Integration of heterogeneous expression data sets extends the role of the retinol pathway in diabetes and insulin resistance

The MIT Faculty has made this article openly available. Please share how this access benefits you. Your story matters.

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>As Published</td>
<td><a href="http://dx.doi.org/10.1093/bioinformatics/btp559">http://dx.doi.org/10.1093/bioinformatics/btp559</a></td>
</tr>
<tr>
<td>Publisher</td>
<td>Oxford University Press</td>
</tr>
<tr>
<td>Version</td>
<td>Final published version</td>
</tr>
<tr>
<td>Accessed</td>
<td>Mon Apr 25 00:42:00 EDT 2016</td>
</tr>
<tr>
<td>Citable Link</td>
<td><a href="http://hdl.handle.net/1721.1/72972">http://hdl.handle.net/1721.1/72972</a></td>
</tr>
<tr>
<td>Terms of Use</td>
<td>Creative Commons Attribution Non-Commercial</td>
</tr>
<tr>
<td>Detailed Terms</td>
<td><a href="http://creativecommons.org/licenses/by-nc/2.5">http://creativecommons.org/licenses/by-nc/2.5</a></td>
</tr>
</tbody>
</table>
Integration of heterogeneous expression data sets extends the role of the retinol pathway in diabetes and insulin resistance

Peter J. Park1,2,3, Sek Won Kong1,4, Toma Tebaldi5,6, Weil R. Lar3, Simon Kasif1,7,8 and Isaac S. Kohane1,2,3,∗

1Children’s Hospital Informatics Program at the Harvard-MIT Division of Health Sciences and Technology, 2Brigham and Women’s Hospital, 3Center of Biomedical Informatics, Harvard Medical School, 4Department of Cardiology, Children’s Hospital, Boston, MA 02115, USA, 5Centre for Integrative Biology, 6Department of Information Engineering and Computer Science, University of Trento, Italy, 7Center for Advanced Genomic Technology, Boston University and 8Department of Biomedical Engineering, Boston University, Boston, MA 02215, USA

Received on June 14, 2009; revised on September 7, 2009; accepted on September 22, 2009

ABSTRACT

Motivation: Type 2 diabetes is a chronic metabolic disease that involves both environmental and genetic factors. To understand the genetics of type 2 diabetes and insulin resistance, the Diabetes Genome Anatomy Project (DGAP) was launched to profile gene expression in a variety of related animal models and human subjects. We asked whether these heterogeneous models can be integrated to provide consistent and robust biological insights into the biology of insulin resistance.

Results: We perform integrative analysis of the 16 DGAP data sets that span multiple tissues, conditions, array types, laboratories, species, genetic backgrounds and study designs. For each data set, we identify differentially expressed genes compared with control. Then, for the combined data, we rank genes according to the frequency with which they were found to be statistically significant across data sets. This analysis reveals RetSat as a widely shared component of mechanisms involved in insulin resistance and sensitivity and adds to the growing importance of the retinol pathway in diabetes, adipogenesis and insulin resistance. Top candidates obtained from our analysis have been confirmed in recent laboratory studies.

Contact: Isaac_kohane@harvard.edu

1 INTRODUCTION

Type 2 diabetes mellitus is a chronic, progressive metabolic disorder and is one of the fastest-growing public health problems. Given an increased prevalence of obesity and aging population, recent estimates suggest that the worldwide prevalence will grow from 2.8% in 2000 to 4.4% in 2030, affecting 171 million in 2000 to 366 million in 2030 (Wild et al., 2004). The primary characteristics of type 2 diabetes are insulin resistance, relative insulin deficiency and hyperglycemia, and it can be easily diagnosed based on chronic elevated blood glucose concentration. While there is a strong heritable component, this has not been defined in the vast majority of cases.

∗To whom correspondence should be addressed.

To understand the interface between insulin action, insulin resistance, obesity and the genetics of type 2 diabetes, the Diabetes Genome Anatomy Project (DGAP) was initiated in 2003 to use a multi-dimensional genomic approach to characterize the relevant set of genes and gene products as well as the secondary changes in gene expression that occur in response to the metabolic abnormalities present in diabetes. Over the course of the project, a variety of data sets were collected through expression profiling studies on the Affymetrix platform, both from human and mouse tissues. In human studies, gene expression data were collected from case-control studies involving normal, insulin resistant, obese and diabetic subjects; in mouse studies, expression patterns were obtained before and after insulin stimulation in normal and various knock-out models, and adipogenic diets. An open question was whether there were common mechanisms in insulin resistance or sensitivity that could be identified by integrating results across this highly heterogeneous corpus.

In this work, we carry out an integrative analysis of the ∼450 arrays from the 16 data sets collected in this project. Analysis of the aggregate data presents complications due to the multiple sources of heterogeneity, such as species, platforms, laboratories, sample sizes and experimental design. The data set, for instance, includes several array types including Hu6800 and U133 (human) and U74, U74v2 and MOE430 (mouse). Few are simple two-group comparisons of clinical samples, while others involve strain, age, tissue comparisons in multi-factorial designs. A few of the data sets have been studied extensively already but in isolation. We aim to carry out a comprehensive analysis of the aggregate data focusing on the commonalities between the data sets. There are two important underlying assumptions in our analysis. The first is that the individual experiments were appropriately designed to capture a transcriptome signature relevant to insulin resistance whether in a mouse model of IRS-1 ‘knock outs’ versus wild-type mice or in a comparison of obese diabetic humans versus obese non-diabetics. Second, given the well known heterogeneity of measurement across different platforms (Kuo et al., 2002), even from the same manufacturer (Nimgaonkar et al., 2003), only robustly shared molecular processes pertaining to several models...
of insulin resistance, obesity and/or diabetes will be detectable. That is, regardless of the multiplicity of etiologies, we assume that there exists a small number of common pathophysiological mechanisms across diabetes, insulin resistance and obesity. Based on our computations with these two assumptions, we have been able to confirm previous findings that implicates the retinol pathway (Yang et al., 2005) in insulin sensitivity/resistance and adipogenesis, as well as reconfirming the well known dysregulation of oxidative phosphorylation (Lowell and Shulman, 2005; Mootha et al., 2003; Patti et al., 2003) and the JAK-STAT pathway (Schwartz and Porte, 2005).

2 METHODS

2.1 Data availability, normalization and quality control

Both raw data (CEL files) and processed data are available from the DGAP website (http://www.diabetesgenome.org). The total number of data sets in this database is 19. Three data sets were excluded for the following reasons: dataset 1 was generated on MG-U74A array, which was later found to have a large fraction of incorrectly labeled probes; dataset 3 was a time-course experiment on adipocyte differentiation that did not have a proper control to be informative for this study; dataset 9 was excluded because it was the only one generated on the Affymetrix hu6800 platform. Including this early-model platform would have reduced the total number of genes common to the platforms significantly.

Given the large number of arrays generated in multiple laboratories, it is inevitable that some hybridizations failed or had experimental biases that require special attention. Visual inspection of array images revealed several arrays with spatial artifacts. But because each probe set on Affymetrix arrays consists of many probes distributed randomly on the array, a small amount of spatial artifacts observed were unlikely to affect the expression values significantly. We also calculated the distribution of expression values, the number of Present/Absent calls, and other statistics for each array to ensure that only high quality arrays are used. We found two arrays that failed entirely and were not therefore included in the analysis. As expression levels in different data sets are often derived from raw data using different algorithms, we recalculated the expression levels for all arrays with the same PLIER algorithm (Affymetrix, 2005). Data were normalized by setting the trimmed mean of all arrays to be the same.

2.2 Identification of differentially expressed genes

Several different experimental designs are present in the data, but in each case, a control group was present and two-group comparisons were possible. To identify differentially expressed genes, the t-test was used. When the sample size was small, a regularized form of the test was used to guard against false positives that may appear due to under-estimated gene-specific variance.

To determine statistical significance in a genome-wide study, adjustment for multiple hypotheses is usually applied. To adjust for multiple hypothesis testing correctly, three layers of multiplicity must be considered. The first is due to the large number of genes within each data set; the second is due to the multiple groups within each data set; and the last is due to the multiple data sets that are involved. In this study, we applied a liberal, relaxed criterion for each data set and used the multiplicity of data sets to filter the gene list. Thus, we used standard $P=0.05$ for each data set and applied a Bonferroni-type correction only for the number of comparisons in certain study designs. For instance, the data set on muscle insulin receptor knockout (MIRKO) and control mice (Yechoor et al., 2004) requires four separate comparisons: WT versus Stz (insulin resistant) and WT versus Stz insulin, on both Lox strain and MIRKO strain. Therefore, the threshold for this data was set at $P=0.05/4=0.0125$. The final significance was computed by a permutation approach.

3 RESULTS

3.1 Combining data from multiple studies

The data sets in this project were highly diverse, as shown in Table 1. For unbiased analysis, variables to consider in analysis include organism (human, mouse), array type (MG-U74Av2, MOE430, HG-U133A), tissue type (adipocyte, brown preadipocyte, fibroblast, hepatocyte, myocyte, pancreas, skeletal), mouse strain (B6 versus 129), mouse genotype (WT versus IRS-1 KOs), insulin sensitivity, treatment (DM2, IGT, NGT), laboratory in which the experiment was conducted, and study design (case versus control, time series). After mapping the human and mouse genes using the Homologene database (release 43.1), we displayed all usable arrays as points in a 3D principal component space (Figure 1A). To examine possible artefactual variations, we examined the distribution of the points by each variable including species, tissue type, array type, and laboratory (Figures 1B–E). In Figure 1C, for instance, we show the example of variation due to tissue type for mouse data. The figure illustrates that the variation due to tissue type is greater than that within each data set. In Figure 1E, we show the clustering effect for the laboratory setting. The data from the four laboratories in which the experiments were performed were clearly separated. Some of the variation is due to the differences in array type or species used in each laboratory, but there is clear effect of experimental protocols or technician at each site. These data sets were generated some time ago, and we expect that some of these variations have been reduced in more recent data sets due to improved technology; however, these variations still do exist and it is important to avoid confounding effects due to these factors when analyzing the data.

Given the clear separation between the samples in different categories, it is clear that expression levels cannot be combined directly among different studies. Our previous studies have suggested that even the data generated on different generations of the same manufacturer’s platforms introduce significant variability due to the changes in probe locations so that they cannot be directly combined (Hwang et al., 2004). This suggests that the data should be combined in a different manner. One such method is at the level of P-values. Fisher’s statistic, for instance, is based on the summation of negative log of the P-values across the studies. For this study, however, we found this to be too sensitive to a small subset of studies in which a gene may be highly significant. An alternative is a non-parametric approach to rank the genes from highest to lowest in significance in each study and to sum the rank of each gene across the studies. A gene that is significant in multiples should have a lower combined rank. In contrast to the Fisher’s method, we found this method to be too sensitive to the case in which a gene is strongly insignificant. A compromise is the product of ranks, or the sum of logarithm of ranks, which reduces the impact of low-scoring genes (Breitling et al., 2004).

Our approach is to determine whether a gene is statistically significant in each of the studies and to measure the overall significance of each gene by the total number of data sets in which it was significant, regardless of its exact P-value. We assumed that even though the underlying data sets are heterogeneous, the more often the gene is found to be significant, the more likely it is to be an important element of insulin signaling, obesity and/or diabetes. When there were multiple comparisons within a data set, statistical significance in any of the comparisons (adjusted for multiple testing) was sufficient to classify the data set as significant. Given the
Table 1. List of experiments in DGAP

<table>
<thead>
<tr>
<th>ID</th>
<th>Sample size</th>
<th>Array type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>27</td>
<td>MG_U74A/B/C</td>
<td>3T3-L1 fibroblast cells, 3T3-L1 adipocyte cells and mouse skeletal</td>
</tr>
<tr>
<td>2</td>
<td>28</td>
<td>MG_U74Av2</td>
<td>Brown preadipocyte IRS knockout profiling</td>
</tr>
<tr>
<td>3</td>
<td>22</td>
<td>MOE430A/B/</td>
<td>3T3-L1 adipocyte differentiation—time course</td>
</tr>
<tr>
<td>4</td>
<td>14</td>
<td>MG_U74Av2</td>
<td>Low versus high fat diet on mice of two genetic backgrounds (B6 versus 129)</td>
</tr>
<tr>
<td>5</td>
<td>16</td>
<td>MG_U74Av2</td>
<td>Low versus high fat diet on mice of two genetic backgrounds (B6 versus 129)</td>
</tr>
<tr>
<td>6</td>
<td>17</td>
<td>MG_U74Av2</td>
<td>Low versus high fat diet on mice of two genetic backgrounds (B6 versus 129)</td>
</tr>
<tr>
<td>7</td>
<td>18</td>
<td>MG_U74Av2</td>
<td>Isolated adipocytes from normal and fat insulin receptor KO (FIRKO) mice</td>
</tr>
<tr>
<td>8</td>
<td>6</td>
<td>MG_U74Av2</td>
<td>Liver—ob/ob mice</td>
</tr>
<tr>
<td>9</td>
<td>21</td>
<td>Hu6800</td>
<td>Human skeletal muscle—type 2 diabetes and family history positive individuals</td>
</tr>
<tr>
<td>10</td>
<td>9</td>
<td>MG_U74Av2</td>
<td>Mouse skeletal muscle—controls, streptozotocin diabetes and insulin treated</td>
</tr>
<tr>
<td>11</td>
<td>12</td>
<td>HG-U133A/B</td>
<td>Human pancreatic islets from normal and Type 2 diabetic subjects</td>
</tr>
<tr>
<td>12</td>
<td>21</td>
<td>MG_U74Av2</td>
<td>Transcription profiling of wild type and PGC-1alpha KO liver and skeletal muscle</td>
</tr>
<tr>
<td>13</td>
<td>12</td>
<td>MG_U74Av2</td>
<td>Effect of PGC-1alpha and PGC-1beta on gene expression in myocytes and hepatocytes</td>
</tr>
<tr>
<td>14</td>
<td>17</td>
<td>MG_U74Av2</td>
<td>IR and IRS-1, single/double het KO—age and genetic background—epididymal white fat</td>
</tr>
<tr>
<td>15</td>
<td>55</td>
<td>MG_U74Av2</td>
<td>IR and IRS-1, single/double het KO—age and genetic background—liver</td>
</tr>
<tr>
<td>16</td>
<td>52</td>
<td>MG_U74Av2</td>
<td>IR and IRS-1, single/double het KO—age and genetic background—skeletal muscle</td>
</tr>
<tr>
<td>17</td>
<td>12</td>
<td>MG_U74Av2</td>
<td>Effect of insulin infusion on skeletal muscle</td>
</tr>
<tr>
<td>18</td>
<td>44</td>
<td>MG_U74Av2</td>
<td>Skeletal muscle—muscle IR KO and control mice—control, streptozotocin diabetic and insulin treated</td>
</tr>
<tr>
<td>19</td>
<td>54</td>
<td>HG-U133A</td>
<td>Human skeletal muscle—type 2 diabetes—Swedish males</td>
</tr>
</tbody>
</table>

All datasets except 1, 3 and 9 were used (see ‘Methods’ section) in the meta-analysis.

Fig. 1. Sample characteristics and systematic differences in principal component spaces for DGAP experiments. All ~450 samples are shown in (A), colored differently for the 17 studies in the combined data set. Systematic differences include differences across murine tissue types (B), species (C), expression measurement platforms (D), laboratories where the measurements were made (E) and patient phenotypes (F).
Table 2. A list of significant genes in the meta-analysis

<table>
<thead>
<tr>
<th>Number of datasets</th>
<th>Gene name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>8/16</td>
<td>RETSAT (FLJ20296)</td>
<td>All-trans-retinol 13,14-reductase</td>
</tr>
<tr>
<td>7/16</td>
<td>KPNB1</td>
<td>Karyopherin (importin) beta 1</td>
</tr>
<tr>
<td></td>
<td>SDHB</td>
<td>Succinate dehydrogenase complex, subunit B, iron sulfur (lp)</td>
</tr>
<tr>
<td></td>
<td>MRPL34</td>
<td>Mitochondrial ribosomal protein L34</td>
</tr>
<tr>
<td></td>
<td>GPX3</td>
<td>Glutathione peroxidase 3 (plasma)</td>
</tr>
<tr>
<td></td>
<td>PAM</td>
<td>Peptidylglycine alpha-amidating monoxygenase</td>
</tr>
<tr>
<td>6/16</td>
<td>ACTN3</td>
<td>Actinin, alpha 3</td>
</tr>
<tr>
<td></td>
<td>CPT1A</td>
<td>Carnitine palmityltransferase 1A (liver)</td>
</tr>
<tr>
<td></td>
<td>RFX1</td>
<td>Regulatory factor X, 1 (influences HLA class II expression)</td>
</tr>
<tr>
<td></td>
<td>TSTA3</td>
<td>Tissue specific transplantation antigen P55</td>
</tr>
<tr>
<td></td>
<td>UQRC1</td>
<td>Ubiquinol–cytochrome c reductase core protein I</td>
</tr>
<tr>
<td></td>
<td>DDX1X</td>
<td>DEAD (Asp-Glu-Ala–Asp) box polypeptide 3, X-linked</td>
</tr>
<tr>
<td></td>
<td>DCTN6</td>
<td>Dynactin 6</td>
</tr>
<tr>
<td></td>
<td>TRAPP C4</td>
<td>Trafficking protein particle complex 4</td>
</tr>
<tr>
<td></td>
<td>TGFBI14</td>
<td>Transforming growth factor beta 1 induced transcript 4</td>
</tr>
<tr>
<td></td>
<td>HNRPA2</td>
<td>Heterogeneous nuclear ribonucleoprotein A/B</td>
</tr>
<tr>
<td></td>
<td>IFRD1</td>
<td>Interferon-related developmental regulator</td>
</tr>
<tr>
<td></td>
<td>SNX3</td>
<td>Sorting nexin 3</td>
</tr>
<tr>
<td></td>
<td>GSTM2</td>
<td>Glutathione S-transferase M2 (muscle)</td>
</tr>
<tr>
<td></td>
<td>TXN2</td>
<td>Thioredoxin 2</td>
</tr>
<tr>
<td></td>
<td>TXN2</td>
<td>Thioredoxin 2</td>
</tr>
<tr>
<td></td>
<td>NDUFA8</td>
<td>NADH dehydrogenase (ubiquinone) I alpha subcomplex, 8, 19kDa</td>
</tr>
<tr>
<td></td>
<td>GABARAPL1</td>
<td>GABA(A) receptor-associated protein like 1</td>
</tr>
<tr>
<td></td>
<td>SCD</td>
<td>Stearyl-CoA desaturase (delta-9 desaturase)</td>
</tr>
<tr>
<td></td>
<td>TXNIB</td>
<td>Tenascin X</td>
</tr>
<tr>
<td></td>
<td>LFITM3</td>
<td>Similar to Interferon-induced transmembrane protein 3</td>
</tr>
</tbody>
</table>

The three columns show the number of datasets in which that gene was deemed significant, the human gene name and a brief description, respectively. The top gene is the all-trans-retinol 13,14-reductase (RETSAT), which was significant in 8 of the 16 datasets.

3.2 A list of common differentially expressed genes

A total of 68 comparisons were carried out in the 16 datasets. For each dataset, a liberal threshold was applied to define statistical significance (see Methods for details); if a gene reached statistical significance in any of the comparisons in a dataset, it was deemed significant in the dataset. Table 2 shows the main result of the analysis. The most frequently significant gene was retinol saturase (all-trans-retinol 13,14-reductase, RETSAT), which was significant in eight of the 16 datasets. Five genes (KPNB1, SDHB, MRPL34, GPX3, PAM) were significant in seven of the 16; another 20 were significant in six of the 16. If we rank the genes not by the number of datasets in which it was significant but by the mean statistic across all comparisons, RETSAT is ranked at number 2, while PAM (which was significant in seven of 16) becomes number 1 (list not shown).

To determine just how unlikely it was to find the number of genes differentially expressed in common across the number of conditions shown in Table 2, we permuted the phenotypic labels of the data sets 30,000 times and calculated the number of differentially expressed genes shared across conditions. This allowed us to calculate P-values for the number of genes shared in conditions in Table 2. These are shown in Table 3.

Table 3. P-values for number of differentially expressed genes shared across DGAP experiments

<table>
<thead>
<tr>
<th>Number of datasets in which a gene is significant</th>
<th>Number of genes</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>1</td>
<td>7.52 × 10⁻⁹</td>
</tr>
<tr>
<td>7</td>
<td>21</td>
<td>2.25 × 10⁻⁷</td>
</tr>
<tr>
<td>6</td>
<td>122</td>
<td>5.59 × 10⁻⁶</td>
</tr>
<tr>
<td>5</td>
<td>371</td>
<td>9.75 × 10⁻⁵</td>
</tr>
<tr>
<td>4</td>
<td>1072</td>
<td>0.0013</td>
</tr>
<tr>
<td>3</td>
<td>1958</td>
<td>0.050</td>
</tr>
<tr>
<td>2</td>
<td>2061</td>
<td>0.41</td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

These P-values were estimated from the distributions obtained from 30,000 permutations. In each permutation, the phenotypic labels within each of the 16 experiments were randomized, lists of differentially expressed genes were generated, and the results were combined across data sets to generate the null distribution.

When we first described these results to the DGAP External Advisory Board, they were quite intrigued given the then-recent publication by Moise et al. (2008) of RBP4 which had revealed the retinol signaling pathway as important in modulating insulin diversity of study designs and underlying models, this approach is robust to the varying qualities of the data sets. If one data set is not informative, its effect on the overall conclusion is minimal.
sensitivity in several in vitro and in vivo models. As shown in Figure 2, the expression of RETSAT is also very consistently up-regulated across a variety of mouse and human models of insulin resistance and down-regulated across models of active adipogenesis. Subsequent genome-wide analyses by others (see ‘Discussion’ section) have further supported the role of RetSat as another member of the retinol pathway responsible in part for insulin sensitivity and adipogenesis (Schupp et al., 2009).

The next most widely differentially expressed genes are KPNB1, SDHB, MRPL34, GPX3, PAM. Of these genes, variants have been implicated in several pathological processes but are not implicated in processes dysregulated in obesity, insulin resistance and diabetes (see ‘Discussion’ section). However, GPX3 has recently been identified as reducing extracellular hydrogen peroxide levels causing insulin resistance in skeletal muscle cells (Chung et al., 2009). In these experiments, GPX3 expression prevented the antioxidant effects of the thiazolidine oral hypoglycemic agents on insulin action.

3.2.1 Pathway analysis The marked significance of RetSat up-regulation is not directly obvious from a pathway analysis. Indeed, in the original Gene Set Enrichment Analysis publication (Mootha et al., 2003), the retinol metabolism pathway was the lowest ranked pathway. To obtain a perspective on which processes are most shared across the DGAP experiments, we took the set of 520 genes differentially expressed in four or more data set and calculated the enrichment of Gene Ontology (GO) labels using the DAVID program (Dennis et al., 2003) with the results shown in Table 4. The table highlights the well-known perturbation of oxidative phosphorylation and energetics in insulin resistant states,
As the amount of data from expression profiling studies has increased to the one in Rhodes (Ramaswamy to extract a signature, which was then validated in other data sets. In other cases, analysis of combined data used one dataset as well as various mouse models in addition to other variables. here, which are more heterogeneous, encompassing human samples context of cancer studies in an attempt to identify a set of genes that 2007). Analysis using multiple data sets has been done mostly in the increasingly important, particularly in the context of a multiplicity of experiments with non-overlapping results (Ioannidis, & Porte, 2005).

Table 4. Pathway analysis—pathways implicated by the 420 genes differentially expressed in at least four experiments

<table>
<thead>
<tr>
<th>GO term</th>
<th>Count</th>
<th>Set size</th>
<th>P</th>
</tr>
</thead>
<tbody>
<tr>
<td>BP GO:0006091 generation of precursor metabolites and energy</td>
<td>47</td>
<td>649</td>
<td>1.97E-09</td>
</tr>
<tr>
<td>BP GO:0051186 cofactor metabolic process</td>
<td>23</td>
<td>236</td>
<td>4.44E-07</td>
</tr>
<tr>
<td>BP GO:0006732 coenzyme metabolic process</td>
<td>20</td>
<td>197</td>
<td>1.63E-06</td>
</tr>
<tr>
<td>BP GO:0009060 aerobic respiration</td>
<td>9</td>
<td>41</td>
<td>1.14E-05</td>
</tr>
<tr>
<td>BP GO:0057272 regulation of cell cycle</td>
<td>33</td>
<td>529</td>
<td>1.76E-05</td>
</tr>
<tr>
<td>BP GO:0022402 cell cycle process</td>
<td>41</td>
<td>749</td>
<td>2.96E-05</td>
</tr>
<tr>
<td>BP GO:0006119 oxidative phosphorylation</td>
<td>13</td>
<td>115</td>
<td>6.32E-05</td>
</tr>
<tr>
<td>BP GO:0007259 JAK-STAT cascade</td>
<td>8</td>
<td>43</td>
<td>1.37E-04</td>
</tr>
<tr>
<td>BP GO:0044248 cellular catalytic process</td>
<td>33</td>
<td>596</td>
<td>1.70E-04</td>
</tr>
<tr>
<td>BP GO:0007243 protein kinase cascade</td>
<td>25</td>
<td>393</td>
<td>1.78E-04</td>
</tr>
<tr>
<td>BP GO:0044262 cellular carbohydrate metabolic process</td>
<td>23</td>
<td>350</td>
<td>2.17E-04</td>
</tr>
<tr>
<td>BP GO:0006118 electron transport</td>
<td>28</td>
<td>480</td>
<td>2.66E-04</td>
</tr>
<tr>
<td>BP GO:0006084 acetyl-CoA metabolic process</td>
<td>7</td>
<td>38</td>
<td>4.91E-04</td>
</tr>
<tr>
<td>BP GO:0009059 macromolecule biosynthetic process</td>
<td>43</td>
<td>913</td>
<td>4.97E-04</td>
</tr>
<tr>
<td>BP GO:0045786 negative regulation of progression through cell cycle</td>
<td>16</td>
<td>209</td>
<td>5.59E-04</td>
</tr>
<tr>
<td>BP GO:0051187 cofactor catalytic process</td>
<td>7</td>
<td>39</td>
<td>5.68E-04</td>
</tr>
<tr>
<td>CC GO:0005739 mitochondrion</td>
<td>83</td>
<td>963</td>
<td>1.46E-22</td>
</tr>
<tr>
<td>CC GO:0044429 mitochondrial part</td>
<td>56</td>
<td>523</td>
<td>2.71E-19</td>
</tr>
<tr>
<td>CC GO:0005740 mitochondrial envelope</td>
<td>40</td>
<td>381</td>
<td>1.74E-13</td>
</tr>
<tr>
<td>CC GO:00031966 mitochondrial membrane</td>
<td>39</td>
<td>363</td>
<td>1.79E-13</td>
</tr>
<tr>
<td>CC GO:0019866 organelle inner membrane</td>
<td>35</td>
<td>303</td>
<td>4.90E-13</td>
</tr>
<tr>
<td>CC GO:0031967 organelle envelope</td>
<td>45</td>
<td>559</td>
<td>3.90E-11</td>
</tr>
<tr>
<td>CC GO:0031975 envelope</td>
<td>45</td>
<td>561</td>
<td>4.35E-11</td>
</tr>
<tr>
<td>CC GO:0044455 mitochondrial membrane part</td>
<td>17</td>
<td>115</td>
<td>4.03E-08</td>
</tr>
<tr>
<td>CC GO:0005759 mitochondrial matrix</td>
<td>20</td>
<td>171</td>
<td>8.94E-08</td>
</tr>
<tr>
<td>CC GO:00031980 mitochondrial lamina</td>
<td>20</td>
<td>171</td>
<td>8.94E-08</td>
</tr>
<tr>
<td>CC GO:0003279 ribosomal subunit</td>
<td>15</td>
<td>141</td>
<td>1.66E-05</td>
</tr>
<tr>
<td>CC GO:0042579 microbody</td>
<td>12</td>
<td>92</td>
<td>2.37E-05</td>
</tr>
<tr>
<td>CC GO:0005777 peroxisome</td>
<td>12</td>
<td>92</td>
<td>2.37E-05</td>
</tr>
</tbody>
</table>

Shown are the top ranked (by P-value) pathways based on the 520 genes differentially expressed in common across four or more DGAP experiments. Also shown are the number of genes in that GO set measured by each microarray platform ('Set size') and the overlap between the GO category genes and the differentially expressed genes ('Count'). BP and CC denote 'Biological Processes' and 'Cellular Components' in the GO classification. Gene sets with more than 1000 genes were considered non-specific and were eliminated from the list.

4 DISCUSSION

As the amount of data from expression profiling studies has increased in recent years, meta-analysis of multiple data sets has become increasingly important, particularly in the context of a multiplicity of underpowered experiments with non-overlapping results (Ioannidis, 2007). Analysis using multiple data sets has been done mostly in the context of cancer studies in an attempt to identify a set of genes that are consistently dysregulated (Rhodes et al., 2002) across similar datasets. In other cases, analysis of combined data used one dataset to extract a signature, which was then validated in other data sets (Ramaswamy et al., 2003). The approach in this article is similar to the one in Rhodes et al. (2004), but it is more robust to the data here, which are more heterogeneous, encompassing human samples as well as various mouse models in addition to other variables.

The top ranked gene in this integrative analysis, RetSat, is another member of a growing number of genes in the retinol pathway implicated in insulin sensitivity and resistance. Mouse RetSat catalyzes the saturation of the C13–C14 double bond of all-trans-retinol to produce all-trans-13,14-dihydroretinol. RetSat is expressed in adipose tissue and therefore may result in conversion of an inhibitor of adipose differentiation, all-trans-retinol, into a much weaker inhibitor of differentiation (Moise et al., 2008). Furthermore, this year, Schupp et al. (2009) independently demonstrated through a genome-wide Chromatin Immunoprecipitation on chip (ChIP-chip) assay of PPARγ an important target in intron 1 of RetSat in an adipocyte in vitro system. Furthermore, PPARγ [repeatedly implicated in obesity and diabetes (Bell et al., 2005; Zeggini et al., 2007)] was shown to regulate RetSat expression in adipocytes, and loss of RetSat impairs adipocyte differentiation. Schupp et al. (2009) found that, contrary to their expectations, there was decreased expression of RetSat in obese mice possibly related to the increased insulin sensitivity of adipocytes during expansion of adipose tissue (as compared to older hypertrophic adipocytes). These findings are mirrored in the results.
shown in Figure 2, and were captured by the metric of shared differential expression across multiple experiments. As shown in Table 3, one gene being differentially expressed across eight of the DGAP experiments by chance was extremely unlikely (P-value of 7.52 × 10^{-12}).

The next most widely differentially regulated genes across the various DGAP conditions include KPNB1, SDHB, MRPL34, GPX3 and PAM (in 7 of 16 conditions). One of these, GPX3 (glutathione peroxidase), is highly correlated in expression with RETSAT across multiple tissues in the Gene Expression Omnibus (GEO) (Barrett et al., 2005) as measured on the Affymetrix HG-U133 plus 2.0 platform (calculations not shown). Whether this implicates GPX3 in the retinol pathway remains to be determined. As noted previously GPX3 was nonetheless implicated this year in the handling of oxidative stress in muscle cells leading to insulin resistance (Chung et al., 2009). Although these top-ranked genes appear to hit the mark, they are differentially expressed in no more than half the DGAP experiments.

Determining the extent to which various mouse models correctly capture the features of the diseases they are supposed to mimic is difficult. Comparison of expression profiles between a murine model and human tumors has been used to resolve this issue previously for lung cancer (Sweet-Cordero et al., 2005). In the instance of insulin resistance and diabetes, our results here indicate the presence of some of the common features between human samples and mouse models. That is, assumptions made here regarding the existence of common end-point of a multiplicity of etiologies of diabetes and obesity across organisms have made the triangulation of molecular signatures across heterogeneous experiments a productive effort.

ACKNOWLEDGEMENTS

The authors thank Dr M.E. Pati and Dr R. Kahn for their leadership of the DGAP activities that made this work possible.

Funding: National Institutes of Health Roadmap for Medical Research, grant U54LM008748 to P.P. and I.K.

Conflict of Interest: none declared.

REFERENCES


