Supplementary of

Network Stratification Analysis for Identifying Function-specific Network Layers

Chuanchao Zhang1,2,#, Jiguang Wang2,3,6,#, Chao Zhang3, Juan Liu1,\*, Dong Xu3,\*, Luonan Chen2,4,5,\*

1School of Computer, Wuhan University, Wuhan 430072, China; 2Key Laboratory of Systems Biology, Innovation Center for Cell Signaling Network, Institute of Biochemistry and Cell Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200233, China; 3Department of Computer Science and Christopher S. Bond Life Sciences Center, University of Missouri, Columbia, MO 65211, USA; 4School of Life Science and Technology, ShanghaiTech University, Shanghai 201210, China; 5Institute of Industrial Science, University of Tokyo, Tokyo153-8505, Japan; 6Department of Systems Biology and Department of Biomedical Informatics, Columbia University, New York, NY10032, USA

#These authors contributed equally to this work.

\*Corresponding authors

**S0: Parameter setting of the case**

We construct a network consisting of ten-node, with the weights [adjacent](app:ds:adjacent)of this network (W). W=[0, 0, 0.53, 0.45, 0, 0, 0, 0.23, 0, 0; 0, 0, 0, 0.48, 0.34, 0, 0, 0, 0, 0.39; 0.53, 0, 0, 0, 0, 0, 0, 0, 0, 0;0.45, 0.48, 0, 0, 0.71, 0.48, 0.57, 0, 0, 0;0, 0.34, 0, 0.71, 0, 0, 0.57, 0.72, 0, 0;0, 0, 0, 0.48, 0, 0, 0, 0, 0, 0;0, 0, 0, 0.57, 0.57, 0, 0, 0, 0.33, 0;0.23, 0, 0, 0, 0.72, 0, 0, 0, 0, 0;0, 0, 0, 0, 0, 0, 0.33, 0, 0, 0;0, 0.39, 0, 0, 0, 0, 0, 0, 0, 0].The label of nodes (S) is provided by the vector S, S=[1, 1, 0.09, 0.02, 1, 1, 0.34, 0.29, 0.17, 0.40].

**S1: True positive rate and Negative positive rate of NetSA, SANDY and GO annotation**

**Table S1.** The true positive rate and the negative positive rate of NetSA, SANDY and GO annotation based on the KEGG pathway of cell cycle as an objective criterion. Table S1 is the other manifestation format of Fig 2 (B).

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Methods | GO annotation | SANDY | NetSA (λ=1) | NetSA (λ=0.9) | NetSA (λ=0.8) | NetSA (λ=0.7) | NetSA (λ=0.6) | NetSA (λ=0.5) | NetSA (λ=0.4) | NetSA (λ=0.3) | NetSA (λ=0.2) | NetSA (λ=0.1) | NetSA (λ=0) |
| TPR | 0.6941 | 0.5059 | 0.7765 | 0.7765 | 0.8118 | 0.8235 | 0.8235 | 0.8353 | 0.8353 | 0.8353 | 0.8471 | 0.8471 | 0.8588 |
| NPR | 0.0561 | 0.0748 | 0.0674 | 0.0825 | 0.1149 | 0.1505 | 0.1835 | 0.2061 | 0.2179 | 0.2304 | 0.2405 | 0.2482 | 0.3572 |

**TPR** represents the true positive rate; **NPR** represents the negative positive rate.

**S2: Effectiveness of our method by using Reactome pathway of cell cycle as an objective criterion**

We also selected the cell cycle pathway from the Reactome pathways including 1321 interactions and 108 nodes, and compared our method with other methods again by using Reactome pathway of cell cycle as an objective criterion. The accuracy of the predicted genes associated with the known cell cycle decreases. But when SANDY and NetSA methods have similar false positive rates by adjusting the parameter λ, SANDY identified 34% genes reported in Reactome pathways with 8% false positive rate while NetSA contained almost 40% Reactome genes with 6% false positive rate. As a reference case, we also conducted GO annotations, which identified 40% with only 6% false positive rate, similar to our result. In addition, as shown in Figure S1, with different parameters, NetSA can have various results on cell cycle pathway. Thus, the NetSA is better than SANDY (the results are shown in Figure S1). A reason / cause for this decline of the accuracy on Reactome pathway than KEGG pathway would be the overlapping genes from two different databases. The numbers of the genes related to the cell cycle in KEGG pathways and Reactome pathways are 125 and 108, respectively. And the number of the overlapping genes, which can be found in both two pathway databases, is only 28 (overlapping ratios are 22% and 25% for KEGG pathways and Reactome pathways, respectively); but, the overlapping ratios of the genes related to cell cycle between GO annotations and Reactome pathways / KEGG pathways are 40% / 70%. Obviously, in the case of cell cycle, the Reactome pathways and KEGG pathways are very different. Since our method is supervised by GO annotations, our method would be better to use the KEGG pathways.

Note that, Fig S1 is not the ROC curve, and we only used it to compare SANDY and NetSA according to true positive rate and false positive rate of the selected sub-networks in the Reactom pathway of cell cycle. To illustrate that we evaluate these methods according to true positive rate and negative positive rate, we use Table S2 to further summarize the true positive rate and the negative positive rate of these methods in Fig 2 (B).

**Table S2.** The true positive rate and the negative positive rate of NetSA, SANDY and GO annotation based on the Reactom pathway of cell cycle as an objective criterion. Table S2 is the other manifestation format of Fig S1.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Methods | GO annotation | SANDY | NetSA (λ=1) | NetSA (λ=0.9) | NetSA (λ=0.8) | NetSA (λ=0.7) | NetSA (λ=0.6) | NetSA (λ=0.5) | NetSA (λ=0.4) | NetSA (λ=0.3) | NetSA (λ=0.2) | NetSA (λ=0.1) | NetSA (λ=0) |
| TPR | 0.4058 | 0.3478 | 0.4058 | 0.4348 | 0.4783 | 0.5507 | 0.5652 | 0.5652 | 0.5652 | 0.5797 | 0.5797 | 0.5942 | 0.6667 |
| NPR | 0.0650 | 0.0801 | 0.0783 | 0.0928 | 0.1250 | 0.1593 | 0.1918 | 0.2145 | 0.2264 | 0.2385 | 0.2488 | 0.2562 | 0.3635 |

**TPR** represents the true positive rate; **NPR** represents the negative positive rate.

**S3:The development of human T1D indicated by NetSA**

## To study T1D progression, we applied NetSA to 104 human samples from 24 normal, 43 newly diagnosed, 20 one-month and 20 four-month T1D patients [[1](#_ENREF_1)]. The disease genes associated with T1D had 3976 genes and were downloaded from the Genecards database (http://www.genecards.org)

***Comparison between NetSA and traditional network-partition***

Similar to above HCC study, we compared NetSA with network-partition method and showed the result in Fig S2. The average disease genes enrichment ratios of these methods in four stages of T1D are shown in Table S3, which evaluates the performance of these methods. And the increased ratio indicates the improvement of NetSA compared to the network-partition method. As shown in Figure 3 and Table S3, clearly GO terms in NetSA are significantly enriched of the known T1D associated genes in all four stages.

**Table S3**. Performance of NetSA and the network-partition method in the disease genes enrichment ratios based on the T1D analysis. And the increased ratio indicates the improvement of NetSA compared to the network-partition method in the disease genes enrichment ratios

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Methods | T1D (normal) | T1D (new ) | T1D (1-month) | T1D (4-month ) |
| NetSA | 0.543 | 0.570 | 0.586 | 0.600 |
| network-partition method | 0.404 | 0.412 | 0.500 | 0.547 |
| increased ratio | 34.4% | 38.2% | 17.3% | 9.5% |

***T1D progression characterized by NetSA***

We also constructed the function spectrum of T1D progression based on this gene expression data. As shown in Supplementary Table S2, for instance, we found that cellular response to bacterial lipopeptide (GO:0071727), detection of diacyl bacterial lipopeptide (GO:0042496) and detection of triacyl bacterial lipopeptide (GO:0042495) play roles on five disease stages by the reciprocal of the variation coefficient (CV) of the state vector.  Actually, bacterial lipopeptide is known to activate Toll-like receptor 2 (TLR2), which is found to be expressed on cultured podocytes and glomerular endothelial cells [[2](#_ENREF_2)], and TLR2 through innate immune receptors to reduce T1D incidence under the conditions of low hygiene, and the potential of targeting them for treating T1D [[3](#_ENREF_3)]; and hence, the result selected by our analysis could be validated. Based on the Standard Deviation (SD) of the state vector, we also found the Wnt signaling pathway involved in somitogenesis (GO: 0090244) plays important roles on disease stages (new and one-month stages). Because of the Wnt pathway is involved in lipid metabolism and glucose homeostasis, and mutations in LRP5 may lead to the development of diabetes and obesity[[4](#_ENREF_4)]. Hence, these facts validated our results.

Moreover, we show the clustering results between the consistent BPs and samples by the SN in Figure S3, which demonstrates the ability of the consistent BPs to distinguish the samples.

***Additional comparison between NetSA and SPIA***

Based on the above strategy, we selected normal stage and four-month stage in T1D data to run NetSA and SPIA. We got the 23 super pathways association with T1D in MalaCards database [[5](#_ENREF_5)] and found NetSA and SPIA identified 4 and 3 pathways of these pathways, respectively. The identified pathways are summarized in Fig S4. The two methods both can find T1D associated pathways, and NetSA can obviously obtain more complete information associated with the diseases.

**S4: Influence of the independent GO terms**

The independent GO terms in this article are those terms in a GO term set without ancestor relationship and thus are considered to have independent functional relationships.

In order to make the GO terms clear on independence, we marked whether or not a GO term is an independent GO term in the results for the two methods, i.e., Supplementary Table S3 and Supplementary Table S4 represent the results of NetSA and the network-partition method including the four disease phases of HCC, and Supplementary Table S5 and Supplementary Table S6 are the results of NetSA and the network-partition method including the four disease phases of T1D.

We found that most of the GO terms in NetSA are independent GO terms; on the contrary, the GO terms in the network-partition method are generally not independent. For instance, in HCC, the enrichment ratio of the independent GO terms in NetSA for the four stages, including normal, cirrhotic, dysplastic and early stages, are 82.5%, 85.7%, 84.8% and 85.4%, and the module-based method are 8.7%, 4.8%, 8.5% and 3.9%; in T1D, the enrichment ratio of the independent GO terms in NetSA for the four stages, including normal, new, 1-month and 4-month, are 81.3%, 75.5%, 86.2%, and 87.5%, and the network-partition method are 8.3%, 5.5%, 6.8% and 10.6%. Therefore, the improved performance of NetSA is due to the independent GO terms.

**S5: Classify the samples by using average gene expressions of those sub-networks corresponding to the consistent BPs selected by NetSA**

We clustered the samples and the consistent BPs by the average gene expressions of those sub-networks for the HCC and T1D progression data, respectively. And the result was showed in Figure S5, Figure S6, respectively. We found that the sub-network set corresponding to the consistent BPs can distinguish the samples well.

**S6: Application of NetSA on T1D analysis based on the integrated networks**

The integrated networks from the multiple sources have more human proteomes and more genetic related relationships. Actually, NetSA could easily extract the specific sub-network and achieve better performance because this specific network of one biological function would have more genetic related relationships in the integrated networks than in the PPI network. Meanwhile, the integrated networks also contain some predicted/inconvincible genetic related relationships, and thus the conventional network-partition methods may be under-estimated by using those networks.

We compared NetSA and network-partition method on the T1D analysis by applying the integrated networks, which is obtained from the STRING database (<http://string-db.org/>). The results are shown in Fig S7. The average enrichment ratios of disease genes by these methods in four stages of T1D are shown in Table S4. The increased ratio indicates the improvement of NetSA compared to the conventional network-partition method. The results show that NetSA actually has a better performance than the network-partition method on these integrated networks.

On the T1D analysis, we also compared the performances of NetSA/ network-partition method by using the PPI network or the integrated network (shown in Fig S2 and Table S3). And the performance of NetSA increases as 6.63% when using the integrated network instead of the PPI network, meanwhile, the performance of network-partition method descends as 4.07%. Thus, NetSA would have a better performance by using the integrated network, but the network-partition method has a worse performance.

Although the PPI network contains the less human proteome, it provides us the reliable genetic regulated relationships and has less predicted genetic regulated relationships. That means it would provide a fair evaluation on the effectiveness of NetSA and the network-partition method.

**Table S4.** Performance of NetSA and network-partition method in the disease genes enrichment ratios on the T1D analysis by using the integrated networks. And the increased ratio indicates the improvement of NetSA compared to the network-partition method in the average disease genes enrichment ratios

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Methods | T1D (normal) | T1D (new) | T1D (1-month) | T1D (4-month) |
| NetSA | 0.604 | 0.621 | 0.615 | 0.609 |
| network-partition method | 0.356 | 0.437 | 0.518 | 0.470 |
| increased ratio | 69.4% | 42.0% | 18.7% | 29.4% |

**References**

[1] E. C. Kaizer, C. L. Glaser, D. Chaussabel, J. Banchereau, V. Pascual, and P. C. White, "Gene expression in peripheral blood mononuclear cells from children with diabetes," *J Clin Endocrinol Metab,* vol. 92, pp. 3705-11, Sep 2007.

[2] R. D. Pawar, L. Castrezana-Lopez, R. Allam, O. P. Kulkarni, S. Segerer, E. Radomska*, et al.*, "Bacterial lipopeptide triggers massive albuminuria in murine lupus nephritis by activating Toll-like receptor 2 at the glomerular filtration barrier," *Immunology,* vol. 128, pp. e206-21, Sep 2009.

[3] S. Karumuthil-Melethil, N. Perez, R. Li, and C. Vasu, "Induction of innate immune response through TLR2 and dectin 1 prevents type 1 diabetes," *J Immunol,* vol. 181, pp. 8323-34, Dec 15 2008.

[4] T. Jin, "The WNT signalling pathway and diabetes mellitus," *Diabetologia,* vol. 51, pp. 1771-1780, Oct 2008.

[5] N. Rappaport, N. Nativ, G. Stelzer, M. Twik, Y. Guan-Golan, T. I. Stein*, et al.*, "MalaCards: an integrated compendium for diseases and their annotation," *Database-the Journal of Biological Databases and Curation,* Apr 12 2013.

**Figure Legend**

**Fig S1**. Effectiveness of NetSA to identify cell cycle pathway. Reactome pathway of cell cycle process is used as objective criterion. The results of SANDY and also GO annotations are also given. Different results of NetSA are obtained by setting various . False positive rate and true positive rate are calculated by comparing genes in a candidate method and genes in Reactome pathway. Clearly, when SANDY and NetSA methods have similar false positive rates by adjusting the parameterλ, SANDY identified 34% genes reported in Reactome pathways with 8% false positive rate while NetSA contained almost 40% Reactome genes with 6% false positive rate. As a reference case, we also conducted GO annotations, which identified 40% with only 6% false positive rate, similar to our result. In addition, with different parameters, NetSA can have various results on cell cycle pathway. Thus, the NetSA is better than SANDY.

**Fig S2.** Comparison between NetSA and the network-partition method based on the T1D analysis, which is evaluated by the relevance to T1D associated genes. The blue curve represents the performance of NetSA. The red curve represents the performance of the network-partition method. The X-axis represents the sequence of biological process groups, where a group contains 20 terms. The Y-axis represents the ratio of the known T1D associated genes enriched in a GO term (corresponding to a sub-network from NetSA or a module from the network-partition method). Note that, for a biological process group, its value is the mean of the known disease gene enrichment ratio of all 20 terms in this group.

**Fig S3.** Clustering the samples by the consistent BPs as features. These are 1151 consistent BP terms and 106 samples in four stages. We show the result in a heat map style, where four colors in bar correspond to the samples in the four stages. Clearly, the consistent BPs can accurately distinguish the samples as features with the accuracy as 100%.

**Fig S4.** The figure shows different capacity of two methods to identify T1D related pathways. red bar indicates that the pathway is identified by NetSA or SPIA, while black means it not identified. each column represents a pathway and each row shows a method.

**Fig S5.** Clustering the samples by using the average gene expressions of the sub-networks corresponding to the consistent BPs for the HCC progression data. These are 1023 consistent BP terms and 54 samples in four stages. We show the result by heat map. In the heat map, there are four colors in bar corresponding to the samples in the four stages. Clearly, based on the average gene expression of the sub-network, the consistent BP set can accurately distinguish the sample set as features well.

**Fig S6.** Clustering the samples by using the average gene expressions of the sub-networks corresponding to the consistent BPs for the T1D progression data. These are 1151 consistent BPs and 106 samples which contain four stages. We show the result in a heat map style. In the heat map, there are four colors in bar corresponding to the samples in the respective four stages. Clearly, based on the average gene expression of the sub-network, the consistent BPs can distinguish the samples well.

**Fig S7.** Comparison between NetSA and the network-partition method based on the T1D analysis by using the integrated networks, which is evaluated by the relevance to T1D associated genes. The blue curve represents the performance of NetSA. The red curve represents the performance of the network-partition method. The X-axis represents the sequence of biological process groups, where a group contains 20 terms. The Y-axis represents the ratio of the known T1D associated genes enriched in a GO term (corresponding to a sub-network from NetSA or a module from the network-partition method). Note that, for a biological process group, its value is the mean of the known disease gene enrichment ratio of all 20 terms in this group.

**Supplementary Table Legend**

**Supplementary Table S1:** Biological function spectrum for the HCC progression by integrating the most relevant term levels in all stages, computing and merging the significance values of the results in each stage. The significance values of the results in each stage can reveal deeper molecular pathogenesis of HCC development and progression.

**Supplementary Table S2:** Biological function spectrum for the T1D progression by integrating the most relevant term levels in all stages, computing and merging the significance values of the results in each stage. The significance value of the result in each stage can reveal deeper molecular pathogenesis of T1D development and progression.

**Supplementary Table S3:** The GO terms results and its independent GO terms of NetSA in the four disease phases of HCC.

**Supplementary Table S4:** The GO terms results and its independent GO terms of the module-based method in the four disease phases of HCC

**Supplementary Table S5:** The GO terms results and its independent GO terms of NetSA in the four disease phases of T1D.

**Supplementary Table S6:** The GO terms results and its independent GO terms of the module-based method in the four disease phases of T1D