An enhanced disease network with robust cross-phenotype relationships via variant frequency-inverse phenotype frequency

Perelman School of Medicine, University of Pennsylvania

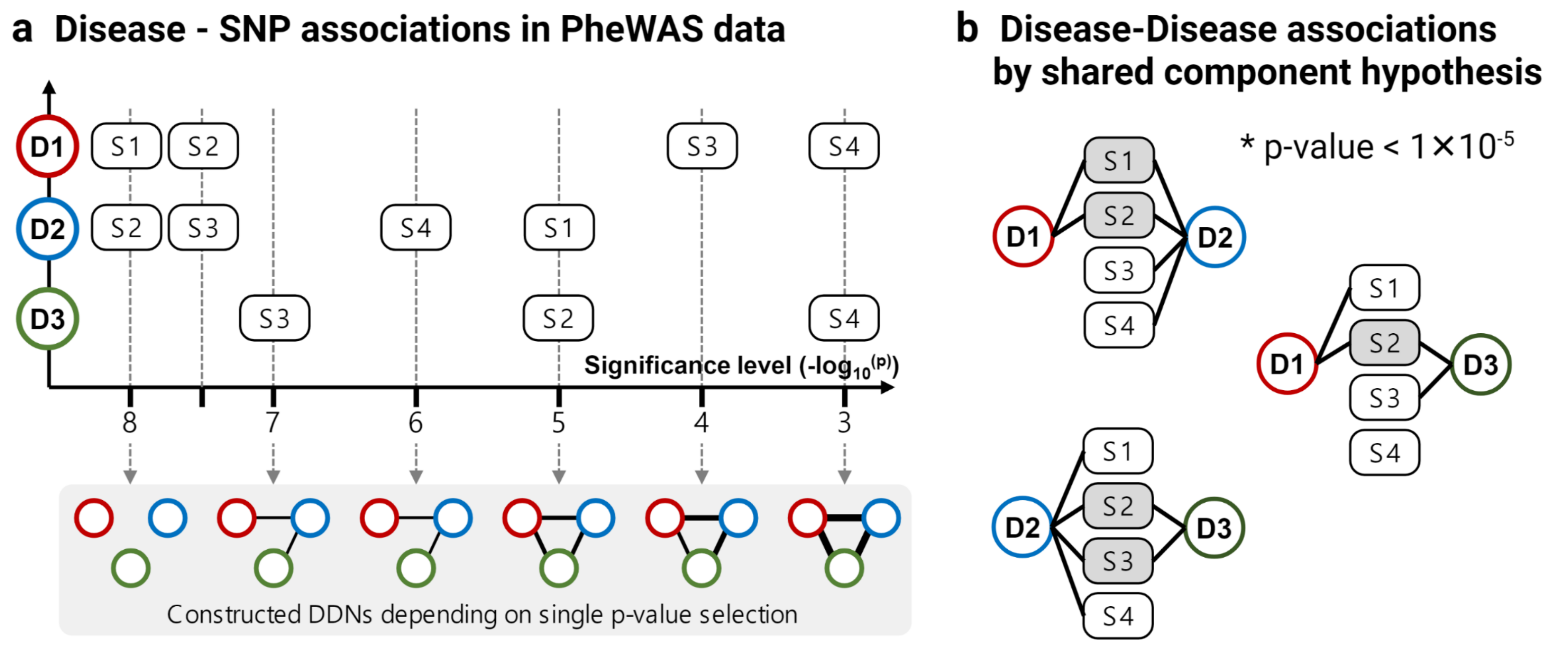
Yonghyun Nam, Vivek Sriram, Dokyoon Kim

Department of biostatistics, epidemiology & informatics

* Contact: Yonghyun Nam, Ph.D.,(Yonghyun.Nam@pennmedicine.upenn.edu)

Abstract

Motivation: Biobank-scaled phenome-wide association studies (PheWAS) identify associations between multiple phenotypes and common genetic variants. Disease-disease networks (DDNs) can intuitively observe relationships across many phenotypes. However, DDNs constructed from PheWAS summary data will highlight different disease associations depending on the chosen significance threshold. Thus, we proposed a novel method, variant frequency-inverse phenotype frequency (VF-IPF), to increase significance and decrease uncertainty in associations when constructing DDNs. **Results:** We constructed an enhanced DDN (eDDN) among 421 phenotypes from UK Biobank (UKBB) PheWAS summary data. We then used VF-IPF to enhance connections across diseases. To validate the enhanced cross phenotype relationships in eDDN, graph-based semi-supervised learning was applied to predict scores of disease co-occurrence. Ground truth co-occurrences were generated from UKBB inpatient data. Compared to conventional DDNs built from individual significance levels, the eDDN showed improved performance in identifying cross-phenotype associations. To demonstrate its clinical significance, we further validated our predictions using myocardial infarction as an index disease of interest against co-occurrence information from the Penn Medicine Biobank.

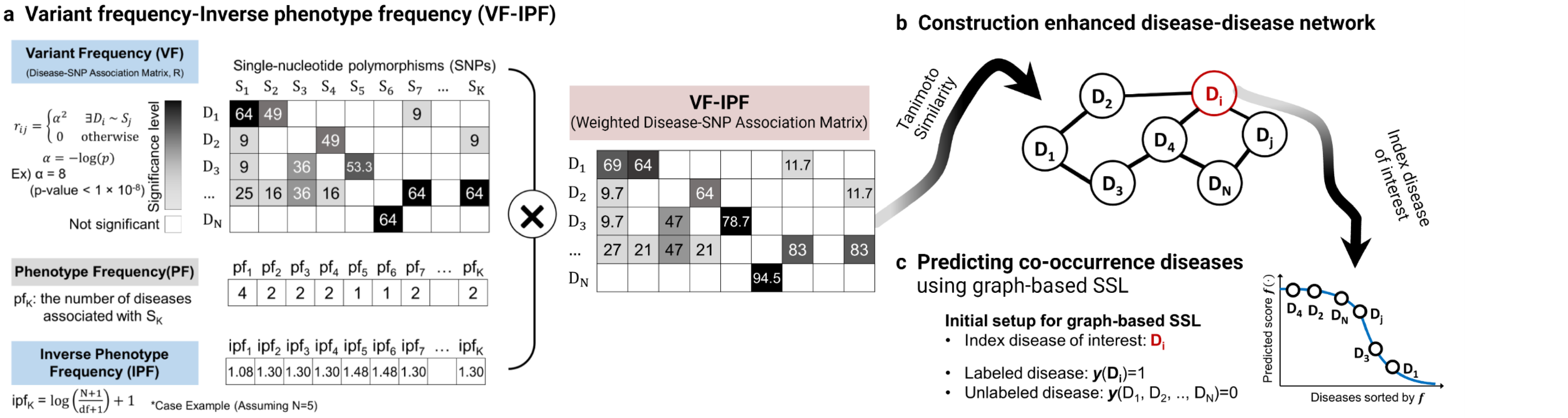

a Disease - SNP associations in PheWAS data

b Disease-Disease associations by shared component hypothesis

D1
D2
D3
S1
S2
S3
S4
Significance level ($-\log_{10}(p)$)
8
7
6
5
4
3
Constructed DDNs depending on single p-value selection
* p-value < 1×10^{-5}

Proposed Method: Enhanced DDN with VF-IPF

- We develop a novel method to **increase the significance of associations** and **decrease uncertainty of associations** simultaneously in disease-SNP associations regardless of the arbitrary selection of significance levels
- We propose **the enhanced disease-disease network (eDDN)**, which can discover more scalable and robust relationships across multiple phenotypes from biobank-scaled PheWAS data


a Variant frequency-Inverse phenotype frequency (VF-IPF)

Variant Frequency (VF) (Disease-SNP Association Matrix, R)

$$r_{ij} = \begin{cases} \alpha^2 & \exists D_i \sim S_j \\ 0 & \text{otherwise} \end{cases}$$

$$\alpha = -\log(p)$$

Ex) α = 8 (p-value < 1×10^{-8})

Single-nucleotide polymorphisms (SNPs)

	S_1	S_2	S_3	S_4	S_5	S_6	S_7	…	S_K
D_1	64	49					9		
D_2	9			49					9
D_3	9		36		53.3				
…	25	16	36	16			64		64
D_N						64			

Phenotype Frequency(PF)

pf_K: the number of diseases associated with S_K

pf_1	pf_2	pf_3	pf_4	pf_5	pf_6	pf_7	…	pf_K
4	2	2	2	1	1	2		2

Inverse Phenotype Frequency (IPF)

$$ipf_K = \log\left(\frac{N+1}{df+1}\right) + 1$$

ipf_1	ipf_2	ipf_3	ipf_4	ipf_5	ipf_6	ipf_7	…	ipf_K
1.08	1.30	1.30	1.30	1.48	1.48	1.30		1.30

*Case Example (Assuming N=5)

VF-IPF (Weighted Disease-SNP Association Matrix)

D_1	69	64					11.7		
D_2	9.7			64					11.7
D_3	9.7		47		78.7				
…	27	21	47	21			83		83
D_N						94.5			

b Construction enhanced disease-disease network

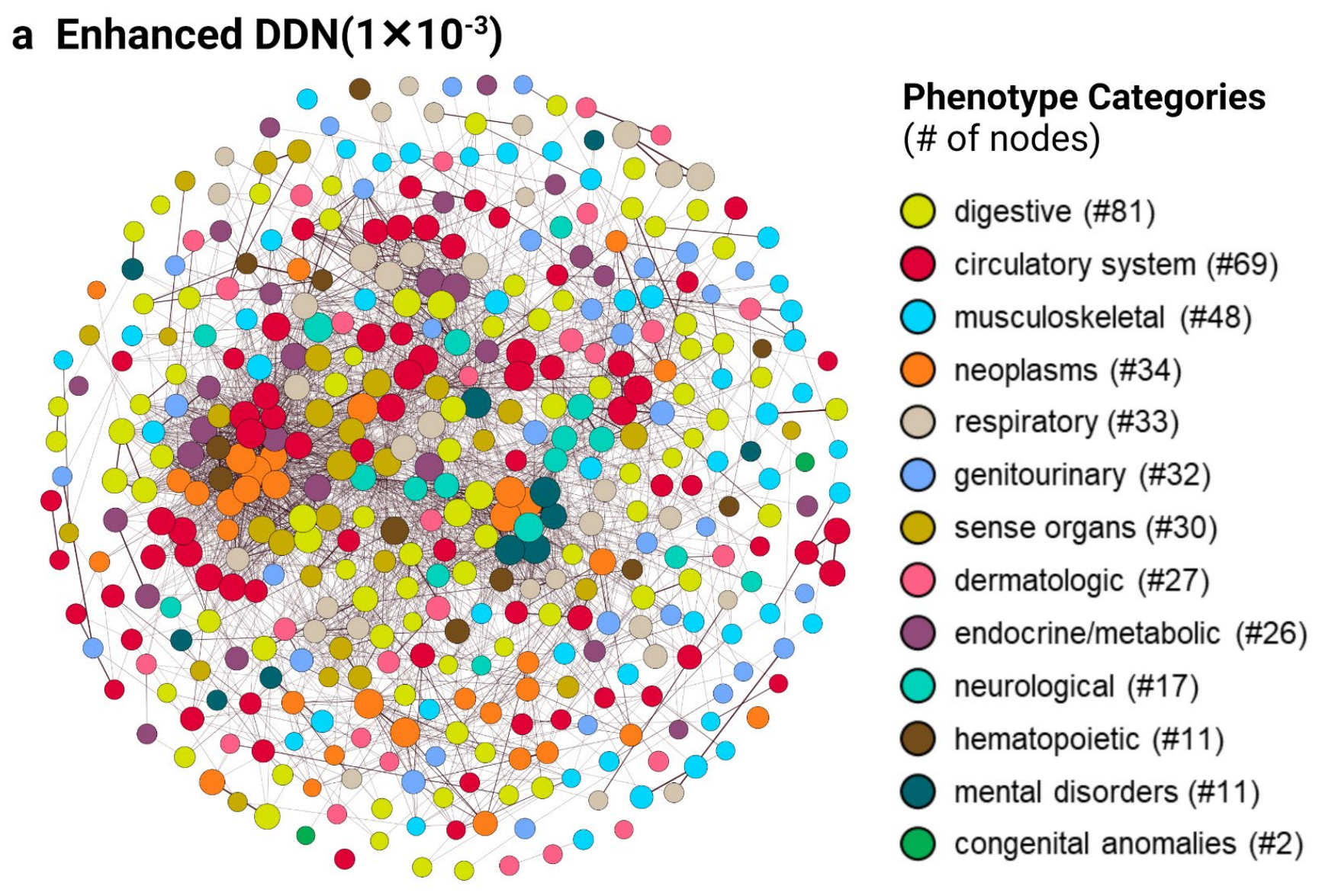
c Predicting co-occurrence diseases using graph-based SSL

Initial setup for graph-based SSL

- Index disease of interest: D_i
- Labeled disease: $y(D_i)=1$
- Unlabeled disease: $y(D_1, D_2, .., D_N)=0$

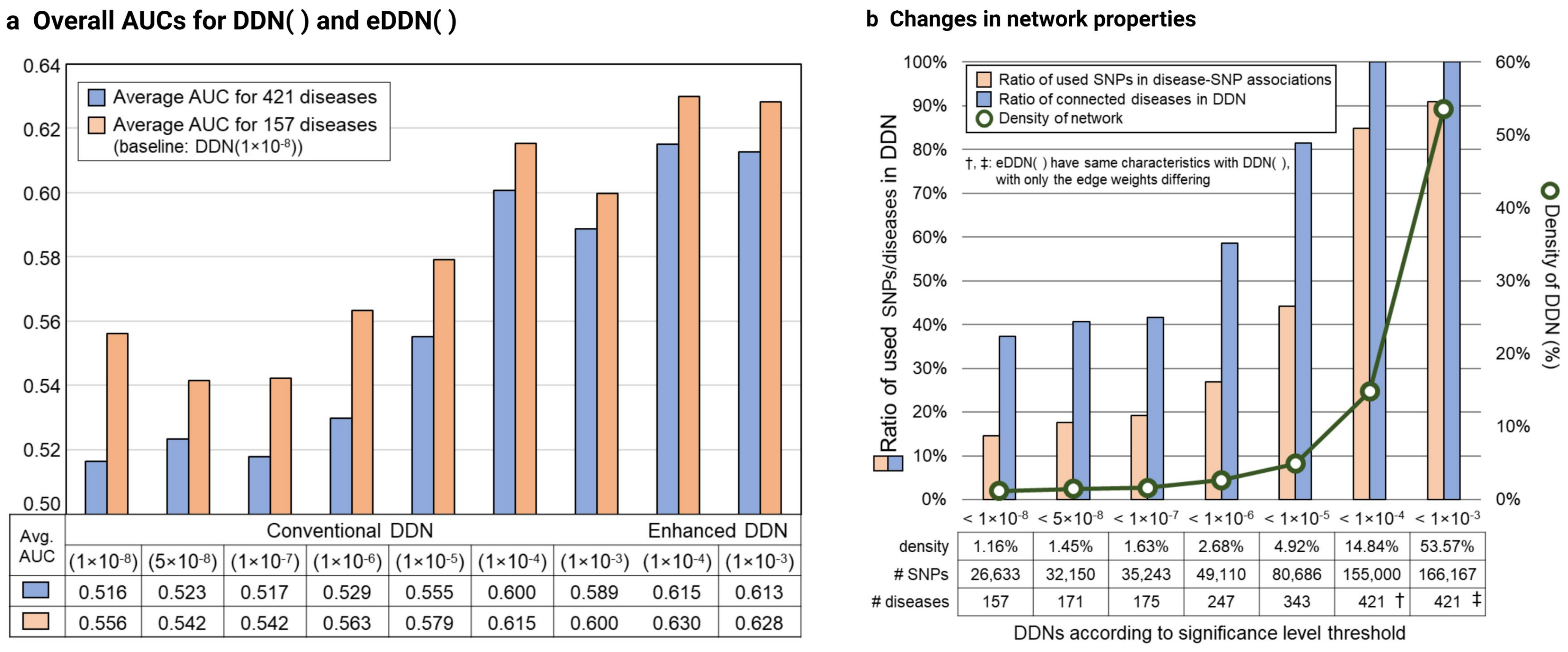
Result (1): Construction of eDDN

- The eDDN were constructed using UKBB PheWAS summary statistics
  - Number of Phenotypes: 421


a Enhanced DDN(1×10^{-3})

Result (2): Performance Comparison

- **Task:** Predicting co-occurrence scoring
  - Ground truth: Comorbidity measurements calculated by UKBB inpatient EHR data
  - Performance measurement: AUC
- The eDDN showed the highest average AUC of 0.630 in co-occurrence prediction when compared with the other eight DDNs


a Overall AUCs for DDN( ) and eDDN( )

Avg. AUC	Conventional DDN							Enhanced DDN	
	(1×10^{-8})	(5×10^{-8})	(1×10^{-7})	(1×10^{-6})	(1×10^{-5})	(1×10^{-4})	(1×10^{-3})	(1×10^{-4})	(1×10^{-3})
🟦	0.516	0.523	0.517	0.529	0.555	0.600	0.589	0.615	0.613
🟧	0.556	0.542	0.542	0.563	0.579	0.615	0.600	0.630	0.628

b Changes in network properties

†, ‡: eDDN( ) have same characteristics with DDN( ), with only the edge weights differing

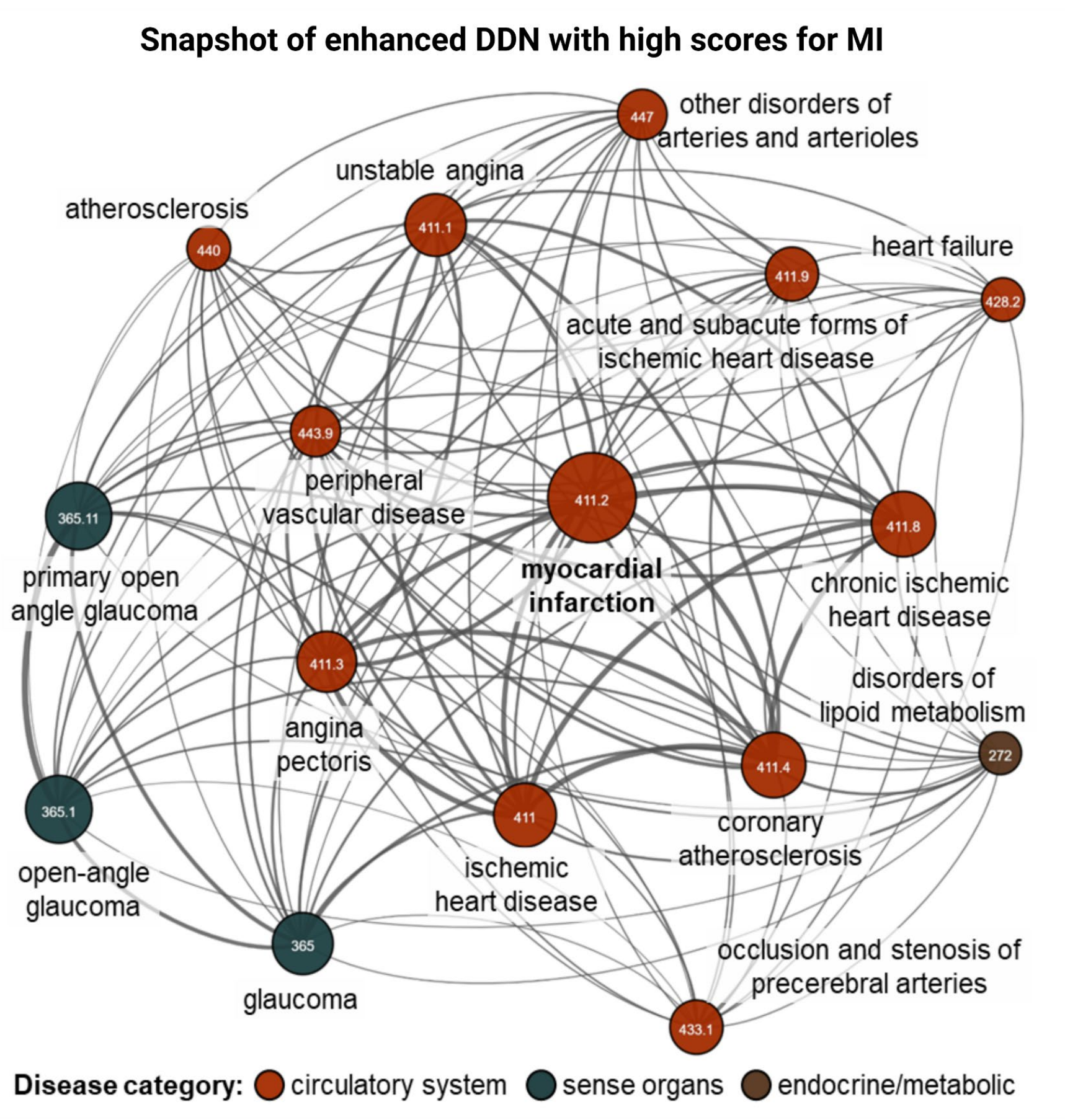
	< 1×10^{-8}	< 5×10^{-8}	< 1×10^{-7}	< 1×10^{-6}	< 1×10^{-5}	< 1×10^{-4}	< 1×10^{-3}
density	1.16%	1.45%	1.63%	2.68%	4.92%	14.84%	53.57%
# SNPs	26,633	32,150	35,243	49,110	80,686	155,000	166,167
# diseases	157	171	175	247	343	421 †	421 ‡

DDNs according to significance level threshold

0.64
0.62
0.60
0.58
0.56
0.54
0.52
0.50
Average AUC for 421 diseases
Average AUC for 157 diseases (baseline: DDN(1×10^{-8}))
Ratio of used SNPs/diseases in DDN
Density of DDN (%)
Ratio of used SNPs in disease-SNP associations
Ratio of connected diseases in DDN
Density of network
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
60%
50%
40%
30%
20%
10%
0%

Result (3): Clinical implementation of co-occurrence scoring

- This result shows a proof-of-concept study for demonstrating the utility of eDDN.
- Transferring the discovered knowledge in genetic information to clinical significance is essential in translational bioinformatics.
- The results can give significance to the discovered information and provide evidence that can be used for clinical implementation.
- Predicted results were verified using UKBB EHR in the analysis as well as comorbidity information obtained from Penn Medicine Biobank (PMBB), supplementing the evidence that can be used to utilize the clinical implications.


Snapshot of enhanced DDN with high scores for MI

other disorders of arteries and arterioles
447
unstable angina
411.1
atherosclerosis
440
heart failure
428.2
411.9
acute and subacute forms of ischemic heart disease
443.9
peripheral vascular disease
365.11
primary open angle glaucoma
411.2
myocardial infarction
411.8
chronic ischemic heart disease
411.3
angina pectoris
disorders of lipoid metabolism
272
411.4
coronary atherosclerosis
365.1
open-angle glaucoma
411
ischemic heart disease
365
glaucoma
occlusion and stenosis of precerebral arteries
433.1
Disease category: circulatory system sense organs endocrine/metabolic

Phenotype description (PheCode)	Score (f)	Relative Risk [95% CI]	
		UKBB	PMBB
Coronary atherosclerosis (411.4)	0.9884	13.014 [13.011, 13.017]	3.439 [3.437, 3.442]
Chronic ischemic heart disease (411.8)	0.9883	13.542 [13.539, 13.545]	6.400 [6.390, 6.410]
Open-angle glaucoma (365.1)	0.9904	1.848 [1.807, 1.890]	1.133 [1.084, 1.183]
Primary open angle glaucoma (365.11)	0.9901	1.857 [1.815, 1.899]	1.555 [1.386, 1.745]
Ischemic Heart Disease (411)	0.9867	11.283 [11.281, 11.284]	3.488 [3.485, 3.490]
Unstable angina (411.1)	0.9852	14.569 [14.558, 14.580]	5.164 [5.146, 5.181]
Glaucoma (365)	0.9844	1.943 [1.931, 1.956]	1.331 [1.315, 1.348]
Angina pectoris (411.3)	0.9832	9.448 [9.445, 9.451]	4.255 [4.243, 4.267]
Occlusion and stenosis of precerebral arteries (433.1)	0.9723	6.116 [6.068, 6.165]	2.809 [2.799, 2.819]
acute and subacute forms of ischemic heart disease (411.9)	0.9721	21.922 [21.885, 21.958]	5.104 [5.036, 5.174]
Peripheral vascular disease (443.9)	0.9656	8.273 [8.255, 8.292]	3.227 [3.219, 3.234]
Other disorders of arteries and arterioles (447)	0.9640	5.584 [5.493, 5.676]	2.259 [2.249, 2.269]
Atherosclerosis (440)	0.9422	10.366 [10.084, 10.656]	3.053 [3.041, 3.065]
Heart failure (428.2)	0.9376	12.325 [12.314, 12.335]	3.948 [3.915, 3.981]
Disorders of lipoid metabolism (272)	0.9325	5.301 [5.300, 5.303]	1.718 [1.717, 1.720]

Conclusion

- The eDDN can have more reliable explanatory power for disease-disease associations
- It can be robust even in the presence of false positive disease-SNP associations according to less stringent thresholds (p-value < 1×10^{-3}).

*This study was submitted to 21st European Conference on Computational Biology (Under consideration)