A Network-Based Meta-analysis Strategy for the Selection of Potential Gene Modules in Type 2 Diabetes

Ronnie Alves[1], Marcus Mendes[2], and Diego Bonnato[2]

[1] Vale Institute of Technology, Belém, Brazil
ronnie.alves@vale.com
[2] Federal University of Rio Grande do Sul, Porto Alegre, Brazil
diego@cbiot.ufrgs.br, cla_atm_milo@hotmail.com

Abstract. We propose an integrative network-based meta-analysis strategy to enable the selection of potential gene markers for one of the most prevalent diseases worldwide, Type 2 diabetes (T2D), formally known as the non-insulin dependent diabetes mellitus. Comprehensive elucidation of the genes regulated through this disorder and their wiring will provide a more complete understanding of the overall gene network topology and their role in disease progression and treatment. The proposed strategy was able to find conservative gene modules which play interesting role in T2D, pointing to gene markers such as NR3C1, ADIPOR1 and CDC123. Network-based meta-analysis by enumerating conserved gene modules pave a practical approach to the identification of candidate gene markers across several related transcriptomic studies. The NEMESIS *R* pipeline for network-based meta-analysis is also provided.

Keywords: gene co-expression network analysis, candidate gene markers, type 2 diabetes, meta-analysis, system biology.

1 Introduction

Type 2 diabetes (T2D), formally known as the non-insulin dependent diabetes mellitus, is the most common type of diabetes. It has been taking the quality of an endemic disease, affecting more than 170 millions of people around the world. In fact, there is tragic trend that by the end of 2030 more than 300 millions of people will develop T2D.

In T2D, either the body does not produce enough insulin or the cells do not respond to the insulin.The effects of insulin, insulin deficiency and insulin resistance vary according to the physiological function of the organs and tissues concerned, and their dependence on insulin for metabolic processes. Those tissues defined as insulin dependent, based on intracellular glucose transport, are mainly adipose tissue and muscle. Although insulin resistance occurs in most obese individuals, diabetes is usually forestalled through compensation with increased insulin. This increase in insulin occurs through an expansion of beta-cell mass

J.C. Setubal and N.F. Almeida (Eds.): BSB 2013, LNBI 8213, pp. 160–169, 2013.

and/or increased insulin secretion by individual beta-cells. Failure to compensate for insulin resistance leads to T2D [1–3].

T2D is a multifactorial disease caused by both oligo- and polygenic genetic factors as well as non-genetic factors that result from a lack of balance between the energy intake and output and other life style related factors. There is a plenty of data related to the genetics of T2D. Though, many genes and gene products as well as their interactions with the environment at the molecular, cellular, tissue, and the whole organism levels are still unknown. Understanding of diabetes pathogenesis is critical to the development of new strategies for effective prevention and treatment of this disease [4–6].

With the advance of high-throughput technologies, it is now possible to get deep insight into the orchestration of the complex biological functions either activated or not during disease development. Genome-wide association studies (GWASs) have discovered association of several loci with Type 2 diabetes. In such studies, gene expression profiles are evaluated from a set of several associated studies where genes presenting stable modulation patterns across these studies could be potential markers. However in most cases they do not take into account the network interaction of the genes involved in the correlated T2D pathways [7].

In the present work we propose a network-based meta-analysis strategy for the selection of potential gene regulatory modules in T2D. The basis to retrieve such potential gene modules relies on the proper inference of weighted gene co-expression networks from associated T2D transcriptomic studies. Next, gene modules identified among these studies are evaluated for functional enrichment of the conserved gene modules (*consensus cliques*).

The main contributions of the proposed strategy are:

- An empirical evaluation of weighted gene co-expression network on T2D studies;
- A new method for the localization of functional consensus gene modules through frequent pattern mining;
- The NEMESIS *R* pipeline for the selection of potential gene regulatory modules, revealing biological functions correlated to T2D pathogenesis.

The remainder of this paper is organized as follows. In Section 2 we present the transcriptomic data sets as well as the methods used by the proposed strategy. Next, in Section 3 we present the pipeline developed and the main results obtained. Conclusions and future work are provided in Section 4.

2 Materials and Methods

2.1 Transcriptomic Data Sets

We carefully selected four transcriptomic studies from the Gene Expression Omnibus (GEO) which were properly elaborated to measure T2D progression and

treatment. In the first dataset (GSE12389) diverse roles were demonstrated regarding interferon-gamma (IFN-γ) in the induction and regulation of immune-mediated inflammation using a transfer model of autoimmune diabetes [8]. The second dataset (GSE2253) investigated the molecular mechanism by which extracellular hIAPP mediates pancreatic beta-cell apoptosis [9]. It is known that extracellular hIAPP oligomers are toxic to pancreatic beta-cells and associated with apoptosis. The third study (GSE12639) highlights genetic regulatory mechanisms in the remote zone of left ventricular (LV) free wall in order to partly explain the more frequent progression to heart failure after acute myocardial infarction (AMI) in diabetic rats [10]. And finally, the fourth dataset (GSE13270) studied Type 2 diabetes progression and the development of insulin resistance in two animal models with and without a high fat diet superimposed on these models [11].

2.2 Preprocessing Affymetrix Data

The Affymetrix expression data (CEL files) were preprocessed using the Robust Multi-array Average (RMA) normalization approach through the rma() function in the *affy* R package. RMA employs quantile normalization and smooths technical sources of variability across samples. Next, only expressed genes were selected for further analysis. A gene (probe) was considered expressed if it was called (P)resent or (M)arginal in at least 75% of all samples in a given dataset. Present and Marginal calls were determined by the mas5calls() function in the *affy* R package. A detection call answers the question: *Is the transcript of a particular gene Present or Absent?* In this context, absent means that the expression level is below the threshold of detection. That is, the expression level is probably not different from zero. In the case of an uncertainty, we can get a marginal call. It is important to note that some probe-sets are more variable than others, and the minimal expression level provably different from zero may range from a small value to very large value (for a noisy probe-set). The advantage of asking the question in this way without actual expression values is that the results are easy to filter and to interpret. For example, we may only want to look at genes whose transcripts are detectable in a particular experiment. Given that co-expressed gene networks are created based on correlation metrics over the study, such fllter strategy allow us to remove potential inconsistencies in the Chip. Thus, it was used a cutoff of 75% with only one particular exception for the GSE12389 study (50%). Table 1 presents the results of the preprocessing step over all selected studies.

2.3 Calculating Candidate Gene Modules

The network topology of each study was calculated through the application of the WGCNA R package [12], and various soft-thresholding powers were properly applied to find a good fitness of the scale-free topology. The soft-thresholding strategy, adopted by WGCNA, keeps all possible links and raises the original

Table 1. Data preprocessing of Affymetrix datasets

Study	Samples	Genes before	Genes after	Affy.Chip
GSE12389	8	45101	5316	Mouse430_2
GSE2253	20	22690	11686	Mouse430A
GSE12639	12	31099	14998	Rat230_2
GSE13270	101	31099	13935	Rat230_2

coexpression values to a power "beta" so that the high correlations are emphasized at the expense of low correlations. An example of the scale-free topology calculated for the GSE13270 study is presented in Figure 1.

Once the network has been constructed, module inference is the next step. Modules are defined as clusters of densely interconnected genes. WGCNA detects gene modules using unsupervised clustering, i.e. without the use of a priori defined gene sets. In fact, modules are calculated based on a topological overlap measure [13] that has been applied successfully in several applications. The user has a choice of several module detection methods. The default method is hierarchical clustering using the standard R function hclust, branches of the hierarchical clustering dendrogram correspond to modules and can be identified using one of a number of available branch cutting methods, for example the constant-height cut or two Dynamic Branch Cut methods. One drawback of hierarchical clustering is that it can be difficult to determine how many (if any) clusters are present in the data set. Although the height and shape parameters of the Dynamic Tree Cut method provide improved exibility for branch cutting and module detection, it remains an open research question how to choose optimal cutting parameters or how to estimate the number of clusters in the data set [12].

The final power threshold p calculated to each data set was defined as follows: GSE12389 (p=18), GSE2253 (p=9), GSE12639 (p=12) and GSE13270 (p=12). Next, modules for each study were extracted, and only significant intramodular (hub) genes at each module were selected for further analysis. We use the intramodular connectivity measure to define the most highly connected intramodular hub gene as the module representative. In fact, intramodular hub genes are highly correlated with the module eigengene. A gene in a module is considered significant if it has a strong p-value (< 0.001) membership, i.e., correlation between module eigengenes and expression values in the Chip (Figure 2). In Table 2 the effects of the calculation of the candidate genes are presented to each corresponding study.

2.4 Functional Enrichment Analysis

We determined the specific biological processes relevant for each candidate gene module by calculating GO terms and pathway enrichment. Furthermore, each module is also a generalized clique [13]. We obtained significant (p-value$<$0.05)

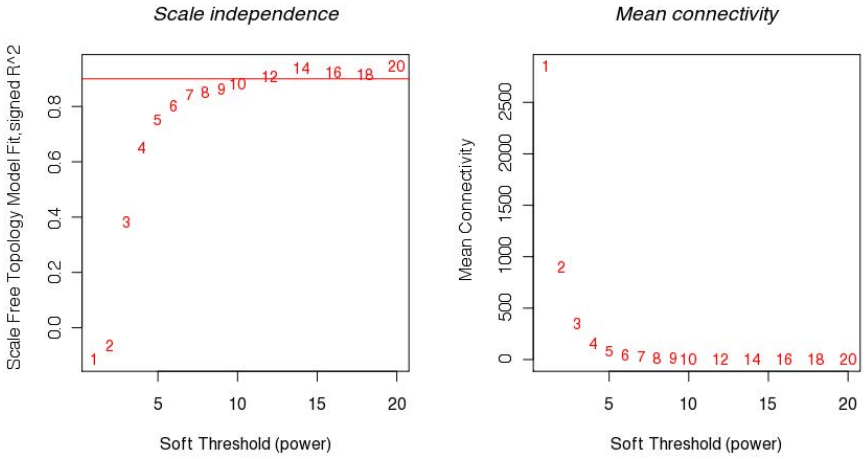

Fig. 1. Analysis of network topology for various soft-thresholding powers for the GSE13270 study. The left panel shows the scale-index (y-axis) as a function of the soft-thresholding power (x-axis). The right panel displays the mean connectivity degree (y-axis) as a function of the soft-thresholding power (x-axis).

Table 2. Data preprocessing of Affymetrix datasets

Study	Modules	Genes before	Genes after	Affy.Chip
GSE12389	12	5316	1566	Mouse430_2
GSE2253	17	11686	5570	Mouse430A
GSE12639	34	14998	5660	Rat230_2
GSE13270	16	13935	9836	Rat230_2

GO and pathway enrichment for all modules. The respective *Entrez* gene identification was obtained through the *biomaRt* R package. Next, we make use of the *GOstats* R package as well as the related *Affymetrix Chip Expression Set* annotation data to each associated organism.

2.5 Finding Consensus Modules

Consensus modules were detected by exploring the retrieved functional annotations and evaluating the co-occurrence of these annotations across the related T2D studies. Thus, whether a significant annotation is shared among distinct modules, across several networks, such observation could be seen as good indication of consensus. The intuition of exploring co-occuring gene sets has been explored broadly by the data mining community in several gene association studies [14]. However, as far as we are concerned, there is no direct reference of its utilization in network-based meta-analysis. In this work, we say that we

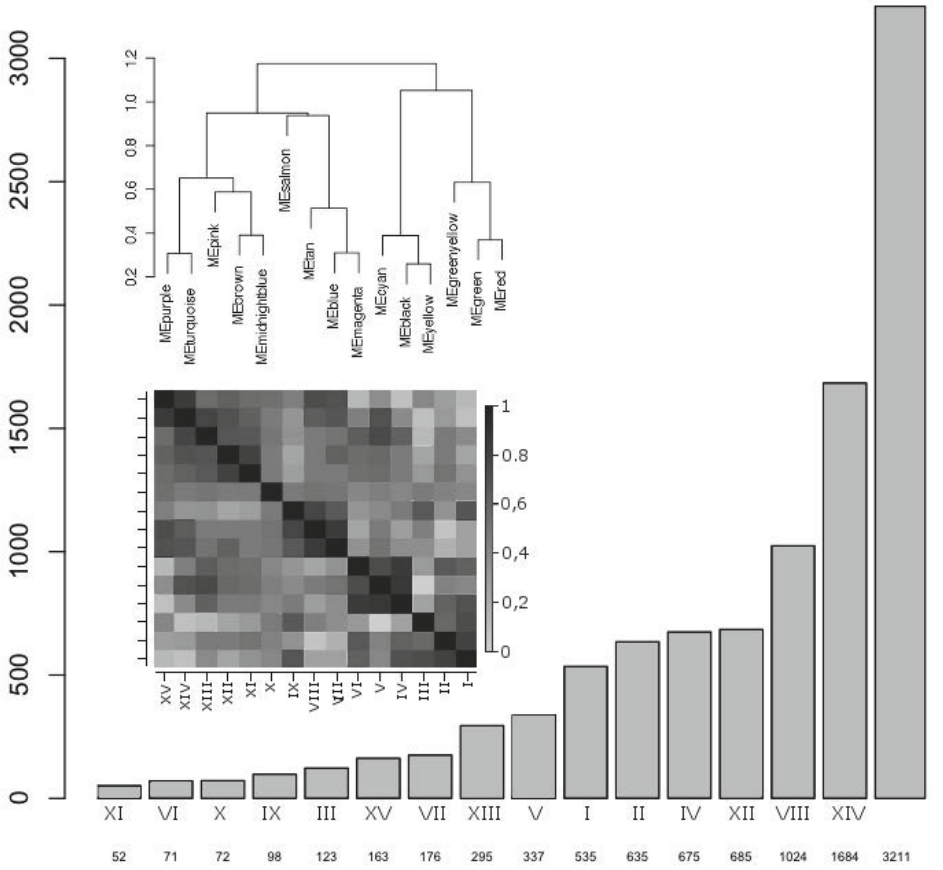

Fig. 2. Heatmap plot of the adjacencies in the eigengene network including the trait weight build on the GSE13270 study. Each row and column in the heatmap corresponds to one module eigengene (labeled by color) or weight. In the upper dendogram (heatmap), white color represents low adjacency (negative correlation), while black represents high adjacency (positive correlation). Squares of black color along the diagonal are the gene modules (Bar plots in the bottom). Genes that were not assigned to any modules were assigned to the largest Bar plot on the right.

have a consensus module when its associated annotation is shared across different studies. The consensus significance is measured by metrics like *support* and *confidence* of the co-occurring annotations. Therefore, before exploring such patterns we have to introduce two concepts called *Transactions* and *Item set* . Each transcriptomic study is related to one *transaction_id* and it is composed by several gene modules (i.e, either GO or KEGG annotations). An *item set* is an annotation (or a set of annotation) that appears in more that one study. Thus, if a particular annotation has a support of 75%, it does mean that this functional behavior is observed in three out of four related studies (see Transcriptomic data

sets). By using such consensus strategy we avoid the hard task to conciliate all different gene names in all distinct Affymetrix platforms and organisms, focusing on the search of functional gene modules closely related to T2D pathology.

2.6 Selection of Potential T2D Genes

Potential genes are those ones that are significantly covered by the enriched consensus modules. Since only the most conserved annotations are selected for further analysis, it was necessary to devise a reverse engineering approach for the identification of the consensus genes. Thus, we first selected all associated genes to the most relevant *GO terms* with its *EntrezGene* information. Further, this gene set was matched with the gene list obtained by the consensus analysis. For the GO information we used the functional annotation *GO.db* database, and for each organism its associated species annotation database. For instance to the *mus musculus* we selected the *org.Mm.eg.db.* We have also used the *Phenopedia* database [15] to evaluate the correspondence of the selected potential genes with the well-known T2D (*human*) genes induced by this database.

3 Results and Discussion

3.1 NEMESIS: The *NE*twork-Based *ME*ta-analy*SIS* Pipeline

Meta-analysis has been applied broadly in several disease studies to improve the search for potential gene markers. Network-based strategies highlight important gene regulatory modules, but it cannot ensure that such module(s) could be conserved along with other related studies. Therefore, one pontential alternative, as proposed here, is the combination (or meta-analysis) of several gene co-expression networks to pulling out gene modules providing relevant functional association to the disease phenotype. Once having all selected transcriptomic data sets for analysis the first step is the proper normalization procedure. Next, gene modules are enumerated by exploring network functions available in the WGCNA *R* package. The following task is then the functional enrichment analysis to retrieve significant gene annotations related to network modules. After getting these annotations, it is possible to search for the conserved biological functions by exploring frequent patterns on the annotated modules. Finally, candidate gene markers are evaluated through the mapping of the biological functions associated with pathology under investigation. The proposed pipeline is summarized in Figure 3. The *R* scripts and additional material are free available online at `https://sites.google.com/site/alvesrco/nemesis`.

3.2 Potential T2D Gene Markers

The presented strategy was able to enumerate potential gene markers correlated to T2D. For instance, the NR3C1 gene, being also a well-known product of a transcription factor highly associated to T2D.

Next, we highlight the main candidate genes retrieved by the NEMESIS pipeline applied on T2D transcriptomic data sets:

- The gene ADIPOR1 encodes a protein which acts as a receptor for adiponectin, a hormone secreted by adipocytes which regulates fatty acid catabolism and glucose levels. Binding of adiponectin to the encoded protein results in activation of an AMP-activated kinase signaling pathway which affects levels of fatty acid oxidation and insulin sensitivity. Patients who developed T2D present a low activity of this gene when compared with normal ones [16];
- The gene CDC123 encodes proteins highly associated to the production of insulin. Variations of this gene are also related to a low production of the hormone [17];
- The gene SERPINE1 encodes a member of the serine proteinase inhibitor (serpin) superfamily. This member is the principal inhibitor of tissue plasminogen activator (tPA) and urokinase (uPA), and hence is an inhibitor of fibrinolysis. Comparative proteomic profiling of plasma from individuals with either diabetes or obesity and individuals with both obesity and diabetes revealed SERPINE 1 as a possible candidate protein of interest, which might be a link between obesity and diabetes [3].

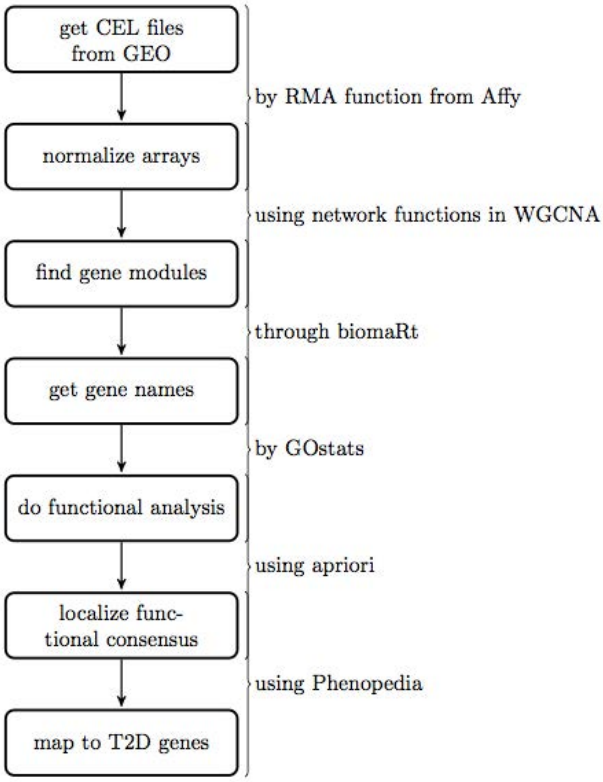

Fig. 3. The NEMESIS *R* pipeline to explore network-based meta-analysis on transcriptomic data

4 Conclusions

We have introduced a network-based meta-analysis approach to discover potential gene modules, biologically associated, to type 2 diabetes pathology. Though, we also envisage its application on other complex diseases, such as those ones with large collection of transcriptomic data available in the Phenopedia database. The NEMESIS *R* pipeline developed could be easily extended to explore gene markers on other diseases.

In the present study four transcriptomic studies were selected for the experimental analysis. Despite the good results, it would be interesting to explore the pipeline with more T2D data sets to increase significance of the discovered patterns. However, the more data we use the more preprocessing efforts are necessary in order to reduce data dimensionality, smoothing the creation of the associated gene co-expression networks. And consequently, the search for consensus gene modules.

There are plenty of open challenges with the network-based meta-analysis. We list the following directions to pursue in near future: i) a more compact and semi-automatic way to identify gene network modules (*cliques*), ii) an optimization procedure to calculate the thresholds for the most relevant annotation across studies, and iii) extend the pipeline to deal with RNA-Seq data.

Acknowledgements. We would like to thank reviewers for helpful suggestions and criticisms that have contributed to improve substantially the article. We also thanks Wallace Lira for the preparation of Figure 2. This work is partially supported by the Brazilian National Research Council (CNPq – *Universal calls*) under the BIOFLOWS project [475620/2012-7].

References

1. Liu, M., Liberzon, A., Kong, S.W., Lai, W.R., Park, P.J., Kohane, I.S., Kasif, S.: Network-based analysis of affected biological processes in type 2 diabetes models. PLoS Genet. 3(6), e96 (2007)
2. Stumvoll, M., Goldstein, B.J., van Haeften, T.W.: Type 2 diabetes: principles of pathogenesis and therapy. Lancet 365(9467), 1333–1346 (2005)
3. Kaur, P., Reis, M.D., Couchman, G.R., Forjuoh, S.N., Greene, J.F., Asea, A.: Serpine 1 links obesity and diabetes: A pilot study. J. Proteomics Bioinform. 3(6), 191–199 (2010)
4. Goh, K.I., Cusick, M.E., Valle, D., Childs, B., Vidal, M., Barabási, A.L.: The human disease network. Proc. Natl. Acad. Sci. U. S. A. 104(21), 8685–8690 (2007)
5. Keller, M.P., Choi, Y., Wang, P., Davis, D.B., Rabaglia, M.E., Oler, A.T., Stapleton, D.S., Argmann, C., Schueler, K.L., Edwards, S., Steinberg, H.A., Chaibub Neto, E., Kleinhanz, R., Turner, S., Hellerstein, M.K., Schadt, E.E., Yandell, B.S., Kendziorski, C., Attie, A.D.: A gene expression network model of type 2 diabetes links cell cycle regulation in islets with diabetes susceptibility. Genome Res. 18(5), 706–716 (2008)
6. Park, K.S.: Prevention of type 2 diabetes mellitus from the viewpoint of genetics. Diabetes Res. Clin. Pract. 66(suppl. 1), S33–S35 (2004)

7. Jain, P., Vig, S., Datta, M., Jindel, D., Mathur, A.K., Mathur, S.K., Sharma, A.: Systems biology approach reveals genome to phenome correlation in type 2 diabetes. PLoS One 8(1), e53522 (2013)
8. Calderon, B., Suri, A., Pan, X.O., Mills, J.C., Unanue, E.R.: Ifn-gamma-dependent regulatory circuits in immune inflammation highlighted in diabetes. J. Immunol. 181(10), 6964–6974 (2008)
9. Casas, S., Gomis, R., Gribble, F.M., Altirriba, J., Knuutila, S., Novials, A.: Impairment of the ubiquitin-proteasome pathway is a downstream endoplasmic reticulum stress response induced by extracellular human islet amyloid polypeptide and contributes to pancreatic beta-cell apoptosis. Diabetes 56(9), 2284–2294 (2007)
10. Song, G.Y., Wu, Y.J., Yang, Y.J., Li, J.J., Zhang, H.L., Pei, H.J., Zhao, Z.Y., Zeng, Z.H., Hui, R.T.: The accelerated post-infarction progression of cardiac remodelling is associated with genetic changes in an untreated streptozotocin-induced diabetic rat model. Eur. J. Heart Fail 11(10), 911–921 (2009)
11. Almon, R.R., DuBois, D.C., Lai, W., Xue, B., Nie, J., Jusko, W.J.: Gene expression analysis of hepatic roles in cause and development of diabetes in goto-kakizaki rats. J. Endocrinol. 200(3), 331–346 (2009)
12. Langfelder, P., Horvath, S.: Wgcna: an r package for weighted correlation network analysis. BMC Bioinformatics 9, 559 (2008)
13. Zhang, B., Horvath, S.: A general framework for weighted gene co-expression network analysis. Stat. Appl. Genet. Mol. Biol. 4, Article17 (2005)
14. Alves, R., Rodriguez-Baena, D.S., Aguilar-Ruiz, J.S.: Gene association analysis: a survey of frequent pattern mining from gene expression data. Brief. Bioinform. 11(2), 210–224 (2010)
15. Yu, W., Clyne, M., Khoury, M.J., Gwinn, M.: Phenopedia and genopedia: disease-centered and gene-centered views of the evolving knowledge of human genetic associations. Bioinformatics 26(1), 145–146 (2010)
16. Tomas, E., Tsao, T.S., Saha, A.K., Murrey, H.E., Zhang, C.C., Itani, S.I., Lodish, H.F., Ruderman, N.B.: Enhanced muscle fat oxidation and glucose transport by acrp30 globular domain: acetyl-coa carboxylase inhibition and amp-activated protein kinase activation. Proc. Natl. Acad. Sci. U. S. A. 99(25), 16309–16313 (2002)
17. Grarup, N., Andersen, G., Krarup, N.T., Albrechtsen, A., Schmitz, O., Jørgensen, T., Borch-Johnsen, K., Hansen, T., Pedersen, O.: Association testing of novel type 2 diabetes risk alleles in the jazf1, cdc123/camk1d, tspan8, thada, adamts9, and notch2 loci with insulin release, insulin sensitivity, and obesity in a population-based sample of 4,516 glucose-tolerant middle-aged danes. Diabetes 57(9), 2534–2540 (2008)