New Data Analysis and Mining Approaches Identify Unique Proteome and Transcriptome Markers of Susceptibility to Autoimmune Diabetes*

Ivan C. Gerling†‡§¶, Sudhir Singh§, Nataliya I. Lenchik‡, Dana R. Marshall¶, and Jian Wu‡

Non-obese diabetic (NOD) mice spontaneously develop autoimmunity to the insulin producing beta cells leading to insulin-dependent diabetes. In this study we developed and used new data analysis and mining approaches on combined proteome and transcriptome (molecular phenotype) data to define pathways affected by abnormalities in peripheral leukocytes of young NOD female mice. Cells were collected before mice show signs of autoimmunity (age, 2–4 weeks). We extracted both protein and RNA from NOD and C57BL/6 control mice to conduct both proteome analysis by two-dimensional gel electrophoresis and transcriptome analysis on Affymetrix expression arrays. We developed a new approach to analyze the two-dimensional gel proteome data that included two-way analysis of variance, cluster analysis, and principal component analysis. Lists of differentially expressed proteins and transcripts were subjected to pathway analysis using a commercial service. From the list of 24 proteins differentially expressed between strains we identified two highly significant and interconnected networks centered around oncogenes (Myc and Mycn) and apoptosis-related genes (Bcl2 and Casp3). The 273 genes with significant strain differences in RNA expression levels created six interconnected networks with a significant over-representation of genes related to cancer, cell cycle, and cell death. They contained many of the same genes found in the proteome networks (including Myc and Mycn). The combination of the eight, highly significant networks created one large network of 272 genes of which 82 had differential expression between strains either at the protein or the RNA level. We conclude that new proteome data analysis strategies and combined information from proteome and transcriptome can enhance the insights gained from either type of data alone. The overall systems biology of prediabetic NOD mice points toward abnormalities in regulation of the opposing processes of cell renewal and cell death even before there are any clear signatures of immune system activation. Molecular & Cellular Proteomics 5:293–305, 2006.

The effort to sequence mammalian genomes has spurred a rapid development of research tools that allow comprehensive evaluations of molecular phenotypes to study systems biology rather than just focusing on single molecules or pathways. Microarrays allow comprehensive characterization of transcriptomes, and 2D1 gel or gel-free proteome technology allows evaluation of expression of thousands of proteins in a single procedure. It is now possible to determine to what extent genetic or external manipulation of a biological system alters the expression of any of thousands of genes and proteins. In contrast to the traditional approach, the hypothesis that explains the connections observed is not created until after data have been collected. Although these methods are not always successful given the inevitable limitations and pitfalls inherent to all technology, they have certainly created exciting new insights in many fields. The ability of this discovery approach to make completely unpredictable and novel discoveries by application of a systematic process has been authenticated by results from many investigators (1–4).

Microarray technology currently offers the fastest and most comprehensive molecular evaluations. The estimated total number of genes in humans is 20–35,000, suggesting that currently available gene chips cover a very large proportion of all expressed genes (5). However, physiology is based on the level of proteins and their activation or deactivation by post-translational modifications such as glycosylation, acetylation, phosphorylation, myristoylation, or proteolytic cleavage. Changes in these parameters are not revealed by measuring mRNA levels. Use of protein-based assays is the obvious and easiest way to ensure that these biologically important changes are detected. Indeed the combined information from gene- and protein-based assays can complement each other

---

The abbreviations used are: 2D, two-dimensional; NOD, non-obese diabetic; PCA, principal component analysis; MTC, multiple test correction; ANOVA, analysis of variance; SSP, standard spot.
and produce more complete information than can be gained from either alone. Unfortunately this combined approach is infrequently done and hardly ever presented.

2D PAGE is a widely used technique for the separation of proteins and comparative proteomics. This method separates proteins based on two molecular properties. The first dimension separates proteins based on their pI. The second dimension separates proteins based on their size. After staining, each protein is found in a spot at coordinates that represent its unique size and pI (6). Depending on the sample size, composition, and gel parameters, several hundred up to 10 thousand individual protein spots can be detected on a gel (7).

In comparative proteomic studies, the 2D PAGE images are analyzed to compare the protein spot patterns and identify differences between samples or groups of samples. Computer programs have been developed to aid the detection of protein features (8–10). Matching algorithms in these programs ensure that comparable protein spots on different gels are correctly matched and obtain quantitative measurements for each spot based on pixel intensity and volumes. The large amount of data generated from each sample requires the use of appropriate statistical methods to extract the information of interest. Traditional analysis of proteome data has consisted primarily of univariate analysis, for example, of treated versus control populations using Student's t test. Yet if the changes in the expression of several proteins combined creates a consistent pattern of variation that is related to a specific biological mechanism, a univariate approach may not be sufficient. Fortunately DNA microarray analysis programs were designed specifically to handle large datasets generated from biological systems. They have built-in functions with good graphical interfaces and do not require the user to have programming skills. Clustering algorithms and PCA in these programs identify patterns of expression that may suggest correlation. In this study we describe an approach to export 2D gel proteome data from an image analysis program, import it into a gene analysis program, customize files, normalize datasets, and conduct basic statistical analyses as well as clustering algorithms and PCA on the proteome data. This approach produced substantial additional information and insights beyond what the simple t test in the image analysis software could produce.

Many investigators are characterizing both transcriptomes and proteomes of their samples to gain a more complete molecular phenotype. We have previously described our protocols to isolate both proteins and RNA from the same samples to be processed for either proteome or transcriptome analysis (11, 12). By conducting analysis at both the protein and RNA expression levels in a process we call “molecular phenotyping,” the two types of data can be used to confirm and supplement insights drawn from the other.

In this study we describe for the first time the application of molecular phenotyping to gain new insights into the development of autoimmune diabetes in mice. The NOD mouse is a model of autoimmune (type 1) diabetes mellitus. The insulin producing beta cells in the islet of Langerhans of NOD mice become the targets of a spontaneously developing autoimmune attack at around 5 weeks of age. The accumulation of leukocytes around the islets becomes increasingly more invasive and destructive until eventually the total beta cell mass has been reduced to a level where the animals cannot produce enough insulin to maintain normal blood glucose levels (13–15). The prepathological stages before 5 weeks of age when leukocytes infiltrate the islet target tissue are not well understood (16, 17). We hypothesized that the aggressive phenotype of leukocytes manifested by the invasion of islets at 5 weeks of age may be developing in part because of underlying defects manifested at the molecular level much earlier. To characterize these defects we collected leukocytes at both 2 and 4 weeks of age. Systems and pathways appearing abnormal at both of these ages would represent “basic defects” driven much more by underlying genetics than by the developing pathology. For the comparison we decided to use the C57BL/6 mouse strain that has been extensively used for immunology research and has shown no indication of being prone to development of autoimmune diseases or of suffering from any other major immune response abnormalities. The design of studying two strains at two ages allows us to focus on consistent basic defects and avoid artifacts of unique short lived events at a specific age. The use of two-way ANOVA on these data still allows us to reveal strain differences even if there is an underlying shift in expression due to age/development. The proteome and transcriptome data were analyzed to identify both proteins and genes that were differentially expressed between the two strains. The resulting lists of proteins and genes were analyzed for the presence of members of known molecular networks. These networks showed substantial overlap and connectivity with the common biological “theme” of cell proliferation and cell death.

EXPERIMENTAL PROCEDURES

Sample Preparation and 2D PAGE Analysis of Proteomes—Proteins and RNA were extracted from splenic leukocytes of 2- and 4-week-old NOD and C57BL/6 mice using a trisreagent (n = 5 for each strain and each age). Precipitated proteins were separated by 2D PAGE using a pI range 4–7 strip and stained with SYPRO Ruby stain. The specific methods have been described previously (11, 12, 18, 19). The gels were scanned, and TIFF images were imported into PDQuest image analysis software (version 7.1, Bio-Rad) where the protein spots were detected using the automated spot detection feature. The images were visually inspected at high resolution and corrected for undetected or incorrectly detected spots/artifacts. We used a stringent definition of spots to avoid collecting data from very faint spots because in our experience data from those spots are subject to substantial intersample variability and often suffer from lack of reproducibility. Although this approach leaves us with a relatively low number of protein spots, it improves the overall quality of the dataset. Upon completion of the image analysis, data were stored both as a 2D image and as a Gaussian image.

Image Analysis of Gels—The dataset consists of 20 samples total
that can be divided into four groups of five if both strain (NOD and C57BL/6) and age (2 and 4 weeks) are considered. These groups can be combined to two groups of 10 (either by age or by strain) to allow data analysis by Student’s t test. Within the “MatchSet” produced by the image analysis software, 251 spots were identified as being expressed in at least one of the 20 gels. We used the default normalization of data in the MatchSet that is based on total pixel quantity in valid spots. This normalization method assumes that few protein spots change within the experiment and that the changes average out across the whole gel. PDQuest software includes three statistical tests for data analysis, the Student’s t test (a parametric test) and the Mann-Whitney rank sum test and Wilcoxon signed rank test (both non-parametric tests). The Student’s t test was performed, comparing protein expression in the two mouse strains \( (n = 10, p < 0.05) \).

**Export of Data and Preparation of Formatted Expression Files**—The intensity values that are generated for each spot along with the spot identification number can be easily exported to a Microsoft Excel spreadsheet. After the MatchSet was loaded, “export MatchSet” from the export submenu (in the file menu) was selected. A new dialogue window opened with options to either export data for each gel in a MatchSet (Spot Data by Gel) or for each replicate group in the MatchSet (Spot Data by Group). The “spot data by gel” option was selected, and then the data for each gel/spot to be exported were selected. For this analysis all but the “normalized quantity” was deselected because that was the only information to be analyzed. The SSP number for each spot was included, and the file was exported. The file type for the exported data automatically defaults to an .xls spreadsheet that can be opened with Microsoft Excel or other spreadsheet software. From within the spreadsheet software, spot IDs were ordered alphanumerically as rows, and spot-normalized intensities were ordered alphanumerically as columns. For each gel there is a value given if a spot is not found on that gel (the background value). In the 20 gels of our dataset this value was between 4 and 119. Although this is much smaller than the values of even faint spots (usually above 1000), on a log scale it can represent a substantial difference, which in log-based graphic presentations will appear as differences in “spot intensities” (although there are no spots). To avoid misrepresentation, all background values were increased to 119. Although this was expected to reduce the number of differentially expressed spots marginally, in our dataset the effect was so small that we got identical lists of differentially expressed spots with or without the background value correction. The final step was to save the spreadsheet as a tab-delimited text file (.txt).

**Analysis of Proteome Data in Gene Expression Analysis Software**—We have used a software called GeneSpring (version 6.2, Silicon Genetics, Redwood, CA) to analyze our expression array data and therefore used that program in the current protein expression analysis. Similar analyses can be done using other commercially available or free gene expression analysis software (20). The software provides option to use Student’s and Welch’s t tests as well as one- and two-way ANOVA for data analysis. Those tests allow user-defined parameters and provide selection of a number of posthoc tests to correct for multiple testing. There is also a variety of clustering algorithms including K-means and hierarchical clustering, again with user-defined parameters. Principal component analysis may be done on “genes” (the term assigned to proteins in this analysis) or on samples. We used all of these tests as described under “Results.”

The MatchSet expression levels spreadsheet (in tab-delimited ASCII file format) was imported into GeneSpring using “import data” from the file menu. The proteome file is classified as a “custom array,” therefore one must create a template to identify the columns in the spreadsheet to be used for analysis. This option requires the user to “create a custom genome.” Therefore, File > Import Data > Create New Genome > Next was selected. For this analysis the “spot ID” is the “gene identifier,” and the “normalized spot intensity” is the “signal.” The standard procedure for “creating an experiment” was followed (in this software that refers to grouping all of the samples to be analyzed and defining them as an “experiment”). The next step involves defining normalization and sample groupings for this dataset. There are many options for normalization. In this experiment we only selected “per gene normalization” because this is more appropriate for a small number of data points. The next step is to define the experiment parameters for each sample, in this case age and strain. The final step was to select those parameters that will be considered in those analyses that require groups to be defined. In this experiment the mice can be classified by strain without regard to age (NOD or C57BL/6, \( n = 10 \), by age without regard to strain (2 or 4 weeks, \( n = 10 \)), or by age and strain (2- or 4-week-old NOD, 2- or 4-week-old C57BL/6, \( n = 5 \)).

**Spot Identification**—Protein spots that were found to have interesting patterns of differential expression were subjected to identification of protein(s) present in the spot using peptide mass fingerprinting as described previously (11).

**Transcriptome Analysis**—The RNA isolated from the trireagent extract of the cells was subjected to gene expression analysis on Affymetrix MOE430A and MOE430B expression arrays using procedures described previously (12). These expression arrays each contain sequences from 22,000 probe sets (genes) whose expression levels are evaluated using the MAS 5.0 software (for details see www.affymetrix.com). The expression data produced by the MAS 5.0 software were imported into GeneSpring software and analyzed. In GeneSpring we conducted two-way ANOVA using the Westfall and Young multiple test correction on the 26,530 probe sets in which at least one of the 20 samples had produced a present or marginal call by the MAS 5.0 software.

**Analysis of Molecular Networks**—To investigate whether lists of differentially expressed proteins belong to specific pathways, we conducted “Ingenuity Pathways Analysis” on a subscription, web-delivered application that enables biologists to discover, visualize, and explore networks relevant to their experimental results such as gene expression array datasets. We used this service for both our lists of proteins and RNA expression data. For a detailed description of Ingenuity Pathways Analysis visit www.ingenuity.com.

The files containing gene identifiers (Swiss-Prot identifiers for protein lists and Affymetrix probe set identifiers for gene lists) were uploaded as a tab-delimited text file. Each gene identifier was mapped to its corresponding gene object in the Ingenuity Pathways Knowledge Base. The genes on the uploaded lists, called Focus Genes, were then used as the starting point for generating biological networks. To start building networks, the application queries the Ingenuity Pathways Knowledge Base for interactions between Focus Genes and all other gene objects stored in the knowledge base and generates a set of networks with a network size of no more than 35 genes/proteins. Ingenuity Pathways Analysis then computes a score for each network according to the fit of the user’s set of significant genes. The score is derived from a p value and indicates the likelihood of the Focus Genes in a network being found together due to random chance. A score of 2 indicates that there is a 1 in 100 chance that the Focus Genes are together in a network due to random chance. Therefore, scores of 2 or higher have at least a 99% confidence of not being generated by random chance alone.

The networks are displayed graphically as nodes (genes/gene products) and edges, or lines (the biological relationships between the nodes). The node symbols used by Ingenuity are derived from the symbol of the gene encoding the corresponding protein/mRNA. Human, mouse, and rat orthologs of a gene, although stored as separate objects in the knowledge base, are represented as a single node in the network. Nodes are displayed using various shapes that represent
the functional class of the gene product. Lines with arrows indicate induction, whereas other lines indicate different types of relationships such as phosphorylation, enzymatic action, etc.

Biological functions were assigned to each gene network by using the findings that have been extracted from the scientific literature and stored in the Ingenuity Pathways Knowledge Base. The biological functions assigned to each network are ranked according to the significance of that biological function to the network. A Fischer’s exact test is used to calculate a p value determining the probability that the biological function assigned to that network is explained by chance alone. We listed only the three most significant biological functions for each network.

RESULTS

t Test and ANOVA—Fig. 1 show examples of the 2D gel images collected from leukocyte samples. All 20 images were aligned and analyzed using PDQuest software as described previously (11, 18). This analysis collects spot intensity values for each spot on each gel and assigns the same identity (SSP number) to spots on different gels if they by their placement (x, y coordinates) on the gels are deemed to contain the same protein(s). To identify 2D gel spot intensity differences between two groups, the t test (assuming equal variances and not selecting log transformation of the data) is in our opinion an appropriate statistical test (and it is the default setting for the image analysis software). This test is available on both the image and gene analysis software, and the same seven spots (listed previously) were identified by both programs as different between the two strains at the p < 0.05 level. For easy visualization of expression patterns, the seven spots that were identified by t test as being differentially expressed between the two strains of mice were grouped using a hierarchical clustering algorithm and displayed as a heat map of a “gene tree” (Fig. 2). This organization of data groups the spots according to how similar they are in patterns of expression. For example the two main branches of the gene tree represent spots that are more intense in NOD mice (top five spots) or less intense in NOD mice (last two spots). The spots that are paired at the lowest level have the most similar expression levels and are more likely to be co-regulated proteins (e.g. 2007 and 3506 or 1508 and 3503).

Multiple test corrections (MTCs) can be helpful because they address the consistency of the replicates and therefore the “tightness” of the dataset. As such they can reduce the number of type 1 errors (false positives). They are also somewhat flawed in their use for the analysis of transcriptomes and proteomes because almost all of them assume that the signal value for each spot is independent of all others in that sample (e.g. like sequential flips of a coin), yet this is certainly not true for gene expression data because biological pathways are highly interrelated and interdependent. Changes in intensity of one protein spot is not independent of the changes in another particularly when the phosphorylated and non-phosphorylated version of a protein create separate spots. The choice of whether to use MTCs depends on how important it is to avoid type 1 errors and whether the results will be confirmed by other independent experiments. The gene analysis software allows for the selection of different types of MTCs. We did not use MTCs on our proteome data because the data will be subjected to pathway analysis and compared with the results from transcriptome analysis as a downstream approach to reduce the impact of false positives on the final conclusions drawn from the data.

In contrast to the image analysis program, the gene expression analysis program allowed us to conduct a two-way ANOVA. This analysis was conducted on the spot intensity data to evaluate age effects, strain effects, and interaction between the two parameters. In that analysis we found that eight spots had significant (p < 0.05) strain effects, four spots had age effects, and 12 spots showed significant interaction between the two parameters. Three spots belonged to all
three of those groups (had significant strain, age, and interaction effects). The remaining spots had only one of the three types of effect (Fig. 3). As expected, although there is some overlap between the lists of proteins showing strain differences by ANOVA and by Student’s t test, there are also differences.

**K-means Clustering**—The gene analysis program provides clustering algorithms such as K-means, self-organizing maps, and hierarchical clustering. Clustering is a multivariate analysis technique that seeks to organize information about variables so that groups of similarly expressed genes can be formed (21). Used on 2D gel proteome data, this approach may reveal otherwise unnoticed correlations between certain spots via their expression patterns. This type of analysis does not include statistical significance. Therefore, the groups following a specific pattern may contain spots for which the magnitude of that pattern is not statistically significant. K-means clustering requires the user to define the number of groups their whole dataset (all 251 spots) will be divided into. Generally the first “gene pattern” entered into each set is determined randomly, and a centroid is calculated. Genes continue to be added to the sets with centroids continuously recalculated and genes continuously moved to different sets for the number of iterations requested. Ultimately each set contains the genes closest to the calculated centroid for that set (22, 23). This does not necessarily mean that all genes in a set will have the same pattern across samples, but if one has strong group-defining patterns of expression, these will be obvious in some of the sets. There are no clear rules for the number of sets to begin with, and it is advisable to try a range of set numbers. After trying a range of set numbers from 5 to 25 we selected 14 sets, 100 iterations, and five additional random starts, resulting in the sets displayed in Fig. 4. Of the five spots that the previous Student’s t test analysis had identified to be more intense in NOD mice, four were found in one K-means cluster together with 11 other spots (following a similar expression pattern). Because those proteins show an expression pattern similar to those proteins selected by the t test, we consider them worthy of further investigation as a group of co-regulated proteins with potential expression differences between strains.

**Principal Component Analysis**—PCA is often used to reduce the size of datasets of high dimensionality by identifying genes that contribute most to the variation in the dataset and plotting them in a multidimensional format (24–26). In the case of this experiment, the set consists of 251 measure-
ments for 20 samples for a total of 5,020 individual values. The typical genomic microarray experiment has up to 400,000 values or more. Rather than attempt to work with these types of numbers in two dimensions, the contribution to the variation across the sample set is calculated for each gene, and it is assigned to a group, or component, that is represented in multidimensional space as a vector of specific direction and length (26). By convention, there are the same numbers of groups as there are samples, but only a completely random dataset would use all possible vectors equally. The greater the

Fig. 4. K-means clustering of 251 protein spots into 14 different groups using change correlation. One set marked by a box (Set 7) consists of 15 proteins of which four were also selected by the Student’s t test as having strain differences with higher expression levels in NOD mice. For each cluster the relative expression levels of proteins normalized to a mean of 1 is given on a log scale on the y axis. The x axis represents each of the 20 samples in the following order: NOD 2 weeks (1–5), NOD 4 weeks (6–10), C57BL/6 2 weeks (11–15), and C57BL/6 4 weeks (16–20).

Fig. 5. Principal component analysis of proteome data represented in a three-dimensional graph with components 1–3 displayed on the three axes. Eight protein spots forming a cluster (having similar values within each component) are marked by a circle.
variability in the set, the more components are likely to contribute in a statistically identifiable way. Generally most of the variation in any experiment will be accounted for by three, and possibly four, components. The most common way to view the results of a PCA is in a three-dimensional scatter plot as seen in Fig. 5. All data points are represented, but depending on the combination of components plotted, the shape of the “cloud” of data points changes.

In evaluating this type of data presentation, the focus should be on regions that result in the cloud of data points not being a sphere. Genes in such a region are selected, and it can be determined whether the variation they contribute represents a pattern of expression that correlates with the already defined sample groups. If groups are not already defined, then this type of analysis can be used to define sample groupings. In Fig. 5 we show the results of PCA analysis on our proteome data. Principal component 1 eigenvalue 2100.74 accounted for 53.81% of the variance, principal component 2 eigenvalue 657.2 accounted for 16.83% of the variance, and principal component 3 eigenvalue 348.9 accounted for 8.94% of the variance. All other principal components combined account for the remaining 20.42% of the variance.

### TABLE I
Identity of proteins in spots from data analysis

Each spot in the gel matched-set was assigned an identifying SSP number. Proteins in the spots were identified by peptide fingerprinting of tryptic fragments. For identified spots the protein identifier from the Swiss-Prot database is given. For proteins that are represented in the Ingenuity database their symbol is given with the network number in parentheses. Finally the commonly used name of the protein is given. ER, endoplasmic reticulum.

| Spot SSP no. | Swiss-Prot accession no. | Ingenuity (network no.) | Name |
|--------------|--------------------------|-------------------------|------|
| 1508         | P20029                   | HSPA5 (1)*               | Heat shock 70-kDa protein 5 |
| 2707         | Q01853                   | VCP (2)                 | Transitional ER ATPase |
| 3503         | P63038                   | HSPD1 (2)               | Chaperonin (HSP60, mitochondrial) |
| 3506         | P50516                   | ATP6V1A (5)             | Vacular ATPase A subunit 1 |
| 1302, 2007, 2709 | No ID                 |                         |      |
| Two-way ANOVA strain | | | |
| 1302, 1508, 2709, 3503, 3506 | | | See t test |
| 4001         | P67778                   | PHB (2)                 | Prohibitin (mitochondrial) |
| 4104         | P47757                   | CAPZB (1)               | F-actin capping protein β subunit |
| 6103         | No ID                    |                         |      |
| Two-way ANOVA strain and age interaction | | | |
| 2213         | P60710                   | ACTB (1)                | β-Actin |
| 2503         | P14733                   | LMNB1 (2)*              | Lamin B1 |
| 2506         | P14733                   | LMNB1 (2)*              | Lamin B1 |
| 5207         | Q91WD5                   | NDUFS2 (6)              | NADH-ubiquinone oxidoreductase 49-kDa subunit (mitochondrial) |
| 5406         | P27773                   | GRP58 (1)               | Protein-disulfide isomerase A3 |
| 7105         | Q93092                   | TALDO1 (3)              | Transaldolase |
| 7106         | P70227                   | ITPR3 (2)               | Type 3 inositol trisphosphate receptor |
| 3103, 6611   | No ID                    |                         |      |
| K-means       | | | See t test |
| 1508, 2007, 2709, 3503, 2709 | | | See ANOVA strain and age interaction |
| 2506         | Q61599                   | ARHGDB (2)              | Rho GDP dissociation inhibitor 2 |
| 1311         | P56480                   | ATP5B (1)*              | ATP synthase β chain (mitochondrial) |
| 1606         | P20029                   | HSPA5 (1)*              | Heat shock 70-kDa protein 5 |
| 2002         | Q99PT1                   | ARHGDA (2)              | Rho GDP-dissociation inhibitor 1 |
| 2215         | Q9D1Q6                   | Tioredoxin domain-containing P.4 |
| 2301         | P05213                   | K-α-1 (1)               | Tubulin α-1 chain |
| 2407         | P61979                   | HNRPK (1)               | Heterogeneous nuclear ribonucleoprotein K |
| 2509         | Q61233                   | LCP1 (4)                | L-plastin (lymphocyte cytosolic P.1) |
| 3204         | P60843                   | EIF4A1 (2)              | Eukaryotic initiation factor 4A-I |
| 3501         | P17156                   | HSPA2 (7)               | Heat shock 70-kDa protein 2 |
| PCA          | | | |
| 1208         | P14206                   | LAMR1 (1)               | 40 S ribosomal protein SA |
| 1316         | P56480                   | ATP5B (1)*              | ATP synthase β chain (mitochondrial) |
| 4507         | P28843                   | DPP4 (1)                | Dipeptidyl peptidase IV (CD26) |
| 7413         | P80314                   | CCT2 (1)                | T-complex protein 1, β subunit |
| 1105, 3112, 6206, 7202 | No ID                    |                         |      |

* Indicates proteins that are identified in more than one spot.
We evaluated the three-dimensional plot of principal components 1, 2, and 3 from multiple angles and found a group of protein spots (marked by a circle) that all clearly separate from the rest in the view shown in Fig. 5. These spots had SSP numbers 1105, 1208, 1316, 3112, 4507, 6206, 7202, and 7413. This group represents spots that have a similar type of variance between the 20 gels.

Identification of Proteins in Differentially Expressed Spots—In the above sections we have described a number of different data analysis approaches that identified groups of protein spots whose intensity appeared to be associated with interesting differences in group parameters. Each of these spots were subjected to tryptic digestion and peptide mass fingerprinting to identify protein(s) in the spot. Table I shows the identity of the spots from the t test, ANOVA strain effects, additional spots with strain and age interaction by the ANOVA, the K-means cluster that followed the strain difference pattern, and finally the PCA group. These data indicated that in the PCA group was a protein (ATPS5B) that also was identified in the K-means cluster, suggesting the potential that the PCA group also may associate with a strain difference pattern.

Identification of Protein Networks Associated with Differential Expression—Our proteome data analysis identified 37 differentially expressed spots, and we identified proteins in 27 of these spots. The list of affected proteins was only 24 because there were three sets of spots containing the same protein in different spots (likely due to different post-translational modifications of the same protein). Although the suggestion of strain effects for all of these 24 proteins was not supported by statistical analysis we decided to conduct network analysis on them as a group. The rationale that allows for use of this “big net” data mining strategy is that no conclusions are draw until those data have been further sorted and confirmed by other methods. Network analysis will sort out proteins that do not belong to a molecular network together with the other proteins (helping remove any “false positives” that do not belong with the rest). Furthermore the risk of reaching invalid conclusions will be reduced by the next step where we compare the resulting proteome networks to networks created by transcriptome data that has been analyzed for strain effect with a very stringent statistical method.

The Ingenuity database had information about 23 of the 24 proteins, and two networks with highly significant scores of 18 and 14, respectively, were created. Located within these two networks (of 35 genes each) were 10 and eight focus genes, respectively. The remaining five proteins did not fall into larger networks, although the database contained some information about them (Fig. 6). In Table I we have indicated both the gene symbol produced by Ingenuity as well as (in parentheses) the specific network to which each protein belongs.

Further information in the database about the genes in each protein network indicated that although the server limits the size of networks to 35 genes in each, the two proteome networks could easily be connected based on the server information about the individual proteins in each of them. Fig. 7 shows the graphic presentation of the two proteome networks and the connections within each network as provided by the server. In addition we have added a few connections between networks (broken lines) that exist according to information in the database. In this figure the gray icons indicate that a protein is a focus gene, that is it comes from our list of differentially expressed proteins (Table I). If the database (for computational reasons) had not limited the networks to a size of maximum 35 genes, we believe that a single network of these genes would have been created.

Transcriptome Networks Associated with Strain Differences—The transcriptome data (produced with RNA from the same samples that produced the proteome dataset) were also subjected to analysis using the GeneSpring software. On these data we first filtered out probe sets that did not have at least one sample with significant or marginal expression. Data from the resulting 26,540 probe sets were subjected to a two-way ANOVA using the Westfal and Young MTC. This produced a list of 273 probe sets that with a very high level of confidence can be declared as differentially expressed between strains. This list of genes was also uploaded for Ingenuity pathway analysis. The Affymetrix chips contain a large number of probe sets that represents genes (or expressed sequence tags) for which there is almost no knowledge. Therefore, it was not a surprise that only 96 probe sets, representing 89 different genes, had a match with information in the Ingenuity database. Of those 89 different genes, 64 belonged to six different networks. The remaining 25 genes could not be connected to any other focus genes to create significant networks based on the information in the database.

Function and Interconnection of Networks—The Ingenuity server provides a list with information regarding all networks that have been calculated from an uploaded list of focus genes. The server provides information about the number of focus genes in each network, the likelihood of these focus networks.
genes being in a network due to random chance (the score), and the biological functions significantly over-represented in each network. In Table II we show the score, number of focus genes, and top three biological functions for each of the two proteome networks (P1 and P2) and the six transcriptome networks (T1–T6). Within these eight networks, cancer is represented in three (P1, T2, and T4), cell cycle is represented in three (T4, T5, and T6), cell death is represented in two (P2 and T1), cell signaling is represented in two (T1 and T6), and cell-to-cell signaling and interaction is represented in two (P1 and T3). No other biological function was among the top three for more than one network. The overall theme for these networks (representing differences in leukocytes from diabetes-susceptible NOD and control C57BL/6 mice) appears to focus...
on cell proliferation and death with the expected association of inter- and intracellular signaling.

The networks identified by the proteome and by the transcriptome data had substantial overlap and known interactions. Fig. 8 shows the 272 genes that belong to either a proteome or a transcriptome network. Members of each network are either within the main box indicating the name of the network and the number of focus genes/number of genes in the network, or they are connected to that main box with a triple line. Focus (differentially expressed) genes are listed in bold font. Prdx2 is both a member of proteome network 1 and transcriptome network 2 but only a focus gene at the transcriptome level. Based on information from the database we have indicated some of the connections known to exist between members of different networks by a broken line between boxes with the names of the connected genes.

**FIG. 8.** All members of the two proteome networks and six transcriptome networks and their interconnections. Members of a network are either within the main box indicating the name of the network and the number of focus genes/number of genes in the network, or they are connected to that main box with a triple line. Focus (differentially expressed) genes are listed in bold font. Prdx2 is both a member of proteome network 1 and transcriptome network 2 but only a focus gene at the transcriptome level. Based on information from the database we have indicated some of the connections known to exist between members of different networks by a broken line between boxes with the names of the connected genes.

**FIG. 9.** Genes differentially expressed at the protein (square boxes) and mRNA (oval boxes) levels and their direct network connections to the genes Prdx2, Myc, Mycn, and Gapd.

posttranslational modification) level, the data suggest a role for these four genes in explaining strain differences.

In addition to all the shared genes between the eight networks, Fig. 8 also indicates (with broken lines) some of the known connections between genes in different networks. As an example Casp3 is a member of proteome network 2. The Ingenuity database indicates that Vegf can modify expression of Casp3 and that Casp3 can induce cleavage of Eif2s1.
Therefore, we have placed broken lines from Casp3 to Vegf and Eif2s1. The figure does not show connection between members of the two proteome networks because they were shown in Fig. 7. It should be noted that Fig. 8 only shows a few selected connections between all the networks to demonstrate the interconnectivity of the networks yet retain a readable image. The large combined network of all six transcriptome and two proteome networks contained 272 genes of which 82 were differentially expressed focus genes.

**DISCUSSION**

Biological systems utilize highly complex, interrelated metabolic and signaling pathways to function. Whether examining the transcriptome or the proteome profiles of specific cells or tissues, one acquires large datasets that are inherently complex. As a result we consider it beneficial to analyze proteome datasets using as many tools as possible. Traditional software for analysis of 2D gel proteome data only offers very limited and univariate data analysis tools, and development of customized programs can be time-consuming. As an alternative, we developed a protocol that allows us to export the protein expression data from a 2D gel image dataset and then import it for analysis in a popular program used for analysis of transcriptome data. This gave us the opportunity to not only conduct Student’s t test on the data but also ANOVA, cluster analysis, and PCA. Clustering algorithms and PCA have a different statistical emphasis allowing the identification of patterns of expression without predefining groups and/or using statistical cut-off values in the analysis.

Using a mixture of statistical, clustering, and PCA analysis on a 2D gel proteome dataset from NOD and C57BL/6 leukocytes, we produced a list of 37 spots that were likely to be differentially expressed between the two strains of mice. Admittedly because of the methods used to produce this list, it was also likely that some of these 37 spots had no expression differences between strains. The lack of statistical significance is acceptable in a discovery project where type 1 errors are much less of a concern than in traditional hypothesis testing because the data are not used to draw any conclusions before additional validation approaches in the downstream process have been applied. In our case these processes included narrowing the list of interest with network analysis and comparison with transcriptome data collected by a much more statistically stringent process. This study not only demonstrates the use of gene expression software for analysis of proteome data generated from 2D gels, it also demonstrates that this approach allows investigators to extract much more information from a proteome dataset. The t test available in traditional image analysis software produced a list of only seven spots differentially expressed between strains.

Unfortunately proteins in 10 of the 37 differentially expressed spots could not be identified, and only 24 different proteins were identified because some spots were found to contain the same protein. The identification of the same protein in two spots is likely due to the fact that many transla- tional modifications create separate new protein spots on a 2D gel due to significant shifts of the pI of the proteins. It is important to note that data on spot intensity in a 2D gel are not the same as a protein assay for that protein. In most cases a 2D gel spot only reflects the amount of the protein that is or is not modified by a specific post-translational modification.

Network analysis using the Ingenuity server indicated that of the 24 proteins identified six did not belong to a network that could be produced from the information available in their database. One of those six had no information available in the database; the remaining five had information in the database, but it could not connect them to each other or to the two networks within the limitations on network size of no more than 35. Each of these six proteins could be false positives or simply reflect that the database or the scientific community does not know about connections that actually exist. The latter problem is particularly large for the use of network analysis on our transcriptome data because the Affymetrix arrays contains a large number of expressed sequence tags whose gene products have not yet been studied. Finally it is also possible that the non-connected proteins could represent false identification of a differentially expressed protein such as may occur if more than one protein is present in a spot.

The Ingenuity server restricts the size of networks to a maximum of 35 genes and attempts to create networks that contain the highest possible number of genes from the uploaded list (called focus genes). Our proteome data produced two networks containing 10 and eight focus genes, respectively. With scores of 18 and 14 the probability of creating these networks by chance was very low. A quick look at the information in the Ingenuity database clearly indicated that many genes in either network were connected to genes in the other network and that it would be appropriate to consider this a single network of 70 genes (18 of which were differentially expressed). The reasons that we do not detect all the genes in these networks as differentially expressed at the protein level could be that they are not expressed in high enough amounts to be detected on a 2D gel, or perhaps any differential expression is simply not large enough to be determined by this technology.

It is interesting that the biological functions of some of the central genes in this proteome network are centered around the opposing biological processes of cell proliferation (oncogenes Myc and Mycn) and cell death (Bcl2 and Casp3). Based on our proteome data one could hypothesize that a struggle between these two processes may lead to the abnormal expression we observe for a number of genes at the protein level. As with any other discovery dataset the hypothesis produced will have to be tested by traditional scientific methods. Indeed more work is needed just to create a specific and testable hypothesis or model that may involve mutual inter-
actions and regulatory effects of some or all of the 70 genes in this proteome network.

To further support the conclusions regarding biological processes and specific networks that differ between NOD and C57BL/6 mice, we also conducted analysis of the gene expression at the mRNA level in the same 20 samples on which we had conducted proteome analysis. The network analysis of a gene list produced by stringent statistical analysis of these data produced six networks. The proteome networks shared one or more genes with all of these transcriptome networks except for transcriptome network 1. This sharing included that both transcriptome and proteome networks contained the oncogenes Myc and Mycn. However, the differentially expressed proteins and transcripts pointing toward those central genes were different. Although expression of a gene at the mRNA and protein level does not always correlate, other reasons may explain why we have so little overlap between lists of differentially expressed proteins and transcripts: 1) spots on 2D gels often only represents a single post-translationally modified version of a protein, not the total expression of a gene at the protein level, 2) the analysis approach and statistical threshold used on the two types of data are very different, 3) our proteome analysis is limited to proteins that have pl of 4–7, and 4) the fraction of the total proteome covered is much smaller than the fraction of the total transcriptome covered. A final point regarding the proteome and transcriptome based networks is that genes from each of the eight networks have known connections to multiple genes from other networks. It is clear that the whole system of eight networks has such a degree of connection to each other that they can be viewed as a single network containing 272 genes of which 82 are differentially expressed.

Each of these eight networks is assigned functions based on the known functions of their members and the statistical likelihood that a specific function would be represented with that frequency by chance. We compared the three highest ranking (most significant) functions over-represented in these networks and found that most of them centered on the processes of cell proliferation and death. Signaling between and within cells was the second systems level process that could be seen in those data. In contrast the process of immune response was only listed as second rank in a single network, suggesting that at this young age the excessive activation of the immune response in NOD mice is still rather minimal.

Defects in the process of apoptosis have been observed in NOD mice, and it has been hypothesized to be important in explaining why autoreactive leukocytes in this mouse persist (27–29). Indeed one report indicated that defects in apoptosis could be a common theme for many autoimmune diseases (30, 31). The oncogenes Myc and Mycn are central to both proteome and transcriptome networks created by our data. The resistance of lymphocytes from NOD mice to glucocorticoid-induced apoptosis was associated with an up-regulation of Myc in contrast to C57BL/6 mice where Myc was down-regulated (32), and treatment of NOD mice with the streptococcal wall component OK432 prevents diabetes, restores dexamethasone-induced apoptosis, and was shown to be associated with down-regulation of Myc (33). Our systems biology data suggest that these apoptosis defects may be basic defects that are manifested even before the processes of immune response and immune activation become dominant. Our data also point to an association of these apoptosis defects with the Myc/Mycn oncogenes and demonstrate a large number of new genes that may be affected via these oncogenes. Furthermore many of the differentially expressed genes that are directly connected to these oncogenes are translation factors and ribosomal proteins. One of these, the ribosomal protein RPL8, appears to be expressed at very low levels in NOD mouse leukocytes, suggesting that basic cellular functions could be affected probably via Myc/Mycn.

Our data support the idea that the balance between cell death and cell survival is defective in NOD mice and suggest that this is an early and central defect that is manifest in peripheral leukocytes even before the immune system is activated. Our data also point to specific networks that may be worth further investigation to gain a deeper insight into the origins of this dysfunction. We cannot rule out that some of the differences between NOD and C57BL/6 mice found in this study could represent genetic diversity not connected to autoimmunity at all. However, many of the differentially expressed genes are clearly connected to processes and networks known to be associated with development of autoimmunity. Comparison of NOD to other control strains should help determine, for each specific network/process, whether it is more connected to benign genetic divergence than it is to susceptibility to autoimmunity.

Acknowledgments—We gratefully acknowledge research facilities made available by the Veterans Affairs Research Service, Memphis, TN, and Dr. John Stuart.

REFERENCES
1. Evans, C.-O., Young, A. N., Braun, M. R., Brat, D. J., Parks, J. S., Neish, A. S., and Oyesiku, N. M. (2001) Novel patterns of gene expression in pituitary adenomas identified by complementary deoxyribonucleic acid microarrays and qualitative reverse transcription-polymerase chain reaction. J. Clin. Endocrinol. Metab. 86, 3097–3107
2. Jungblut, P. R., Zimony, A. U., Zeindl-Eberhart, E., Stuhl, J., Koupiola, K., Pleibner, K. B., Otto, A., Muller, E. C., Sokolowska-Kohler, W., Grabber, G., and Stoffler, G. (1999) Proteomics in human disease: cancer, heart and infectious diseases. Electrophoresis 20, 2100–2110
Molecular Phenotyping of NOD Mice

3. Stepan, C., Bailey, S., Bhat, S., Brown, E., Bannere, R., Wright, C., Patel, H., Ahima, R., and Lazar, M. (2001) The hormone resistin links obesity to diabetes. Nature 409, 307–312

4. Gerling, I. C., Sun, Y., Ahokas, R. A., Wodi, L. A., Bhattacharya, S. K., Warrington, K. J., Postlethwaite, A. E., and Weber, K. T. (2003) Allosteronism. An immunostimulatory state precedes the proinflammatory/fibrogenic cardiac phenotype. Am. J. Phys. 285, H813–H821

5. International Human Genome Sequencing Consortium (2004) Finishing the euchromatic sequence of the human genome. Nature 431, 931–945

6. Gorg, A., Pastew, W., and Gunter, S. (1988) The current state of two-dimensional electrophoresis with immobilized pH gradients. Electrophoresis 9, 531–543

7. Jungblut, P., Thiede, B., Zimny, A. U., Muller, E., Scheler, C., Wittmann, L. B., and Otto, A. (1998) Resolution power of two-dimensional electrophoresis and identification of proteins from gels. Electrophoresis 17, 839–847

8. Raman, B., Cheug, A., and Marten, M. R. (2002) Quantitative comparison and evaluation of two commercially available, two-dimensional electrophoresis image analysis software packages, Z3 and Melanie. Electrophoresis 23, 2194–2202

9. Rosengren, A. T., Salmi, J. M., Aittokallio, T., Westerholm, J., Lahesmaa, R., Nyman, T. A., and Nevalainen, O. S. (2003) Comparison of PDQuest and Progenesis software packages in the analysis of two-dimensional electrophoresis gels. Proteomics 3, 936–946

10. Gharbi, S., Gaffney, P., Yan, A., Zelb, M. J., Cramer, R., Waterfield, M. D., and Timms, J. F. (2002) Evaluation of two dimensional differential gel electrophoresis for proteomic expression analysis of a model breast cancer cell system. Mol. Cell. Proteomics 1, 91–98

11. Lenchik, N., Desiderio, D. M., and Gerling, I. C. (2005) Two-dimensional gel electrophoresis characterization of the mouse leukocyte proteome, using a tri-reagent for protein extraction. Proteomics 5, 2202–2209

12. Gerling, I. C., Ali, C., and Lenchik, N. (2003) Characterization of early developments in the splenic leukocyte transcriptome of NOD mice. Ann. N Y Acad. Sci. 1005, 157–160

13. Chatenoud, L., and Bach, J. F. (2005) Regulatory T cells in the control of autoimmune diabetes: the case of the NOD mouse. Int. Rev. Immunol. 24, 247–267

14. Anderson, M. S., and Bluestone, J. A. (2005) The NOD mouse: a model of immune dysregulation. Annu. Rev. Immunol. 23, 447–485

15. Leiter, E. H. (2005) Nonobese diabetic mice and the genetics of diabetes susceptibility. Curr. Diabetes Rep. 5, 141–148

16. Sarvetnick, N. (2000) Etiology of autoimmunity. Immunol. Res. 21, 357–362

17. Yoon, J. W., Jun, H. S., and Santamaria, P. (1998) Cellular and molecular mechanisms for the initiation and progression of beta cell destruction resulting from the collaboration between macrophages and T cells. Autoimmunity 27, 109–122

18. Wu, J., Lenchik, N., Pabst, M. J., Solomon, S. S., Shull, J., and Gerling, I. C. (2005) Functional characterization of two-dimensional gel-separated proteins using sequential staining. Electrophoresis 26, 225–237

19. Solomon, S. S., Buss, N., Shull, J., Monnier, S., Majumdar, G., Wu, J., and Gerling, I. C. (2005) Proteome of H-411E (liver) cells exposed to insulin and tumor necrosis factor-α: analysis of proteins involved in insulin resistance. J. Lab. Clin. Med. 145, 275–283

20. Dresen, I. M., Husing, J., Kruse, E., Boes, T., and Jockel, K. H. (2003) Software packages for quantitative microarray-based gene expression analysis. Curr. Pharm. Biotechnol. 4, 417–437

21. Stein, R. C., and Zvelebil, M. J. (2002) The application of 2D gel-based proteomics methods to the study of breast cancer. J. Mammary Gland Biol. Neoplasia 7, 385–393

22. MacQueen, J. B. (1967) Some methods for classification and analysis of multivariate observations. Proc. Berkeley Symp. Math. Stat. Probab. 1, 281–297

23. Sherlock, G. (2000) Analysis of large-scale gene expression data. Curr. Opin. Immunol. 12, 201–205

24. Verhoeckx, K. C., Bijlsma, S., de Groene, E. M., Witkamp, R. F., van der Greef, J., and Rodenburg, R. J. (2004) A combination of proteomics, principal component analysis and transcriptomics is a powerful tool for the identification of biomarkers for macrophage maturation in the U937 cell line. Proteomics 4, 1014–1028

25. Raychaudhuri, S., Stuart, J. M., and Altman, R. B. (2000) Principal components analysis to summarize microarray experiments: application to sporation time series. Pac. Symp. Biocomput. 455–466

26. Yeung, K. Y., and Ruzzo, W. L. (2001) Principal component analysis for clustering gene expression data. Bioinformatics 17, 763–774

27. Kishimoto, H., and Sprent, J. (2001) A defect in central tolerance in NOD mice. Nat. Immunol. 2, 1025–1031

28. Arreaza, G., Salojin, K., Yang, W., Zhang, J., Gill, B., Mi, Q. S., Gao, J. X., Meagher, C., Cameron, M., and Delovitch, T. L. (2003) Deficient activation and resistance to activation-induced apoptosis of CD8+ T cells is associated with defective peripheral tolerance in nonobese diabetic mice. Clin. Immunol. 107, 103–115

29. Decallonne, B., van Etten, E., Giulietti, A., Casteels, K., Overbergh, L., Bouillon, R., and Mathieu, C. (2003) Defect in activation-induced cell death in non-obese diabetic (NOD) T lymphocytes. J. Autoimmun. 20, 219–226

30. Aune, T. M., Maas, K., Moore, J. H., and Olsen, N. J. (2003) Gene expression profiles in human autoimmune disease. Curr. Pharm. Des. 9, 1905–1917

31. Hayashi, T., and Faustman, D. L. (2003) Role of defective apoptosis in type 1 diabetes and other autoimmune diseases. Recent Prog. Horm. Res. 58, 131–153

32. Martins, T. C., and Aguas, A. P. (1998) Involvement of c-myc in the resistance of non-obese diabetic mice to glucocorticoid-induced apoptosis. Immunol. Cell. Biol. 76, 377–382

33. Decallonne, B., Overbergh, L., Casteels, K. M., Gysseman, C., Bouillon, R., and Mathieu, C. (2000) Streptococcal wall component OK432 restores sensitivity of non-obese diabetic (NOD) thymocytes to apoptotic signals. Diabetologia 43, 1302–1308