Integrating siRNA and protein-protein interaction data to identify an expanded insulin signaling network

Zhidong Tu[1], Carmen Argmann[1], Kenny K. Wong[2], Lyndon J. Mitnaul[2], Stephen Edwards[1], Iliana C. Sach[1], Jun Zhu [1], Eric E. Schadt [1§]

[1]Rosetta Inpharmatics, a whole subsidiary of Merck & Co., Inc., 401 Terry Ave N, Seattle, Washington
[2]Department of Cardiovascular Disease, Merck Research Laboratories, Rahway, New Jersey

[§]Corresponding author

Corresponding email address
Eric Schadt: eric.schadt@gmail.com

Supplementary Materials

1. Supplementary Box
2. Supplementary Figures
3. Supplementary Tables
4. Supplementary Methods

Supplementary Box 1 - Pruning algorithm

1. Set all non-hit nodes in the current network as unvisited
2. If there is one or more unvisited non-hit node, randomly select one and mark it as visited. Otherwise, stop the pruning.
3. Check the node selected in step 2. If at least one neighbor is a hit, go to step 2. Otherwise, go to step 4.
4. Remove this node from network. Check if all the hit genes are still connected, if not so, undo the remove and go to step 2. Otherwise, go to step 5.
5. Check through all the unvisited non-hit genes and find those that are no longer connected to positive hit genes. Remove these nodes from the network if there are any. Go to step 2.

Supplementary Figures

Causal gene selection

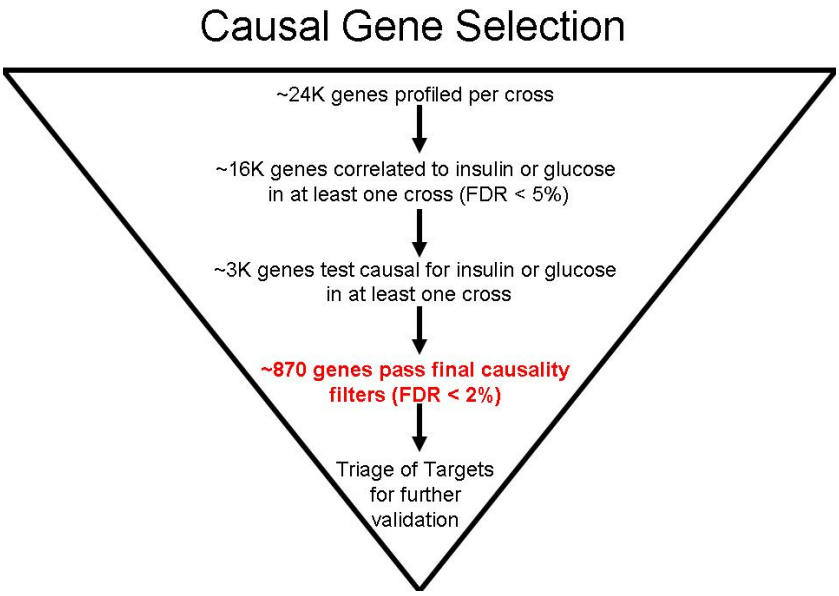


Figure S-1a. The strategy of selecting ~870 causal genes from BxH mouse cross.

Figure S-1b

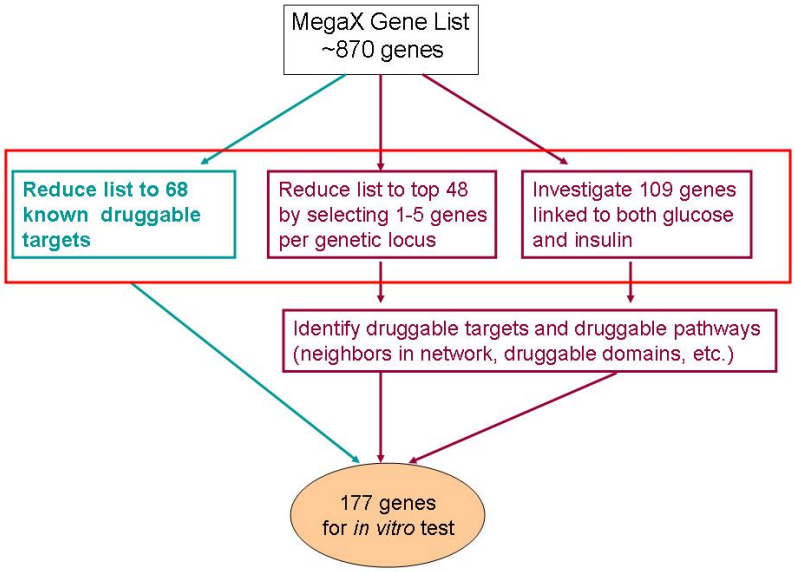
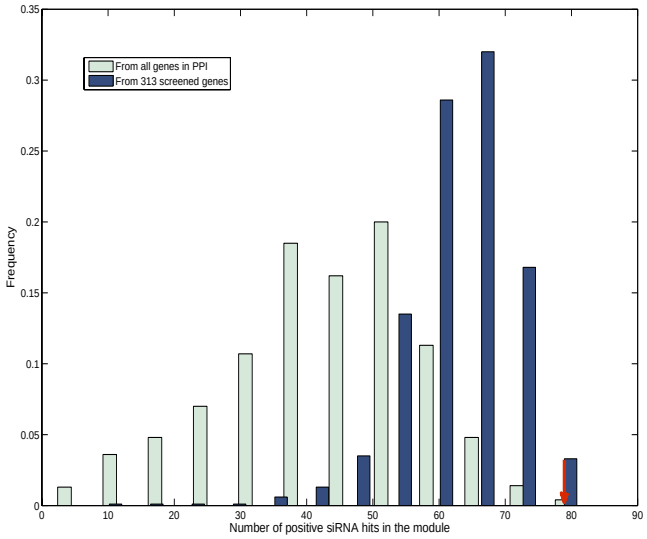
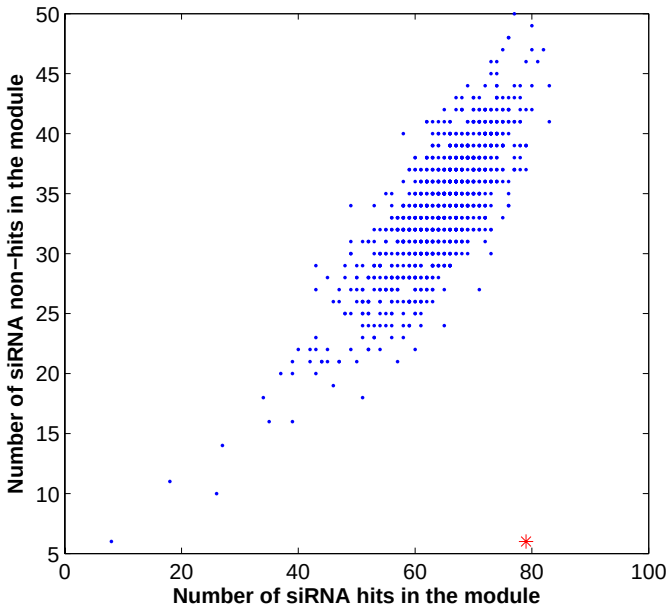


Figure S-1b. The filtering used to select 177 genes from the ~870 causal gene list.

a)

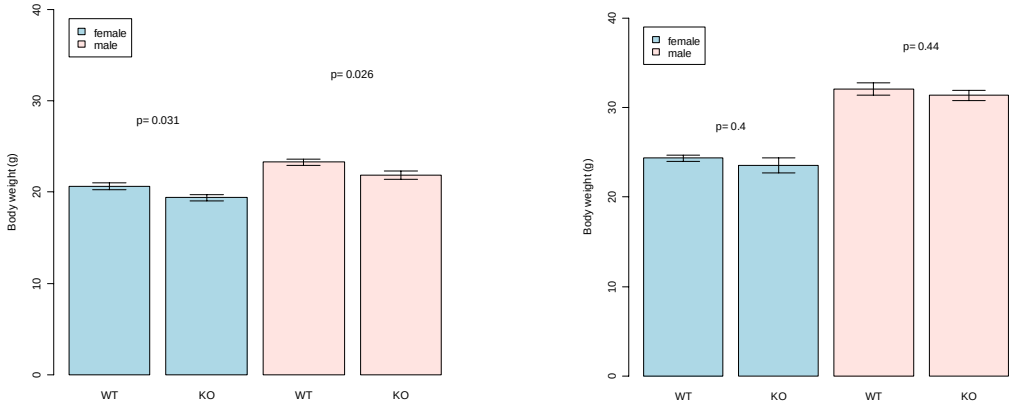


b)

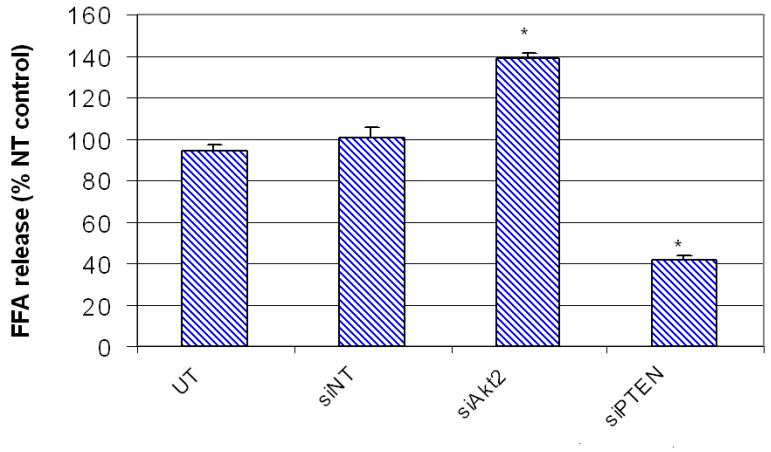


Supplementary Figure 2. Permutation tests assess the significance of the PEXA module. **A)** Two types of permutation tests were carried out to assess whether the PEXA module was enriched for siRNA hit genes, as described in the text. In the first test the null distribution was estimated by randomly selecting 313 genes from the set of genes comprising the PPI network (green bars). In the second test the null distribution was

estimated by randomly sampling the 126 siRNA hit list genes from the set of 313 genes that was screened (blue bars). In both cases the observed number of siRNA hit list genes in the PEXA module (red arrow) was significantly more than observed in the permuted data. **B)** Scatter plot of the distribution of the number of siRNA non-hit nodes contained in the PEXA module vs. the number of siRNA hit nodes in the network. The blue dots were generated from the second permutation test described in **A).** The red asterisk indicates the results from the PEXA module applied to the observed data. These results highlight that the siRNA hits are significantly supported by the independently generated KEGG and PPI data, confirming their coherence for insulin signaling pathways.



Supplementary Figure 3. Body weights of wildtype and *S1pr2*-/- mice at a) age of wk 8 and b) age of wk 20.



Supplementary Figure 4. Free fatty acid release measured from adipocytes which were untransfacted (UT), transfacted with scrambled siRNA (siNT), with *Akt2* siRNA and with *Pten* siRNA.

Supplementary Tables

Gene selection for siRNA screen

Table S1. We selected 313 genes for siRNA screen. 876 genes were obtained as causal for diabetes related traits using an integrative causality test as previously described (Schadt 2005). 177 were further selected from these genes based on certain internal criteria such as protein functionality, druggability, etc. The rest of the list was taken from several sources, which included manually selected genes by Merck scientists, genes related to fatty acid beta-oxidation (FAO), orphan peptidases (OPI) based on mouse genetics, historical microarray data and external information.

	Numb. of Genes	siRNA hits	Fraction (%)
Causality test	177	78	44.1
FAO	18	6	33.3
OPI	50	14	28.0
Manually picked	68	28	41.2
Total*	313	126	40.3

Test upon expression signature gene sets

Table S2. List of the four knock out mouse gene expression signature sets as measured from adipose tissue. These expression signature sets were performed with PEXA module and KEGG insulin signaling pathway gene set. For all the KO experiments, expression signature gene sets were significantly more enriched in PEXA modules genes than KEGG insulin pathway genes. For the KO phenotype, we only listed a representative one and omitted the rests.

KO Gene	Sig size[1]	OL Ins PW[2]	E-value[3]	OL PEXA[4]	E-value	KO Phenotype[5]
Alox5	4947	44	5.4×10^{-3}	72	1.9×10^{-7}	abnormal cholesterol level
Cbr1	7493	60	7.0×10^{-3}	93	1.5×10^{-6}	abnormal circulating insulin level
Lrp5	1075	11	0.19	26	2.1×10^{-6}	impaired glucose tolerance
Mc3r/Mc4r	5898	52	1.8×10^{-3}	85	6.5×10^{-9}	obese

[1] The gene expression signature size. Differentially expressed genes were defined by ANOVA with cutoff value of significance set to be <0.01.
[2] The number of overlapped genes between expression signature genes and the KEGG insulin signaling pathway genes.
[3] Corrected *P*-values based on Fisher Exact Test. The background was set to be 22,770 which is the number of all the genes represented on Rosetta/Merck Mouse 25k v1.4 microarray.
[4] The number of overlapped genes between expression signature genes and PEXA module genes.
[5] Knock out phenotypes were obtained from Mouse Genome Informatics database(Eppig et al. 2007).

Supplementary Methods

Considerations on siRNA hit rate

Hit rate for our screen was higher than several whole genome screen results reported elsewhere by other groups investigating different biological processes in different cell lines/organisms. One main reason is certainly the selection criteria for genes entered into the screen, but other factors likely contribute to this as well. These factors include the cellular process under investigation, the design of the assay system, e.g., cell line, siRNA vendor, assay protocol, readout measurement, and the threshold chosen for making hit calls, etc. As these factors are different for different experiments, these hit rates are not directly comparable (these single factors can easily cause several fold difference in hit rates).

On the other hand, we don't have gold standard sets of positive and negative control genes that do or do not modulate insulin dependent FFA release in 3T3-L1 adipocyte. Recent publications of our own and others point out that there are a lot more genes contribute to diabetes and obesity related phenotypes (Chen et al. 2008; Emilsson et al. 2008).

Given all the above, we did not explicitly estimate the false positive rate of the siRNA screening result. Instead, we first carefully checked our knockdown experiments to

ensure that for most genes, their mRNAs were knocked down to at least 55% of their normal levels using Taqman measurement. Second, we did perform several positive controls for the assay. As shown in Supplementary Figure 4, the FFA release significant changed in expected directions for *Akt2* and *Pten* where their mRNAs were knocked down by 57% and 69% as measured by Taqman analysis. FFA release did not change for untransfected cells or cells transfected with scrambled siRNA. For genes functions as insulin signaling activators (e.g., *Akt2*), the KD increases FFA release, and vice versa (e.g., *Pten*). Although we do not expect every screen result to be accurate, for those well annotated genes (other examples are *Insr*, *Atgl*), the siRNA screen results meet our expectation. Third, we tested and demonstrated that siRNA hits are informative based on permutation test and KEGG pathway enrichment test (see main text, Table 1). Fourth, we observed different hit rate for different groups of genes selected based on different criteria, shown in (Supplementary Table 1). The gene set selected based on causality has the highest positive hit rate, while the gene set selected based on orphan peptides and correlation with clinical traits has the lowest positive rate. The hit rate increases as the selection criteria gets more stringent. All of these suggest that the siRNA result contains valuable information and is suitable for further analysis.

Implementation of PEXA

The algorithm of PEXA was described in the main text and we list a few implementation details here.

1. Not every KEGG interaction is directed. If an edge does not have direction, PEXA treats them as bidirectional edges pointing in both directions in identifying seeding paths. We do not disallow such long tails in the seeding step, since there might be chance that siRNA hits interact with these long tails during expansion and we cannot tell in advance. However, if no siRNA hits interact with these long tails, they will be removed by the pruning.

2. In the graphical displays of pathways from KEGG database website, they frequently use a single symbol to represent an array of genes. For example, in the insulin signaling pathway, *AKT* actually represents *AKT1*, *AKT2*, and *AKT3*. (This can be seen by following the link http://www.kegg.com/dbget-bin/show_pathway?hsa04910 and then click on *AKT*, *AKT1*, *2*, and *3* will be displayed instead of one gene). When we drew the seeding paths (and the rest two networks), we displayed only one symbol for one node, since otherwise, some nodes names were too long to display well. When selecting symbols, we use the positive result if results are conflicting. However, internal to PEXA, we include all the genes and treat each gene as a single node.

Alox5-/- mouse construction and expression profiling

Alox5-/- mice were constructed and maintained at UCLA as previously described(Mehrabian et al. 2002). 20 female wild type and 5-lipoxygenase knockout mice

had been fed on a normal chow diet for 16 weeks and 5 animals from each group were sacrificed. Adipose tissues were collected and subjected to expression profiling. 5 normal mice's RNAs were pooled together as control, and were hybridized with each *Alox5*-/- mouse RNA extract on Rosetta Mouse 25k microarrays.

Cbr1-/- mouse construction and expression profiling

This experiment compares gene expression in iWAT from mice fed high fat chow and against a pool of normal mice also fed high fat chow. Fourteen week old C56Bl/6J mice maintained on high fat diet for six weeks (plus a vehicle control arm with animals maintained on regular chow) were sacrificed. iWAT were harvested for profiling.

Lrp5-/- mouse construction and expression profiling

Lrp5 knockout mice construction was previously reported(Fujino et al. 2003). *Lrp5* -/- mice were compared to *Lrp5* (+/-) mice. Samples were prepared separately for each tissue (liver, colon and fat) and sex; 2 pools of 3 animals for each sex and each of the 3 genotypes. A reference pool was prepared by pooling RNA from 6 heterozygotes.

Mc3r/M4r double knockout mouse construction and expression profiling

Both wildtype and *Mcr3/Mcr4* knockout male mice of ~6 month old were under diet induced obesity (DIO) for 2.5 months (weight of WT about 30g and knockout about 55g). iWAT tissues were collected at the time of sacrifice. Wildtype DIO mice were compared with DIO *Mc3r/Mc4r* knockout mice.

Expression signature gene set calculation

Differentially expressed genes were calculated using ANOVA test. Signature lists were selected by applying a cutoff *P*-value of < 0.01.

Reference:

Chen, Y., J. Zhu, P.Y. Lum, X. Yang, S. Pinto, D.J. MacNeil, C. Zhang, J. Lamb, S. Edwards, S.K. Sieberts, A. Leonardson, L.W. Castellini, S. Wang, M.-F. Champy, B. Zhang, V. Emilsson, S. Doss, A. Ghazalpour, S. Horvath, T.A. Drake, A.J. Lusis, and E.E. Schadt. 2008. Variations in DNA elucidate molecular networks that cause disease. *Nature* **452:** 429-435.

Emilsson, V., G. Thorleifsson, B. Zhang, A.S. Leonardson, F. Zink, J. Zhu, S. Carlson, A. Helgason, G.B. Walters, S. Gunnarsdottir, M. Mouy, V. Steinthorsdottir, G.H. Eiriksdottir, G. Bjornsdottir, I. Reynisdottir, D. Gudbjartsson, A. Helgadottir, A. Jonasdottir, A. Jonasdottir, U. Styrkarsdottir, S. Gretarsdottir, K.P. Magnusson, H.

Stefansson, R. Fossdal, K. Kristjansson, H.G. Gislason, T. Stefansson, B.G. Leifsson, U. Thorsteinsdottir, J.R. Lamb, J.R. Gulcher, M.L. Reitman, A. Kong, E.E. Schadt, and K. Stefansson. 2008. Genetics of gene expression and its effect on disease. *Nature* **452:** 423-428.

Eppig, J.T., J.A. Blake, C.J. Bult, J.A. Kadin, J.E. Richardson, and G. the Mouse Genome Database. 2007. The mouse genome database (MGD): new features facilitating a model system. *Nucl. Acids Res.* **35:** D630-637.

Fujino, T., H. Asaba, M.-J. Kang, Y. Ikeda, H. Sone, S. Takada, D.-H. Kim, R.X. Ioka, M. Ono, H. Tomoyori, M. Okubo, T. Murase, A. Kamataki, J. Yamamoto, K. Magoori, S. Takahashi, Y. Miyamoto, H. Oishi, M. Nose, M. Okazaki, S. Usui, K. Imaizumi, M. Yanagisawa, J. Sakai, and T.T. Yamamoto. 2003. Low-density lipoprotein receptor-related protein 5 (LRP5) is essential for normal cholesterol metabolism and glucose-induced insulin secretion. *Proceedings of the National Academy of Sciences of the United States of America* **100:** 229-234.

Mehrabian, M., H. Allayee, J. Wong, W. Shih, X.-P. Wang, Z. Shaposhnik, C.D. Funk, and A.J. Lusis. 2002. Identification of 5-Lipoxygenase as a Major Gene Contributing to Atherosclerosis Susceptibility in Mice. *Circ Res* **91:** 120-126.

Schadt, E.E. 2005. An integrative genomics approach to infer causal associations between gene expression and disease. *Nature Genet.* **37:** 710-717.