experimental challenge. Protein abundance, subcellular localization, and turnover are highly dynamic, while biological patterns are dictated by overlapping developmental signals, physiological cues, environmental constraints, and disease perturbations. Classical gel-based proteomic screening techniques, such as two-dimensional PAGE, have proven to be ineffective for monitoring the proteome, particularly lower-abundance proteins, and are biased against the identification of small, basic, or membrane-bound proteins (39, 67). In contrast, gel-free shotgun protein profiling strategies, such as the one used in this study, allow for far more comprehensive characterization of complex biological samples (39, 41, 67). Since its introduction (39, 41, 67), the MudPIT technique has proven to be a very capable method (reviewed in Refs. 39, 40, 68). Our group routinely applies MudPIT to monitor the protein patterns of various organelles isolated from rodent tissues (*i.e.* mouse organs) and cultured mammalian cell lines with the objective of elucidating the biochemical properties associated with these various cell types (42, 57). Although the results of our pilot efforts to define the proteome of endocrine cell lines such as islet  $\alpha$ TC-1 are still quite limited (3, 5), based on the results reported here we would argue that proteomic and functional genomic approaches are highly complementary and indeed synergistic when combined. Although a broad correlation was observed between the mRNA and protein patterns, the corre-

spondence proved to be lower than might be expected.

Of course, assuming analytical error is not a cause, genes whose expressed cognate transcription and translation products do not correlate are of special interest because they may be indicative of protein regulation by posttranslational mechanisms, for instance by targeted protein degradation.<sup>6</sup> On the other hand, outliers whose protein expression was significantly higher than expected based on the corresponding mRNA levels (*i.e.* not detected) might include long-lived proteins and, as a class, might be expected to be less common than outliers whose mRNA expression was higher (due to incomplete proteomic sampling). This latter class may contain proteins subject to proteolytic regulation, or perhaps regulation by transport into or out of the nucleus, to another organelle or alternatively to the plasma membrane (insoluble membrane proteins were largely missed by the MudPIT profiling procedure used in this study). These subsets merit additional scrutiny because they may represent factors important in determining or regulating cell-type-specific biological functions.

The issue of proteomic detection coverage is an important consideration. Part of the current limitations associated with shotgun profiling stems from a failure to properly (and optimally) interpret the vast collection of acquired spectra, leading to both many false-negatives (missed identifications) and false-positives (incorrect identifications). To this end, several groups have been trying to develop appropriate computational and statistical tools and methods for evaluating, validating, and mining large-scale protein expression datasets (45, 57, 58, 69). We have argued here that prior knowledge of gene expression patterns can, in fact, be used as supporting confirmatory evidence in favor of tentative (marginal preliminary scoring) protein identifications. Based on our preliminary findings, we propose, as a second heuristic rule (albeit one still requiring definitive formal proof), that any tentative protein gene product deemed to be expressed by microarray analysis and having a high confidence score ( $\geq 85\%$  likelihood by STATQUEST or a similarly derived statistical measure) should be accepted as correctly identified regardless of the total number of supporting spectra.

Data integration is also complementary to alternate biochemical methods aiming for sample simplification or enrichment to improve proteomic detection, such as subcellular fractionation and/or nondenaturing conventional chromatography, which have proven to be fruitful for extending detection limits as well as providing a more biologically informative context for interpreting profiles (37, 42, 57). Of course, the introduction of new generations of high-performance instruments with much greater performance in terms of sensitivity, dynamic range, and scan speeds, will also surely help to surmount the under-sampling problem.

<sup>6</sup> Another important form of regulation, reversible posttranslational modification with phosphate, acetyl, glycosyl, or lipid groups, for example, was not addressed in this study.

*Acknowledgments*—We would like to thank Thomas Kislinger for technical assistance, Alexandr Ignatchenko and Pingzhao Hu for advice with computing, and both Grace Flock and Xiemin Cao for help with the cell culture and RT-PCR.

\* This study was supported by grants from National Science and Engineering Research Council of Canada, Genome Canada, the Ontario Genome Institute, the McLaughlin Centre for Molecular Medicine, and the Ontario Research and Development Challenge Fund (to A. E.), and by an operating grant from the Canadian Institute of Health Research (to D. J. D.). D. J. D. is supported by a Canada Research Chair in Regulatory Peptides.

§ The on-line version of this manuscript (available at <http://www.mcponline.org>) contains supplemental material.

¶ M. M. and C. C. contributed equally to this study.

‡‡ To whom correspondence should be addressed: CH Best Institute, Room 402, 112 College Street, Toronto, Ontario, Canada M5G 1L6. Tel.: 416-946-7281; Fax: 416-978-8528; E-mail: [andrew.emili@utoronto.ca](mailto:andrew.emili@utoronto.ca).

#### REFERENCES

- Cardozo, A. K., Berthou, L., Kruhoffer, M., Orntoft, T., Nicolls, M. R., and Eizirik, D. L. (2003) Gene microarray study corroborates proteomic findings in rodent islet cells. *J. Proteome Res.* **2**, 553–555
- Hui, H., Wang, C., Li, H., Bulotta, A., D'Amico, E., Khoury, N., Nguyen, E., Di Mario, U., Chen, I. Y., and Perfetti, R. (2004) Gene expression profiling of cultured human islet preparations. *Diabetes Technol. Ther.* **6**, 481–492
- Mizusawa, N., Hasegawa, T., Ohigashi, I., Tanaka-Kosugi, C., Harada, N., Itakura, M., and Yoshimoto, K. (2004) Differentiation phenotypes of pancreatic islet  $\beta$ - and  $\alpha$ -cells are closely related with homeotic genes and a group of differentially expressed genes. *Gene* **331**, 53–63
- Shalev, A., Pise-Masison, C. A., Radonovich, M., Hoffmann, S. C., Hirshberg, B., Brady, J. N., and Harlan, D. M. (2002) Oligonucleotide microarray analysis of intact human pancreatic islets: Identification of glucose-responsive genes and a highly regulated TGF $\beta$  signaling pathway. *Endocrinology* **143**, 3695–3698
- Wang, J., Webb, G., Cao, Y., and Steiner, D. F. (2003) Contrasting patterns of expression of transcription factors in pancreatic alpha and beta cells. *Proc. Natl. Acad. Sci. U. S. A.* **100**, 12660–12665
- Flock, G., Cao, X., and Drucker, D. J. (2005) Pdx-1 is not sufficient for repression of proglucagon gene transcription in islet or enteroendocrine cells. *Endocrinology* **146**, 441–449
- Flock, G., and Drucker, D. J. (2002) Pax-2 activates the proglucagon gene promoter but is not essential for proglucagon gene expression or development of proglucagon-producing cell lineages in the murine pancreas or intestine. *Mol. Endocrinol.* **16**, 2349–2359
- Nian, M., Drucker, D. J., and Irwin, D. (1999) Divergent regulation of human and rat proglucagon gene promoters *in vivo*. *Am. J. Physiol.* **277**, G829–G837
- Laser, B., Meda, P., Constant, I., and Philippe, J. (1996) The caudal-related homeodomain protein Cdx-2/3 regulates glucagon gene expression in islet cells. *J. Biol. Chem.* **271**, 28984–28994
- Ritz-Laser, B., Estreicher, A., Gauthier, B., and Philippe, J. (2000) The paired homeodomain transcription factor Pax-2 is expressed in the endocrine pancreas and transactivates the glucagon gene promoter. *J. Biol. Chem.* **275**, 32708–32715
- Powers, A. C., Efrat, S., Mojsos, S., Spector, D., Habener, J. F., and Hanahan, D. (1990) Proglucagon processing similar to normal islets in pancreatic  $\alpha$ -like cell line derived from transgenic mouse tumor. *Diabetes* **39**, 406–414
- Brown, P. O., and Botstein, D. (1999) Exploring the new world of the genome with DNA microarrays. *Nat. Genetics* **21**, (suppl.) 33–37
- Schena, M., Shalon, D., Davis, R. W., and Brown, P. O. (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270**, 467–470
- Southern, E., Mir, K., and Shchepinov, M. (1999) Molecular interactions on microarrays. *Nat. Genet.* **21**, 5–9
- Lipshutz, R. J., Fodor, S. P., Gingeras, T. R., and Lockhart, D. J. (1999) High density synthetic oligonucleotide arrays. *Nat. Genet.* **21**, 20–24

16. Hughes, T. R., Mao, M., Jones, A. R., Burchard, J., Marton, M. J., Shannon, K. W., Lefkowitz, S. M., Ziman, M., Scheelter, J. M., Meyer, M. R., Kobayashi, S., Davis, C., Dai, H., He, Y. D., Stephanian, S. B., Cavet, G., Walker, W. L., West, A., Coffey, E., Shoemaker, D. D., Stoughton, R., Blanchard, A. P., Friend, S. H., and Linsley, P. S. (2001) Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. *Nat. Biotechnol.* **19**, 342–347
17. Epstein, J. R., and Walt, D. R. (2003) Fluorescence-based fibre optic arrays: A universal platform for sensing. *Chem. Soc. Rev.* **32**, 203–214
18. Affymetrix (2004) *GeneChip Expression Platform: Comparison, Evolution, and Performance*, Affymetrix, Santa Clara, CA
19. Affymetrix (2001) *GeneChip Arrays Provide Optimal Sensitivity and Specificity for Microarray Expression Analysis*, Affymetrix, Santa Clara, CA
20. Irizarry, R. A., Bolstad, B. M., Collin, F., Cope, L. M., Hobbs, B., and Speed, T. P. (2003) Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res.* **31**, e15
21. Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, Y. D., Antonellis, K. J., Scherf, U., and Speed, T. P. (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Bio-statistics* **4**, 249–264
22. Cope, L. M., Irizarry, R. A., Jaffee, H. A., Wu, Z., and Speed, T. P. (2004) A benchmark for Affymetrix GeneChip expression measures. *Bioinformatics* **20**, 323–331
23. Stein, L. D. (2003) Integrating biological databases. *Nat. Rev. Genet.* **4**, 337–345
24. Quackenbush, J. (2002) Microarray data normalization and transformation. *Nat. Genet.* **32**, (suppl.) 496–501
25. Cronin, M., Ghosh, K., Sistare, F., Quackenbush, J., Vilker, V., and O'Connell, C. (2004) Universal RNA reference materials for gene expression. *Clin. Chem.* **50**, 1464–1471
26. Hartemink, A. J., Gifford, D. K., Jaakkola, T. S., and Young, R. A. (2001) Maximum-likelihood estimation of optimal scaling factors for expression array normalization. *Proc. SPIE, Microarrays: Optical Technologies and Informatics* **4266**, 132–140
27. Holloway, A. J., van Laar, R. K., Tothill, R. W., and Bowtell, D. D. (2002) Options available—from start to finish—for obtaining data from DNA microarrays II. *Nat. Genet.* **32**, (suppl.) 481–489
28. Chuaqui, R. F., Bonner, R. F., Best, C. J., Gillespie, J. W., Flaig, M. J., Hewitt, S. M., Phillips, J. L., Krizman, D. B., Tangrea, M. A., Ahram, M., Linehan, W. M., Knezevic, V., and Emmert-Buck, M. R. (2002) Post-analysis follow-up and validation of microarray experiments. *Nat. Genet.* **32**, (suppl.) 509–514
29. Curk, T., Demsar, J., Xu, Q., Leban, G., Petrovic, U., Bratko, I., Shaulsky, G., and Zupan, B. (2005) Microarray data mining with visual programming. *Bioinformatics* **21**, 396–398
30. Frank, E., Hall, M., Trigg, L., Holmes, G., and Witten, I. H. (2004) Data mining in bioinformatics using Weka. *Bioinformatics* **20**, 2479–2481
31. Wang, J., Myklebost, O., and Hovig, E. (2003) MGraph: Graphical models for microarray data analysis. *Bioinformatics* **19**, 2210–2211
32. Huber, W., von Heydebreck, A., Sultmann, H., Poustka, A., and Vingron, M. (2002) Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics* **18**, (Suppl. 1) S96–S104
33. Li, C., and Wong, W. H. (2001) Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. *Proc. Natl. Acad. Sci. U. S. A.* **98**, 31–36
34. Wu, Z., Irizarry, R. A., Gentleman, R., Murillo, F. M., and Spencer, F. (2003) *A Model Based Background Adjustment for Oligonucleotide Expression Arrays*, Johns Hopkins University, Department of Biostatistics Working Papers, Baltimore, MD
35. Dudoit, S., Gentleman, R. C., and Quackenbush, J. (2003) Open source software for the analysis of microarray data. *BioTechniques*, (suppl.) 45–51
36. Quackenbush, J. (2001) Computational analysis of microarray data. *Nat. Rev. Genet.* **2**, 418–427
37. Mootha, V. K., Bunkenborg, J., Olsen, J. V., Hjerrild, M., Wisniewski, J. R., Stahl, E., Bolouri, M. S., Ray, H. N., Sihag, S., Kamal, M., Patterson, N., Lander, E. S., and Mann, M. (2003) Integrated analysis of protein composition, tissue diversity, and gene regulation in mouse mitochondria. *Cell* **115**, 629–640
38. Aebersold, R., and Mann, M. (2003) Mass spectrometry-based proteomics. *Nature* **422**, 198–207
39. Kislinger, T., and Emili, A. (2003) Going global: Protein expression profiling using shotgun mass spectrometry. *Curr. Opin. Mol. Ther.* **5**, 285–293
40. Washburn, M. P., Ulaszek, R., Decui, C., Schieltz, D. M., and Yates, J. R., 3rd. (2002) Analysis of quantitative proteomic data generated via multi-dimensional protein identification technology. *Anal. Chem.* **74**, 1650–1657
41. Washburn, M. P., Wolters, D., and Yates, J. R., 3rd (2001) Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat. Biotechnol.* **19**, 242–247
42. Pan, Y., Kislinger, T., Gramolini, A. O., Zvaritch, E., Kranias, E. G., MacLennan, D. H., and Emili, A. (2004) Identification of biochemical adaptations in hyper- or hypocontractile hearts from phospholamban mutant mice by expression proteomics. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 2241–2246
43. Eng, J. K., McCormack, A. L., and Yates, J. R. I. (1994) An approach to correlate tandem mass-spectral data of peptides with amino-acid-sequences in a protein database. *J. Am. Soc. Mass Spectrom.* **11**, 976–989
44. Tabb, D. L., McDonald, W. H., and Yates, J. R., 3rd. (2002) DTASelect and Contrast: Tools for assembling and comparing protein identifications from shotgun proteomics. *J. Proteome Res.* **1**, 21–26
45. Liu, H., Sadygov, R. G., and Yates, J. R., 3rd (2004) A model for random sampling and estimation of relative protein abundance in shotgun proteomics. *Anal. Chem.* **76**, 4193–4201
46. Robinson, M. D., Grigull, J., Mohammad, N., and Hughes, T. R. (2002) FunSpec: A web-based cluster interpreter for yeast. *BMC Bioinformatics* **3**, 35
47. Affymetrix (2001) *GeneChip Muring Genome U74Av2 Set*, Affymetrix, Santa Clara, CA
48. Affymetrix (2004) *GeneChip Expression Analysis Data Analysis Fundamentals*, Affymetrix, Santa Clara, CA
49. Affymetrix (2002) *Statistical Algorithms Description Document*, Affymetrix, Santa Clara, CA
50. Bolstad, B. M., Irizarry, R. A., Astrand, M., and Speed, T. P. (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* **19**, 185–193
51. Affymetrix (2001) *New Statistical Algorithms for Monitoring Gene Expression on GeneChip Probe Arrays*, Affymetrix, Santa Clara, CA
52. Affymetrix (2001) *Fine Tuning Your Data Analysis: Tunable Parameters of the Affymetrix Expression Analysis Statistical Algorithms*, Affymetrix, Santa Clara, CA
53. Affymetrix (2001) *Statistical Algorithms Reference Guide*, Affymetrix, Santa Clara, CA
54. Seo, J., Bakay, M., Chen, Y. W., Hilmer, S., Shneiderman, B., and Hoffman, E. P. (2004) Interactively optimizing signal-to-noise ratios in expression profiling: Project-specific algorithm selection and detection  $p$ -value weighting in Affymetrix microarrays. *Bioinformatics* **20**, 2534–2544
55. Link, A. J., Eng, J., Schieltz, D. M., Carmack, E., Mize, G. J., Morris, D. R., Garvik, B. M., and Yates, J. R., 3rd. (1999) Direct analysis of protein complexes using mass spectrometry. *Nat. Biotechnol.* **17**, 676–682
56. Steen, H., and Mann, M. (2004) The ABC's (and XYZ's) of peptide sequencing. *Nat. Rev. Mol. Cell. Biol.* **5**, 699–711
57. Kislinger, T., Rahman, K., Radulovic, D., Cox, B., Rossant, J., and Emili, A. (2003) PRISM, a generic large-scale proteomic investigation strategy for mammals. *Mol. Cell. Proteomics* **2**, 96–106
58. Sadygov, R. G., Liu, H., and Yates, J. R. (2004) Statistical models for protein validation using tandem mass spectral data and protein amino acid sequence databases. *Anal. Chem.* **76**, 1664–1671
59. Han, J., and Kamber, M. (2000) Data mining: Concepts and techniques. *The Morgan Kaufmann Series in Data Management Systems*, Morgan Kaufmann Publishers, San Francisco, CA
60. Tabb, D. L., MacCoss, M. J., Wu, C. C., Anderson, S. D., and Yates, J. R., 3rd. (2003) Similarity among tandem mass spectra from proteomic experiments: Detection, significance, and utility. *Anal. Chem.* **75**, 2470–2477
61. Durr, E., Yu, J., Krasinska, K. M., Carver, L. A., Yates, J. R., Testa, J. E., Oh, P., and Schnitzer, J. E. (2004) Direct proteomic mapping of the lung microvascular endothelial cell surface in vivo and in cell culture. *Nat. Biotechnol.* **22**, 985–992
62. Aebersold, R. (2003) Quantitative proteome analysis: Methods and applications. *J. Infect. Dis.* **187**, (Suppl. 2) S315–S320



63. Pomeroy, S. L., Tamayo, P., Gaasenbeek, M., Sturla, L. M., Angelo, M., McLaughlin, M. E., Kim, J. Y., Goumnerova, L. C., Black, P. M., Lau, C., Allen, J. C., Zagzag, D., Olson, J. M., Curran, T., Wetmore, C., Biegel, J. A., Poggio, T., Mukherjee, S., Rifkin, R., Califano, A., Stolovitzky, G., Louis, D. N., Mesirov, J. P., Lander, E. S., and Golub, T. R. (2002) Prediction of central nervous system embryonal tumour outcome based on gene expression. *Nature* **415**, 436–442
64. Li, C., and Wong, W. H. (2001) Model-based analysis of oligonucleotide arrays: Model validation, design issues and standard error application. *Genome Biol.* **2**, RESEARCH0032
65. Shedden, K., Chen, W., Kuick, R., Ghosh, D., Macdonald, J., Cho, K. R., Giordano, T. J., Gruber, S. B., Fearon, E. R., Taylor, J. M., and Hanash, S. (2005) Comparison of seven methods for producing Affymetrix expression scores based on false discovery rates in disease profiling data. *BMC Bioinformatics* **6**, 26
66. Wu, Z., and Irizarry, R. A. (2004) Preprocessing of oligonucleotide array data. *Nat. Biotechnol.* **22**, 656–658
67. Aebersold, R., and Cravatt, B. F. (2002) Proteomics—Advances, applications and the challenges that remain. *Trends Biotechnol.* **20**, S1–S2
68. Washburn, M. P., Ulaszek, R. R., and Yates, J. R., 3rd. (2003) Reproducibility of quantitative proteomic analyses of complex biological mixtures by multidimensional protein identification technology. *Anal. Chem.* **75**, 5054–5061
69. Colinge, J., Chiappe, D., Lagache, S., Moniatte, M., and Bougueleret, L. (2005) Differential proteomics via probabilistic peptide identification scores. *Anal. Chem.* **77**, 596–606
70. Su, A. I., Cooke, M. P., Ching, K. A., Hakak, Y., Walker, J. R., Wiltshire, T., Orth, A. P., Vega, R. G., Sapinoso, L. M., Moqrich, A., Patapoutian, A., Hampton, G. M., Schultz, P. G., and Hogenesch, J. B. (2002) Large-scale analysis of the human and mouse transcriptomes. *Proc. Natl. Acad. Sci. U. S. A.* **99**, 4465–4470