T2DM-GeneMiner a web resource for meta-analysis and marker identification for type 2 diabetes mellitus

Rasche, Axel; Herwig, Ralf
Max Planck Institute for Molecular Genetics,
Department Vertebrate Genomics, Bioinformatics Group,
Ihnestr. 63-73, D-14195 Berlin
rasche@molgen.mpg.de, herwig@molgen.mpg.de

Abstract

Background

Multiple functional genomics data for complex human diseases have been published and made available by researchers worldwide. Main goal of these studies is the detailed analysis of a particular aspect of the disease. Recently, meta-analysis approaches have been published that try to extract meaningful disease genes and networks by integrating and combining these individual studies using bioinformatics strategies.

Results

Here we report on a meta-analysis approach that combines high-throughput data of heterogeneous origin in the domain of type 2 diabetes mellitus, in particular in connection with obesity as a risk factor. Different data sources such as DNA microarrays, ChIP on chip and qualitative data from multiple tissues from human and mouse are integrated and validated by a scoring system in order to assign disease relevance to the genes. Using a random sampling approach we computed a set of 213 genes most relevant for obesity-induced type 2 diabetes mellitus. Furthermore, we extrapolated functional information on cellular networks associated with these genes such as pathway information, protein-protein interactions and gene regulatory networks. In order to allow users to derive type 2 diabetes mellitus relevance for any given gene we have set up a web interface that allows the screening of the gene in the light of the underlying data.

Conclusion

Using a simple scoring algorithm we computed a core set of 213 genes that show significant disease relevance in the data sets under study. These genes have been further validated in the functional context of networks and exhibit high potential for understanding diabetic pathways and pathway cross-talk. Our web resource allows the user to access the information that was gathered and to assess disease relevance for any human or mouse gene. Thus, we conclude that our study is a valuable resource for diabetes research and a template for meta-analysis studies in other disease domains.

1. Data Assembly and Annotation

The focus of this study is obesity induced type 2 diabetes mellitus. Thus, data sets focus on experimental studies of that background. In our analysis we combine heterogeneous data from diverse sources that target different levels of cellular information. Disease specific sources are binary and transcriptome data. The annotation spans ChIPonChip experiments and databases for pathways, protein-protein-interaction and more.

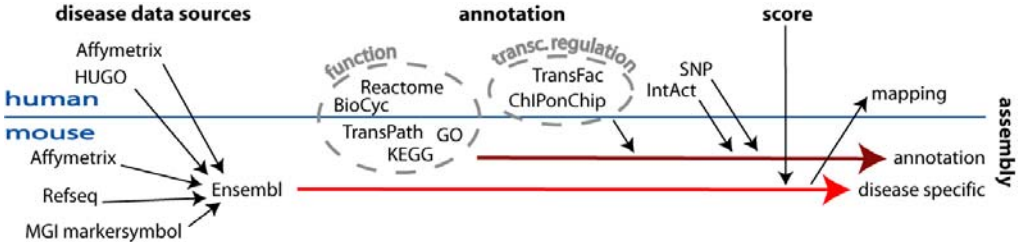

Fig. 1: Initially our study was based on mouse data. However, in order to enhance disease relevance further human data sets were added and mapped via sequence homology. Retrieving knowledge from different sources requires the mapping and synchronization of different gene identifiers (ID). The different IDs are mapped on their mouse Ensembl gene ID. Mapping and merging of the information has been done in R/BioConductor. To ease the access for researchers we add the more informative identifier like HUGO IDs, Entrez and RefSeq.

In each of 21 sources information is bound to an ID. Relate the information from the different sources necessitates the assignment of the diverse IDs. Complexity reduction is achieved by choosing a refence ID, in our case Ensembl mouse gene ID. The assignments are generated for each source ID to the common reference ID avoiding the challenge of the harmonisation of all different IDs. BioMart, a companion of the Ensembl database, is the resource to retrieve the assignments. These are generated by sequence homology. Assigning has to be done along several axes. First axis are the orthologs between the species human, mouse and rat and the second axis gene-transcript-protein. Even in one axis different concepts have to be unified, e.g. the Ensembl or Entrez gene concept and the concept of Affymetrix probe sets.

This assignment leads into technical difficulties. It is no mapping in the mathematical sense, e.g. a gene ID can lead into several transcript ID. We distinguish four cases in the mapping. The first case are indeed the bijective mappings. In the second the source ID ends in two target IDs. The information is replicated. In the third case two source ID join one target ID and we must merge information. The method of merging can for example be pasting, calculating the mean or the maximum, depending on the study goal. The fourth case misses a target ID for the source ID and the corresponding information is lost. In summary consolidating the biochemical information demands assigning and is rarely possible without losses.

2. Scoring Disease Genes

Basically, three types of information were considered for scoring genes – binary counts, gene expression fold changes and information on transcriptional regulation. For each group of information we summarise the scores of the individual experiments.

Gene expression is quantitative and thus is weighted with a formula covering different measures. The score reflects the strength of the effect with the fold change. Up- or downregulation is equalised by taking the logarithm of the fold change. A two-fold change in expression is understood in general to mark a substantial change so we take the logarithm to the basis two. Unsound measurements are downweighted by the standard error of the replicates. Expression changes are only sensitive, if the gene is expressed and the presence tag is considered analogous to the standard error. The total score was computed as the sum across all studies. Using simple arithmetical calculations we derived a robust formula. This smooth approach circumvents the need to generate differentially expressed gene for the single studies and conserves most of the quantitative information. To avoid the assumption of a specific distribution of the score, we sampled a background distribution. The sampled distribution determines a cut off for a candidate gene set, facilitating the subsequent analyses.

A meta-analysis has also the potential to discriminate between general and study specific genes. The score is joined by an entropy derived criterion. The entropy is calculated across the scores and conform scores result in high entropy. Consistently altered genes result in high entropy and strong but specific alteration in low entropy, e.g. by tissue or species specificity.

Our broad approach with a restrictive cut off favors genes with consistent or very strong alteration. The score rates the disease relevance of a gene and the entropy rates the generality.

Proposing new evaluation methods and respective results entails the need of validation. The resulting candidate gene list is compared with the hypergeometric distribution to available T2DM knowledge with three approaches. Medical reviews define first gene sets to compare with the candidate set. The reviews are sources at the same time and we use a leave-one-out way. The second approach is to appraise the overlap with other proposed candidate lists. In third instance the top pathways uprated by the enrichment are ‘*type II diabetes mellitus*’, ‘*Adipocytokine signaling*’ and ‘*Insulin signaling pathway*’ well related to T2DM. Thus we link our results to the well-described actors.

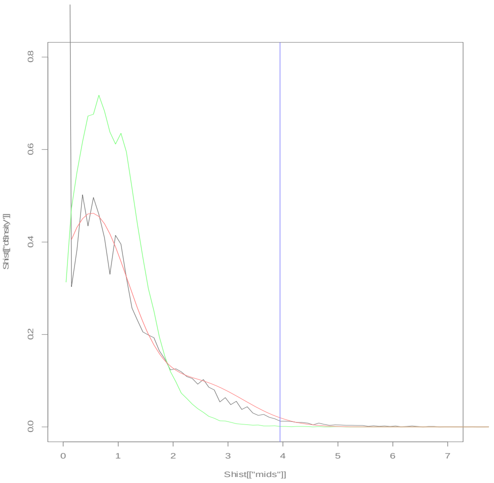

Fig. 2: In order to determine whether a certain score shows significant deviations from a random distribution, a sampled distribution has been generated by drawing from the original data. Those genes were assigned as “significant” that are above the 0.001-quantile of the sampled distribution. This determines a cut-off score value of 3.9. A set of 235 significant genes has a score exceeding the cut-off.

3. Identification of Disease Related Networks

Disease related networks were investigated with three different types of network information – biological pathways, protein-protein interaction networks and transcriptional regulation graphs. These networks define – by annotation – groups of associated genes. Using enrichment analysis based on the hypergeometric distribution (Klipp, Herwig et al. 2005) we were able to assign each annotation item (pathway, transcription target set etc.) a P-value that reflects the enriched occurrence of potential marker genes from our analysis.

P-value	pathway description
9.38E-12	Insulin signaling pathway
6.38E-08	Type II diabetes mellitus
2.86E-06	Adipocytokine signaling pathway
2.97E-05	Focal adhesion

4.03E-05	Starch and sucrose metabolism
5.90E-05	Cell Communication
0.000134286	Metabolism of xenobiotics by cytochrome P450
0.000151979	Glutathione metabolism

Table 1: The table shows the first eleven pathways from the KEGG database (Kanehisa and Goto 2000) with their respective P-values.

Beside the well-described interactions in the pathway databases we use IntAct (Hermjakob, Montecchi-Palazzi et al. 2004) to derive putative type 2 diabetes mellitus subnets.

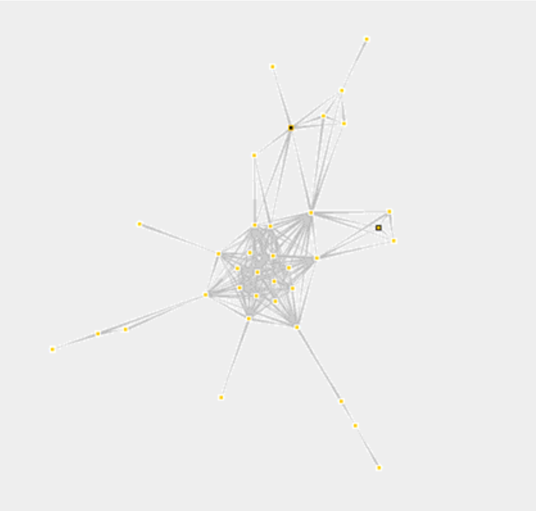

Fig. 3: From 36 marker genes of 235, there are interactions stored in IntAct. The upper picture shows the interaction network of these 36 genes. Apparently there is a highly interconnected network of 18 genes. This cluster stems from a synapse experiment in mouse brain not related to type 2 diabetes mellitus. Protein-protein-interaction information still has a strong bias to the underlying experiments.

A different type of networks has been assessed with the transcriptional regulatory network. Although the information about transcriptional interaction is still sparse, it is already possible to illustrate parts of the transcriptional net.

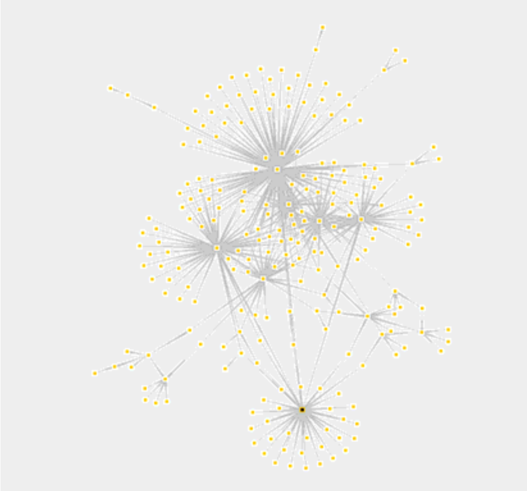

Fig. 4: Transcriptional network of three different transcriptional regulation sets: The liver core regulators (Odom, Dowell et al. 2006) and the intersection of their targets with the significant set of genes, six transcription

factors from TransFac (Matys, Kel-Margoulis et al. 2006) found in the significant set and their targets and the transcription factors with a significant P-value in the enrichment described in the right box with all their targets. The regulators related to the significant genes are running out of the core regulation group with mostly a few targets.

4. References

4.1 acknowledgments

Christoph Wierling and Atanas Kamburov retrieved and edited the pathway data. Thomas Kreitler set up the web-access.
Axel Rasche receives a PhD scholarship from the Max-Planck-Society. Ralf Herwig is paid by the Max-Planck-Society.

4.2 references

Al-Hasani, H. and H. G. Joost (2005). “Nutrition-/diet-induced changes in gene expression in white adipose tissue.” Best Pract Res Clin Endocrinol Metab 19(4): 589-603.
Ashburner, M., C. A. Ball, et al. (2000). “Gene Ontology: tool for the unification of biology.” Nature Genetics 25: 25-29.
Biddinger, S. B., K. Almind, et al. (2005). “Effects of diet and genetic background on sterol regulatory element-binding protein-1c, stearoyl-CoA desaturase 1, and the development of the metabolic syndrome.” Diabetes 54(5): 1314-23.
Birney, E., D. Andrews, et al. (2006). “Ensembl 2006.” Nucleic Acids Res 34(Database issue): D556-61.
Chen, X. a. C., S.W. and Pannell, L.K. and Hess, S. (2005). “Quantitative Proteomic Analysis of the Secretory Proteins from Rat Adipose Cells Using a 2D Liquid Chromatography-MS/MS Approach.” J. Proteome Res. 4(2): 570-577.
Dean, L. and J. McEntyre (2004). The Genetic Landscape of Diabetes, NCBI.
Durinck, S., Y. Moreau, et al. (2005). “BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis.” Bioinformatics 21(16): 3439-3440.
Gentleman, R. C., V. J. Carey, et al. (2004). “Bioconductor: Open software development for computational biology and bioinformatics.” Genome Biology 5: R80.
Gunton, J. E., R. N. Kulkarni, et al. (2005). “Loss of ARNT/HIF1beta mediates altered gene expression and pancreatic-islet dysfunction in human type 2 diabetes.” Cell 122(3): 337-49.
Hermjakob, H., L. Montecchi-Palazzi, et al. (2004). “IntAct: an open source molecular interaction database.” Nucleic Acids Res 32(Database issue): D452-5.
Jackson Labs. (2005). “Human Disease and Mouse Model Detail for NIDDM.”
Joshi-Tope, G., M. Gillespie, et al. (2005). “Reactome: a knowledgebase of biological pathways.” Nucl. Acids Res. 33(suppl 1): 428-432.
Kanehisa, M. and S. Goto (2000). “KEGG: Kyoto Encyclopedia of Genes and Genomes.” Nucleic Acids Research 28(1): 27-30.
Klipp, E., R. Herwig, et al. (2005). Systems Biology in Practice, Wiley-VCH.
Krull, M., S. Pistor, et al. (2006). “TRANSPATH: an information resource for storing and visualizing signaling pathways and their pathological aberrations.” Nucleic Acids Res 34(Database issue): D546-51.
Lan, H., M. E. Rabaglia, et al. (2003). “Gene Expression Profiles of Nondiabetic and Diabetic Obese Mice Suggest a Role of Hepatic Lipogenic Capacity in Diabetes Susceptibility.” Diabetes 52(3): 688-700.

Matys, V., O. V. Kel-Margoulis, et al. (2006). “TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes.” Nucleic Acids Res 34(Database issue): D108-10.

Mootha, V. K., C. M. Lindgren, et al. (2003). “PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes.” Nat Genet 34(3): 267-73.

mult. (2002). Diabetes Genome Anatomy Project.

mult. (2005). Superarray, Bioscience Corporation, Superarray, Bioscience Corporation.

Nadler, S. T., J. P. Stoehr, et al. (2000). “The expression of adipogenic genes is decreased in obesity and diabetes mellitus.” PNAS 97(21): 11371-11376.

Nandi, A., Y. Kitamura, et al. (2004). “Mouse models of insulin resistance.” Physiol Rev 84(2): 623-47.

Odom, D. T., R. D. Dowell, et al. (2006). “Core transcriptional regulatory circuitry in human hepatocytes.” Mol Syst Biol 2: 2006 0017.

Odom, D. T., N. Zizlsperger, et al. (2004). “Control of pancreas and liver gene expression by HNF transcription factors.” Science 303(5662): 1378-81.

OMIM. (2000, 04.10.2005). “Online Mendelian Inheritance in Man, OMIM (TM).” from http://www.ncbi.nlm.nih.gov/omim/

R Development Core Team (2005). R: A Language and Environment for Statistical Computing. Vienna, Austria, R Foundation for Statistical Computing.

Romero, P., J. Wagg, et al. (2004). “Computational prediction of human metabolic pathways from the complete human genome.” Genome Biology 6(1:R2): 17.

Stumvoll, M., B. J. Goldstein, et al. (2005). “Type 2 diabetes: principles of pathogenesis and therapy.” The Lancet 365: 1333-1346.