Supplementary Information

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

Younghoe Lee, PhD1, Jianrong Li, MSc1, Eric Gamazon, PhD1, James L. Chen, MD2, Anna Tikhomirov, PhD1, Nancy J. Cox, PhD1,3,*, and Yves A. Lussier, MD1,4*

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

A. Abbreviation used in the manuscript

| Abbreviation | Description                           |
|--------------|--------------------------------------|
| SNP          | Single Nucleotide Polymorphism       |
| GWAS         | Genome-wide association study        |
| GO           | Gene Ontology                        |
| QTL          | Quantitative Trait Loci              |
| PHG          | Prioritized Host Gene                |
| ITS          | Information Theory Similarity        |
| AODM         | Adult Onset Diabetes mellitus        |
| OR           | Odds Ratio                           |
| WTCCC        | Wellcome Trust Case Control Consortium |
| FUSION       | Finland - United State Investigation of NIDDM Genetics |
| DGI          | Diabetes Genetics Initiative         |
| HUGO         | HUMAN Genome Organizations            |
| GS-GO        | Gold Standard Gene Ontology          |
| HGNC         | Gene Nomenclature Committee          |

B. Supplementary Methods

Intragenic SNP in Genome-Wide Association Study (GWAS) Each intragenic single nucleotide polymorphism (SNP) was annotated to its host gene with the data sources. Genomic coordinates of intragenic SNPs (chromosome and base position) were mapped to the human genome reference assembly (build 36). The SNPs’ RefSeq alleles and host genes were defined using RefSeq genomic coordinates for the gene. As a standard identifier for the host gene of the SNP, the approved gene symbol of core data from HUMAN Genome Organizations (HUGO)’s Gene Nomenclature Committee (HGNC) was used. In the rare instances for which more than one host gene was assigned to a SNP, each gene was counted as an independent entry. The final annotated files include 12,387, 13,442, and 12,185 distinct host genes in Wellcome Trust Case Control Consortium (WTCCC), Finland - United State Investigation of NIDDM Genetics (FUSION), and the Diabetes Genetics Initiative (DGI) respectively.

Odds Ratio and Gold Standard of Biomolecular Systems from Known Diabetes Genes and “Evaluation of Predicted Biomolecular Systems” (Figure 1, Panel III) To evaluate predictions of Gene Ontology (GO) terms associated with Adult Onset Diabetes mellitus (AODM), we developed a biased gold standard with no significant bias of GO terms associated with diabetes genes from 20 AODM genes published by Meigs et al. [1],[2-4]. 19 of these genes were annotated in GO generating 245 distinct GO terms that we call our “gold standard” (GS-GO). GO terms of GS-GO were used to systematically calculate the odds ratio (OR) of the predicted GO biomolecular systems by Information Theory Similarity (ITS). The OR was calculated by 2x2 contingency table with two variables; GS-GO and GO terms
Supplementary Information

predicted by ITS. We further conducted an unbiased control to confirm that GS-GO terms to AODM were more likely to contain biomolecular systems associated with diabetes than GO terms from other host genes. We verified that the distribution of ITS between every combination of GO terms in the GS-GO was significantly different from the ITS found in the 6,000 GO terms found in any of the three GWAS.

**Hierarchical Refinement of Enriched GO Terms** To further refine the results and establish the most significant biomolecular systems among the P-values in an enrichment study, we represented the ontological structure of GO as a directed acyclic graph, composed of nodes (the terms of the GO) and edges (relationships between or among the terms of the GO) [5]. Recent reports demonstrated that enrichment studies conducted over genes in GO can generate falsely significant P-values due to the inheritance of genes in ancestry classes of a significantly enriched class [6, 7]. We provide a novel set-theoretic approach for filtering out false positive signals. Equations 2, 3 and 4 describe the refinement algorithm we developed to retain true positive results. \(V\) is the set of nodes, and \(E\) is the set of edges. The Open Biomedical Ontologies (OBO; http://obo.sourceforge.net/) represent a hierarchical relationship as “is_a”, with \(i\) and \(j\) which stand for a child and a parent respectively. For each node \(v_i \in V\), \(v_i\) contains two elements of its relation with neighboring nodes \(v_j\) namely, \(e_{ij} \in E\) and the P-value from the enrichment study symbolized as \(p_i\). Each node \(v_i\) is defined in terms of a hierarchical relationship such as \(A_i\) for all parents (1st degree Ancestors), \(C_i\) for all children, and \(D_i\) for all descendants. These are represented as \(A_i = \{ v_j \in V \mid \exists e_{ij} \in E \} \), \(C_i = \{ v_j \in V \mid \exists e_{ji} \in E \} \), and \(D_i = \{ v_j \in V \mid \exists e_{k_1,k_2,k_3,\ldots,k_n,i} \in E \}\) to describe the hierarchical relationship. The nodes that have the most statistically significant P-values as compared to their hierarchical neighbors were identified as Regional Minima of P-values (\(V_{RM}\)) and defined by Equation 1. Among Regional Minimum nodes, we further excluded parents that have the same P-values as their children to conserve the most informative nodes to yield the Refined Regional Minimum (\(V_{RRM}\), Equation 2). The subsumed significant associations (Significant Descendants of Refined Regional Minimum or SDRRM) are defined in Equation 3. Finally, Equation 4 defines the subset of retained nodes found in either Equation 2 or 3 excluding falsely significant P-values due to the inherited genes. The algorithm removed 25% false positive GO terms. An example of a false positive ancestor identified by the algorithm is as follows; assume GO term \(j\) is a heretical child of GO term \(J\). A number of genes are annotated in child GO \(j\), while \(b\) genes are annotated in parent GO \(J\). When a disproportionate number of genes annotated in GO \(J\) are also annotated to child GO \(j\) (for example if all genes are the same), then we can assume that the statistical enrichment is inherited as a statistical artifact of our enrichment methods that use inheritance of gene annotations of children to parents. The algorithm uses the dispersion or spread of the P-values in the hierarchies to identify these false positive signals.

**Equation 1** \(V_{RM} = \{ v_i \in V \mid \forall v_j \in A_i \cup C_i, p_j \geq p_i \} \)

**Equation 2** \(V_{RRM} = \{ v_i \in V \mid \exists v_j \in V_{RM}, v_j \in A_i, v_i \in V_{RM} \} \)

**Equation 3** \(V_{SDRRM} = \{ v_i \in V \mid \exists v_j \in V_{RRM}, v_i \in D_j, v_i \not\in V_{RRM}, v_i \not\in V_{RM} \} \)

**Equation 4** \(V_{included} = \{ v_i \in V \mid v_i \in V_{RRM} \cup V_{SDRRM} \} \)

**Theoretical Statistics, Software Implementation and Availability** Non-parametric comparison of a distribution to a theoretical value (Wilcoxon signed-rank test) and non-parametric comparison of medians (Mann-Whitney test) were used to calculate the significance. The ITS was previously implemented in JAVA [8]. Network figure was drawn with Cytoscape ver. 2.5.2 [9].
C. Supplementary Results

**Gold Standard of Biomolecular Systems Associated with AODM** As described in the Methods, the gold standard comprised 245 distinct GO terms (GS-GO) associated with 19 genes. This group of 245 GS-GO terms constitutes an imperfect gold standard as it certainly is incomplete and comprises GO terms related to the function of these genes that are likely unrelated to diabetes mellitus. However we demonstrated that this gold standard is most likely to be enriched in biomolecular system associated with AODM. Additionally, there is no proof that a manually curated, and possibly biased, gold standard of biological systems related to the pathophysiology of diabetes mellitus would be more relevant to the complex inheritance of AODM than one derived from its intragenic SNPs discovered in GWAS using unbiased computational approaches. We compared the median ITS between every combination of the GS-GO terms (ITS\_GS-GO) to those associated with random draws of 245 GO terms which the genes are annotated in GO and repeated this bootstrap 1,000 times to generate an empirical distribution. The observed molecular functions and biological processes of ITS\_GS-GO were respectively 1.4 times and 1.2 times that of the expected values (P<0.0001 in both cases, Wilcoxon Signed Rank test).

**Reproducibility and Validation of Biomolecular Systems in three GWAS Exactly as Predicted** To our knowledge, there has been no study showing either independent replication or validation of a biomolecular system derived from GWAS beyond those systems immediately derived from a single gene polymorphism. As shown in Figure 1, enriched GO terms were further refined to remove about 25% of trivial results considered as false positive signals inherited up the GO hierarchies (Supplementary Methods). Enriched GO terms were stratified according to their respective unadjusted P-value (indeed, very few results meet the Bonferroni correction in a single study as the signal is weak in lists of intragenic SNPs prioritized in a single GWAS. We determined the likelihood of a straightforward replication of a specific enriched GO term in two or three studies. Table 1 (main manuscript) provides the number of GO terms independently enriched in more than one study, and whether these overlaps (reproduced GO terms in independent studies) can be explained by chance or not according to two types of evaluations: I) false discovery rate (FDR) derived from bootstrap (in silico replication) and II) significantly increased OR of finding a GO term of GS-GOs expected. The FDR tends to increase with the increasing number of prioritized host genes (PHG). A lower initial P-value of the enriched GO terms is associated with a lower FDR and predictions matched in three studies are better than those from two studies. However, the latter merits more attention as the bottom panel of the table suggests that GO terms matched in two studies are more numerous and therefore more likely to comprise some noise. While the FDR is low for an initial enrichment of P<0.05 of GO terms in each of the studies, the large number of predicted GO do satisfy statistical significance for the OR of discovering a gold standard gene better than chance. In other words, one needs to be careful when replicating GO terms across two studies only as replicate. GO terms across three studies are significantly enriched in gold standard genes for a broader range of parameters. Using the Bonferroni corrections, we observed that the majority of results in two studies do not meet the criteria. These drawbacks suggest that there is an opportunity for improving the accuracy of predictions of replicated GO terms between two studies using ITS and for increasing the number of accurate predictions between three studies. These results also indicate that at P<0.025 and for PHGs ranging from 300-1000, exactly repeated GO terms between three studies can serve as accurate anchors of biomolecular systems in an ITS conducted over the same dataset or as gold standards of biomolecular systems in future independent studies. Specific names of the GO terms found significant in Table 1 between three studies are provided in Figure 2 (black circles) and Suppl. Tables 1&2.

**Detail Description of Visualization and Analysis of Predicted Biomolecular System (Figure 2, Supplementary Figure 1, Supplementary Tables 1&2).** To demonstrate biomolecular systems predicted by ITS, one specific case was chosen according to the parameters, 1000 PHG with unadjusted P-value of GO<0.01 and with ITS>0.7. This case comprised 69 GO terms, union GO terms of any combinatorial ITS experiments of three GWAS. The visualization was computed automatically by
the Cytoscape software with the default parameter, “Organic Layout” [9]. The initially computed biomolecular systems, which were defined by experts in biology, clustered into 11 biomolecular terms. To visualize the evaluation of predictions, GS-GO terms and GO terms overlap in three studies were also presented in this network. 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). Thickness of all edges between GO represents the level of ITS similarity, range from 0.6 to 1 in order to easily recognize new connection of our predictions to GS-GO by ITS method. **Legend:** 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. 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:** 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. The 69 GO terms and corresponding IDs are listed in **Suppl. Table 2**.

**Funding & Acknowledgements.** This work was supported in part by the NIH/NLM/NCI National Center for Multiscale Analyses of Genomic and Cellular Networks (MAGNET, 1U54CA121852), the NIH/NCRR Clinical & Translational Science Awards (1U54 RR023560-01A1), The Cancer Research Foundation, The UCCRC, ENDGAMe (U01 HL084715), P60 DK20595, 2L01 GM61393, and ADA Mentored Fellow Award (AT). WTCCC and DGI were downloaded when publicly available. We thank Dr. M. Bochnke for graciously providing FUSION and Dr. Yang Liu with the assistance in the programming of the enrichment studies.
Supplementary Figure 1. Biomolecular Systems of Adult Onset Diabetes Mellitus Discovered using Information Theory Similarity. 69 predicted GO terms are visually assembled in 11 distinct "biomolecular systems" using inter GO similarity. For detailed description, see Suppl. Results (above).
### Supplementary Table 1. 11 Biomolecular Systems in Figure 2 and Suppl. Fig. 1.

| ID | Name / description                     | # of GO | # of GS-GO | Known Gene annotated to GS-GO          |
|----|----------------------------------------|---------|------------|----------------------------------------|
| A  | Voltage-gated ion channel activity     | 17      | 1          | KCNJ11                                 |
| B  | Synapse                                | 9       | 0          |                                        |
| C  | GTPase regulator activity              | 7       | 0          |                                        |
| D  | Ion transport                          | 6       | 2          | KCNJ11, SLC30A8                        |
| E  | Membrane                               | 6       | 4          | CDKAL1, KCNJ11, LGR5, MOTCH2, SLC30A8, TSPAN8, VEGFA |
| F  | Receptor activity and neurotransmitter | 6       | 2          | LGR5, NOTCH2, TSPAN8                  |
| G  | Signal transduction                    | 4       | 1          | LGR5, PPARG                            |
| H  | Ras/Rho protein signal transduction    | 4       | 0          |                                        |
| I  | Ion binding                            | 3       | 2          | ADAMTS9, BCL11A, CAMK1D, CDKAL1, JAF1, NOTCH2, PPARG |
| J  | Adhesion                               | 2       | 0          |                                        |
| K  | Glutamate receptor activity            | 2       | 0          |                                        |
Supplementary Table 2. Details of 69 GO Terms found similar in three GWAS in Fig. 2 and Suppl. Fig. 1.

| GO ID       | GO TERM                                           | Biomolecular ID |
|-------------|--------------------------------------------------|-----------------|
| GO:0022843  | voltage-gated cation channel activity            | A               |
| GO:0022838  | substrate specific channel activity              | A               |
| GO:0022836  | gated channel activity                           | A               |
| GO:0022834  | ligand-gated channel activity                    | A               |
| GO:0022832  | voltage-gated channel activity                   | A               |
| GO:0022803  | passive transmembrane transporter activity       | A               |
| GO:0015278  | calcium-release channel activity                 | A               |
| GO:0015276  | ligand-gated ion channel activity                | A               |
| GO:0015267  | channel activity                                 | A               |
| GO:0005267  | potassium channel activity                       | A               |
| GO:0005262  | calcium channel activity                         | A               |
| GO:0005261  | cation channel activity                          | A               |
| GO:0005245  | voltage-gated calcium channel activity           | A               |
| GO:0005244  | voltage-gated ion channel activity               | A               |
| GO:0005234  | extracellular-glutamate-gated ion channel activity| A               |
| GO:0005230  | extracellular ligand-gated ion channel activity  | A               |
| GO:0005216  | ion channel activity                             | A               |
| GO:0044456  | synapse part                                     | B               |
| GO:0045202  | synapse                                          | B               |
| GO:0045211  | postsynaptic membrane                            | B               |
| GO:0042734  | presynaptic membrane                             | B               |
| GO:0016079  | synaptic vesicle exocytosis                      | B               |
| GO:0019226  | transmission of nerve impulse                    | B               |
| GO:0007269  | neurotransmitter secretion                       | B               |
| GO:0003001  | generation of a signal involved in cell-cell signaling| B           |
| GO:0007268  | synaptic transmission                            | B               |
| GO:0005100  | Rho GTPase activator activity                    | C               |
| GO:0019887  | protein kinase regulator activity                | C               |
| GO:0005083  | small GTPase regulator activity                  | C               |
| GO:0005085  | guanyl-nucleotide exchange factor activity       | C               |
| GO:0005096  | GTPase activator activity                        | C               |
| GO:0030695  | GTPase regulator activity                        | C               |
| GO:0005089  | Rho guanyl-nucleotide exchange factor activity   | C               |
| GO:0006811  | ion transport                                    | D               |
| GO:0006813  | potassium ion transport                          | D               |
| GO:0006812  | cation transport                                 | D               |
| GO:0030001  | metal ion transport                              | D               |
| GO:0006816  | calcium ion transport                            | D               |
| GO:0015674  | di-, tri-valent inorganic cation transport       | D               |
### Supplementary Information

| GO:0016020 | membrane | E  |
| GO:0044459 | plasma membrane part | E  |
| GO:0005887 | integral to plasma membrane | E  |
| GO:0005886 | plasma membrane | E  |
| GO:0016021 | integral to membrane | E  |
| GO:0031226 | intrinsic to plasma membrane | E  |
| GO:0030594 | neurotransmitter receptor activity | F  |
| GO:0004888 | transmembrane receptor activity | F  |
| GO:0019199 | transmembrane receptor protein kinase activity | F  |
| GO:0060089 | molecular transducer activity | F  |
| GO:0004871 | signal transducer activity | F  |
| GO:0004872 | receptor activity | F  |
| GO:0007165 | signal transduction | G  |
| GO:0007154 | cell communication | G  |
| GO:0007167 | enzyme linked receptor protein signaling pathway | G  |
| GO:0007242 | intracellular signaling cascade | G  |
| GO:0035023 | regulation of Rho protein signal transduction | H  |
| GO:0046578 | regulation of Ras protein signal transduction | H  |
| GO:0051056 | regulation of small GTPase mediated signal transduction | H  |
| GO:0007265 | Ras protein signal transduction | H  |
| GO:0043167 | ion binding | I  |
| GO:0005509 | calcium ion binding | I  |
| GO:0046872 | metal ion binding | I  |
| GO:0022610 | biological adhesion | J  |
| GO:0007155 | cell adhesion | J  |
| GO:0008066 | glutamate receptor activity | K  |
| GO:0004970 | ionotropic glutamate receptor activity | K  |
| GO:0046658 | anchored to plasma membrane |  |
| GO:0008067 | metabotropic glutamate, GABA-B-like receptor activity |  |
| GO:0030054 | cell junction |  |
References

1. Meigs JB, Shrader P, Sullivan LM, McAteer JB, Fox CS, Dupuis J, Manning AK, Florez JC, Wilson PW, D'Agostino RB, Sr. et al: Genotype score in addition to common risk factors for prediction of type 2 diabetes. *N Engl J Med* 2008, 359(21):2208-2219.

2. Meigs JB, Dupuis J, Herbert AG, Liu C, Wilson PW, Cupples LA: The insulin gene variable number tandem repeat and risk of type 2 diabetes in a population-based sample of families and unrelated men and women. *J Clin Endocrinol Metab* 2005, 90(2):1137-1143.

3. Saxena R, Voight BF, Lyssenko V, Burtt NP, de Bakker PI, Chen H, Roix JJ, Kathiresan S, Hirschhorn JN, Daly MJ et al: Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science* 2007, 316(5829):1331-1336.

4. Zeggini E, Scott LJ, Saxena R, Voight BF, Marchini JL, Hu T, de Bakker PI, Abecasis GR, Almgren P, Andersen G et al: Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nat Genet* 2008, 40(5):638-645.

5. Rhee SY, Wood V, Dolinski K, Draghici S: Use and misuse of the gene ontology annotations. *Nat Rev Genet* 2008, 9(7):509-515.

6. Barry WT, Nobel AB, Wright FA: Significance analysis of functional categories in gene expression studies: a structured permutation approach. *Bioinformatics* 2005, 21(9):1943-1949.

7. Prufer K, Muetzel B, Do HH, Weiss G, Khaitovich P, Rahm E, Paabo S, Lachmann M, Enard W: FUNC: a package for detecting significant associations between gene sets and ontological annotations. *BMC Bioinformatics* 2007, 8:41.

8. Tao Y, Sam L, Li J, Friedman C, Lussier YA: Information theory applied to the sparse gene ontology annotation network to predict novel gene function. *Bioinformatics* 2007, 23(13):i529-538.

9. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T: Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* 2003, 13(11):2498-2504.