Exploring PPI Networks for Disease-Associated Genes

“Walking the Interactome for Prioritization of Candidate Disease Genes”

Köhler, S., Bauer, S., Horn, D. & Robinson, P.N.

The American Journal of Human Genetics **82**, 949–958 (2008).

And

“Associating Genes and Protein Complexes with Disease via Network Propagation”

Vanunu, O., Magger, O., Ruppin, E., Shlomi, T. & Sharan, R. Associating Genes and Protein Complexes with Disease via Network Propagation. *PLoS Comput Biol* **6**, e1000641 (2010).

A Double Bill brought to you by LucasLearning
Stardate 2011.293

Köhler *et al*: Overview

- 1500 OMIM conditions that have no molecular cause listed
- Much disease-gene associations are determined through linkage analysis or association studies
  - Resolution is genomic intervals containing potentially hundreds of genes
- Network-based methods have so far been limited to methods that focus on the local neighborhood
  - Only look at direct neighbors of known disease genes
- Address this by developing methods that use the global network, and compare to effectiveness of previous methods

Methods

New methods use some measure of path connectivity to find putative novel disease genes

- Random Walk with Restart

$$\mathbf{p}^{t+1} = (1-r)\mathbf{W}\mathbf{p}^{t} + r\mathbf{p}^{0}$$

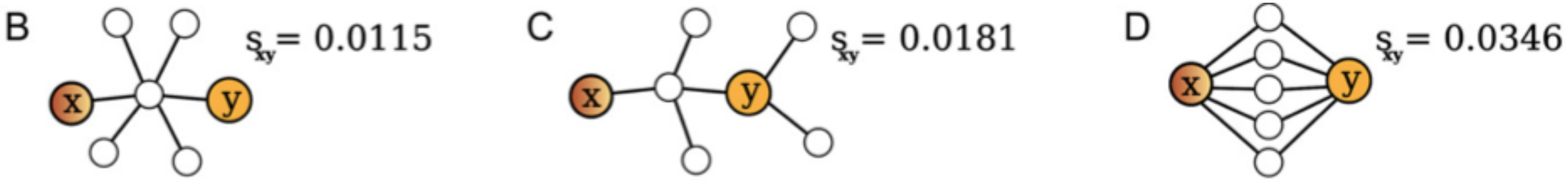
  - $\mathbf{p}^0$ is the probability vector representing the probability of the random walker starting at any of the known disease genes
  - *r* is the restart probability
  - *W* is the column-normalized adjacency matrix
  - $\mathbf{p}^t$ is the probability of the random walker being at any node in the network at time *t*
  - Run this until the change between $\mathbf{p}^t$ and $\mathbf{p}^{t+1}$, measured by the L_1 norm, is less than 10^{-6}

Methods

- Diffusion Kernel
  - $\boldsymbol{K} = e^{-ß\boldsymbol{L}}$
  - ß controls the magnitude of diffusion
  - ***L*** is the Laplacian of the network
  - Score each candidate gene *j* in accordance with its ***K***

$$score(j) = \sum_{i \in disease\_gene\_family} K_{ij}$$

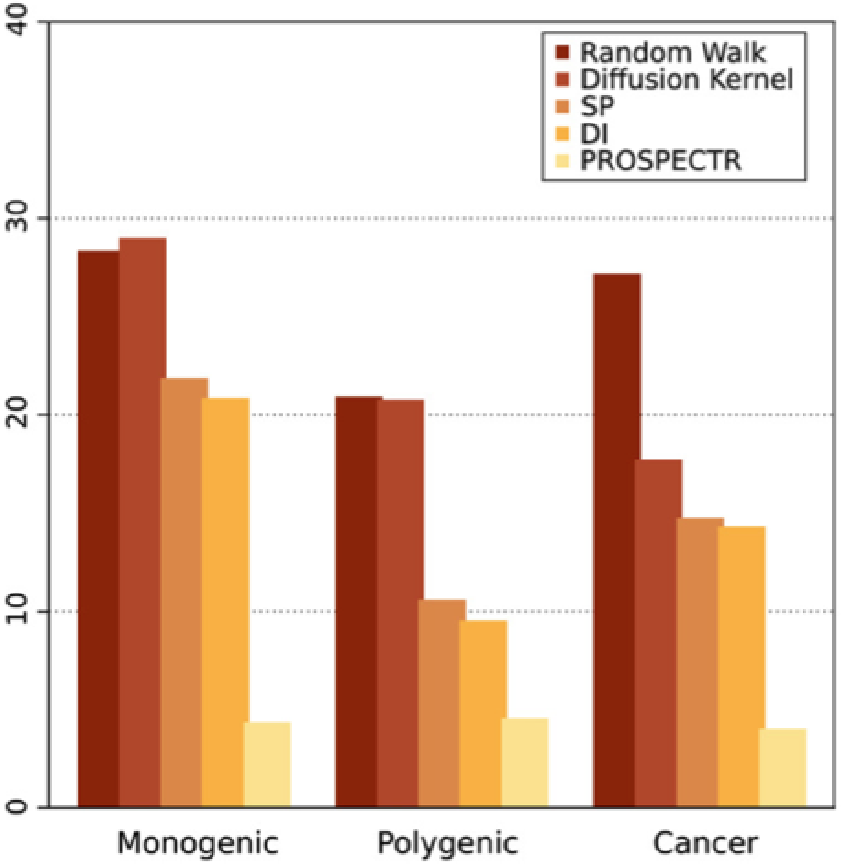
Methods



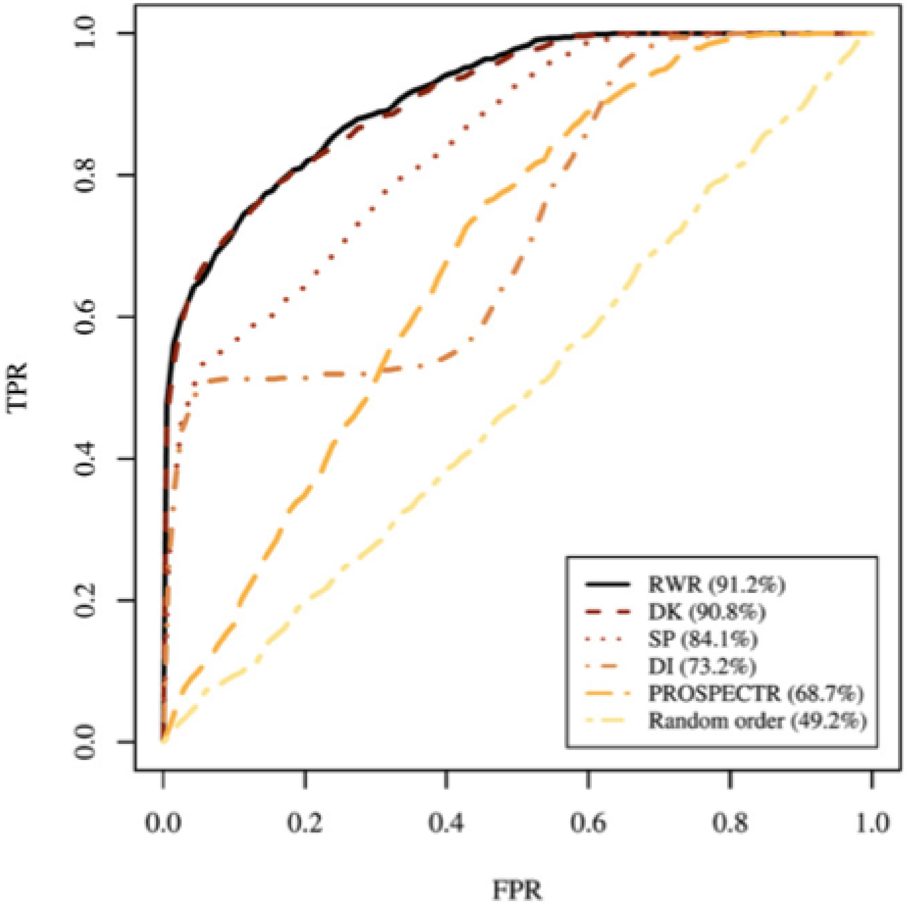
Methods

- Compared global network methods to local network methods
  - Direct Interaction (DI)
  - Shortest Paths (SP)
- Also used PROSPECTR, which uses sequence-based features to rank genes by likelihood of involvement in a particular disease
- Tested on 110 disease-gene families from OMIM
  - 783 genes
  - 86 heterogeneous disorders
  - 12 cancer syndromes
  - 12 complex (polygenic) disorders
- Conducted a leave-one-out cross-validation for each method on this data

Results



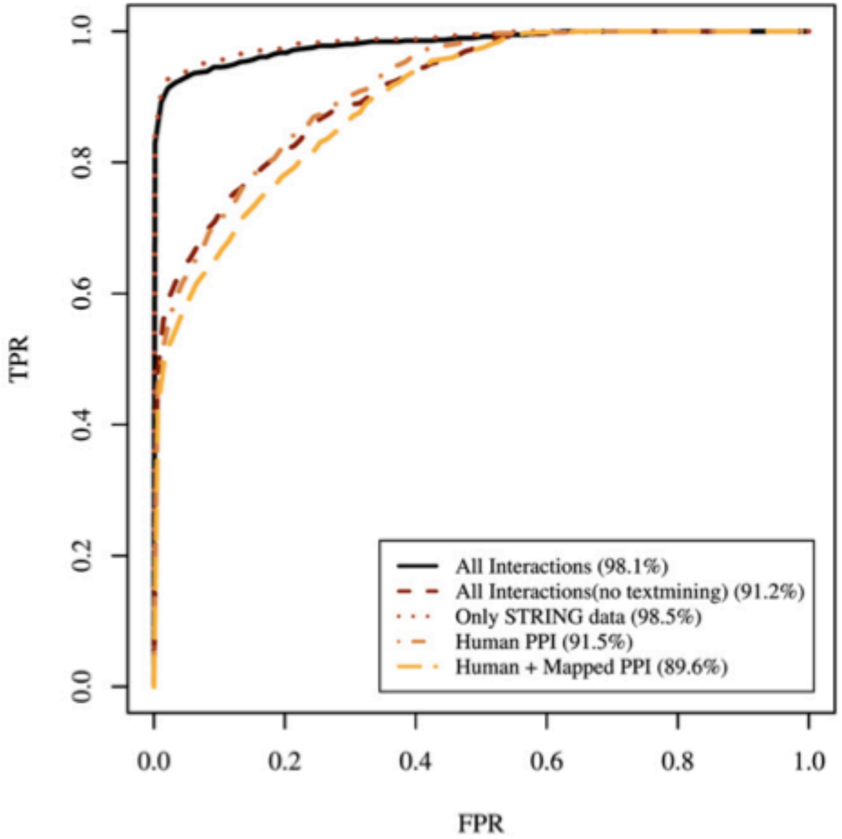
Cross-validation results
(Measures enrichment for true disease genes)



ROC curves of different methods

Comparison of Data Sources

- Interaction data came from a number of sources
  - Five human PPI databases
  - Interologs from four nonhuman species mapped by Inparanoid
  - STRING database: Interaction database based on experimental evidence, comparative genomics, and text mining
- ROC curves constructed when using Random Walk with Restart on subsets of the total interaction data



Conclusions

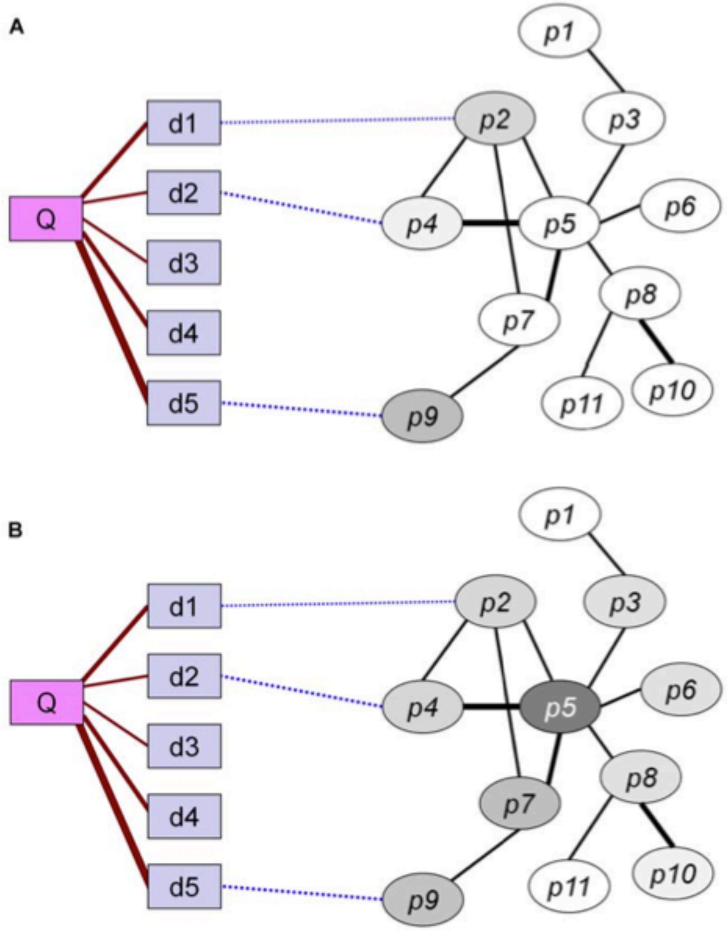
- Global network methods are much more useful for finding genes that may be associated with genetic diseases
- Method is limited by known interaction data
  - Improvements expected as interactome knowledge becomes more complete

Vanunu *et al*: Overview

- Also motivated to improve on network methods that only looked at local portions of the PPI network
- Goal is to both:
  - Prioritize genes for investigation into disease connections, and
  - Find protein complexes and modules involved in the disease of interest

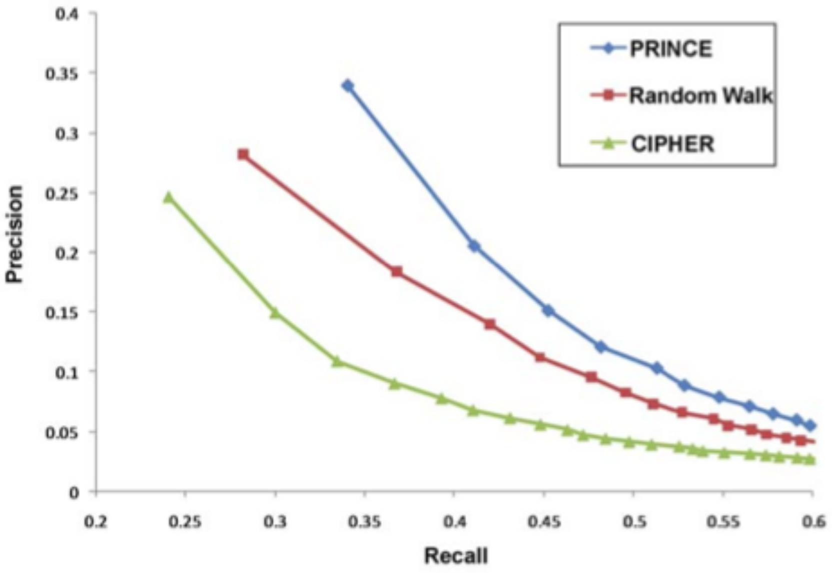
The PRINCE Method

- PRIoritizatioN and Complex Elucidation
- Combines disease similarity (based on MeSH descriptions) with searching a PPI network
- Query disease *Q* is connected to similar diseases (with connections weighted by magnitude of similarity)
- Known genes associated with these diseases are marked as prior information in the PPI network (derived from recent high-throughput experiments and HPRD)
- Compute a smooth scoring function over the network
  - The idea is that, over a number of iterations, there is "flow" from the prior nodes to its neighbors, and from any nodes that received flow on the previous iteration
  - Proceeds until convergence
  - Nodes with a high score after this procedure are prioritized for investigation into disease associations



Comparison to Other Methods

- Compared PRINCE to the Köhler Random Walk and CIPHER, an algorithm for predicting disease-gene associations based on direct interactions
- Conducted leave-one-out cross-validation on all 1,369 OMIM diseases for which at least one known causal gene is on record
- PRINCE consistently outperforms the other methods, even on 2-fold, 5-fold, and 10-fold cross-validation



Identifying Novel Causal Genes

- Used PRINCE on
  - Prostate Cancer
  - Alzheimer's Disease
  - Diabetes Mellitus, type 2
- Found that over 50% of top candidates already had confirmed involvement in these diseases
  - PRINCE provides additional confirmatory evidence
- Rest of top candidates are not previously implicated
  - Novel genes to investigate

Identifying disease-associated protein complexes

- Identified some 700 complexes associated with OMIM diseases
- Tested coherency of these complexes
  - Functional coherency (similar functional annotations)
  - Expression coherency (similar expression patterns under multiple conditions)
  - Conservation coherency (similar phylogenetic profiles)
- Compared PRINCE complexes' coherency to the coherency of:
  - Manually curated GO complexes
  - Computationally predicted PPI complexes (not necessarily disease-associated), and
  - A set of complexes predicted on a phenome-interactome network[1]

[1] Lage K, Karlberg EO, Storling ZM, Olason PI, Pedersen AG, et al. (2007) A human phenome-interactome network of protein complexes implicated in genetic disorders. Nat Biotech 25: 309–316.

Coherency Table

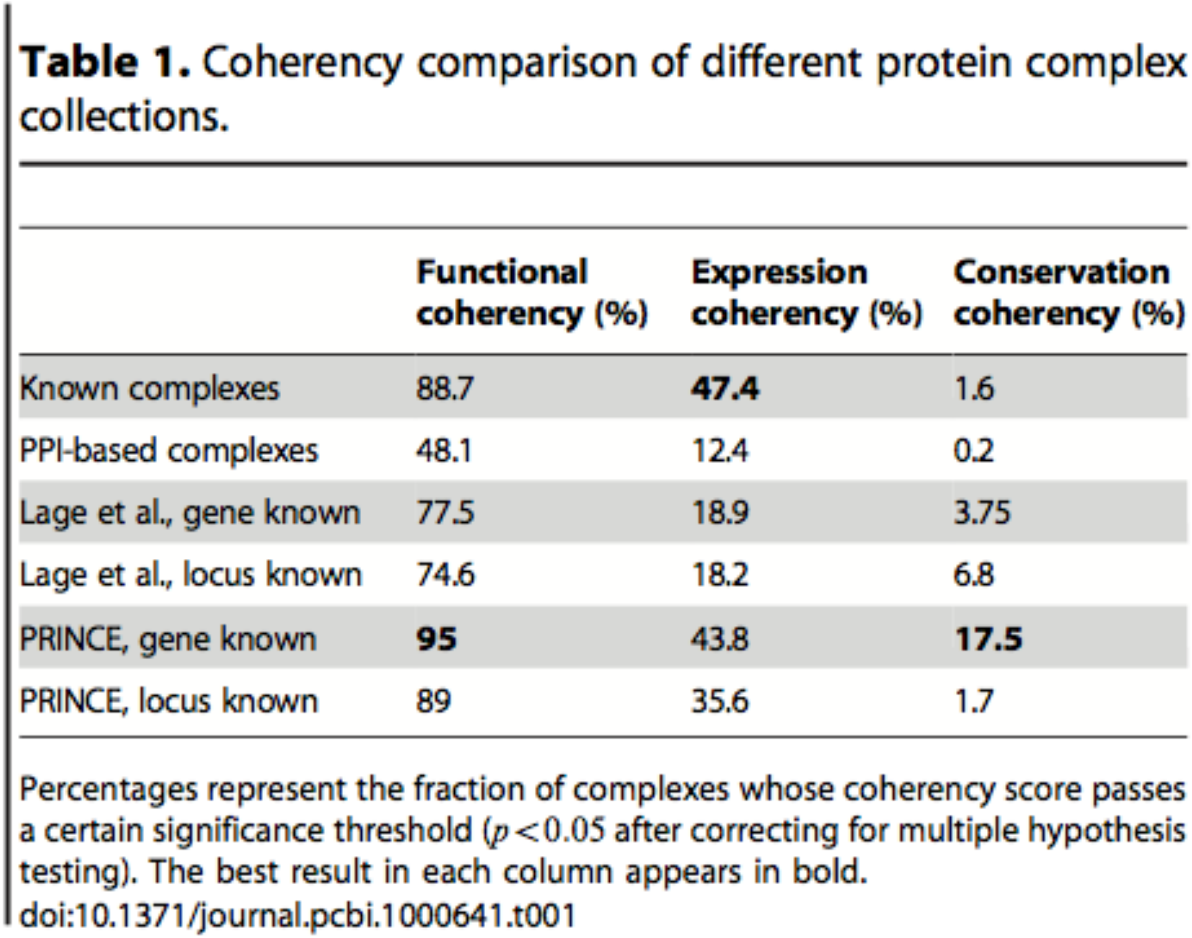


Table 1. Coherency comparison of different protein complex collections.

	Functional coherency (%)	Expression coherency (%)	Conservation coherency (%)
Known complexes	88.7	**47.4**	1.6
PPI-based complexes	48.1	12.4	0.2
Lage et al., gene known	77.5	18.9	3.75
Lage et al., locus known	74.6	18.2	6.8
PRINCE, gene known	**95**	43.8	**17.5**
PRINCE, locus known	89	35.6	1.7

Percentages represent the fraction of complexes whose coherency score passes a certain significance threshold ($p < 0.05$ after correcting for multiple hypothesis testing). The best result in each column appears in bold.

doi:10.1371/journal.pcbi.1000641.t001

Validation Against OMIM

- Checked OMIM entries for mention of proteins in PRINCE complexes not already known to be implicated with their respective diseases
- Found support for members of 61% of PRINCE complexes in this manner, with an average of 3.6 genes/complex mentioned in OMIM
- For random complexes, only 7% of complexes were supported, with an average 1.6 genes/complex mentioned in an OMIM entry

Conclusions

- PRINCE is a powerful method for prioritizing putative disease genes for investigation, and for implicating protein complexes in disease
- Successful at making predictions for complex, polygenic diseases

Limitations

- Relies on prior phenotypic information → useful only for studying phenotypically similar diseases with known genes
- Doesn't incorporate a range of useful information, like expression information
- Dependent on extent of current knowledge of the PPI interactome

The End

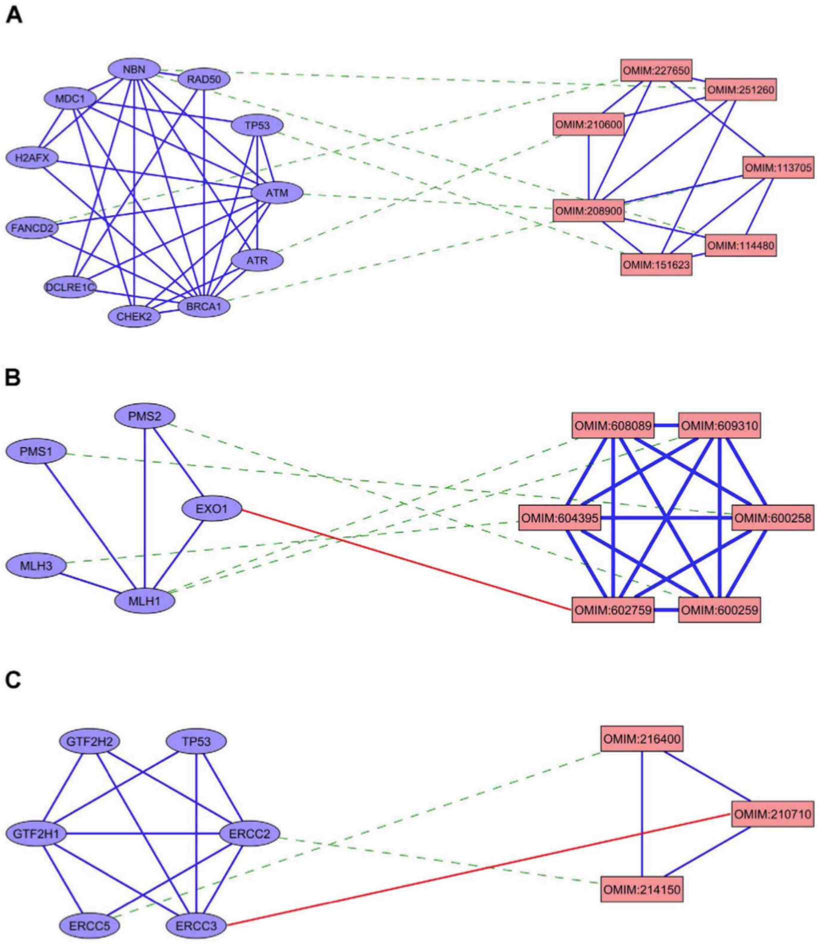
Any Questions?

Supplementary Info: Prioritization Function

- For a node $v \in V$, denote its direct neighborhood by $N(v)$
- Let F:V→R represent a prioritization function
- Let Y:V→[0,1] represent a prior knowledge function
  - Assign 1 to nodes that are known to be related to the given disease q
  - 0 otherwise
- Requirements of F: $F(v) = \alpha\left[\sum_{u \in N(v)} F(u) w'(v,u)\right] + (1-\alpha) Y(v)$
- Computing F iteration by iteration:

$$F^t := \alpha W' F^{t-1} + (1-\alpha) Y$$

Supplementary Info: Case Study of Inferred Complexes


A
NBN
RAD50
MDC1
TP53
H2AFX
ATM
FANCD2
ATR
DCLRE1C
CHEK2
BRCA1
OMIM:227650
OMIM:251260
OMIM:210600
OMIM:113705
OMIM:208900
OMIM:114480
OMIM:151623
B
PMS2
PMS1
EXO1
MLH3
MLH1
OMIM:608089
OMIM:609310
OMIM:604395
OMIM:600258
OMIM:602759
OMIM:600259
C
GTF2H2
TP53
GTF2H1
ERCC2
ERCC5
ERCC3
OMIM:216400
OMIM:210710
OMIM:214150