Biomolecular Systems of Disease Buried Across Multiple GWAS Unveiled by Information Theory and Ontology

Younghee Lee, PhD[1], Jianrong Li, MSc[1], Eric Gamazon, PhD[1], James L. Chen, MD[2], Anna Tikhomirov, PhD[1], Nancy J. Cox, PhD[1,3,*], and Yves A. Lussier, MD[1,4*]

[1]Sect. of Genetic Medicine, [2]Sect. of Hematology/Oncology; Dept. of Medicine; [3]Dept. of Human Genetics; [4]Inst. of Genomics and Systems Biology; Inst. for Translational Medicine; UC Cancer Research Center; Ludwig Center for Metastasis Research and Computational. Inst.; The University of Chicago, Chicago, IL, USA * Corresponding authors

Abstract

A key challenge for genome-wide association studies (GWAS) is to understand how single nucleotide polymorphisms (SNPs) mechanistically underpin complex diseases. While this challenge has been addressed partially by Gene Ontology (GO) enrichment of large list of host genes of SNPs prioritized in GWAS, these enrichment have not been formally evaluated. Here, we develop a novel computational approach anchored in information theoretic similarity, by systematically mining lists of host genes of SNPs prioritized in three adult-onset diabetes mellitus GWAS. The "gold-standard" is based on GO associated with 20 published diabetes SNPs' host genes and on our own evaluation. We computationally identify 69 similarity-predicted GO independently validated in all three GWAS (FDR<5%), enriched with those of the gold-standard (odds ratio=5.89, P=4.81e-05), and these terms can be organized by similarity criteria into 11 groupings termed "biomolecular systems". Six biomolecular systems were corroborated by the gold-standard and the remaining five were previously uncharacterized. http://lussierlab.org/publications/ITS-GWAS

Introduction

Single nucleotide polymorphism (**SNP**) arrays have been extensively used to predict how genetic variants are related to a single phenotype in genome-wide association studies (**GWAS**). Many methods have been developed to confirm these predictions. SNPs predicted by GWAS have been assessed with follow-up studies and with biological models. In contrast, the opposite has not been true. There is a paucity of validation studies exploring the predicted biomolecular functions associated with known SNPs. To properly evaluate these predicted biomolecular functions, a statistically significant number of annotated SNPs using a proper methodology are required. Three important considerations argue for the improvement of the accuracy and validation set of related annotations, or what we are calling "biomolecular systems", predicted by the SNP array. First, biomolecular systems associated with complex diseases are poorly understood and remain largely without computational replication in a different dataset or biological validation. Current enrichment approaches have only been conducted with intragenic SNPs. Second, the existing GWAS can be leveraged at minimal additional expense to unveil additional knowledge. Finally, increasing availability of multiple, independent, and disease-specific SNP array datasets provides an excellent opportunity to analyze across related experiments.

Others have previously conducted straight-forward Gene Ontology (**GO**) [1] enrichment studies over a limited number of host genes annotated to intragenic SNPs. Such single study predictions with an arbitrary number of prioritized SNPs are available through web-based tools [2]. However, these studies do not provide formal evaluations as to the optimal cutoff for the number of prioritized SNPs and their computational validation remains entirely "internal" to a single study, using empirical or theoretical statistics for correction of multiple comparisons. Additionally, others have found a large number of biomolecular systems in the intersection of gene expression and SNPs or interacting quantitative trait loci (QTLs) [3, 4]. Together, these results suggest that long lists of prioritized host genes (**PHG**) annotated to intragenic SNPs in expression arrays contain additional biological function information about diseases beyond that found in a top few SNPs. The limitation to these previous computational validation approaches is that they remain entirely "internal" to a single study and use empirical or theoretical statistics for multiple comparison correction rather than consider any recapitulation in new patient datasets or biological validation. In this paper, we propose that multiple GWAS focusing on the same complex disease phenotype may provide a means for "external" validation of the biomolecular systems predicted by enrichment studies in a single GWAS.

However, to provide this evaluation, biomolecular functions and processes need to be replicated between GWAS. This is challenging as biomolecular functions and processes that are systematically

queried via Gene Ontology contains over 25,000 distinct terms, organized in a directed acyclic graph in various depths. Because of the high dimensionality of GO, the enriched GO terms between two studies may be similar (e.g. parent-child) but not identical. One solution is to use Information Theoretic Similarity (**ITS**), which was originally applied to GO by Lord *et al.* to show consistency between sequence and annotation [5]. Lerman *et al.* have shown high concordance between functions inferred from similarity in GO and measured structure of proteins and functional similarity of proteins. We and others have also shown how biological systems similarity measured by ITS in GO can improve the accuracy of machine learning-based predictions of biological function for inadequately characterized genes [6]. These studies further support the accuracy of ITS and of "*annotation transference*" between sequence, structure, and function [7]

We hypothesize that biomolecular functions and processes discovered in a meta-analysis of several GWAS can formally be "replicated *in silico*" in a separate independent GWAS and validated by a gold standard with no significant bias. This gold standard can be derived from known published disease specific genes and GO annotations. Secondly, we hypothesize that we can increase the number of accurate predictions of disease-related biomolecular functions and processes which are significantly validated in the second GWAS by studying its ITS in GO in addition to its exact GO overlap when compared to other independent GWAS. Third, we also hypothesize that we can organize the resulting specific biomolecular functions and processes into a smaller set of biomolecular systems using the similarity criteria between GO terms to discriminate between uncharacterized systems of a disease and those corroborated ones.

In this proof of concept paper, we chose to focus our analysis on Adult Onset Diabetes Mellitus (**AODM**) as it is a particularly intricate multi-system disease with relatively low odds ratio (**OR**) in individual SNPs in pathologic phenotypes. We also propose an original evaluation based on GO terms associated with known host genes of AODM's SNPs.

Methods

Data Processing and Experimental Design Figure 1 provides a schema and succinct description of the proposed methods and evaluation. **GWAS Datasets and SNP's Host Gene(s)** In this study, we used three independent GWAS: The Wellcome Trust Case Control Consortium (**WTCCC**)[8], Finland - United State Investigation of NIDDM Genetics (**FUSION**)[9], and the Diabetes Genetics Initiative

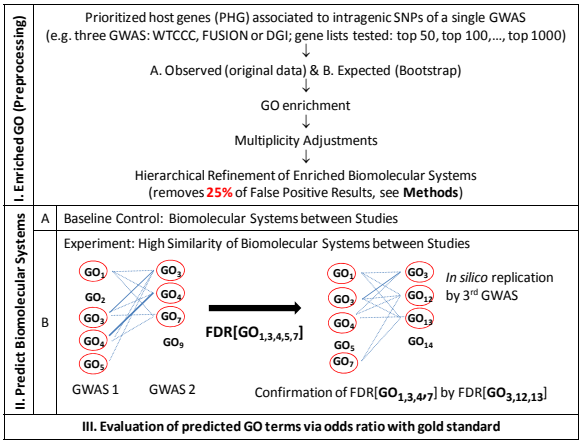

Figure 1. Overall Experimental Design. Molecular functions and biological processes of GO are considered as measurements of biomolecular systems of diseases. The approach is to determine GO terms that can be repeatedly found in three GWAS either by exact match or by similarity to other GO terms. The GO terms enriched in each of three GWAS are extracted (**I**, upper panel); GO terms enriched in both of two GWAS are calculated based on (i) direct overlap of GO terms in both studies (baseline control, **II.A),** and on (ii) information theoretic similarity (ITS) between GO terms of each studies (left side of middle panel, **IIB)**. *In silico* evaluation of the prediction of GO terms enriched in two studies are conducted in the 3rd GWAS (right side of middle panel, **IIB).** Empirical estimates of false discovery rates (**FDR**) are generated via bootstraps (FDR of the *in silico* replication, **II**). Evaluation of predicted GO terms in three GWAS is performed using odds ratio based on a gold standard of Gene Ontology annotations of 20 published AODM SNP's host-genes (**III**, **Suppl. Methods**). GO_{GWAS_i} is the set of GO terms enriched at a certain P-value in the list of prioritized genes in GWAS "*i*".

(**DGI**)[10]. To standardize the gene names, refflat.txt, kgalias.txt, and knowngene.txt were downloaded from the UCSC genome browser (Nov. 2007) [11] as well as the core data from the Human Genome Organizations (**HUGO**)'s Gene Nomenclature Committee (June 30, 2008). Gene Ontology annotation, structural files gene_ontology.obo (July 08, 2008) and gene_annotation.goa_human.gz (Oct. 28, 2008) were downloaded from its website [12]. **SNP's Host Gene(s)** Detailed methods are described in the **Suppl. Methods**.

Functional annotation of GWAS to GO (**Fig. 1, Panel I**) We conducted an enrichment study with GO functional annotations to prioritize biomolecular systems related to AODM using host genes annotated to intragenic SNPs that were prioritized statistically in a GWAS. We systematically evaluated several numbers of prioritized host genes (PHG) as follows, 50, 100, 200, 300, 400, 500, or 1000. The unadjusted *P*-value of the GO enrichment was calculated using

the cumulative hypergeometric distribution provided by an open source Perl API, GO-TermFinder[13]. Bonferroni correction (*P*-value*n, where n=number of GO terms in the test) was applied to control for multiple comparisons. **Hierarchical Refinement of Enriched GO Terms** Recent reports have stated that enrichment studies conducted over genes in GO can generate falsely significant *P*-values due to the inheritance of genes in parent classes which may be highly enriched [14]. To remove such false positive signals inherited in the GO hierarchy during enrichment, we refined the enriched GO terms in each study according to each PHG with a novel set-theoretic method described in the **Supplemental Methods**.

Determining Similarity between GO Enriched GWAS (**Fig. 1, Panel II**) In order to reduce the dimensionality of the predicted GO terms and to increase the precision, we retained only those terms that were either identical in two studies or similar between two studies. We previously implemented Lin's standardized ITS metric [15] that ranges from 0 to 1 to identify similarity between GO terms and have shown that an ITS score ≥ 0.7 was significant and optimal for the prediction of GO function in sparsely annotated genes[6]. GO terms enriched in each GWAS were systematically compared to one another using: (I) simple overlap (e.g. same GO terms or ITS=1, **Fig. 1, Panel IIA**), and with (II) GO terms with ITS ≥ 0.7 (**Fig. 1,** left side of **Panel IIB**), which are each from a distinct GWAS and with a similarity ≥ 0.7 between them. Similar GO terms between two studies were also compared to the 3rd GWAS for calculation of FDR and validation (**Evaluation, *in silico* replication in the 3rd study, Fig. 1,** right side of **Panel IIB**). With three GWAS, there were three possible combinations in which each study serves once as the validation study. Each of these groups of predictions of related GO terms (biological processes and molecular functions) was controlled for with a FDR calculation using a bootstrap of randomly selected genes. 100 bootstraps were conducted for each of the GWAS at each prioritized host gene threshold and subjected to the same analyses (GO enrichment, hierarchical refinements, and ITS between GO of studies), to control for *in silico* replication (**Fig. 1,** right side of **Panel IIB**). Finally, GO terms with a similarity ≥ 0.7 among all three studies were also illustrated in a Cytoscape network[16] (**Fig 2** and **Suppl. Fig.1**).

Generation and Evaluation of Gold Standard (Fig. 1, Panel III) To evaluate our predictions of GO terms associated with AODM, we developed an "gold standard" with no significant bias based on GO annotations of 20 published diabetes genes [17]. Of the 20, 19 genes were annotated in GO generating 245 distinct terms in what we call our "gold standard GO" (**GS-GO**). In addition, we conducted a study confirming that GO terms associated with these genes are more related to one another (and thus to diabetes) than an empirical distribution derived from 100 bootstraps of a random pick of 19 host genes annotated in all three GWAS (*P*<0.0001; details in **Suppl. Results**).

		Unadjusted P	0.001	0.01	0.025	0.05	0.1	0.15	0.2	0.3
		Adjusted P	10^{-6}	0.002	0.038	0.306	1	1	1	1
Intersection of Enriched GO Sets in three GWAS	P H G	50	0	0	0	0	0	0	0	0
		100	0	0	0	0	1	3	4	~~6~~
		200	0	0	0	0	2	9	16	~~31~~
		300	0	0	**5**	**7**	**12**	**19**	23	~~35~~
		400	0	2	**4**	8	**15**	21	30	~~47~~
		500	1	9	**15**	**20**	28	38	48	~~61~~
		1000	7	**14**	**21**	**35**	49	58	67	~~83~~
		Adjusted P	0.002	0.249	1	1	1	1	1	1
Intersection of Enriched GO Sets in two GWAS	P H G	50	0	1	3	6	~~6~~	~~8~~	~~10~~	~~14~~
		100	0	0	1	3	~~9~~	~~16~~	~~19~~	~~25~~
		200	0	0	1	4	~~17~~	~~23~~	~~39~~	~~51~~
		300	3	8	17	23	~~30~~	~~34~~	~~37~~	~~54~~
		400	4	6	14	18	~~32~~	~~43~~	~~59~~	~~81~~
		500	5	12	20	29	~~43~~	~~59~~	~~74~~	~~**106**~~
		1000	8	23	40	51	~~74~~	~~**89**~~	~~**111**~~	~~143~~

Table 1. Reproducibility of Biomolecular Systems predicted in GWAS through Exact GO overlap. This table presents the count of GO terms enriched in each GWAS according to the unadjusted P-value (top row) and that have been replicated in two (FUSION and DGI, lower) or three studies (upper). Further, different rows present results for increasing numbers of prioritized host genes (PHG) derived from GWAS intragenic SNPs. Joint P-values for two and three GWAS were corrected with combinatorial Bonferroni correction ((*P*-value^m)*n, m=number of GWAS, n=count of GO studied~2800). Odds ratios were calculated from a gold standard of biomolecular results (see **Methods**). **Legend**: empty sets (0) are presented in pale grey (see **Methods**, bootstrap); struck, FDR>5% unshaded, meets Bonferroni corrections; bold and red, odds ratio>confidence interval 95% according to our gold standard.

Results and Discussion

Reproducibility and Validation of Biomolecular Systems (Exact Overlap) We estimated the likelihood of a straightforward overlap of a specific enriched GO term in two or three studies. **Table 1** provides the number of overlapping GO terms enriched between FUSION and DGI GWAS (archetypical results, other combinations of GWAS not shown) according to the following parameters: (I) *P*-value of GO enrichment adjusted for multiplicity and reproducibility in two studies, (II) length of PHG. The large number of predicted GO does not satisfy statistical significance as measured by adjusted *P*-value or odds ratio with the gold standard. These drawbacks suggest that there is an opportunity for improving the accuracy of

predictions of replicated GO terms between two studies using similarity (for detail, see **Suppl. Results and Suppl. Tables 1&2**).

Biomolecular Systems Predicted by ITS (Biological Similarity) In a sense, GO annotations serve as a proxy to identify components of biomolecular functions and processes that are then assembled in systems with similarity metric as well. Applying such an approach to uncharacterized SNPs is related to our previous work on predicting GO functions and processes in sparsely annotated genes [6] where we show that similarity could perform optimally in GO for values of ITS ≥0.7. **Table 2** provides the number of similarity-predicted GO terms enriched between three independent GWAS (top) and between two GWAS, FUSION and DGI according to the following parameters: (I) *P*-value of GO enrichment adjusted for multiplicity and reproducibility in multiple studies, (II) length of PHG. As compared to the exact match study (**Table 1**), we demonstrated that ITS has dramatically improved the accuracy of the predictions according to three metrics: (I) adjusted theoretical $P \leq 0.05$ (white zones, non-grey), (II) empirical FDR ≤5% (unstruck data), and (III) statistically significant odds ratio using the GS-GO (**Table 2**, red bolded). The optimal range of prediction with high odds ratio is around 200-300 PHG counts and at an unadjusted *P*-value of the GO enrichment (P=0.025). AS shown in **Table 1**, whether by exact match or by similarity, results between two studies are not statistically significant in most ranges as compared to those between three studies. In contrast, there is a distinct improvement with ITS; the number of accurate predictions tripled and in some cases quintupled as compared to those found in the exact overlap method. These results suggest that ITS provide opportunities to uncover novel biomolecular systems properties overlooked by straightforward overlap methods. We also organized these GO terms as 11 biomolecular system classes by similarity to identify which of the GO terms found by similarity were novel as compared to those found by exact match (**Figure 2**). Cluster C, GTPase regulator activity, does not contain any black circles (exact match) and thus corresponds to a group of GO terms predicted exclusively by similarity. Further, **Suppl. Figure 1** provides a more detailed map of GO terms where 6 "biomolecular systems" are corroborated by the gold standard (red circles) and five may be novel ones. Using ITS, the 69 GO terms that were similar among the three GWAS were also enriched in diabetes signal as they comprised 12 gold standards (GS-GO; P=4.81e-05, cumulative hypergeometric test). There was a significant fourfold increase in the number of predictions and a threefold increase in the recapitulation of our gold standard annotations. At least four biomolecular systems were related to the nervous system, which may indicate a pleiotropy of biomolecular systems involved in complex traits. Although commonly thought as more of an endocrine organ, the endocrine pancreas responsible for secreting insulin and glucagon is a neuroendocrine organ with parasympathethic innervation. Further analysis is described in the **Suppl. Results**.

		Unadjusted P	0.001	0.01	0.025	0.05	0.1	0.15	0.2	0.3
Enriched in GO terms in three GWAS related with ITS		Adjusted P	10^{-6}	0.002	0.044	0.35	1	1	1	1
	P H G	50	0	0	0	0	0	~~2~~	~~2~~	**~~8~~**
		100	0	0	0	0	**10**	~~30~~	~~35~~	**~~54~~**
		200	0	0	**10**	38	53	~~70~~	**~~90*~~**	**~~131*~~**
		300	0	2	**17***	47	64	~~93~~	~~111~~	~~138~~
		400	0	9	35	74	98	~~137~~	~~164~~	~~213~~
		500	3	26	**41***	81	131	~~143~~	~~187~~	~~237~~
		1000	25	**63***	**107***	132	181	~~233~~	~~276~~	~~369~~
Enriched in GO terms in two GWAS related with ITS		Adjusted P	0.002	0.244	1	1	1	1	1	1
	P H G	50	0	**3**	**3**	**6**	**~~8~~**	~~15~~	**~~19~~**	**~~30*~~**
		100	0	0	4	12	~~29~~	**~~45~~**	**~~54~~**	~~79~~
		200	0	**7**	**18**	39	~~61~~	~~87~~	~~116~~	~~134~~
		300	4	**16**	32	53	~~67~~	~~88~~	~~108~~	~~137~~
		400	4	14	42	75	~~103~~	~~138~~	~~173~~	~~221~~
		500	**8**	25	~~37~~	79	~~129~~	~~156~~	~~198~~	~~261~~
		1000	29	62	92	133	~~193~~	~~245~~	~~284~~	~~373~~

Table 2. Reproducibility of Biomolecular Systems Predicted through Information Theoretic Similarity This table present the count of GO terms enriched in each GWAS using information theoretic similarity (ITS) between GO terms (ITS>0.7, **Figure 1, Panel IIB**). The lower table shows the number of GO terms predicted in two studies (FUSION and DGI) by ITS. The upper table is then associated with information theory to those of the WTCCC. Since a similarity of ITS=1 is the same thing as a join sets of GO (intersection of sets) produced in **Table 1**. **Legend:** see **Table 1**; *odds ratio significant in every combination of GWAS at that PHG and P-value.

Future Studies and Limitations We observed several limitations to this study. First, GO does not encompass pathophysiological and higher clinical concepts. Consequently, important associations with complex disease systems cannot be derived. Further, the gold standard we derived from all GO terms associated with diabetes SNPs is far from optimal and likely comprises GO terms not related to diabetes. In future studies, we will explore the use of expression arrays to derive an improved gold standard and also the use of eQTL associations in order to include intergenic SNPs via their "trans"-correlated genes in addition to their intragenic ones. We will use this approach to identify common systems across complex diseases. For example, one can imagine a study of metabolic syndrome by pooling hypertension diabetes, and cardiovascular studies along with obesity to generate the biomolecular systems underlying this complex disease.

Conclusion

We successfully implemented a framework of functional similarity using Gene Ontology to more accurately recapitulate known biomolecular systems associations with complex diseases in GWAS according to a gold standard and to predict uncharacterized ones. We further show that this similarity technique is able to find more associations than a straightforward overlap of GO terms. Currently, a new biological process finds one host gene at a time through GWAS, which requires additional biological validation. To our knowledge, this is the first study that demonstrates reliable novel biological processes repeated across GWAS. Since very little is known of the biomolecular mechanisms relating SNPs to complex diseases and how these processes differ from those of single gene inheritance or from those observed in animal models, this approach could contribute in mapping the SNP phenome in high throughput and at low cost from existing studies, providing an insight into the genetic architecture underpinning the inheritable variants of complex diseases.

(See **Supplementary Information** for Additional References and Acknowledgements)

References

1. Eyre TA, et al. (2006) The HUGO Gene Nomenclature Database. NAR 34: D319-1.
2. Holden M,et al. (2008) GSEA-SNP: applying gene set enrichment analysis to SNP data from genome-wide association studies. Bioinformatics 24: 2784-5.
3. Hubner N, al. (2005) Integrated transcriptional profiling and linkage analysis for identification of genes underlying disease. Nat Genet 37: 243-3.
4. Litvin O, et al. (2009) Modularity and interactions in the genetics of gene expression.PNAS 106: 6441-6.
5. Lord PW, Stevens RD, Brass A, Goble CA (2003) Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. Bioinformatics 19: 1275-83.
6. Lerman G, et al. (2007) Defining functional distance using manifold embeddings of gene ontology annotations. PNAS 104: 11334-9.
7. Tao Y, Sam L, Li J, Friedman C, Lussier YA (2007) Information theory applied to the sparse gene ontology annotation network to predict novel gene function. Bioinformatics 23: i529-38.
8. (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447: 661-78.
9. Scott LJ, et al. (2007) A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. Science 316: 1341-5.
10. Saxena R, et al. (2007) Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. Science 316: 1331-6.
11. Kuhn RM, et al. (2007) The UCSC genome browser database: update 2007. NAR 35: D668-73.

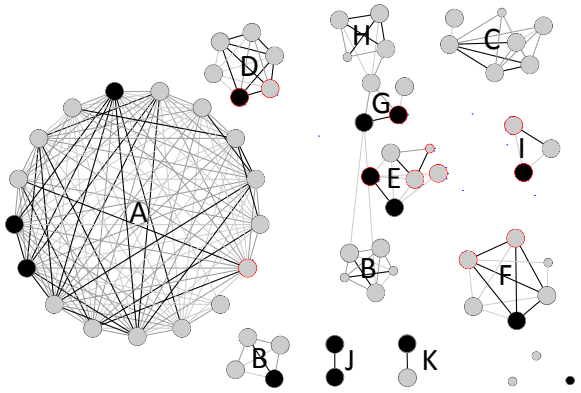

Figure 2. Biomolecular Systems of Adult Onset Diabetes Mellitus Discovered using Information Theoretic Similarity. 69 predicted GO terms are visually assembled in 11 distinct "biomolecular systems" using inter GO similarity. They are also enriched with 12 GO terms of the gold standard (*P*=4.81e-05, cumulative hypergeometric test). Predictions were conducted at ITS>0.7. **Legend**: The node presents GO terms and the edge between nodes indicates ITS relations. Increased line thickness corresponds with increased ITS (ITS of 1 indicates an exact GO match). Grey circles indicate ITS predicted GO terms. Black circles indicate 14 GO terms exactly overlap across three GWAS. Red rimmed circles correspond to 12 gold standard GO terms. Circle size indicates the number of GWAS contributing to the GO terms. **Biomolecules Defined**: **A**, voltage-gated ion channel activity; **B**, synapse (one is from biological process and the other is from molecular function); **C**, GTPase regulator; **D**, ion transport; **E**, membrane; **F**, receptor activity and neurotransmitter; **G**, signal transduction; **H**, Ras/Rho protein signal transduction; **I**, ion binding; **J**, adhesion; **K**, glutamate receptor. (details in **Suppl. Results, Suppl. Fig 1, Suppl. Tables 1&2**)

12. Ashburner M, Ball CA, Blake JA, et al. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.Nat Genet 25: 25-9.
13. Boyle EI, Weng S, Gollub J, Jin H, Botstein D, et al. (2004) GO::TermFinder--open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. Bioinform 20: 3710-5.
14. Rhee SY, et al. (2008) Use and misuse of the gene ontology annotations. NatRev Genet 9: 509-15.
15. Lin D (1998) An information-theoretic definition of similarity. In Proceedings of the Fifteenth International Conference (ICML' 98): 296-304.
16. Shannon P, et al. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 13: 2498-2504.
17. Meigs JB, et al. (2008) Genotype score in addition to common risk factors for prediction of type 2 diabetes. N Engl J Med 359: 2208-19