Uncovering disease-disease relationships through the incomplete interactome

J. Menche et al., Science 347, 1257601 (2015).

linwei@20160419

Outline

- Background
- Disease module hypothesis
  - ➢ The fragmentation of disease modules
  - ➢ Relationship between diseases
- Data Compilation and Analysis

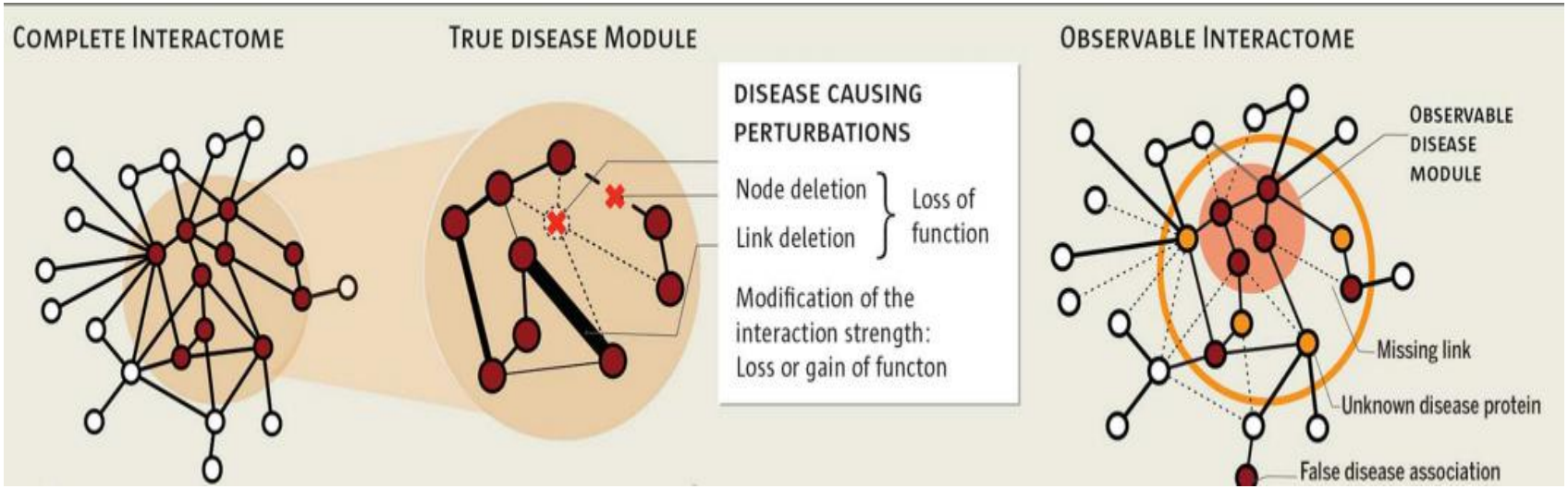
  It works?
  - ✓ the identified disease modules mean something: pathobiological and clinical similarity
  - ✓ Network-based overlap Vs gene-set overlap: examples
- Summary & Discussion

Background

- Disease
  - Abnormality in a single gene(rarely)
  - Interplay of multiple molecular processes
- Interactome
  - A network that integrates all physical interactions within a cell: protein-protein, protein-DNA, metabolic interactions
  - unweighted undirected graph?
- Disease module
  - Disease-associated proteins tend to cluster;
  - Forming a connected subgraph that contains all molecular determinants of a disease

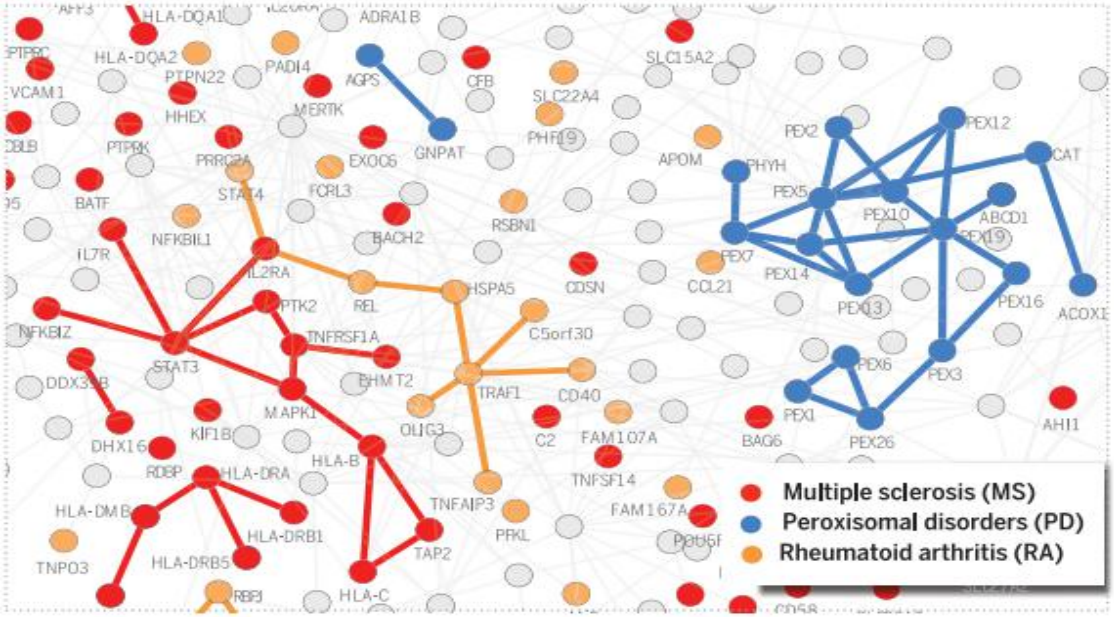
Background

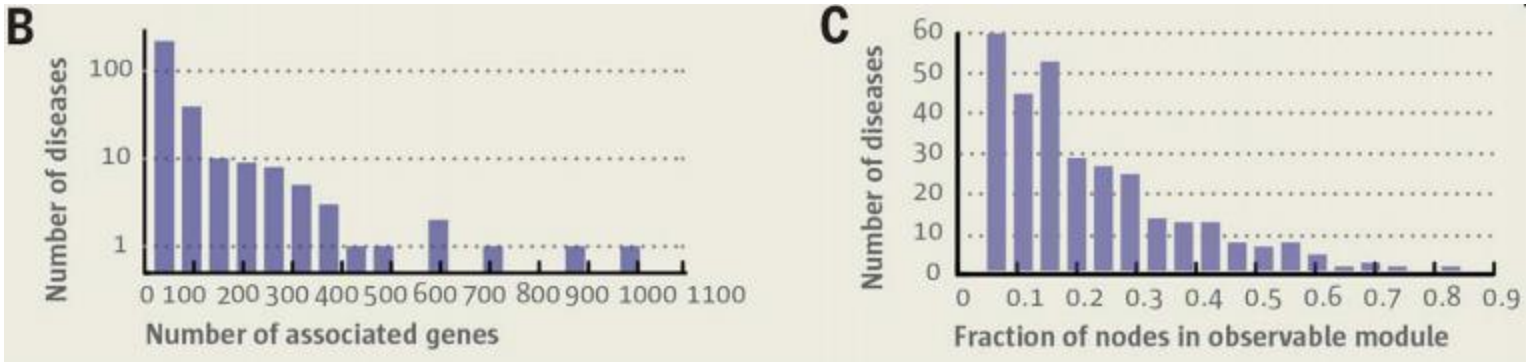
- Interactome & disease module



Background

- Incomplete interactome



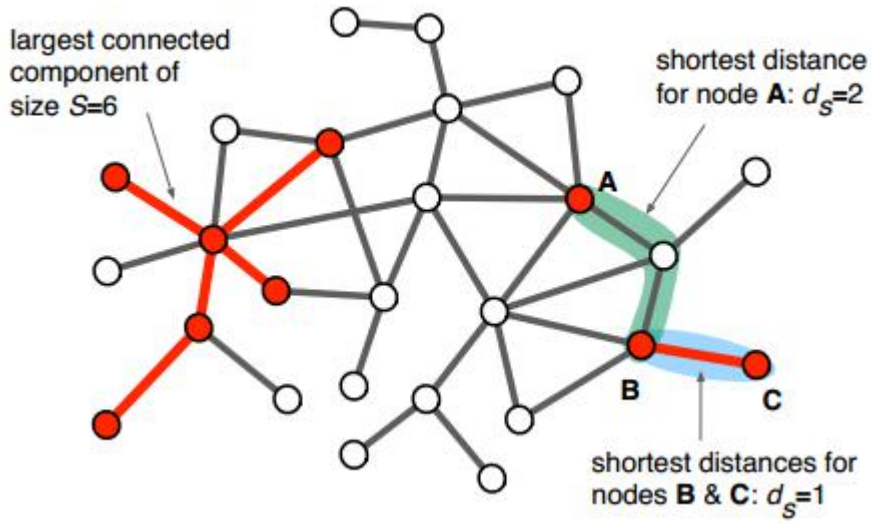


Disease module hypothesis

- The fragmentation of disease modules

two complementary measures to quantify the degree to which disease proteins tend to agglomerate in localized interactome neighborhoods:

- The observable disease module size S;
- Shortest Distance Ds;

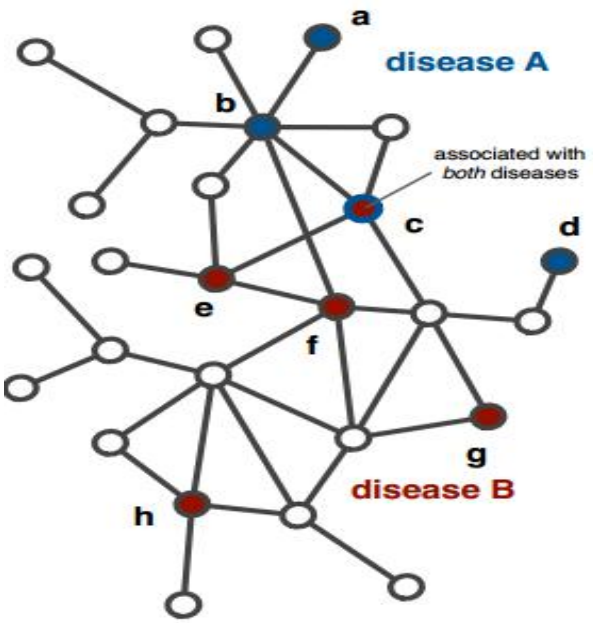


Disease module hypothesis

- Relationship between diseases

We quantify the network-based separation s_{AB} of two diseases A and B by comparing the mean shortest distances $\langle d_{AA} \rangle$ and $\langle d_{BB} \rangle$ *within* the respective diseases, to the mean shortest distance $\langle d_{AB} \rangle$ *between* their proteins:

$$s_{AB} \equiv \langle d_{AB} \rangle - \frac{\langle d_{AA} \rangle + \langle d_{BB} \rangle}{2} . \tag{S9}$$



shortest distances *within* disease **A**:

a: 1
b: 1
c: 1
d: 3

$\langle d_{AA} \rangle = \frac{3}{2}$

shortest distances *within* disease **B**:

c: 1
e: 1
f: 1
g: 2
h: 2

$\langle d_{BB} \rangle = \frac{7}{5}$

shortest distances *between* diseases **A** & **B**:

a: 2
b: 1
c: 0
d: 3

c: 0
e: 1
f: 1
g: 2
h: 3

$\langle d_{AB} \rangle = \frac{13}{9}$

resulting separation:

$s_{AB} = \frac{13}{9} - \frac{1}{2}\left(\frac{3}{2} + \frac{7}{5}\right) = -\frac{1}{180}$

Data Compilation and Analysis

1. Interaction network data

A network of 13,460 proteins that are interconnected by 141,296 physical interactions, involving 299 diseases.

1.1 The Human Interactome(edges)

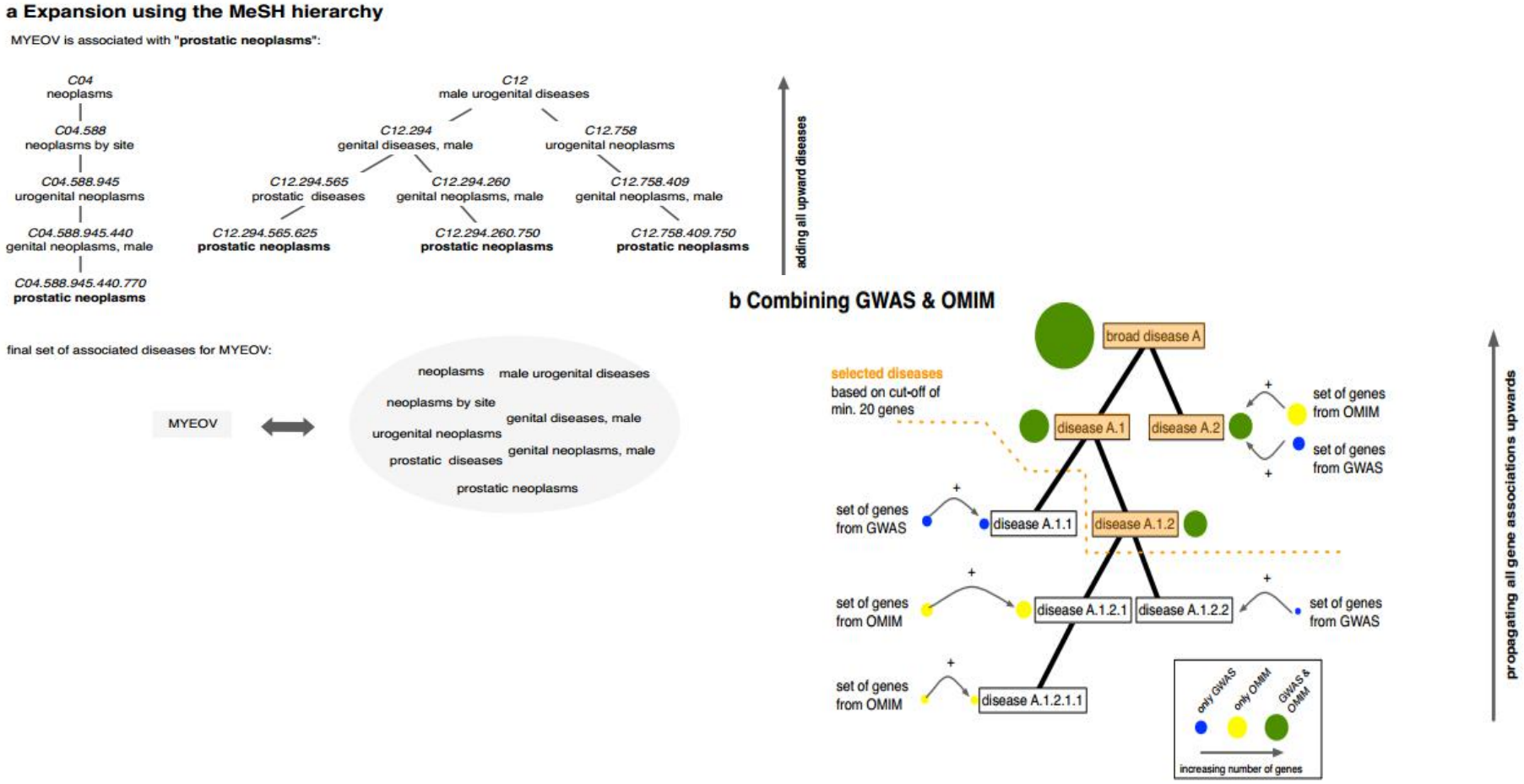
i. Regulatory interaction
ii. Binary interactions
iii. Literature curated interactions
iv. Metabolic enzyme-coupled interactions
v. Protein complexes
vi. Kinase network (kinase-substrate pairs
vii. Signaling interactions

1.2 Disease-gene associations(nodes)

i. OMIM(Online Mendelian Inheritance in Man)
ii. GWAS(Genome-Wide Association Studies)

Data Compilation and Analysis

- Combining GWAS & OMIM data using MeSH(Medical Subject Headings)



Data Compilation and Analysis

2. Gene Ontology (GO) Annotation

The GO ontology is structured as a directed acyclic graph where each term has defined relationships to one or more other terms.

2.1 Three domains:

- Cellular Component
- Biological Process
- Molecular Function

2.2 Similarity of 2 genes (http://go.princeton.edu/GOTermFinder/sampleOutput/query_results.html)

The similarity SGO(a, b) of two proteins a and b is given by the most specific GO term they share:

$$\mathcal{S}_{\mathrm{GO}}(a,b) \equiv \frac{2}{\min(n_i)}\,. \qquad \text{(S1)}$$

2.3 Similarity of a set of disease genes or disease pairs

$$\langle \mathcal{S}_{\mathrm{GO}} \rangle = \frac{1}{n_{\mathrm{pairs}}} \sum_{\{a,b\}} \mathcal{S}_{\mathrm{GO}}(a,b). \qquad \text{(S2)}$$

Data Compilation and Analysis

3. Gene co-expression analysis

3.1 Data source

79 tissues specific gene expression data collected;

5 cancerous tissues and 4 fetal tissues excluded;

3.2 Similarity of 2 genes (Spearman correlation coefficient)

$$r = r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

3.2 Similarity of disease genes/ disease pairs

we calculate the Spearman correlation coefficient $\rho(a, b)$ and the corresponding p-value for each pair of genes a and b across all 70 tissues and average over all n_{pair} pairs:

$$\langle \rho \rangle \equiv \frac{1}{n_{\text{pair}}} \sum_{\{a,b\}} |\rho(a, b)| \,. \qquad \text{(S4)}$$

Data Compilation and Analysis

4. Symptom similarity

4.1 Data source

Symptom-disease associations were extracted from large-scale medical bibliographic records in PubMed and the related MeSH metadata.

4.2 Using symptoms to describe disease

Each disease j is represented by a vector dj = (w1,j,w2,j,...,wn,j) in which wi,j quantifies the PubMed co-occurrence of disease j with symptom i, using tf-idf.

4.3 Similarity of disease pairs

The symptom similarity of two diseases A and B is measured by the cosine similarity cAB o

$$c_{AB} \equiv \frac{\sum_{i=1}^{n} w_{i,A}\, w_{i,B}}{\sqrt{\sum_{i=1}^{n} w_{i,A}^2}\sqrt{\sum_{i=1}^{n} w_{i,B}^2}}, \qquad \text{(S5)}$$

Data Compilation and Analysis

5. Comorbidity

5.1 Data source

A large Medicare patient medical history dataset, containing 13,039,018 patients, to analyse the comorbidity of disease pairs, i.e., the tendency of two diseases to co-occur in the same patients.

5.2 Comorbidity is quantified by the relative risk RR(risk ratio):

$$RR = \frac{n_{AB} n_{\text{tot}}}{n_A n_B}, \text{ where} \qquad \text{(S6)}$$

n_{tot} = total number of patients in the data

n_A, n_B = number of patients diagnosed with disease A and B, respectively

n_{AB} = number of patients diagnosed with both diseases A and B.

RR = (nAB / nA) / (nB / nTotal)

Data Compilation and Analysis

- Gene-set overlap

We use two measures to quantify the overlap between two gene sets A and B:

$$\text{overlap coefficient } C = \frac{|\, A \cap B \,|}{\min(|\, A \,|, |\, B \,|)} \quad , \tag{S13}$$

$$\text{Jaccard-index } J = \frac{|\, A \cap B \,|}{|\, A \cup B \,|} \quad . \tag{S14}$$

$$c_{\text{rand}} = \frac{|\, A \,| \times |\, B \,|}{N} . \tag{S15}$$

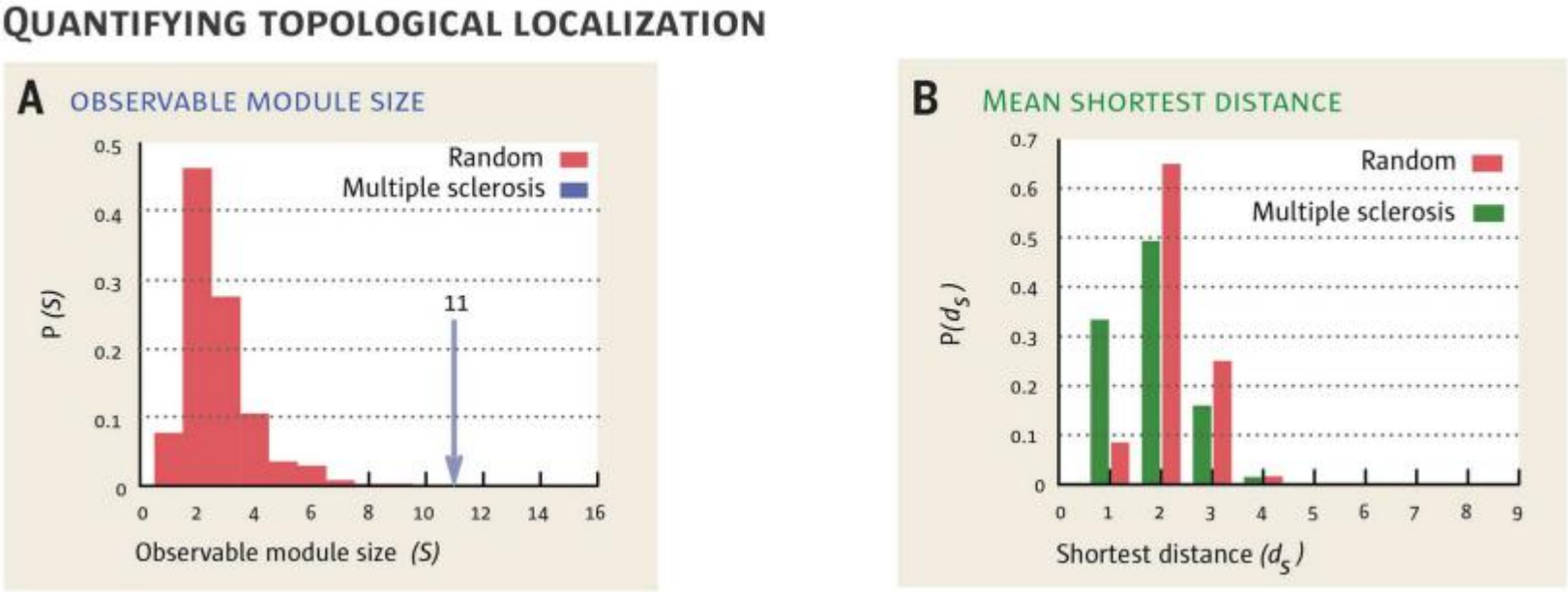
For every observed overlap $c_{\text{obs}} = |\, A \cap B \,|$ we then determine the fold-change

$$\text{fc} = \frac{c_{\text{obs}}}{c_{\text{rand}}} \tag{S16}$$

Data Compilation and Analysis

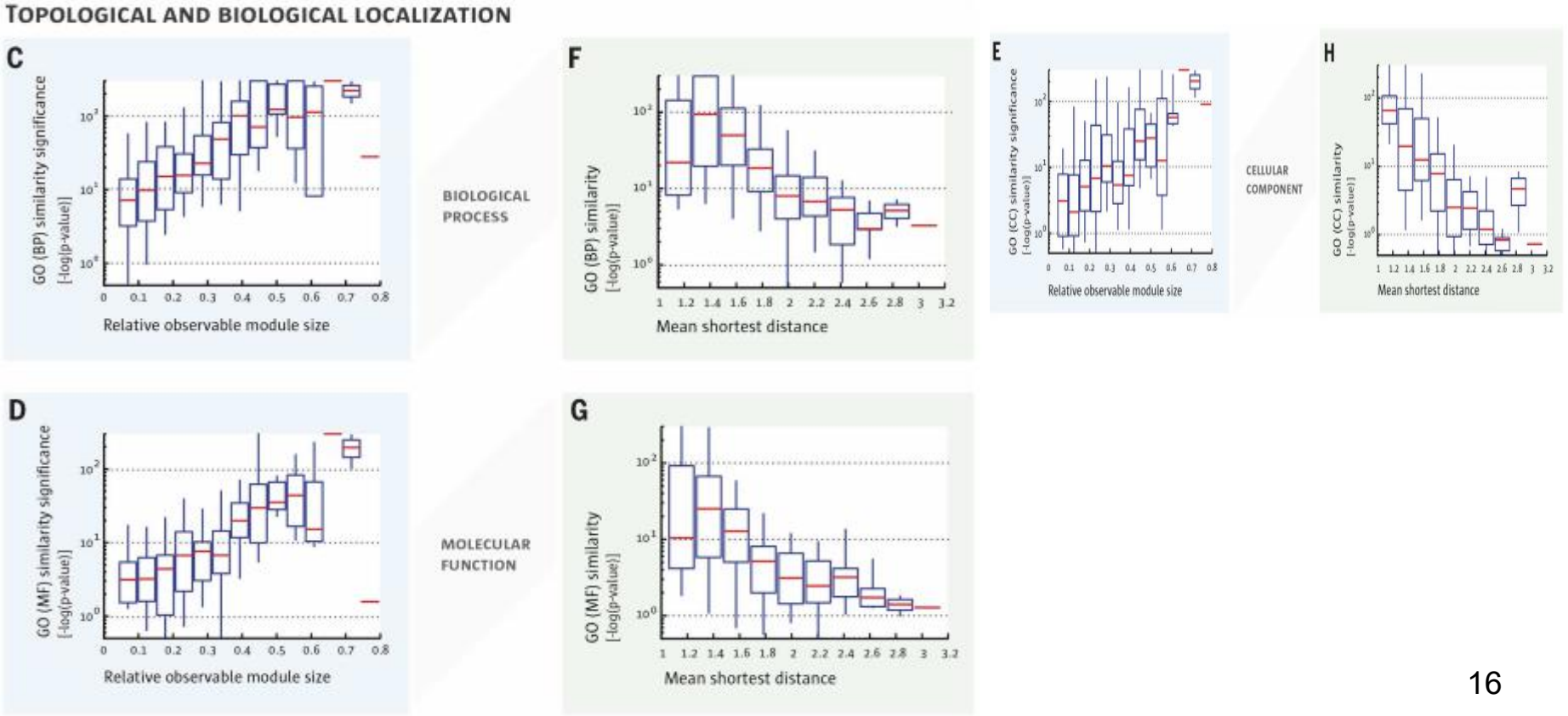
- Disease genes tend to agglomerate in each other's network neighborhood

✓ Module size & Mean shortest distance



Data Compilation and Analysis

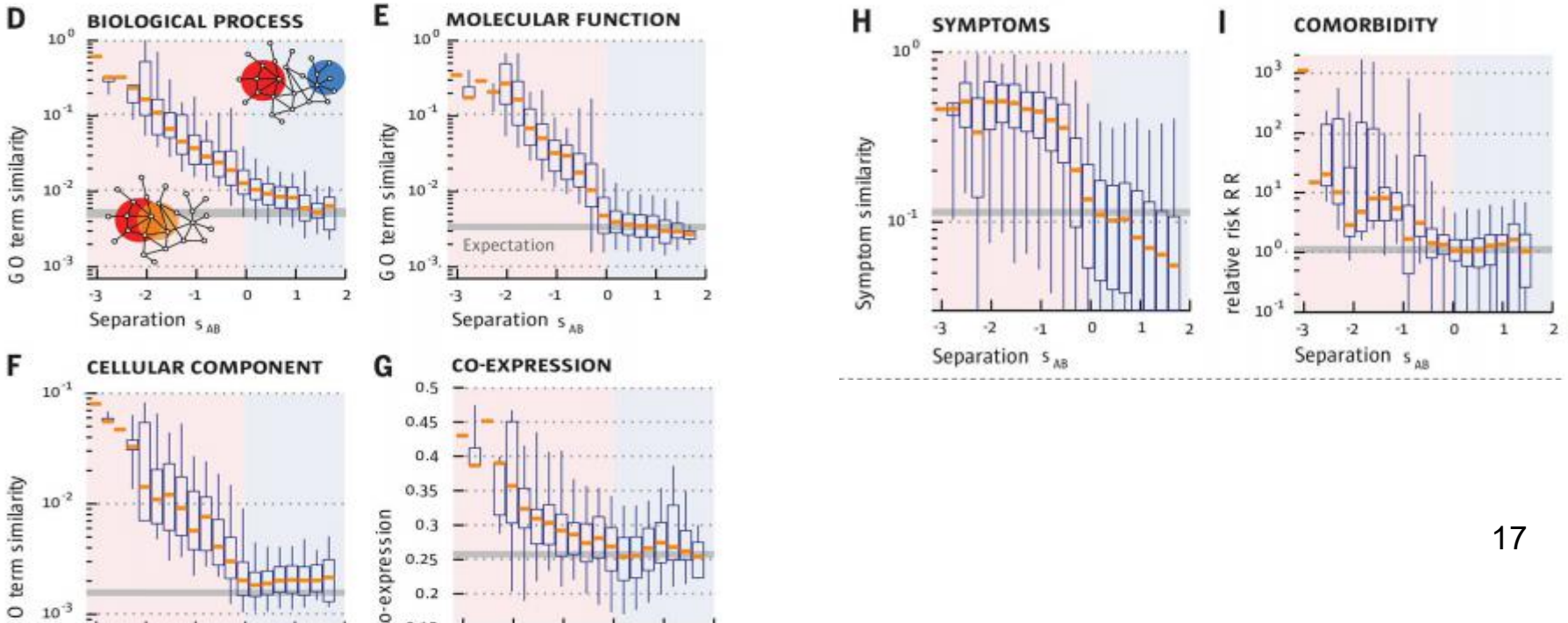
- The more localized a disease is topologically, the higher the significance in the similarity of the associated genes.



Data Compilation and Analysis

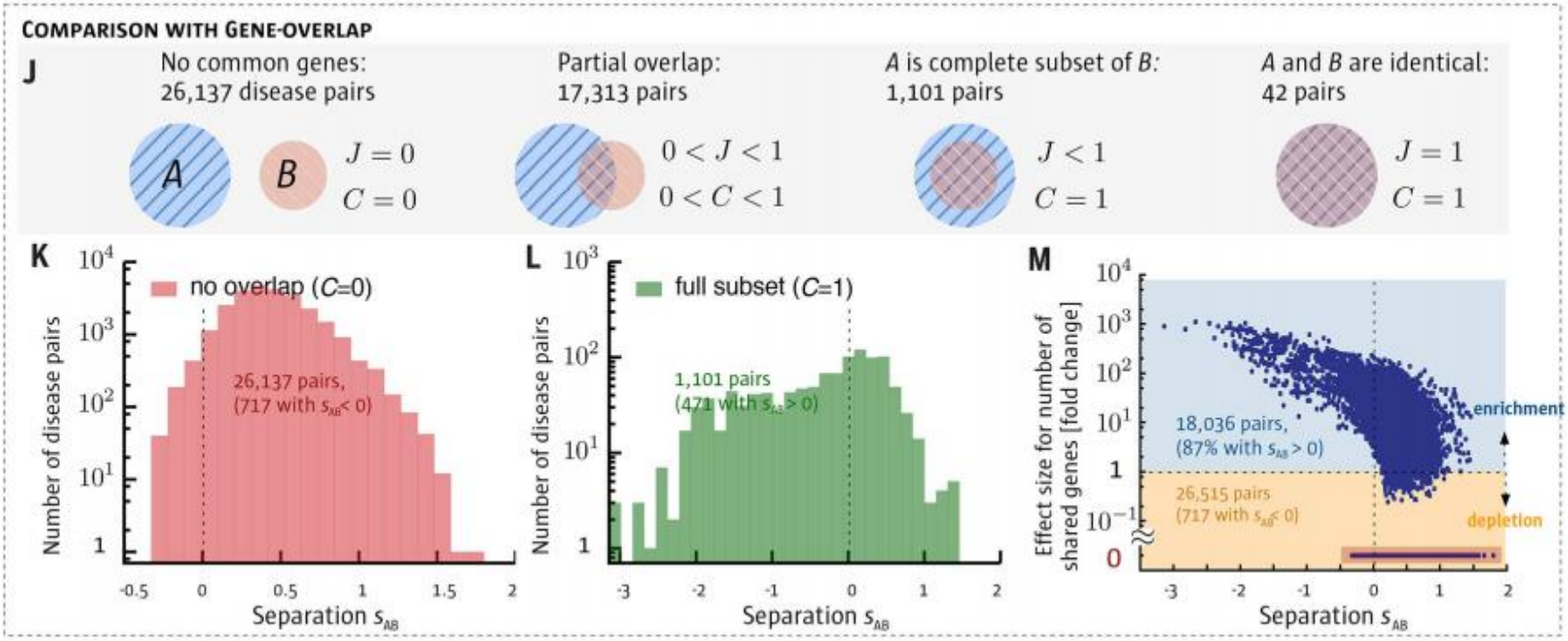
- Topological separation versus biomedical similarity:

✓ The closer of two diseases, the more similar



Data Compilation and Analysis

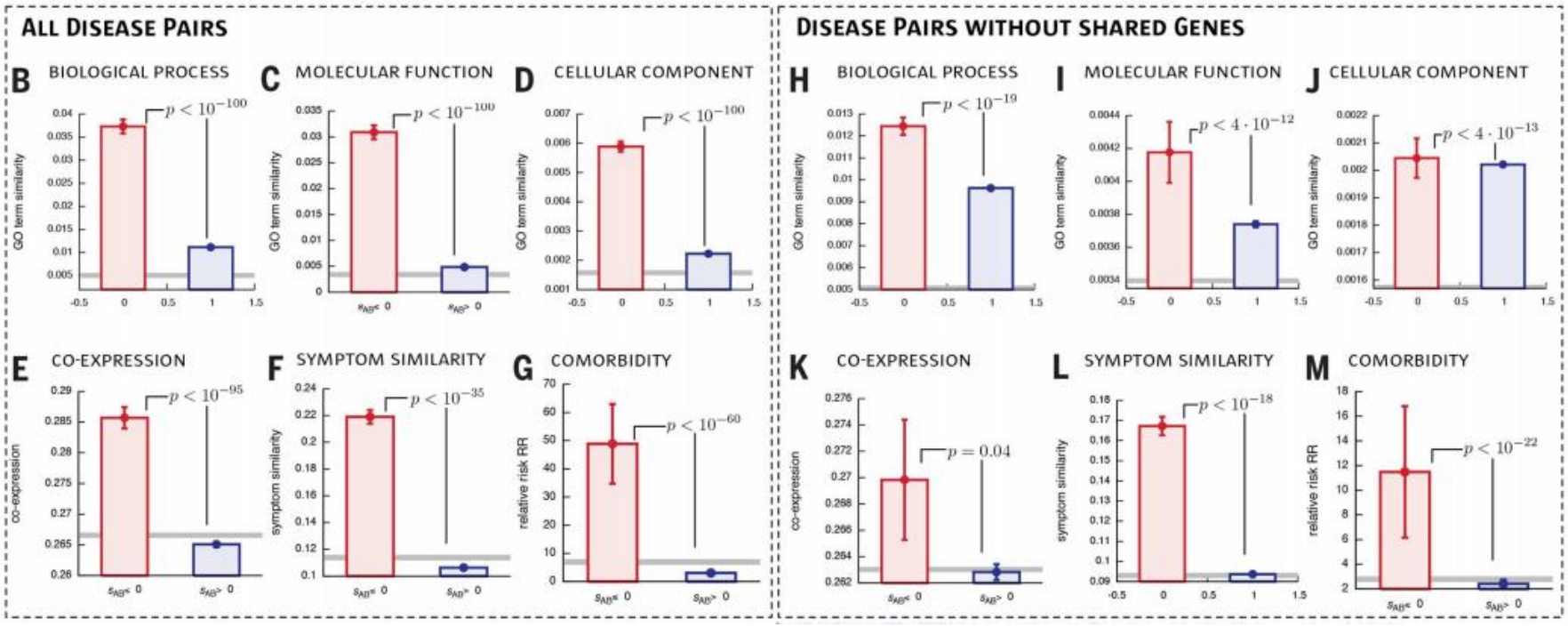
- Comparison with Gene-set overlap



Data Compilation and Analysis

- Biological similarity shown separately for the predicted overlapping and nonoverlapping disease pairs:

✓ Effective and better than Gene-set overlap method



Summary & Discussion

- Generalization for directed networks
- Comparison with unbiased datasets

✓ Using only yeast-two-hybrid (y2h) interaction data.

✓ Using only GWAS disease associations.

- Comparison of disease modules and network communities methods