BIOINFORMATICS ORIGINAL PAPER Vol. 25 no. 23 2009, pages 3121–3127
doi:10.1093/bioinformatics/btp559

Gene expression

Integration of heterogeneous expression data sets extends the

Peter J. Park[1,2,3,], Sek Won Kong[1,4], Toma Tebaldi[5,6], Weil R. Lai[3], Simon Kasif[1,7,8] and Isaac S. Kohane[1,2,3,*]

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

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

Advance Access publication September 28, 2009

Associate Editor: Joaquin Dopazo

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 Dlabetes 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 inheritable component, this has not been defined in the vast majority of cases.

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

*To whom correspondence should be addressed.

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 extend previous findings that implicate 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 unlike 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 are 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 data are integrated.

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 sufficient 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

ID	Sample size	Array type	Description
1	27	MG_U74A/B/C	3T3-L1 fibroblast cells, 3T3-L1 adipocyte cells and mouse skeletal
2	28	MG_U74Av2	Brown preadipocyte IRS knockout profiling
3	22	MOE430A/B, MG-U74Av2/B/C	3T3-L1 adipocyte differentiation—time course
4	14	MG_U74Av2	Low versus high fat diet on mice of two genetic backgrounds (B6 versus 129)—fat
5	16	MG_U74Av2	Low versus high fat diet on mice of two genetic backgrounds (B6 versus 129)—liver
6	17	MG_U74Av2	Low versus high fat diet on mice of two genetic backgrounds (B6 versus 129)—skeletal muscle
7	18	MG_U74Av2	Isolated adipocytes from normal and fat insulin receptor KO (FIRKO) mice sorted into small and large cells
8	6	MG_U74Av2	Liver—ob/ob mice
9	21	Hu6800	Human skeletal muscle—type 2 diabetes and family history positive individuals—Mexican American
10	9	MG_U74Av2	Mouse skeletal muscle—controls, streptozotocin diabetes and insulin treated
11	12	HG-U133A/B	Human pancreatic islets from normal and Type 2 diabetic subjects
12	21	MG_U74Av2	Transcription profiling of wild type and PGC-1alpha KO liver and skeletal muscle
13	12	MG_U74Av2	Effect of PGC-1alpha and PGC-1beta on gene expression in myocytes and hepatocytes
14	57	MG_U74Av2	IR and IRS-1, single/double het KO—age and genetic background—epididymal white fat
15	55	MG_U74Av2	IR and IRS-1, single/double het KO—age and genetic background—liver
16	52	MG_U74Av2	IR and IRS-1, single/double het KO—age and genetic background—skeletal muscle
17	12	MG_U74Av2	Effect of insulin infusion on skeletal muscle
18	44	MG_U74Av2	Skeletal muscle—muscle IR KO and control mice—control, streptozotocin diabetic and insulin treated
19	54	HG-U133A	Human skeletal muscle—type 2 diabetes—Swedish males

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

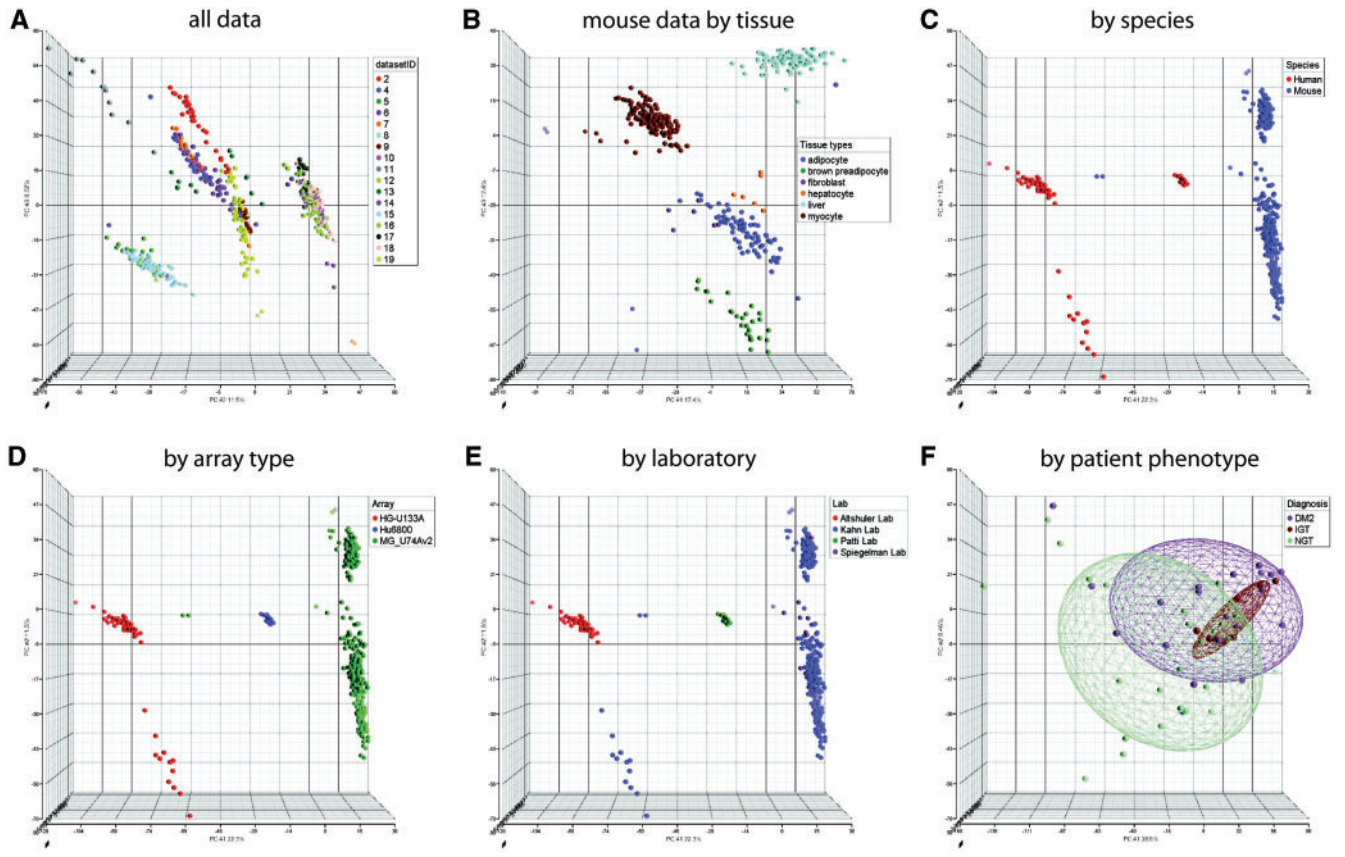

Fig. 1. Sample characteristics and systematic differences in principal component spaces for DGAP experiments. All ~450 samples are shown in (**A**), colored differently for the 17 studies in the combined data set. Systematic differences include differences across murine tissue types (**B**), species (**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

Number of datasets	Gene name	Description
8/16	RETSAT(FLJ20296)	All-*trans*-retinol 13,14-reductase
7/16	KPNB1	Karyopherin (importin) beta 1
	SDHB	Succinate dehydrogenase complex, subunit B, iron sulfur (Ip)
	MRPL34	Mitochondrial ribosomal protein L34
	GPX3	Glutathione peroxidase 3 (plasma)
	PAM	Peptidylglycine alpha-amidating monooxygenase
6/16	ACTN3	Actinin, alpha 3
	CPT1A	Carnitine palmitoyltransferase 1A (liver)
	RFX1	Regulatory factor X, 1 (influences HLA class II expression)
	TSTA3	Tissue specific transplantation antigen P35B
	UQCRC1	Ubiquinol–cytochrome *c* reductase core protein I
	DDX3X	DEAD (Asp–Glu–Ala–Asp) box polypeptide 3, X-linked
	DCTN6	Dynactin 6
	TRAPPC4	Trafficking protein particle complex 4
	TGFB1I4	Transforming growth factor beta 1 induced transcript 4
	HNRPAB	Heterogeneous nuclear ribonucleoprotein A/B
	IFRD1	Interferon-related developmental regulator 1
	SNX3	Sorting nexin 3
	GSTM2	Glutathione S-transferase M2 (muscle)
	TBX2	T-box 2
	TXN2	Thioredoxin 2
	NDUFA8	NADH dehydrogenase (ubiquinone) 1 alpha subcomplex, 8, 19 kDa
	GABARAPL1	GABA(A) receptor-associated protein like 1
	SCD	Stearoyl-CoA desaturase (delta-9-desaturase)
	TNXB	Tenascin XB
	LFITM3	Similar to Interferon-induced transmembrane protein 3

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.

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.

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 in 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

Number of datasets in which a gene is significant	*P*	Number of genes
8	7.52×10^{-9}	1
7	2.25×10^{-7}	5
6	5.59×10^{-6}	21
5	9.75×10^{-5}	122
4	0.0013	371
3	0.013	1072
2	0.090	1958
1	0.41	2061

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

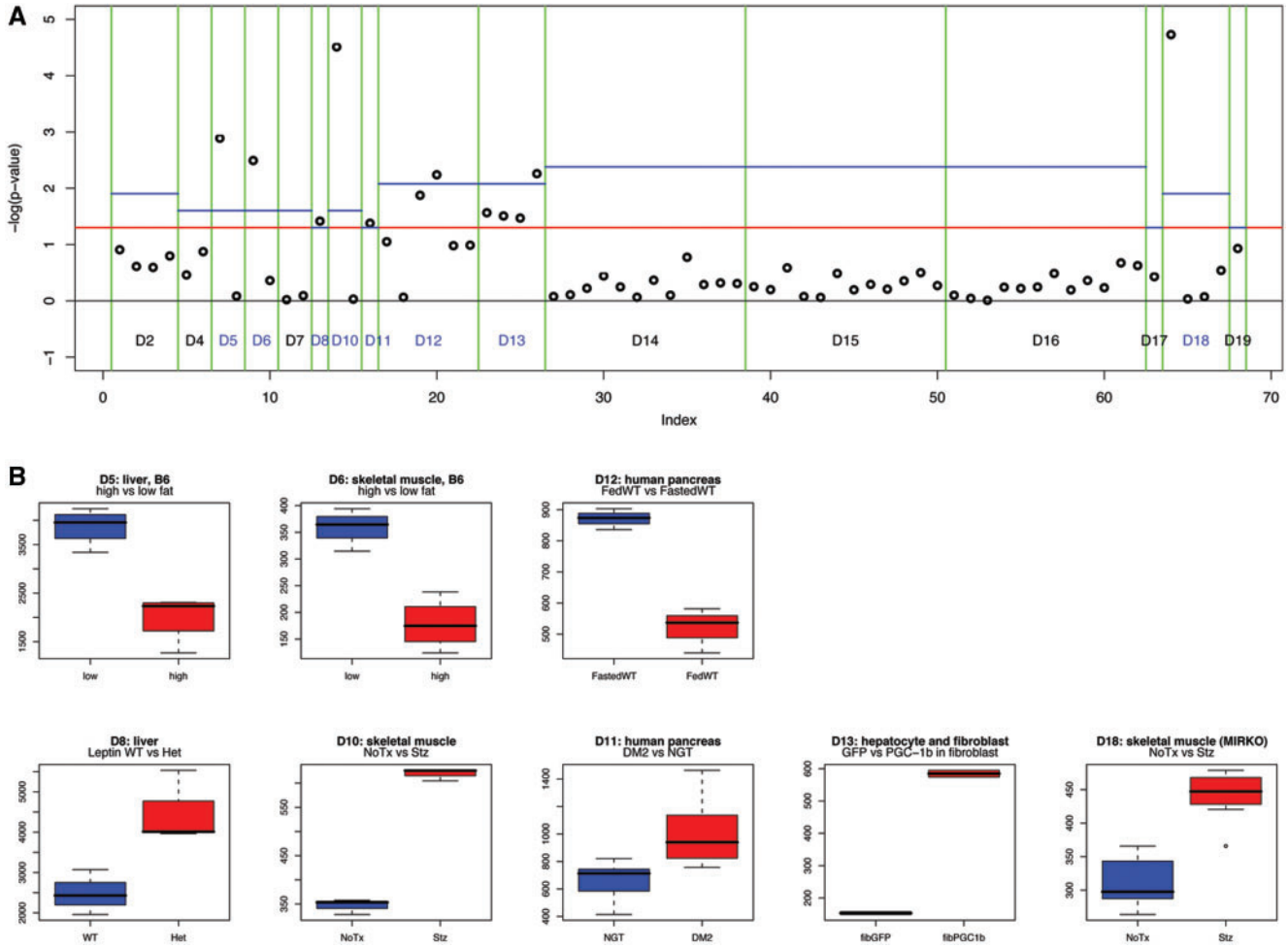

Fig. 2. Profiles of Retinol saturase (all-*trans*-retinol 13,14-reductase) transcript. (**A**) Negative log(*P*-value) for the RetSat transcript (FLJ20296) across all 68 comparisons in 16 data sets (data sets 1, 3, 9 were not included in our analysis—see 'Methods' section). Many studies have a complex design with multiple groups, which results in multiple comparisons. The red horizontal line indicates $p = 0.05$; the blue horizontal line indicates the *P*-value threshold adjusted for multiple comparisons within each data set using the Bonferroni correction; the green vertical lines divide the comparisons into those belonging to different data sets. The data set labels correspond to the experiment numbers in Table 1, with the blue label indicating the data sets in which at least one comparison was statistically significant by the threshold after multiple-testing adjustment. RETSAT is significant in eight data sets. (**B**) Boxplots of gene expression levels in each of the eight data sets with significant differential expression. Insulin resistant states are colored red. The eight data sets were divided into models of adipogenesis (top row) and models of chronic obesity and/or insulin resistance (bottom row).

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,

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

GO term	Count	Set size	*P*
BP GO:0006091 generation of precursor metabolites and energy	47	649	1.97*E*–09
BP GO:0051186 cofactor metabolic process	23	236	4.44*E*–07
BP GO:0006732 coenzyme metabolic process	20	197	1.63*E*–06
BP GO:0009060 aerobic respiration	9	41	1.14*E*–05
BP GO:0051726 regulation of cell cycle	33	529	1.78*E*–05
BP GO:0022402 cell cycle process	41	749	2.96*E*–05
BP GO:0006119 oxidative phosphorylation	13	115	6.32*E*–05
BP GO:0007259 JAK-STAT cascade	8	43	1.37*E*–04
BP GO:0044248 cellular catabolic process	33	596	1.70*E*–04
BP GO:0007243 protein kinase cascade	25	393	1.78*E*–04
BP GO:0044262 cellular carbohydrate metabolic process	23	350	2.17*E*–04
BP GO:0006118 electron transport	28	480	2.66*E*–04
BP GO:0006084 acetyl-CoA metabolic process	7	38	4.91*E*–04
BP GO:0009059 macromolecule biosynthetic process	43	913	4.97*E*–04
BP GO:0045786 negative regulation of progression through cell cycle	16	209	5.59*E*–04
BP GO:0051187 cofactor catabolic process	7	39	5.68*E*–04
CC GO:0005739 mitochondrion	83	963	1.46*E*–22
CC GO:0044429 mitochondrial part	56	523	2.71*E*–19
CC GO:0005740 mitochondrial envelope	40	381	1.74*E*–13
CC GO:0031966 mitochondrial membrane	39	363	1.79*E*–13
CC GO:0019866 organelle inner membrane	35	303	4.90*E*–13
CC GO:0031967 organelle envelope	45	559	3.90*E*–11
CC GO:0031975 envelope	45	561	4.35*E*–11
CC GO:0044455 mitochondrial membrane part	17	115	4.03*E*–08
CC GO:0005759 mitochondrial matrix	20	171	8.94*E*–08
CC GO:0031980 mitochondrial lumen	20	171	8.94*E*–08
CC GO:0033279 ribosomal subunit	15	141	1.66*E*–05
CC GO:0042579 microbody	12	92	2.37*E*–05
CC GO:0005777 peroxisome	12	92	2.37*E*–05

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.

particularly changes localized in mitochondria. It also includes the perturbation of the JAK-STAT pathway that has also been identified as perturbed in diabetes, obesity and insulin resistance (Schwartz and Porte, 2005).

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 Immunoprecipation 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^{-9}).

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 and across organisms have made the triangulation of molecular signatures across heterogeneous experiments a productive effort.

ACKNOWLEDGEMENTS

The authors thank Dr M.E. Patti 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

Affymetrix (2005). Technical Note: Guide to Probe Logarithmic Intensity Error (PLIER) Estimation. Affymetrix, Inc, Santa Clara, CA.

Barrett,T. *et al*. (2005) NCBI GEO: mining millions of expression profiles—database and tools. *Nucleic Acids Res.*, **33**, D562–D566.

Bell,C. *et al*. (2005) The genetics of human obesity. *Nat. Rev. Genet.*, **6**, 221–234.

Breitling,R. *et al*. (2004) Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. *FEBS Lett.*, **573**, 83–92.

Chung,S.S. *et al*. (2009) Glutathione peroxidase 3 mediates the antioxidant effect of peroxisome proliferator-activated receptor gamma in human skeletal muscle cells. *Mol. Cell. Biol.*, **29**, 20–30.

Dennis,G. Jr *et al*. (2003) DAVID: database for annotation, visualization, and integrated discovery. *Genome Biol.*, **4**, P3.

Hwang,K.B. *et al*. (2004) Combining gene expression data from different generations of oligonucleotide arrays. *BMC Bioinformatics*, **5**, 159.

Ioannidis,J.P. (2007) Is molecular profiling ready for use in clinical decision making? *Oncologist*, **12**, 301–311.

Kuo,W.P. *et al*. (2002) Analysis of matched mRNA measurements from two different microarray technologies. *Bioinformatics*, **18**, 405–412.

Lowell,B.B. and Shulman,G.I. (2005) Mitochondrial dysfunction and type 2 diabetes. *Science*, **307**, 384–387.

Moise,A.R. *et al*. (2008) Stereospecificity of retinol saturase: absolute configuration, synthesis, and biological evaluation of dihydroretinoids. *J. Am. Chem. Soc.*, **130**, 1154–1155.

Mootha,V.K. *et al*. (2003) PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat. Genet.*, **34**, 267–273.

Nimgaonkar,A. *et al*. (2003) Reproducibility of gene expression across generations of Affymetrix microarrays. *BMC Bioinformatics*, **4**, 27.

Patti,M.E. *et al*. (2003) Coordinated reduction of genes of oxidative metabolism in humans with insulin resistance and diabetes: potential role of PGC1 and NRF1, *Proc. Natl Acad. Sci. USA*, **100**, 8466–8471.

Ramaswamy,S. *et al*. (2003) A molecular signature of metastasis in primary solid tumors. *Nature Genetics*, **33**, 49–54.

Rhodes,D.R. *et al*. (2002) Meta-analysis of microarrays: interstudy validation of gene expression profiles reveals pathway dysregulation in prostate cancer. *Cancer Res.*, **62**, 4427–4433.

Rhodes,D.R. *et al*. (2004) Large-scale meta-analysis of cancer microarray data identifies common transcriptional profiles of neoplastic transformation and progression. *Proc. Natl Acad. Sci. USA*, **101**, 9309–9314.

Schupp,M. *et al*. (2009) Retinol saturase promotes adipogenesis and is downregulated in obesity. *Proc. Natl Acad. Sci. USA*, **106**, 1105–1110.

Schwartz,M. and Porte,D. (2005) Diabetes, obesity, and the brain. *Science*, **307**, 375–379.

Sweet-Cordero,A. *et al*. (2005) An oncogenic KRAS2 expression signature identified by cross-species gene-expression analysis. *Nature Genetics*, **37**, 48–55.

Wild,S. *et al*. (2004) Global prevalence of diabetes: estimates for the year 2000 and projections for 2030. *Diabetes Care*, **27**, 1047–1053.

Yang,Q. *et al*. (2005) Serum retinol binding protein 4 contributes to insulin resistance in obesity and type 2 diabetes. *Nature*, **436**, 356–362.

Yechoor,V.K. *et al*. (2004) Distinct pathways of insulin-regulated versus diabetes-regulated gene expression: an in vivo analysis in MIRKO mice. *Proc. Natl Acad. Sci. USA*, **101**, 16525–16530.

Zeggini,E. *et al*. (2007) Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. *Science*, **316**, 1336–1341.