**Supplementary materials for gene co-opening network deciphers gene functional relationships**

Wenran Li,1 Meng Wang,1 Jinghao Sun,1 Yong Wang,2,\* Rui Jiang1,\*

1MOE Key Laboratory of Bioinformatics; Bioinformatics Division and Center for Synthetic & Systems Biology, TNLIST; Department of Automation, Tsinghua University, Beijing 100084, China.

2Academy of Mathematics and Systems Science, National Center for Mathematics and Interdisciplinary Sciences, Chinese Academy of Sciences, Beijing 100080, China.

\*Corresponding Authors:

Rui Jiang, Email: ruijiang@tsinghua.edu.cn;

Yong Wang, Email: ywang@amss.ac.cn.

Supplementary Texts

Text S1. Batch effects detection

Batch effects are defined as the systematic non-biological variations between groups of samples or batches [1](#_ENREF_1). Ignoring batch effects can result in serious influence on downstream analysis. Here, we detect the batch effect of DNase-seq data collected from three different projects (see Methods) with guided principle component analysis (gPCA) [2](#_ENREF_2), a common method for identifying batch effects in high-throughput genomic data. The method provides a statistic δ to measure the batch effects. Large value of δ (values near 1) implies that the batch effect is large. To determine whether δ is significantly larger than would be obtained by chance, a *p*-value is estimated using a permutation distribution created by permuting the batch vector for N times (i.e. N=1000). In our case, the value of δ is 0.352 and the *p*-value is 0.447, which indicates that there is no significant batch effect in the DNase-seq data. We further demonstrate the first and second principle components of DNase-seq data on a two dimensional plane (Figure S1), where the DNase-seq profiles from different projects scatter in disorder and all profiles tend to cluster into one group. This also indicates that there is no significant batch effect within DNase-seq data.

Text S2. Construction of a co-expression network

We define a co-expression network by calculating the pairwise correlations between gene expressions. Specifically, we calculate the absolute value of Pearson correlation coefficient (PCC) for all pairwise genes across all RNA-seq samples. Then, we transfer the correlation matrix into a weighted undirected network with a given threshold. Here, the given threshold is set to be 0.8 to make sure the best quality of “scale free” property of the network. The resulting co-expression network contains 10,595 nodes and 190,278 interactions.

Text S3. Network permutation.

We permut the co-opening network by shuffling its nodes with all edges maintained. Specifically, we keep the values of unweighted matrix of the co-opening network unchanged and only shuffle the nodes of the network. In this way, we obtain a permuted network with randomly constructed edges and maintain node degree distribution in the meanwhile.

Text S4. Uniqueness analysis of the co-opening network.

The co-opening network shows different aspects of the biology than PPI networks and co-expression networks in several ways. First, the co-opening network focuses on the consistence of chromatin accessibility of a pair of genes across multiple experiments. This makes it possible for the network to capture the relationship of genes before transcription (say, co-regulation relationships), while co-expression networks and PPI networks mainly concentrate on gene interactions after transcription. Second, when prioritizing disease genes with GWAS data using network-based methods, the coverage of genes in regions related to GWAS is an important characteristic of the network. In our work, we applied the co-opening network to prioritize disease genes for Psoriasis. The disease gene ranked top by the co-opening network is not covered by the co-expression network or the PPI network. This suggests that the co-opening network tend to link two disease-associated genes even if their expression levels are unrelated. Third, the co-opening network and the co-expression network focus on different biological functions. We extract genes which are uniquely covered by the co-opening network or the co-expression network, and identify the functional enrichment of these genes (Figure S5). The results show that genes uniquely covered by the co-opening network are enriched in biological functions different from those covered by the co-expression network.

Some of the unique biological aspects of the co-opening network can be quantified. As shown in Figure 2A and 2B in the main text, the co-opening network uniquely covers 1,860 nodes and 269,086 edges, which are not involved in either the PPI network or the co-expression network.

Since the co-expression network and the co-opening network have similar sizes, we compare the preferences of functional enrichment of these two networks. Specifically, we extract 5,132 genes uniquely covered by the co-opening network and detect the functional enrichment of these genes with GO database [3](#_ENREF_3). Similarly, we detect the functional enrichment of 4,266 genes uniquely covered by the co-expression network. For genes uniquely covered by each network, the top 10 significantly enriched functions are shown in Table S4. We find that functions where genes uniquely covered by the co-opening network enriches are mainly related about membranes, receptor protein signaling pathways and signal transduction, while genes uniquely covered by the co-expression network enriches in functions related to the nuclear, macromolecular complexes and protein complexes. Since genes uniquely covered by different networks enriches in different functions, it is better to choose the co-opening network rather than the co-expression network when studying gene functions related to membranes, receptor protein signaling pathways and signal transduction. This makes the co-opening network unique, and helps others to decide when to use the co-opening network.

Text S5. Details regarding the openness score.

Chromatin accessible regions (peaks) are identified using the tool Hotspot [4](#_ENREF_4). First, enrichment of tags along the genome is gauged in a small window (200–300 bp) relative to a local background model based on the binomial distribution and using the observed tags in a 1Mb surrounding window. Then, each mapped tag is given a z-score (explained below) relative to the surrounding small and background windows centered on the tag. Finally, a peak is defined as a succession of neighboring tags, each of whose z-score was greater than 2.

*z-score calculation*. Suppose *n* observed tags are mapped to the small window, and *N* total tags are mapped to the 1Mb surrounding background window (*N* ≥ *n*). Each tag in the background window is considered an ‘experiment’ with a favorable outcome if it falls in the smaller window. Assuming each base in the 1Mb window is equally likely, the probability of success for each tag is therefore *P* = 250/50,000. Under these assumptions, the binomial distribution applies, and the expected number of tags falling in the smaller window is . The standard deviation of this expected value is .Finally, the *z*-score for the observed number of tags in the smaller window is .

The distribution of peak sizes is shown in Figure S6. In total, 161,918,452 peaks are identified by Hotspot in 628 DNase-seq experiments. 89.47% peak sizes are less than 1kb. The mean and median size of chromatin accessible regions are 500.51bp and 332bp.

The TSS locations are taken from Ensembl release v75 [5](#_ENREF_5). We don’t set a minimum distance between alternative TSSs of the same gene because of several reasons. First, only 39.04% of 25,315 genes have more than one isoform. Second, as shown in Figure S7, 81.00% pairs of alternative TSSs have longer distances than 1kp, which makes it hard for two alternative TSSs to fall into the same peak. Third, we only take the maximum openness score of peaks overlapping with an alternative promoter as the openness score of the promoter, which means that two alternative TSSs falling into one peak may not influence the openness score of the promoter. Therefore, the probability that two alternative TSSs fall into one peak and influence the further network construction is low. More importantly, we calculate the co-opening scores of two alternative promoters as the absolute value of Pearson correlation coefficient and define the co-opening score of two genes as the maximum pairwise co-opening scores of their alternative promoters. That is to say, we only cares about the maximum pairwise co-opening scores of alternative promoters of two genes. It doesn’t matter if two alternative promoters of a gene are the same. Therefore, even if two alternative promoters fall into one peak and happen to have the same openness score, it would make no difference to the final co-opening scores.

Text S6. Random walk with different restart probabilities.

We consider different values for the restart parameter ranging from 0.1 to 0.9 with step size 0.05. For each restart probability, we record the top three genes with the highest stable scores. The result is shown in Table S5.

*TNFSF14* is a disease-associated gene with an insignificant GWAS *p*-value. We focus our analysis on this gene to show the co-opening network prioritizes disease genes. The result shows that *TNFSF14* is stably ranked the first or second whatever the restart probability is, which suggests that the restart probability makes no influence on our further study.

Supplementary Figures



**Figure S1. Batch effects detection for DNase-seq profiles.** The demonstration of the first and second principle components of DNase-seq data from different projects, ROADMAP project (blue), ENCODE project (red) and HapMap project (green).

****

**Figure S2.** **Properties of co-expression network and PPI network.** (A) Distribution of shortest path between pairs of genes in PPI network. On average, any two genes in co-open network are connected via 3.87 links. (B) Distribution of shortest path between pairs of genes in co-expression network. The average shortest path is 6.12. (C) Degree distribution of PPI network. The black line (slope of -1.75) depicts the least-squares fit of the data to a linear line in the log-log plot. (D) Degree distribution of co-expression network. The black line (slope of -1.5) depicts the least-squares fit of the data to a linear line in the log-log plot.

****

**Figure S3. Neighbors of functionally annotated genes.** (A) The subnetwork of FKBPL, containing FKBPL (purple) and its neighbors (grey). Dark blue circles around FKBPL are neighbors annotated with the function of cell cycle by Reactome database. (B) The subnetwork of gene ILF3, containing ILF3 and its neighbors.

****

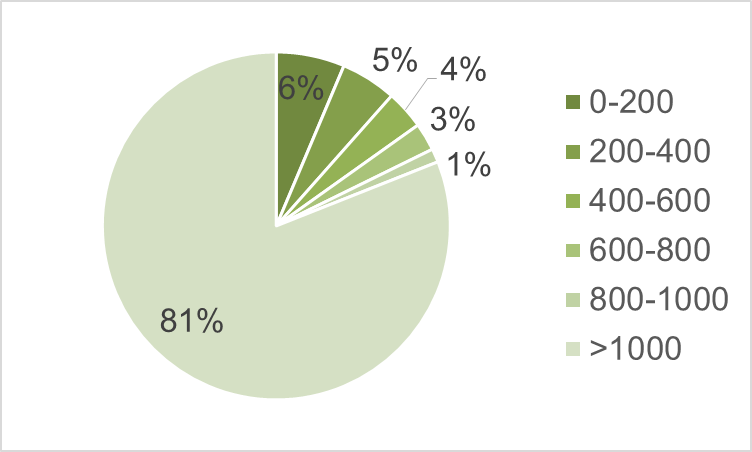
**Figure S4. Disease-associated genes tend to be linked in co-opening network.** The comparison of the link ratio of each disease of KEGG (A), Reactome (B), OMIM (C) and DisGeNet (D) databases in the co-opening network (red dot) with that in permuted networks (grey boxplot, 100 permutations).

****

**Figure S5.** The Venn diagram illustrates the nodes of the co-opening network (blue) and the nodes of the co-expression network (orange).

****

**Figure S6.** The distribution of chromatin accessible region sizes detected by Hotspot.



**Figure S7.** The distribution of distances between pairs of alternative TSSs. Different colors show different distance intervals.

**Supplementary Table**

## Table S1. Comparison of different networks.

|  |  |  |  |
| --- | --- | --- | --- |
| Network | Co-open network | Co-expression network | PPI network |
| Number of Nodes | 11,461 | 10,595 | 13,522 |
| Number of Edges | 271,900 | 190,278 | 72,180 |
| Average Degree [[1]](#footnote-1) | 47.45 | 35.92 | 10.68 |
| Cluster Coefficient [6](#_ENREF_6),[[2]](#footnote-2) | 0.435 | 0.478 | 0.079 |
| Network Diameter [7](#_ENREF_7),[[3]](#footnote-3) | 15 | 42 | 13 |
| Network Heterogeneity [8](#_ENREF_8),[[4]](#footnote-4) | 1.781 | 1.757 | 2.015 |
| R square [9](#_ENREF_9),[[5]](#footnote-5) | 91.18 | 90.29 | 88.89 |

## Table S2. Network components.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Component | Size [[6]](#footnote-6) | Biological function [[7]](#footnote-7) | Genes in component [[8]](#footnote-8) | Enrichment *p*-value [[9]](#footnote-9) | Adjusted *p*-value [[10]](#footnote-10) |
| 1 | 469 | Promoter opening | 26/48 | 4.90E-24 | 8.93E-20 |
| Cell cycle | 47/233 | 2.87E-20 | 1.31E-16 |
| RNA binding | 31/132 | 1.11E-15 | 4.34E-11 |
| 2 | 218 | Immune system process | 32/201 | 7.76E-21 | 6.09E-16 |
| 3 | 115 | Dilated cardiomyopathy | 12/66 | 1.81E-12 | 3.64E-09 |
| 4 | 82 | Ectoderm development | 9/49 | 5.10E-11 | 5.72E-07 |
| 5 | 65 | RNAPI transcription | 8/65 | 2.69E-09 | 2.79E-06 |
| 6 | 65 | Generic transcription pathway | 11/186 | 6.39E-09 | 6.29E-06 |
| 7 | 114 | Cell signaling | 15/254 | 2.98E-08 | 8.07E-05 |
| Receptor protein signaling pathway | 15/205 | 1.61E-09 | 8.45E-06 |
| 8 | 212 | Lipid metabolic process | 18/196 | 2.19E-08 | 6.15E-05 |
| 9 | 74 | Collagen formation | 6/37 | 1.17E-07 | 9.46E-05 |
| 10 | 132 | Nervous system development | 13/230 | 2.38E-06 | 3.40E-03 |
| 11 | 287 | Signaling By hippo | 5/14 | 1.58E-05 | 8.11E-03 |
| 12 | 64 | Metabolism of lipids | 9/292 | 3.22E-05 | 1.52E-02 |
| 13 | 109 | Integral to plasma membrane | 18/611 | 1.66E-05 | 1.63E-02 |
| 14 | 138 | Immune system process | 11/201 | 3.03E-05 | 2.45E-02 |
| 15 | 1157 | Signal transduction | 16/55 | 6.93E-05 | 2.80E-02 |
| 16 | 60 | Female pregnancy | 4/37 | 3.94E-05 | 2.97E-02 |
| 17 | 111 | TGF Beta signaling pathway | 5/56 | 2.01E-04 | 4.38E-02 |

## Table S3. The results of prioritizing disease-associated genes for Psoriasis with the co-opening network.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Gene [[11]](#footnote-11) | *p*-value [[12]](#footnote-12) | *p*-value Rank [[13]](#footnote-13) | Score Rank in Co-opening [[14]](#footnote-14) | Score Rank in Co-expression [[15]](#footnote-15) |
| *TNF* | 4.75E-21 | 1 | 1 | NA |
| ***TNFSF14*** | **0.16** | **1,259** | **2** | NA |
| *PSORS1C2* | 1.00E-15 | 3 | 3 | 4,804 |

**Table S4. The functional enrichment of genes uniquely covered by the co-opening network or the co-expression network.**

|  |  |  |  |
| --- | --- | --- | --- |
| Network | Function | *p*-value | FDR *p*-value |
| Co-opening | Signal transduction | 5.16E-10 | 7.50E-07 |
| Co-opening | Plasma membrane | 1.86E-09 | 1.35E-06 |
| Co-opening | Integral to membrane | 6.68E-09 | 3.24E-06 |
| Co-opening | Intrinsic to membrane | 3.59E-08 | 1.30E-05 |
| Co-opening | Plasma membrane part | 6.73E-08 | 1.96E-05 |
| Co-opening | Transmembrane receptor protein signaling pathway | 2.37E-07 | 4.92E-05 |
| Co-opening | Membrane part | 6.48E-07 | 1.10E-04 |
| Co-opening | Membrane | 2.53E-06 | 3.67E-04 |
| Co-opening | Enzyme linked receptor protein signaling pathway | 3.04E-06 | 4.01E-04 |
| Co-opening | Cell surface receptor linked signal transduction | 5.29E-06 | 6.41E-04 |
| Co-expression | Organelle part | 1.00E-18 | 7.27E-16 |
| Co-expression | Nucleus | 8.45E-19 | 7.27E-16 |
| Co-expression | Intracellular organelle part | 2.78E-18 | 1.35E-15 |
| Co-expression | Nuclear part | 1.06E-15 | 3.11E-13 |
| Co-expression | Macromolecular complex | 1.07E-15 | 3.11E-13 |
| Co-expression | Protein complex | 2.28E-12 | 5.53E-10 |
| Co-expression | Organelle lumen | 9.93E-12 | 1.80E-09 |
| Co-expression | Membrane enclosed lumen | 9.93E-12 | 1.80E-09 |
| Co-expression | RNA binding | 1.19E-10 | 1.92E-08 |
| Co-expression | Nuclear lumen | 2.82E-09 | 3.98E-07 |

**Table S5. The top three genes ranked by the random walk procedure with different restart probabilities.** *TNFSF14* is the gene we highlight in the analysis part of the main text.

|  |  |  |  |
| --- | --- | --- | --- |
| Restart probability | Top 1 | Top 2 | Top 3 |
| 0.10 | **TNFSF14** | TNF | RPS6KA1 |
| 0.15 | **TNFSF14** | TNF | RPS6KA1 |
| 0.20 | **TNFSF14** | TNF | RPS6KA1 |
| 0.25 | TNF | **TNFSF14** | RPS6KA1 |
| 0.30 | TNF | **TNFSF14** | RPS6KA1 |
| 0.35 | TNF | **TNFSF14** | RPS6KA1 |
| 0.40 | TNF | **TNFSF14** | RPS6KA1 |
| 0.45 | TNF | **TNFSF14** | RPS6KA1 |
| 0.50 | TNF | **TNFSF14** | RPS6KA1 |
| 0.55 | TNF | **TNFSF14** | RPS6KA1 |
| 0.60 | TNF | **TNFSF14** | PSORS1C2 |
| 0.65 | TNF | **TNFSF14** | PSORS1C2 |
| 0.70 | TNF | **TNFSF14** | PSORS1C2 |
| 0.75 | TNF | **TNFSF14** | PSORS1C2 |
| 0.80 | TNF | **TNFSF14** | PSORS1C2 |
| 0.85 | TNF | **TNFSF14** | PSORS1C2 |
| 0.90 | TNF | **TNFSF14** | PSORS1C2 |

**REFERENCE**

1. M. Benito, J. Parker, Q. Du, J. Wu, D. Xiang, C. M. Perou and J. S. Marron, *Bioinformatics*, 2004, **20**, 105-114.

2. S. E. Reese, K. J. Archer, T. M. Therneau, E. J. Atkinson, C. M. Vachon, M. De Andrade, J.-P. A. Kocher and J. E. Eckel-Passow, *Bioinformatics*, 2013, btt480.

3. A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander and J. P. Mesirov, *Proceedings of the National Academy of Sciences*, 2005, **102**, 15545-15550.

4. S. John, P. J. Sabo, R. E. Thurman, M.-H. Sung, S. C. Biddie, T. A. Johnson, G. L. Hager and J. A. Stamatoyannopoulos, *Nature genetics*, 2011, **43**, 264-268.

5. P. Flicek, M. R. Amode, D. Barrell, K. Beal, K. Billis, S. Brent, D. Carvalho-Silva, P. Clapham, G. Coates, S. Fitzgerald, L. Gil, C. G. Girón, L. Gordon, T. Hourlier, S. Hunt, N. Johnson, T. Juettemann, A. K. Kähäri, S. Keenan, E. Kulesha, F. J. Martin, T. Maurel, W. M. McLaren, D. N. Murphy, R. Nag, B. Overduin, M. Pignatelli, B. Pritchard, E. Pritchard, H. S. Riat, M. Ruffier, D. Sheppard, K. Taylor, A. Thormann, S. J. Trevanion, A. Vullo, S. P. Wilder, M. Wilson, A. Zadissa, B. L. Aken, E. Birney, F. Cunningham, J. Harrow, J. Herrero, T. J. P. Hubbard, R. Kinsella, M. Muffato, A. Parker, G. Spudich, A. Yates, D. R. Zerbino and S. M. J. Searle, *Nucleic Acids Research*, 2014, **42**, D749-D755.

6. A.-L. Barabasi and Z. N. Oltvai, *Nature reviews. Genetics*, 2004, **5**, 101.

7. J. Xu, A. Kumar and X. Yu, *IEEE Journal on Selected Areas in Communications*, 2004, **22**, 151-163.

8. E. Estrada, *Physical Review E*, 2010, **82**, 066102.

9. A.-L. Barabási, R. Albert and H. Jeong, *Physica A: statistical mechanics and its applications*, 2000, **281**, 69-77.

1. The mean degree of a network, where degree means the number of neighbors of a node. [↑](#footnote-ref-1)
2. The degree to which nodes in a network tend to cluster. [↑](#footnote-ref-2)
3. The length of the longest shortest path in a network. [↑](#footnote-ref-3)
4. The degree of inhomogeneity of a network. [↑](#footnote-ref-4)
5. The metric to evaluate the “scale free” property of a network. [↑](#footnote-ref-5)
6. The number of genes in each component. [↑](#footnote-ref-6)
7. Biological functions from Gene Ontology, KEGG database and Reactome database. [↑](#footnote-ref-7)
8. The number of overlapped genes between a biological term and a component, divided by the total number of genes in biological term as well as in co-opening network [↑](#footnote-ref-8)
9. A *p*-value measures the significance of observed number of overlapped genes. [↑](#footnote-ref-9)
10. *P*-values are adjusted with FDR (false discovery rate). [↑](#footnote-ref-10)
11. Genes rank top by random walk. [↑](#footnote-ref-11)
12. Gene *p*-values obtained from GWAS summary data, which implies the significance of association between a gene and the GWAS catalog. [↑](#footnote-ref-12)
13. The rank of *p*-values in GWAS catalog. [↑](#footnote-ref-13)
14. The rank of stable score obtained with random walking on a co-opening network. [↑](#footnote-ref-14)
15. The rank of stable score obtained with random walking on a co-expression network. [↑](#footnote-ref-15)