Rensselaer

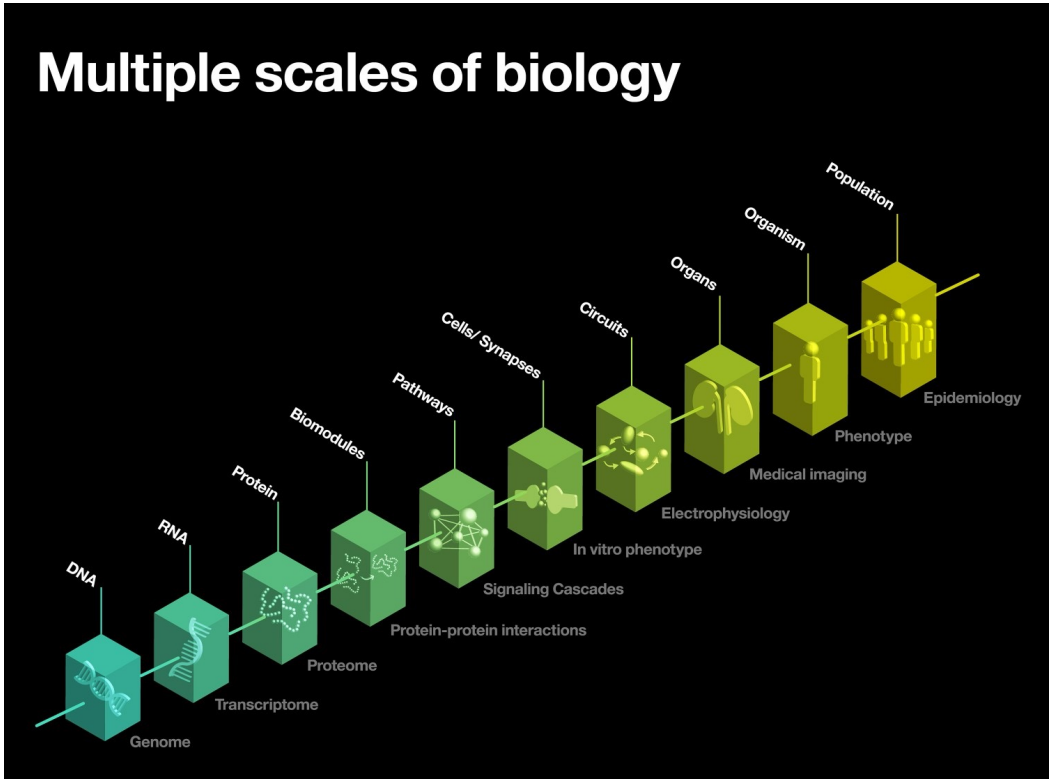
why not change the world?®

Human symptoms-disease network

Zhou, X., Menche, J., Barabási, A. L., & Sharma, A. (2014).

Presented by: Vijay Sadashivaiah (RIN: 662000393) | 11/08/21

- Diseases interact at multiple scales: DNA to Population.
- Understanding the relationship between clinical manifestations (symptoms) of diseases and underlying molecular interactions is relevant.



- Symptoms are crucial in clinical diagnosis and treatment.
  - Ex. Fever, cough, shortness of breath, fatigue, sore throat, runny nose, body aches, vomiting and diarrhea are symptoms of Covid – 19
  - But they could also be symptoms of Influenza
  - These two diseases differ in molecular level interactions

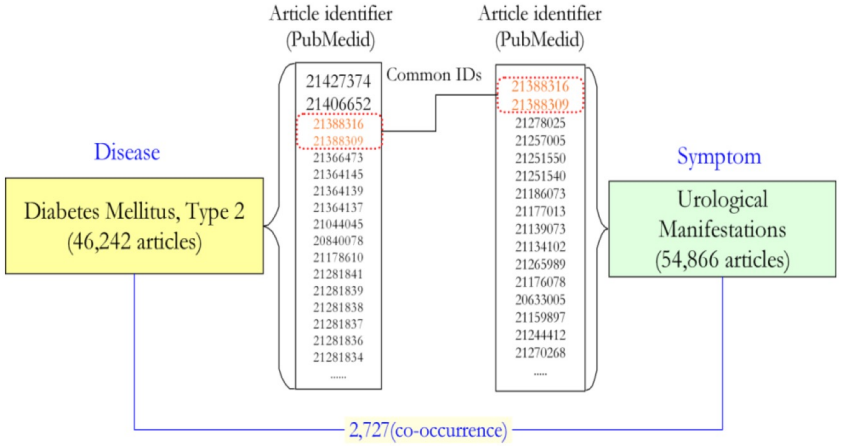
- Strong overlap between diseases and symptoms.
  - Built a network to represent interactions between diseases at multiple levels
  - Integrated genotypic and phenotypic information
    - Phenotype: symptoms
    - Genotype: genes, protein – protein interaction (PPI)

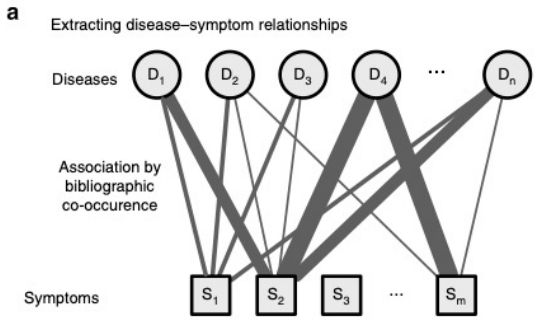
Disease – Symptom Relationships

- Data: PubMed MeSH
  - N = 849,103 (4.2%)
  - Diseases (j) = 4219 (95%)
  - Symptoms (i): 322 (98.5%)
- Network
  - nodes: disease, symptom
  - edges: co-occurance
  - weight: frequency inverse

$$w_{i,j} = W_{i,j} \log \frac{N}{n_i}$$

N: total number of diseases
n_i: number of diseases with symptom i

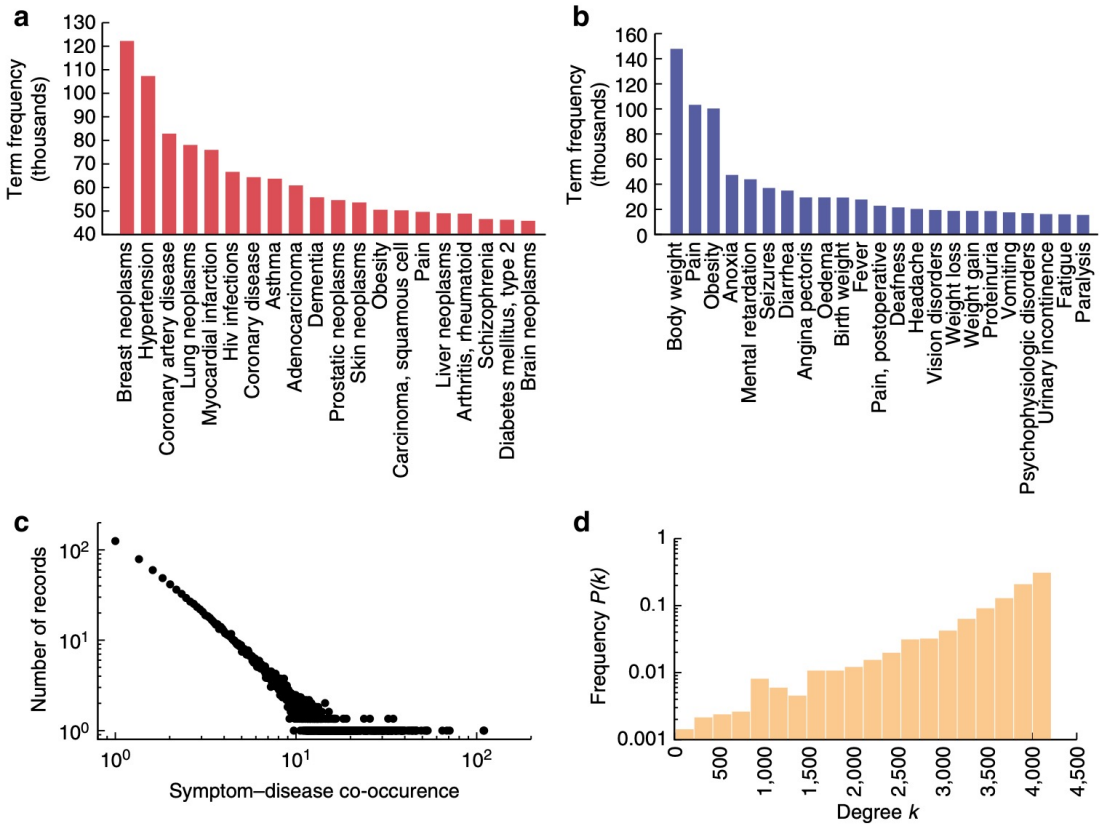




Disease – Symptom Relationships

- Most connected: Hyponatremia
  - 4,214 disease neighbors
  - an electrolyte disorder associated with a number of common symptoms that occur in many diseases, such as headache, nausea and fatigue.

- Least connected: Odontoma
  - eight disease neighbors
  - a tumor originating from teeth.

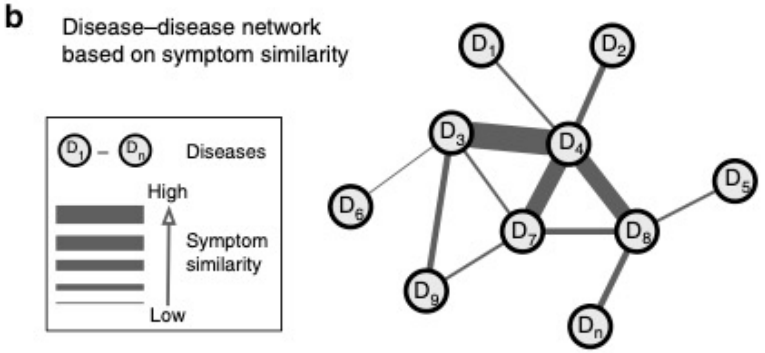


Disease – Disease Relationships

- Network
  - nodes: diseases
  - edges: shared symptoms
  - weight: symptom similarity

$$\cos(d_x, d_y) = \frac{\sum_i d_{x,i}\, d_{y,i}}{\sqrt{\sum_i d_{x,i}^2}\sqrt{\sum_i d_{y,i}^2}}$$

d_x: disease x vector
d_y: disease y vector

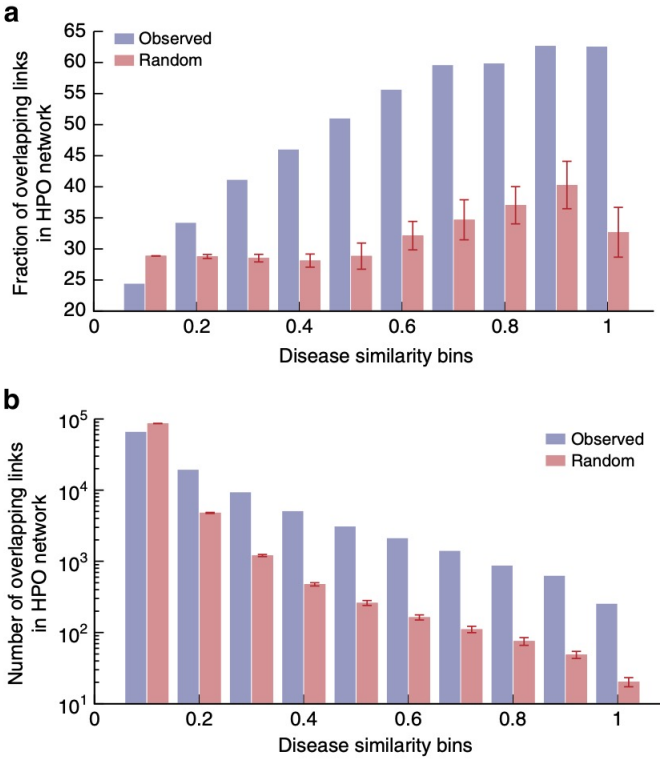


Performance evaluation of HSDN

- Authors extracted 1000 randomly selected samples from PubMed MeSH data
- Manually evaluated the extracted symptom–disease relations with the aid of medical experts
- Findings:
  - Vast majority of relations are medically meaningful
  - Confounding factors were typically drug related side-effects that were mentioned in articles
  - The disease relations in the HSDN are very specific, 57% (1 disease), 28.5% (2 diseases) and only 14.5% (2+)
  - Only 0.8% of the cases contained a negation as in ‘disease X is NOT related to symptom Y’

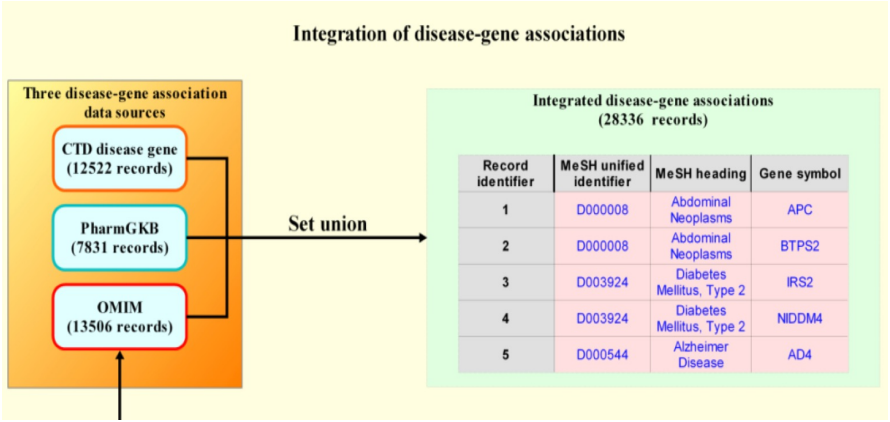
Performance evaluation of HSDN

- To further evaluate HSDN, authors compare with Human Phenotype Ontology
- HPO network is a benchmark network of diseases
  - Disease – disease connection if they share at least one symptom
- Much smaller in comparison to HSDN, but high quality
  - 940 MeSH diseases
  - 121,945 links
- HPO was found to be a subset of HSDN showing that it offers reliable relationships

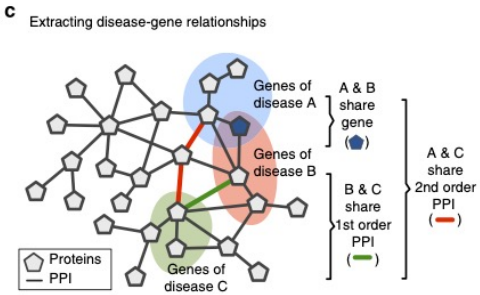


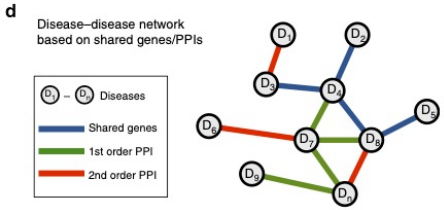
Disease – Gene/PPI Relationships

- Data:
  - Gene: CTD, PharmaGKB, OMIM
    - N = 28,336 records
    - 4,594 genes
  - PPI: HPRD, BioGrid, DIP, IntAct and MINT
    - N = 104,522 interactions
    - 14,212 proteins
- Network
  - Links = 47,410
  - nodes: diseases
  - edges: shared gene/PPI



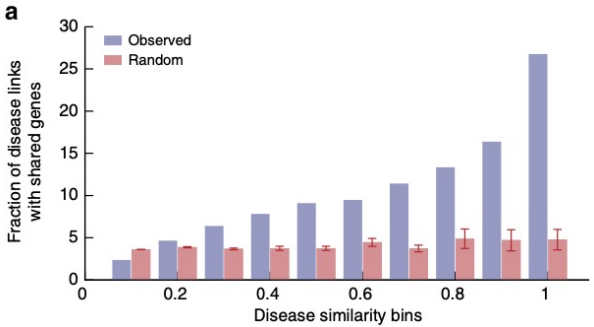
Record identifier	MeSH unified identifier	MeSH heading	Gene symbol
1	D000008	Abdominal Neoplasms	APC
2	D000008	Abdominal Neoplasms	BTPS2
3	D003924	Diabetes Mellitus, Type 2	IRS2
4	D003924	Diabetes Mellitus, Type 2	NIDDM4
5	D000544	Alzheimer Disease	AD4

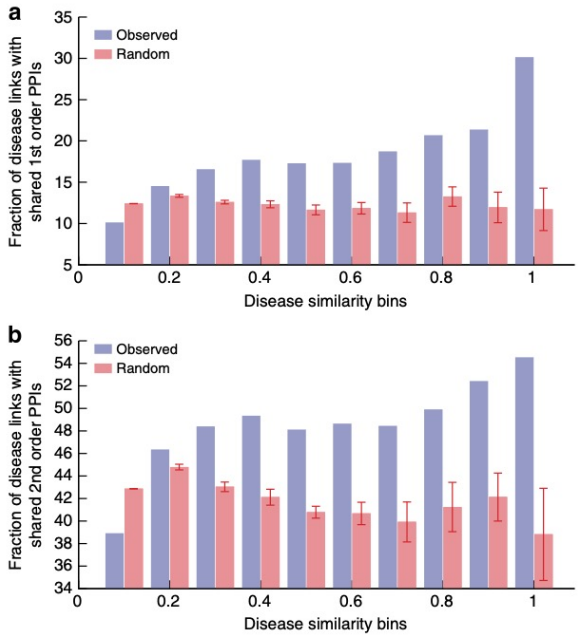




Disease – Gene/PPI Relationships

- Significant overlap between HSDN and Disease – Gene/PPI network
- 41,880 overlapping links
- Examples of high similarity scores:
  - insulin resistance and diabetes mellitus (0.97)
  - duodenal ulcer and stomach ulcer (0.93)

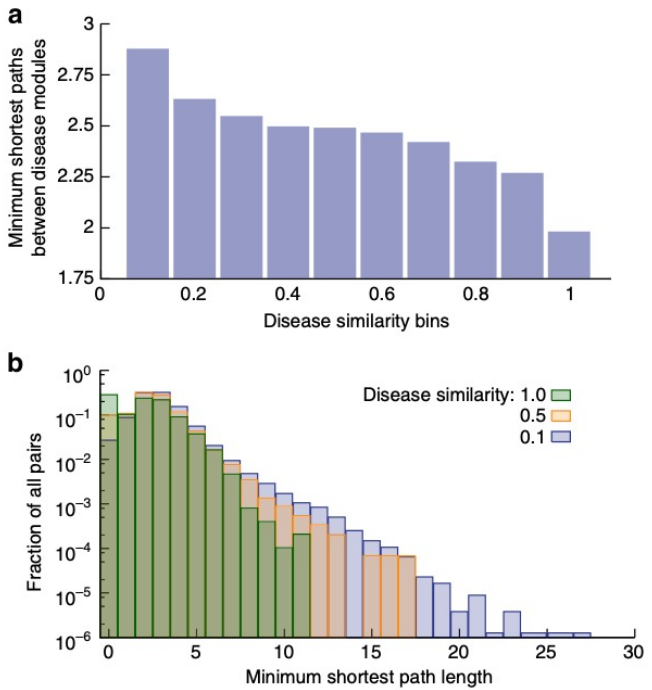




Disease – Gene/PPI Relationships

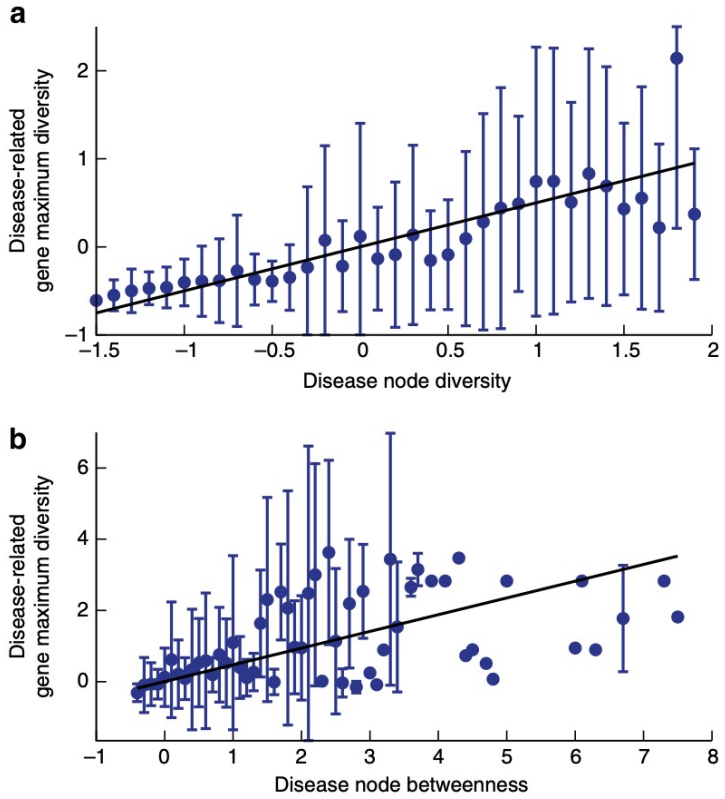
- High symptom similarity strongly correlates with shared genes, as well as with 1st- and 2nd-order protein interactions
- Minimum shortest path length (MSPL)
  - Using Dijkstra's algorithm
  - Strong negative correlation between MSPL and symptom similarity

$$D_{SL}(x,y) = \min_{p_i \in P_x, p_j \in P_y} D(p_i, p_j)$$

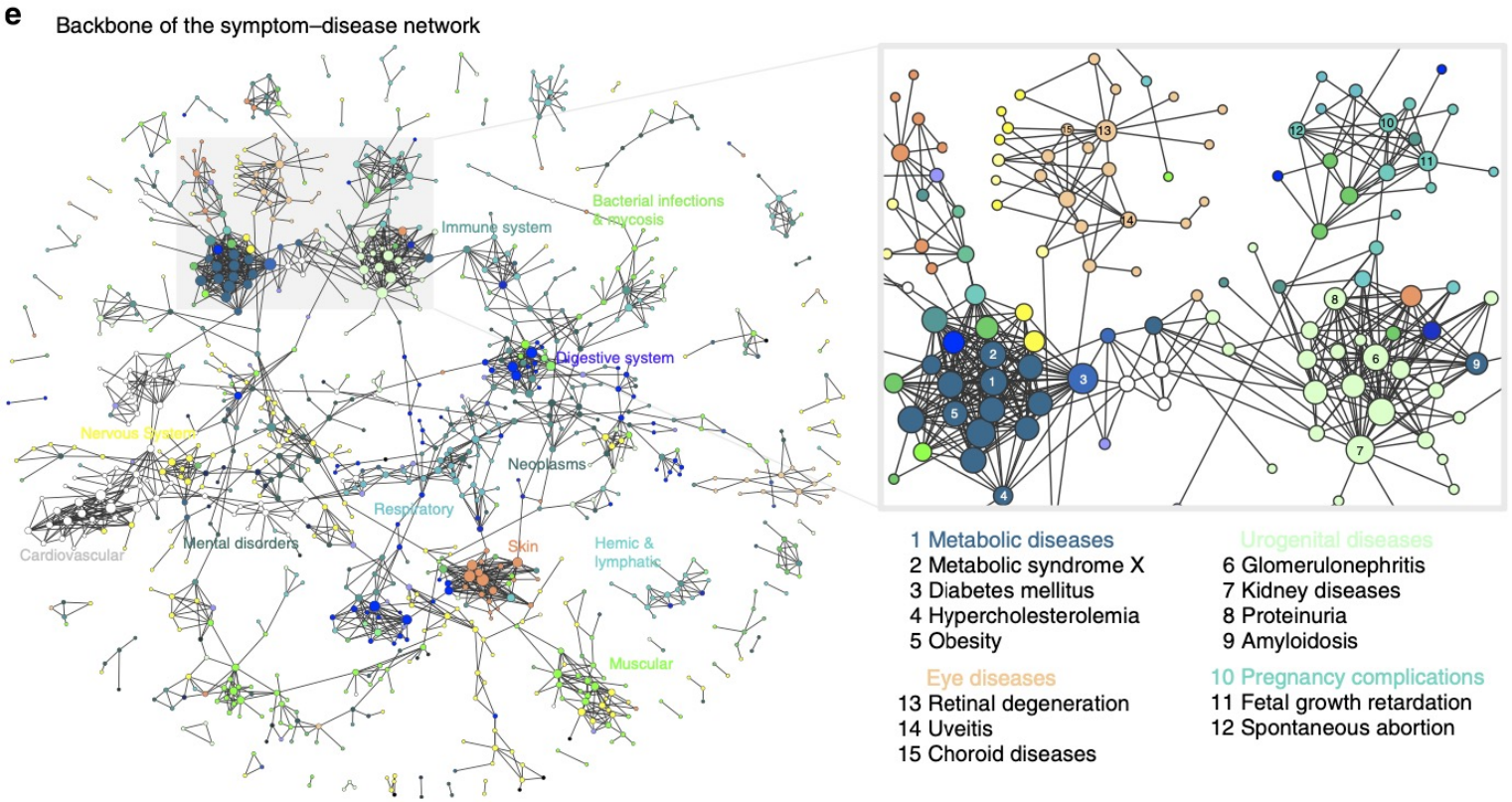


Disease vs molecular diversity

- There is a large discrepancy between clinical manifestations and underlying molecular mechanisms
- Network betweenness
  - large for nodes with many number of shortest paths passing through
- Network diversity
  - large for nodes with many neighbours that have many out-going links themselves.
- Disease with diverse clinical manifestations will typically also have more diverse underlying cellular network mechanisms



Symptom-Gene-PPI Disease Network (SGPDN)



Conclusions

- Symptom based similarity of two diseases, strongly correlate with number of shared genes and PPI interactions.
  - The observed correlations between clinical manifestations and molecular mechanisms of diseases can be highly valuable for functional annotations of genomics and reveal regularities between different disease categories.

Table 1 | The ten symptoms with the highest co-occurrence with Crohn's disease and ulcerative colitis.

Ulcerative colitis		Crohn's disease	
Symptom	**Occurrence**	**Symptom**	**Occurrence**
Diarrhea	214	Diarrhea	228
Psychophysiologic disorders	123	Body weight	141
Body weight	62	Abdominal pain	101
Abdominal pain	34	Pain	63
Pain	31	Psychophysiologic disorders	62
Fever	20	Fever	44
Constipation	18	Weight loss	43
Nausea	17	Oedema	39
Headache	17	Abdomen, acute	26
Weight loss	15	Nausea	24

Symptoms associated with both diseases are shown in red.

Conclusions

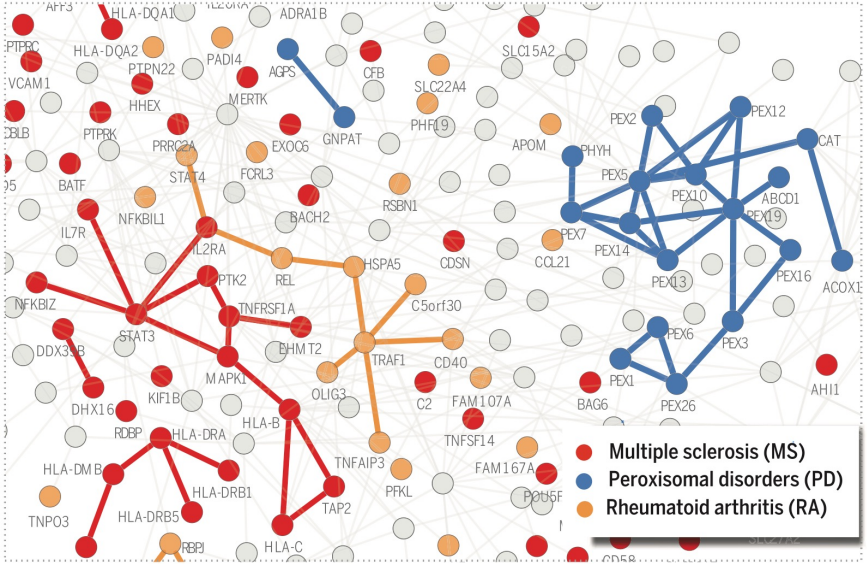
- Drug related research will benefit from such large models combining phenotypic and genotypic information
  - Most drugs approved by the US Food and Drug Administration are merely palliative.
  - i.e., they only treat symptoms rather than targeting disease-specific genes or pathways
- A detailed understanding of how symptoms relate to underlying molecular processes is therefore central effective and individualized treatments
- Symptoms are subjective
  - Symptoms represent the high-level manifestations of a disease that are actually observed by patients and physicians.
  - Objective validation of the patients' experience of major classes of symptoms still remains a pressing challenge in clinical practice

Evaluation/Critique

- MeSH annotations in literature could be inaccurate
  - Symptom not corresponding to disease X might appear due to side effects, contrastive examples etc.
  - Missing information
  - Can be addressed by using sophisticated text-mining algorithms to go through the article and find relevant disease – symptom relationships
- Objective measures of symptoms are not taken into consideration
  - Symptom severity, frequency, prevalence etc.
  - Can be addressed using information rich sources such as Electronic Health Records (EHRs)
- HSDN does not account for differences in symptoms based on demographic features such as age, gender etc.
- Takes a top-down approach to model disease – molecular relationships

Uncovering disease-disease relationships through the incomplete interactome

- Interactome is a network that integrates all physical interactions within a cell
  - PPI, regulatory interactions, metabolic pathway interactions etc
  - Provides better disease-disease relationships than just shared genes or shared PPI
- Disease modules are a connected subgraph that contains all molecular determinants of a disease
- Authors present a network-based framework to identify the location of disease modules within the interactome and study disease-disease relationships using interactome
- Interactomes are typically incomplete due to lack of technological advancements in identifying all possible cellular interactions



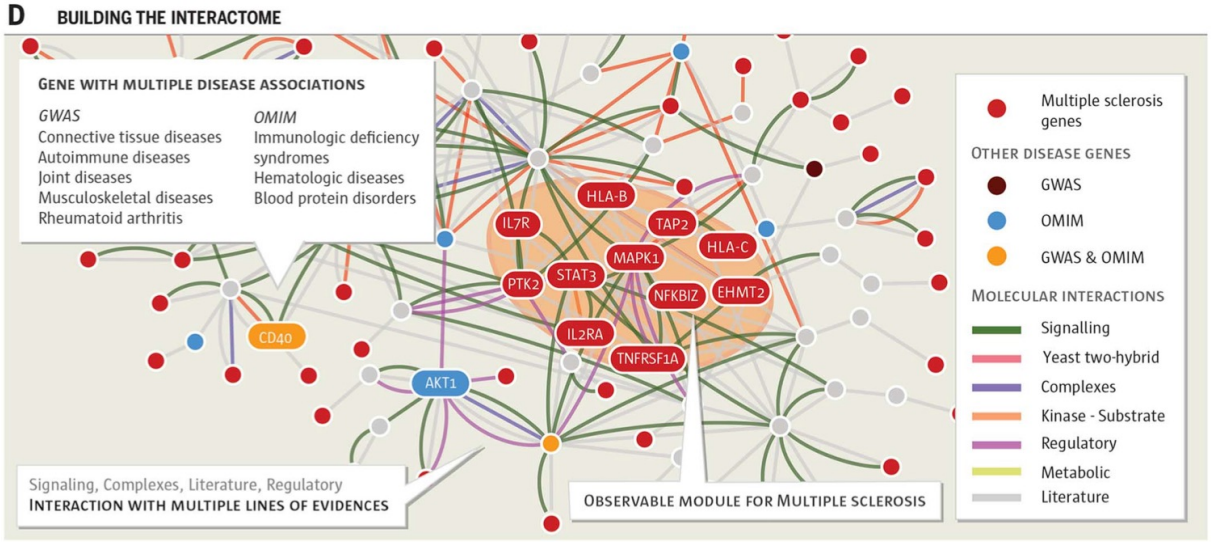
Constructing Interactome

- Data: Union of,
  - Regulatory interactions
  - Binary interactions
  - Kinase-substrate interactions
  - Metabolic interactions
  - Protein complexes

  N = 141,296 interactions

  13,460 proteins

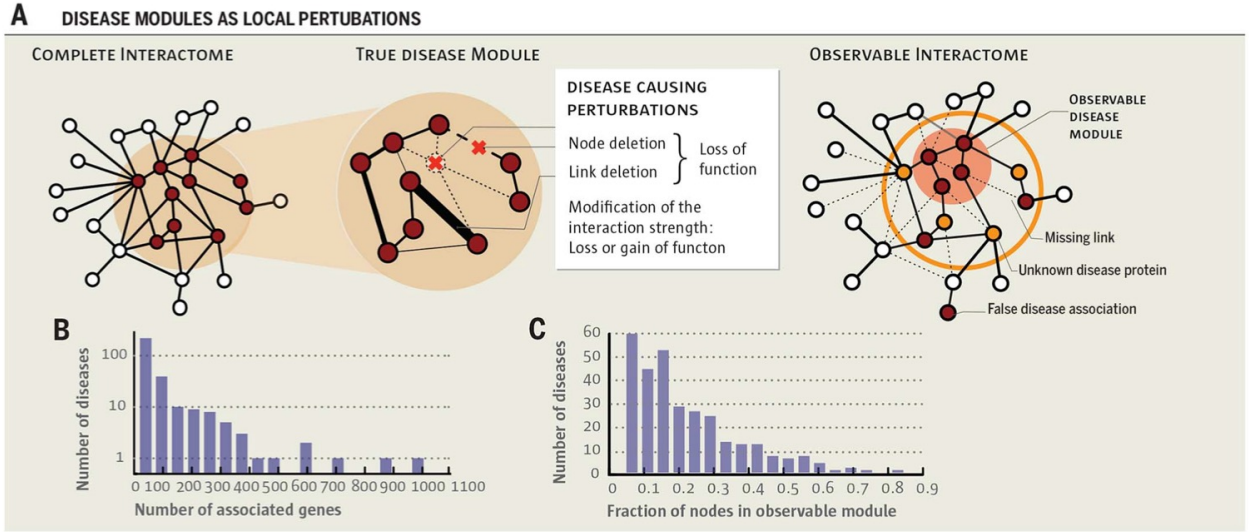
- Network
  - nodes: proteins
  - edges: interaction



What is disease in Interactome?

- According to the disease module hypothesis, a disease represents a local perturbation of the underlying disease-associated subgraph.

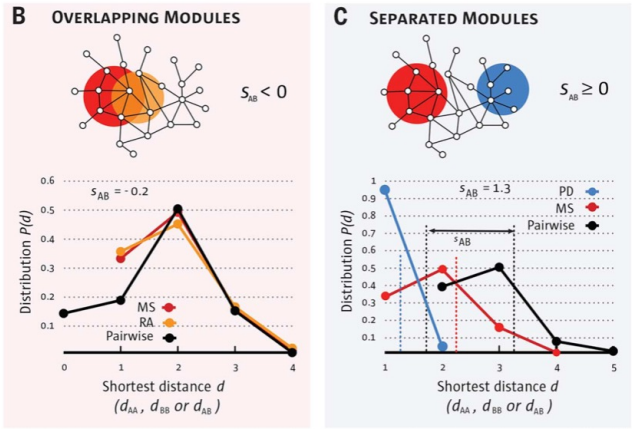
- Such perturbations could represent the removal of a protein (e.g., by a nonsense mutation), the disruption of a protein-protein interaction, or modifications in the strength of an interaction.

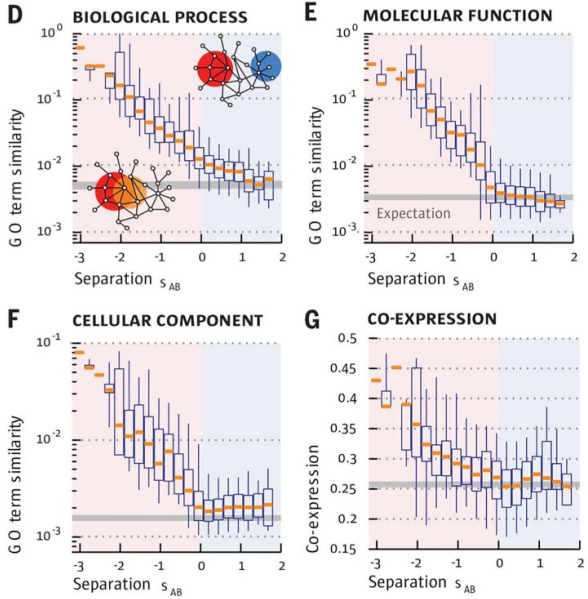


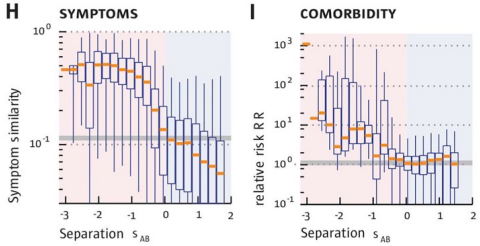
Relationship between diseases

- If two disease modules overlap, local perturbations leading to one disease will likely disrupt pathways involved in the other disease module as well, resulting in shared clinical characteristics

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

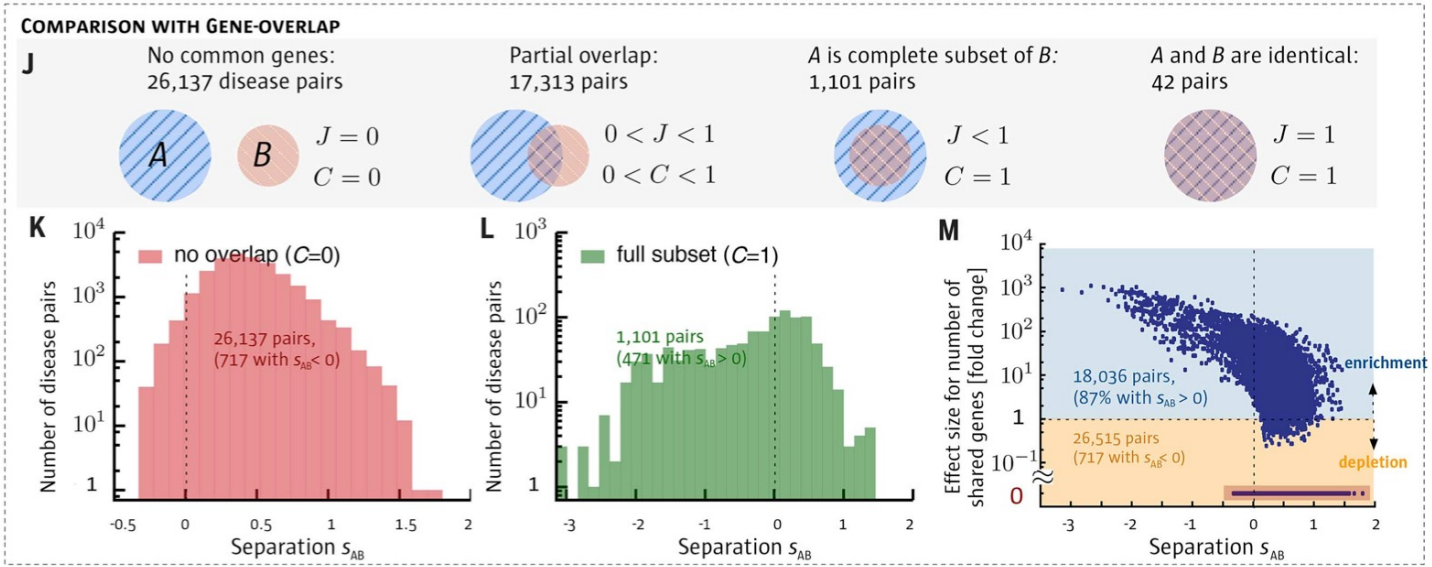






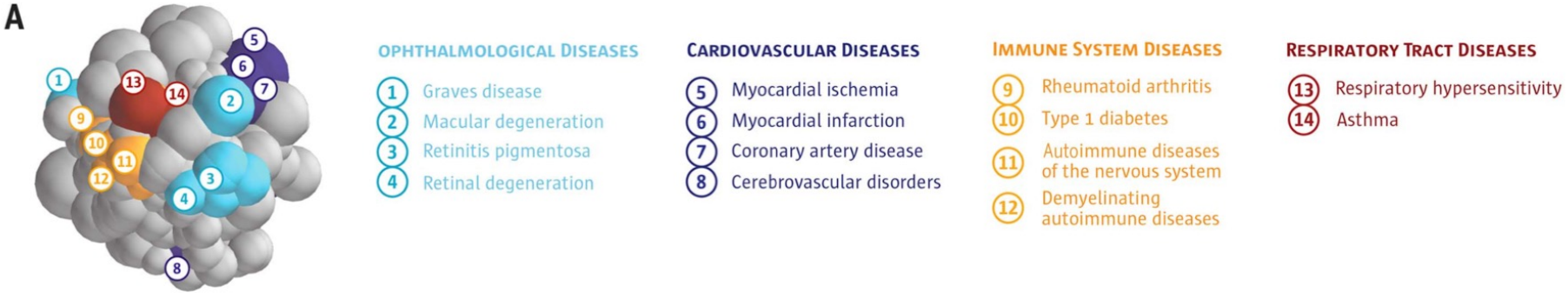
Overlap coefficient: C = |A∩B|/ min(|A|,|B|)

Jaccard-index: J = |A∩B|/|A∪B|



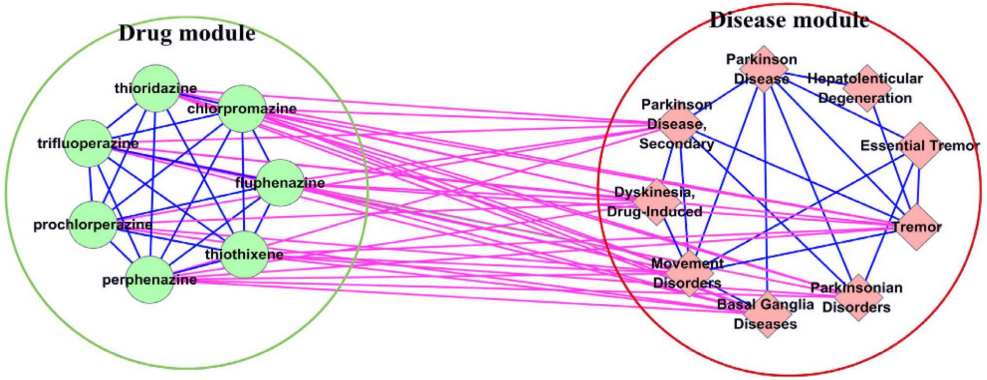
Conclusions

- A complete and accurate map of the interactome could have tremendous impact on our ability to understand the molecular underpinnings of human disease
  - Such a map is yet to be made and authors claim it's at least a decade away
  - Despite the incompleteness, authors show that valuable insights can be drawn using network-based approaches to study human diseases.

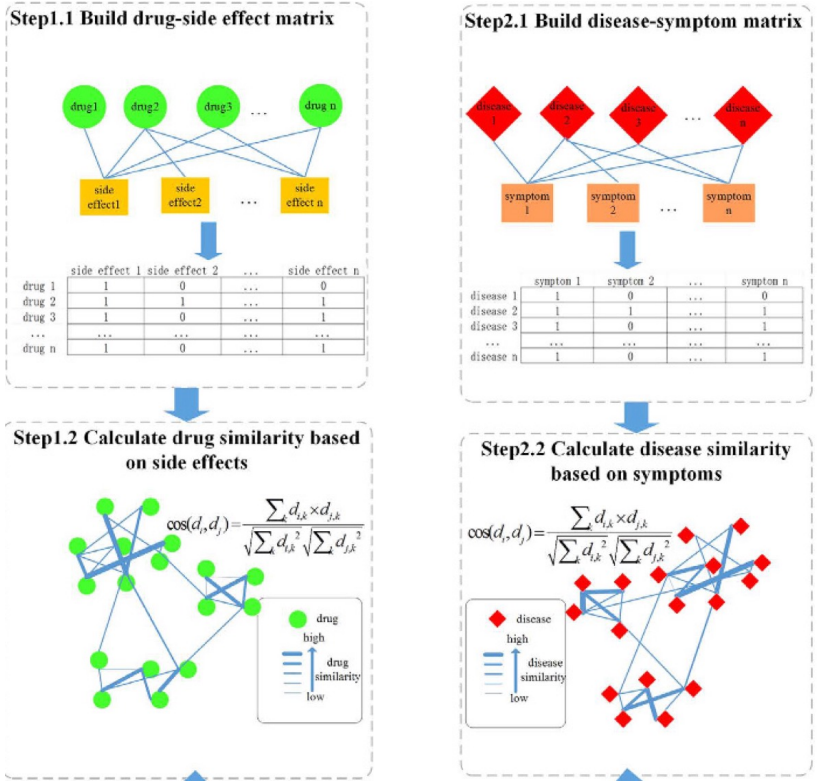


Prediction of new drug indications based on clinical data and network modularity

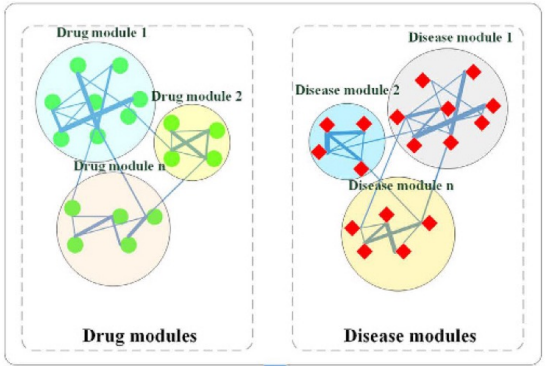
- Traditionally, drug discovery process mainly consists of three stages: discovery, preclinical stage, and clinical development
  - Time consuming and expensive
- Drug repositioning refers to identifying and using known drugs to treat different disease
  - Minoxidil which was originally tested for hypertension now used for hair loss treatment
- Authors create a drug and disease network that captures respective intra-connections and then find associations between them (inter-connections).
  - Data: Disease – Symptom data (HSDN), Drug – Side effects data (SIDER)



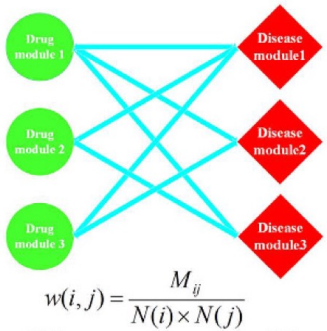
- Data:
  - SIDER: Drug – Side Effect
    - N = 99,423 pairs
    - 996 drugs
  - HSDN:
    - N = 133,106 interactions
    - 1,596 diseases
- Network
  - nodes: drugs/diseases
  - edges: similarity



Step3 Filter constructed networks and cluster them using ClusterONE



Step4 Construct connections between disease and drug modules based on CTD



$$w(i,j) = \frac{M_{ij}}{N(i) \times N(j)}$$

Step5 Rank drug-disease module pairs based on their scores

drug-disease module pairs	rank
drug2-disease2	1
drug1-disease3	2
drug3-disease4	3
...	...

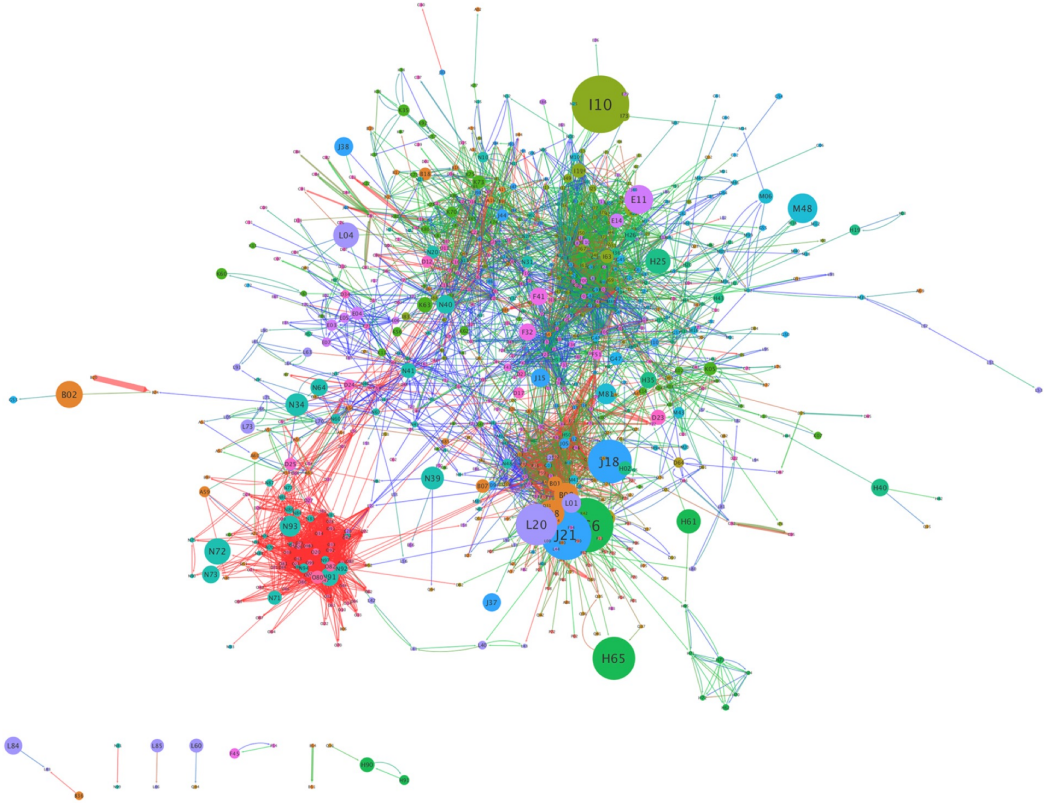
Conclusions

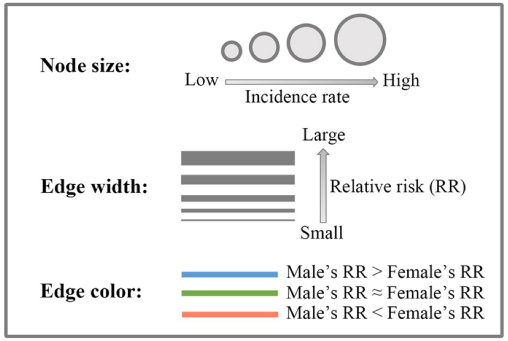
- Based on side effects of drugs and symptoms of diseases, authors construct drug - disease network
- Predictions are validated by testing overlaps with published literatures and Drug-Disease databases.
- Results from the study provide a new approach to complement existing computational methods such as GWAS in identifying candidate drug repositioning targets.

Network-based analysis of diagnosis progression patterns using claims data

- Network models presented so far do not consider demographic risk factors such as age, gender, prior diagnosis, etc.
- Genetic disease networks in which links represent shared genes or protein-protein interactions (PPIs) cannot explain relationships in non-genetic disorders, such as bone fracture.
- The durations (causal-effects) of disease-disease relationships cannot be defined using networks described so far.
- Authors construct a diagnosis progression network of human diseases using large-scale claims data and analyze the associations between diagnoses and it's progression
  - NHIS insurance claims dataset

Data and Network

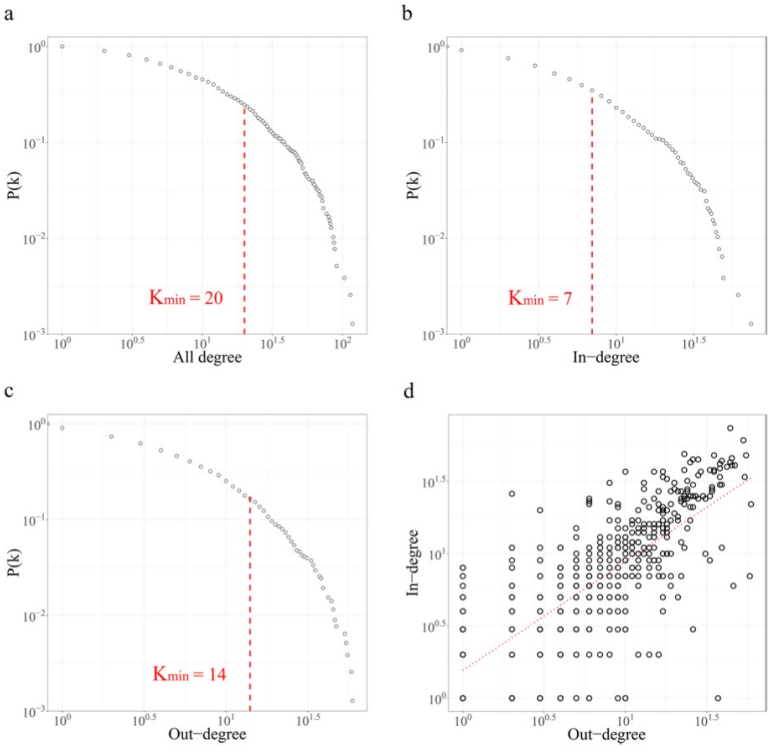


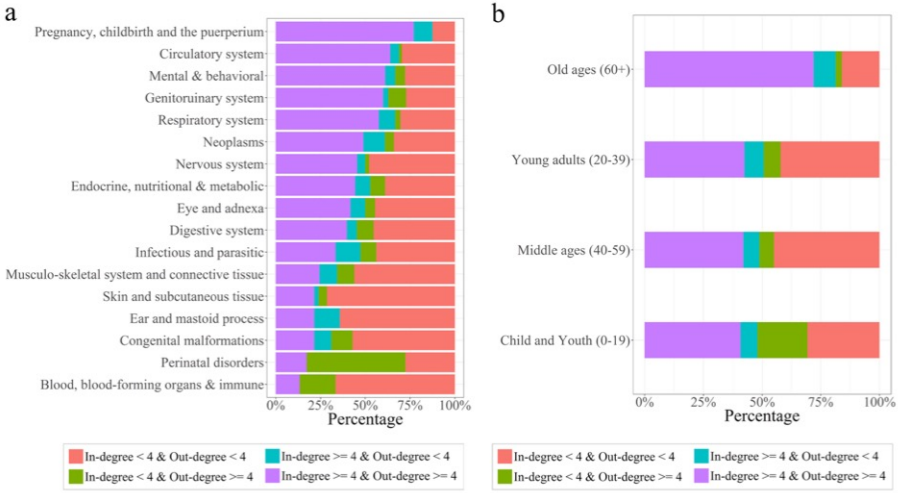


Node color (ICD-10 categories)

A00-B99	Infectious and parasitic
C00-D48	Neoplasms
D50-D89	Blood, blood-forming organs & immune
E00-E90	Endocrine, nutritional & metabolic
F00-F99	Mental & behavioral
G00-G99	Nervous system
H00-H59	Eye and adnexa
H60-H95	Ear and mastoid process
I00-I99	Circular system
J00-J99	Respiratory system
K00-K93	Digestive system
L00-L99	Skin and subcutaneous tissue
M00-M99	Musculo-skeletal system and connective tissue
N00-N99	Genitoruinary system
O00-O99	Pregnancy, childbirth and the puerperium
P00-P96	Perinatal disorders
Q00-Q99	Congenital malformations

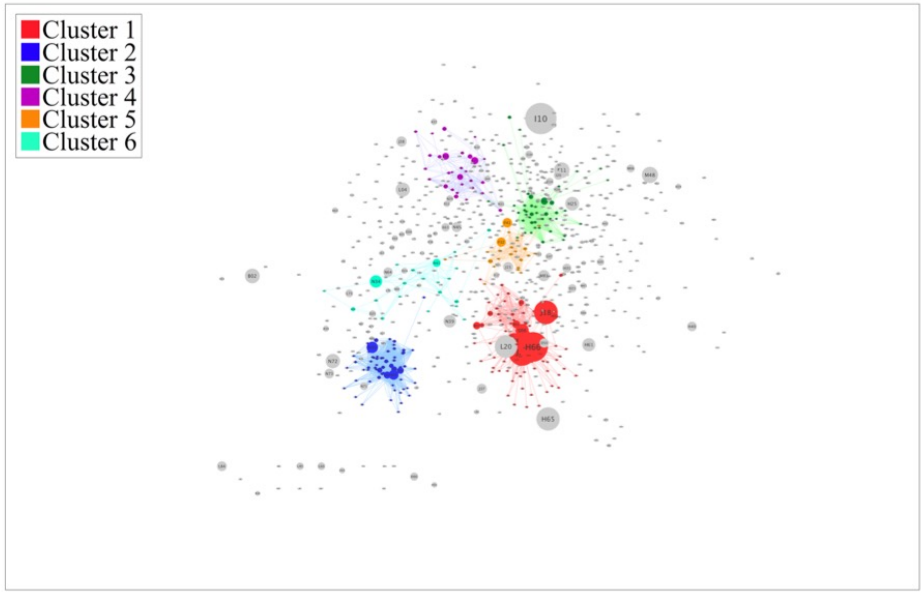
Network properties



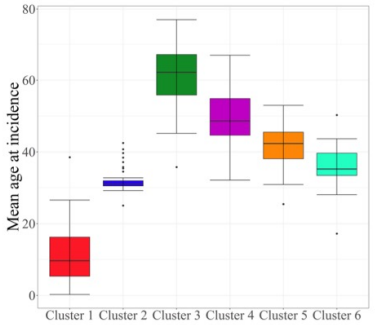


Clustering analysis

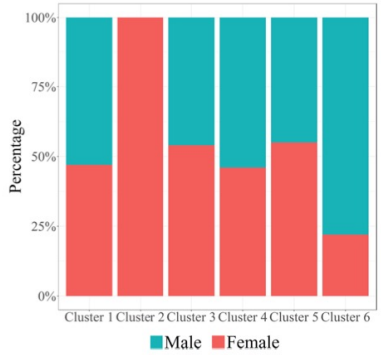
a



b



c



Conclusions

- Unlike existing disease networks based on biological data or clinical data, authors built a directional weighted network taking into account age and gender, and the progression duration.
- Compared to the previous gene-disease networks, this network used claims data that can be rich in information missing in non-genetic disorders.
  - Even in genetic disorders such as cancer, claims data has tissue sample information that's typically missing in lower molecule level datasets
- Network can be useful is early prediction of related disease based on predisposition factors of patients

Thank you!

- Questions?