A DIseAse MOdule Detection (DIAMOnD) Algorithm Derived from a Systematic Analysis of Connectivity Patterns of Disease Proteins in the Human Interactome

Susan Dina Ghiassian, Jörg Menche, Albert-László Barabási

Background & Goal

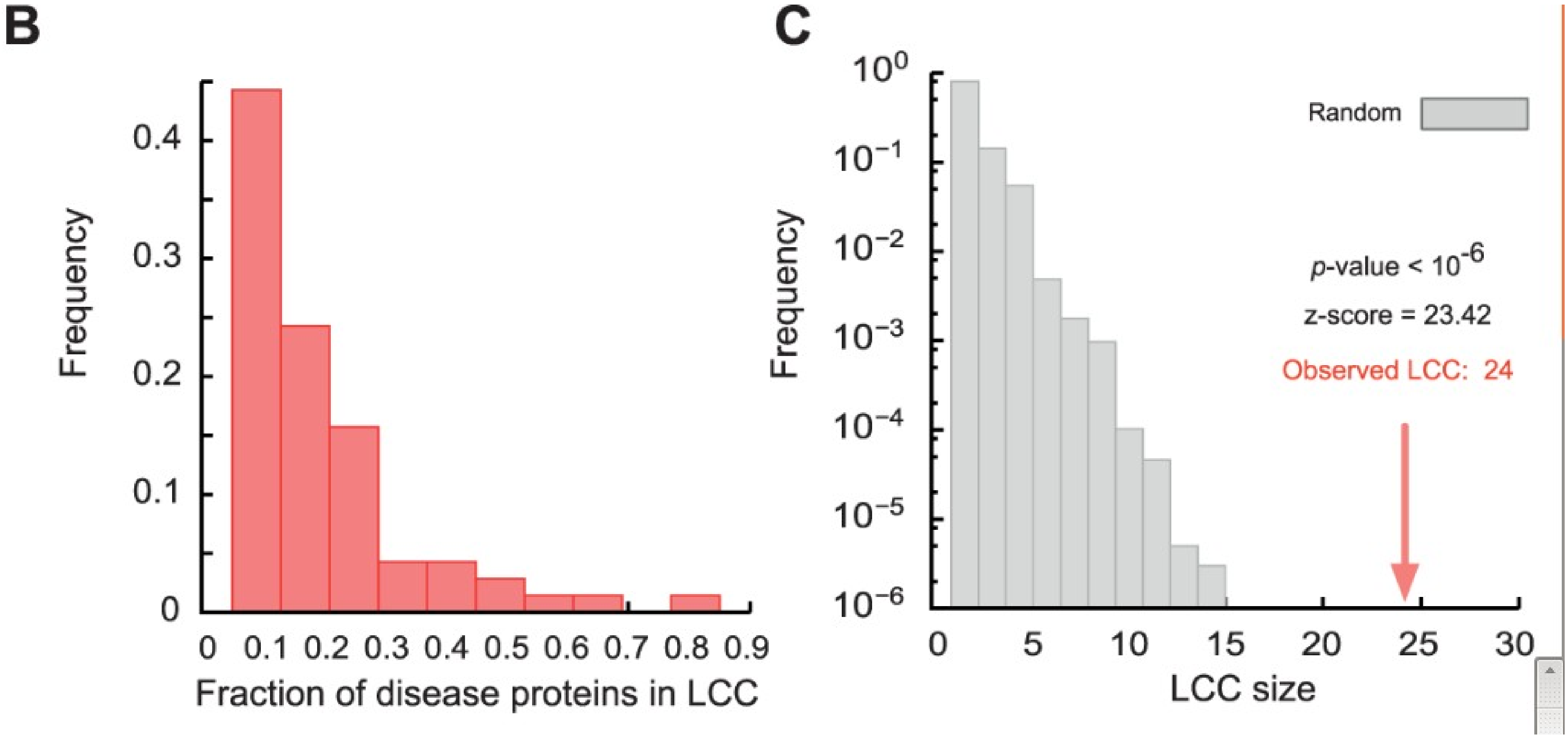
- Diseases are results of many abnormal proteins interacting with each other, disease module exists.
- Studying the underlying connectivity patterns shared among disease modules.
- Disease module detection algorithm to identify full disease module from already known disease proteins.

Preparation: disease protein & PPI

- 70 diseases, with disease related proteins from OMIM and GWAS
- PPI includes regulatory interactions, metabolic interactions, etc. 141,296 interactions among 13,460 proteins.

Interactome maps and disease modules are incomplete

- LCC: largest connected conponent



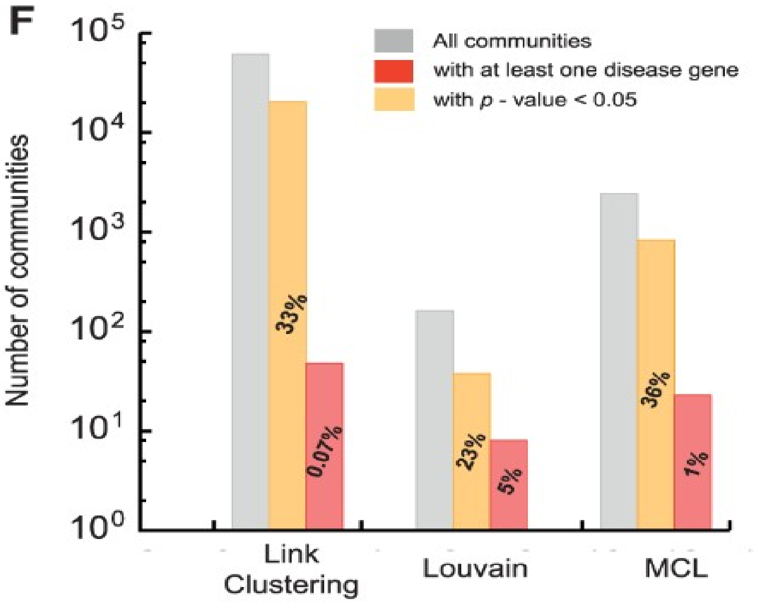
Connectivity pattern—interaction density in module isn't the key

- Result communities from dense subgraph detection algorithm
- Modularity parameter R:

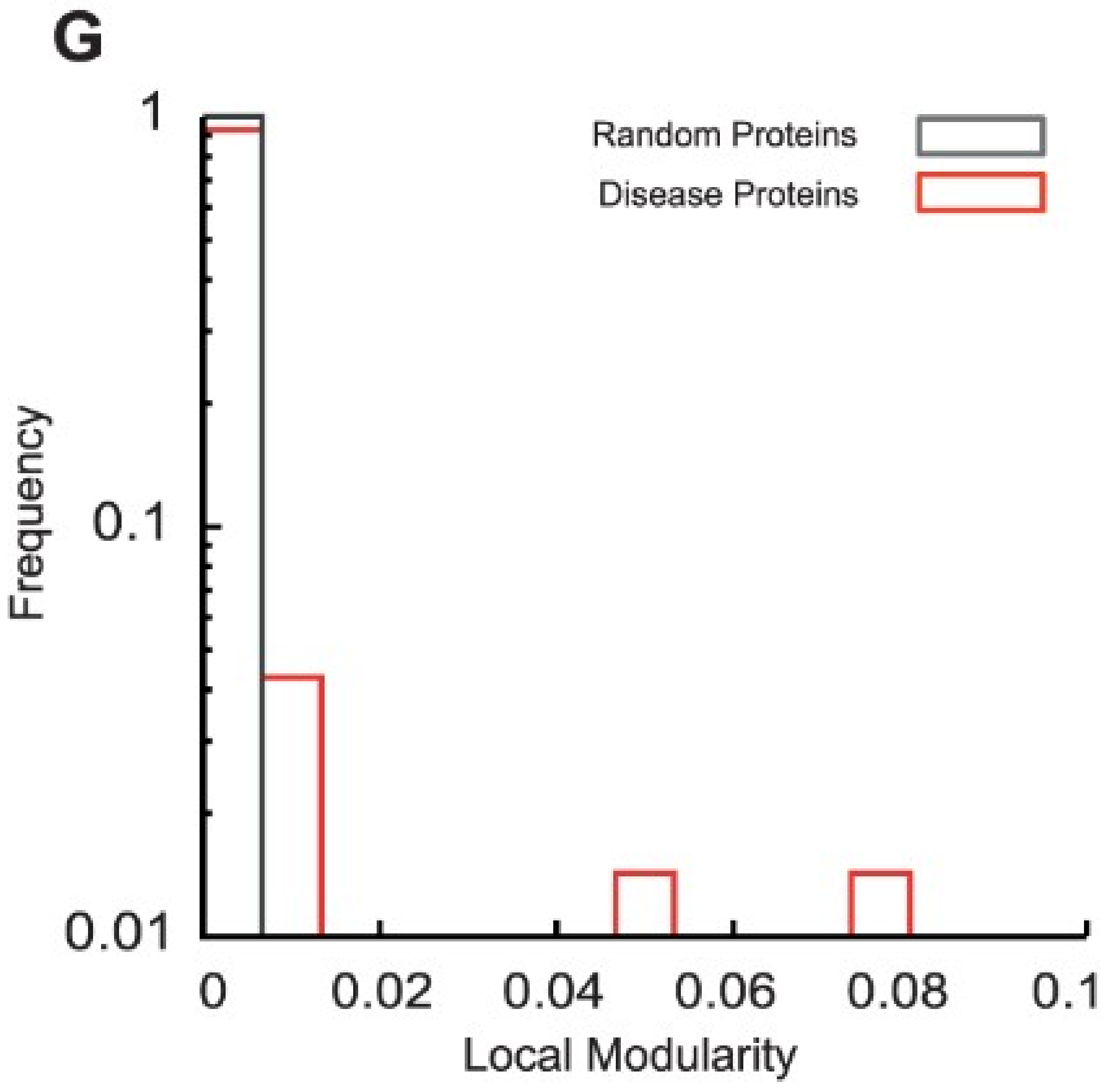
$$R = \frac{number\, of\ links\, of\ boundary\, nodes\, that\, are\, within\, module}{total\, number\, of\ links\, of\ boundary\, nodes}$$

Communities from dense subgraph detection algorithm

- 1%-5% communities are enriched with disease proteins
- These enriched communities only contains ~15%-38% proteins of that kind of disease
- Only 15% diseases has enriched communities



Modularity parameter R



Connectivity pattern—connectivity significance

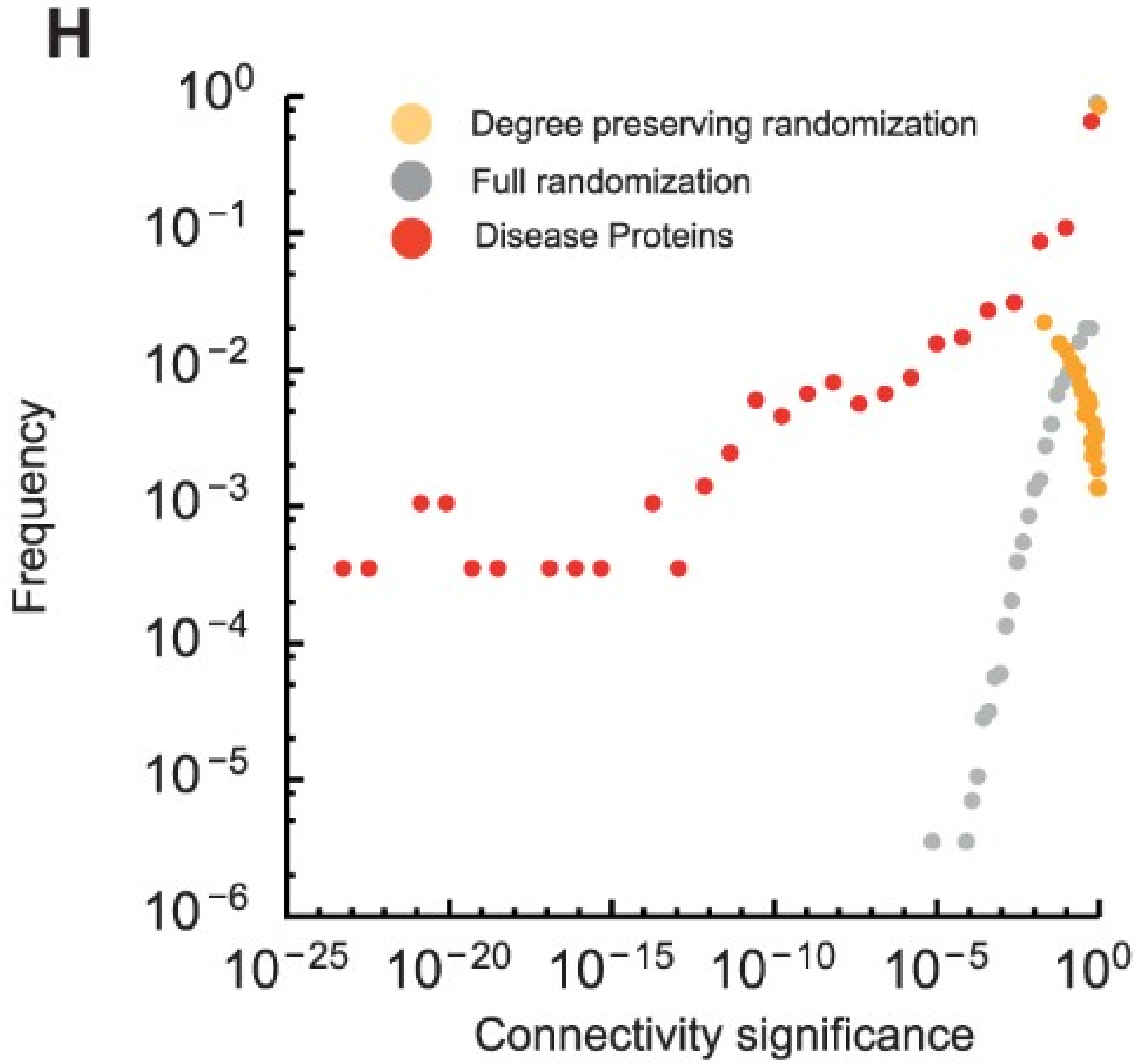
- How significant is a protein interact with seed proteins of disease in a N-node network?

$$p(k,k_s)=\frac{\binom{s_0}{k_s}\binom{N-s_0}{k-k_s}}{\binom{N}{k}}$$

Probability of a protein with total k links having ks links with seed proteins

$$p-value=\sum_{k_i=k_s}^{k} p(k,k_i)$$

Significance of disease proteins

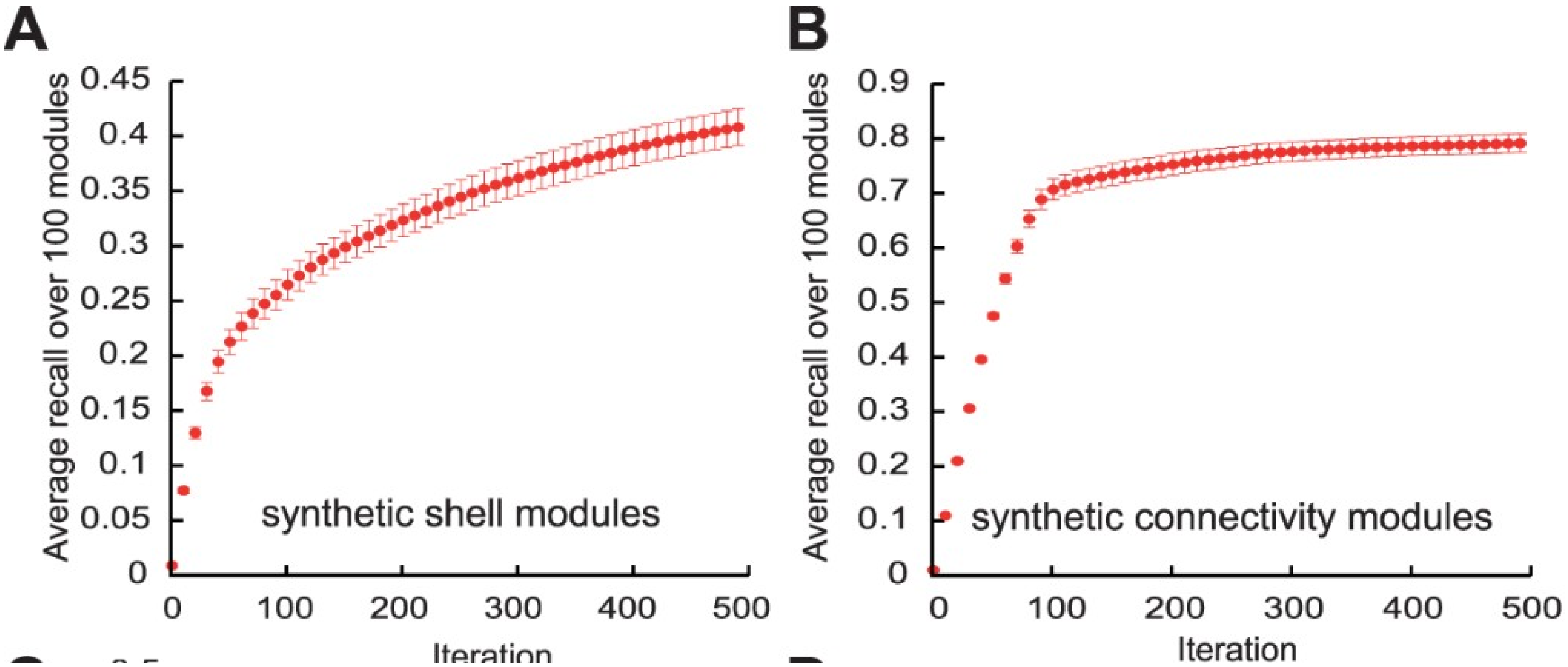


Algorithm to detect disease protein according to significance

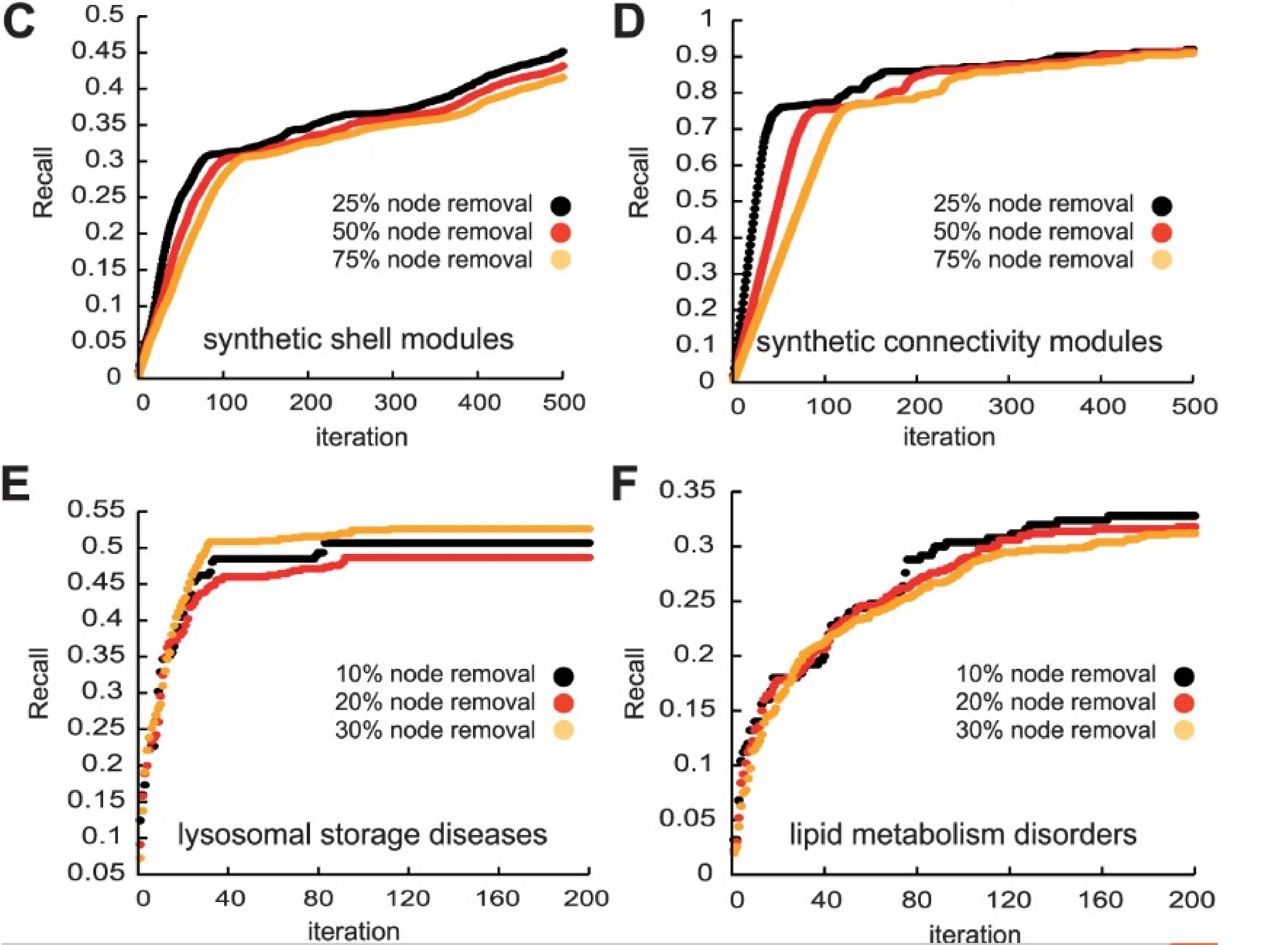
DIAMOnD algorithm

- Determine connectivity significance of any protein to seed proteins
- Rank all proteins based on p-values
- Add the highest rank protein (lowest p-value) to the set of seed proteins
- Repeat the above procedure

Test with synthetic modules

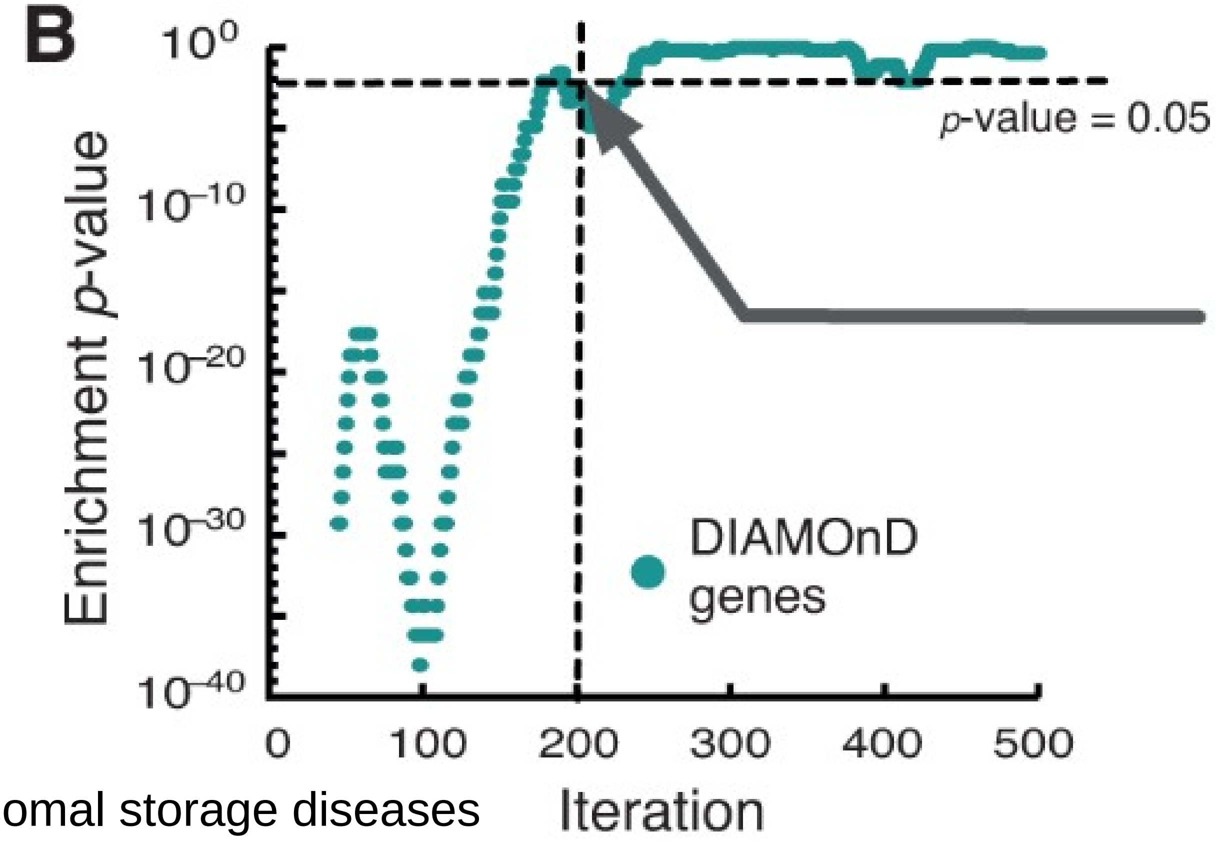


Robust with different seed size



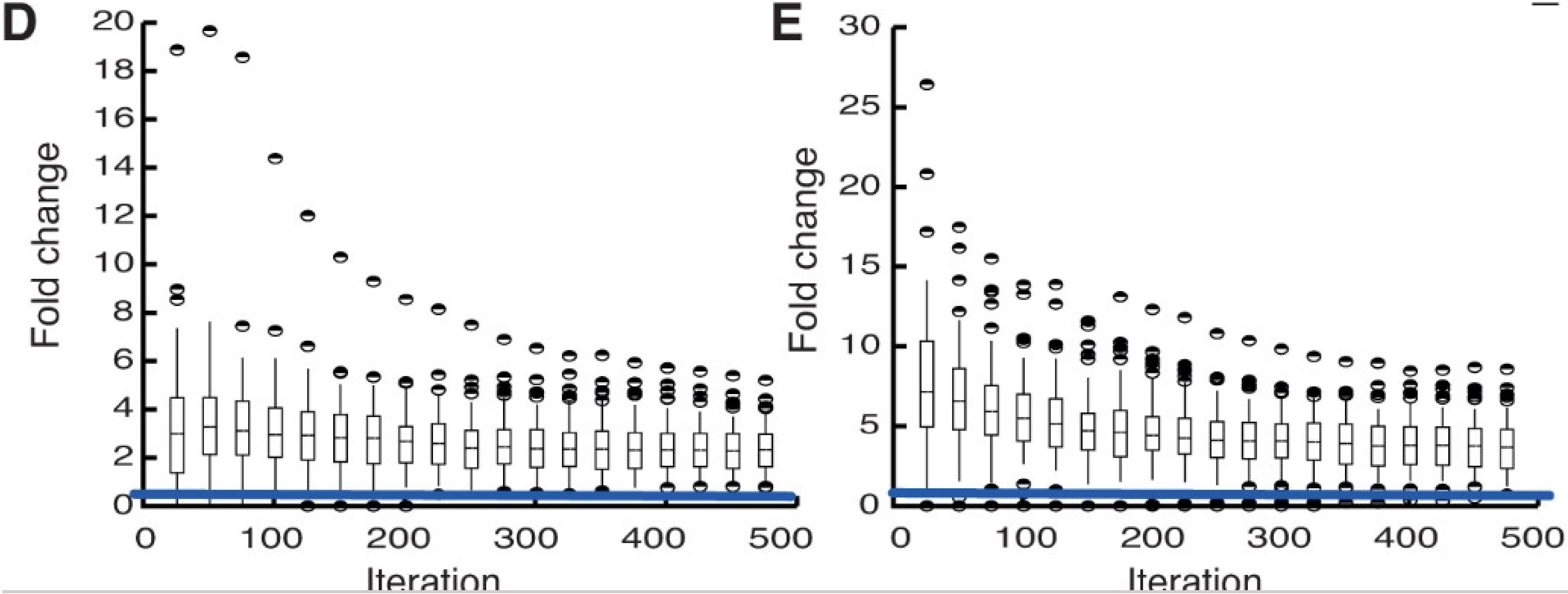
Validating disease module prediction

- Criteria for correctness: similarity of GO annotation and pathway annotation with seed proteins

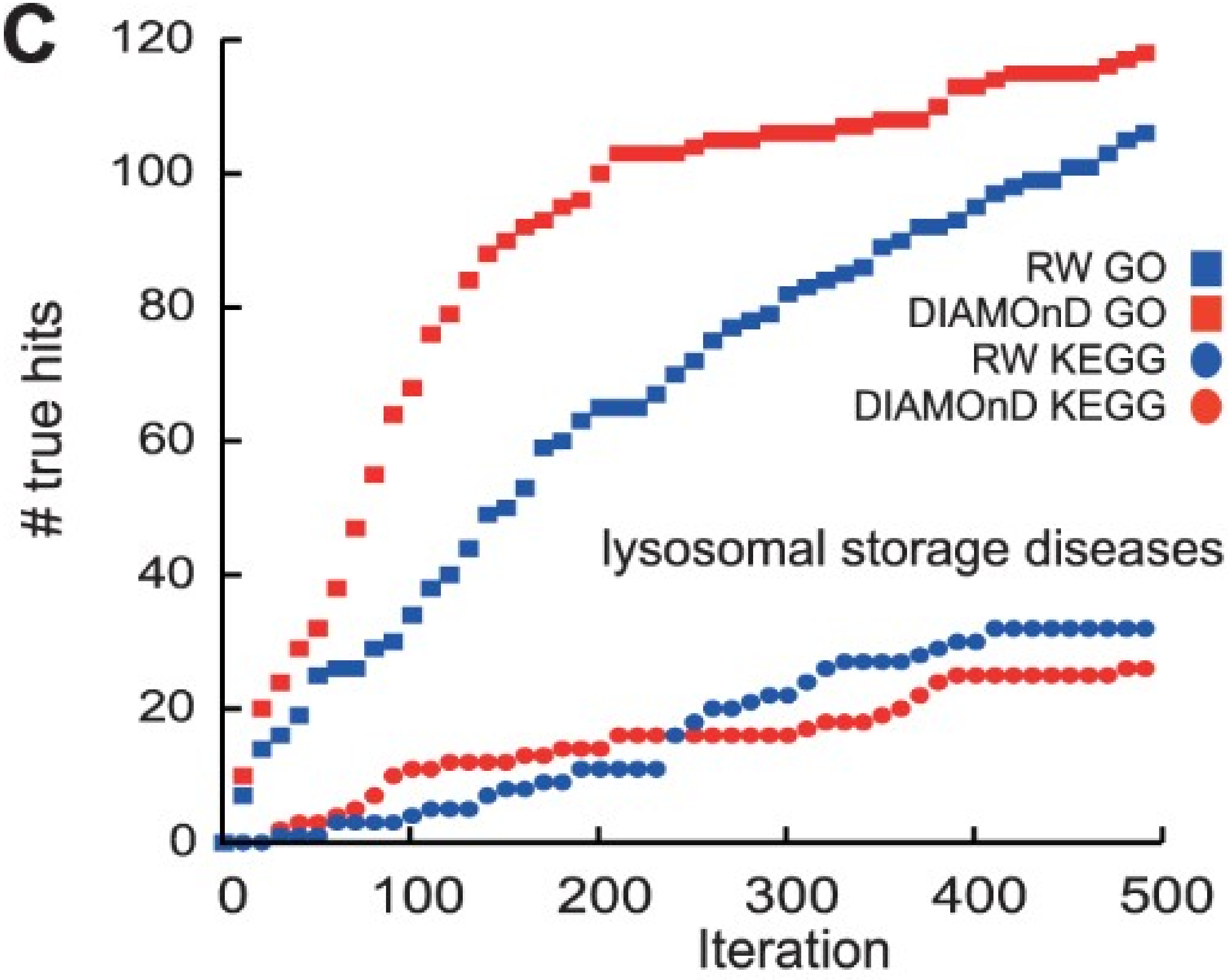


Lysosomal storage diseases

Validation for all 70 diseases



Compare to RW algorithm



Extension—accounting for link weight

- Link to original seed proteins has a weight higher than link to later added seed proteins
- Connectivity probability:

$$p(k, k_s, k_{s_0}) = \frac{\binom{s+(\alpha-1)s_0}{k_s+(\alpha-1)k_{s_0}}\binom{N-s}{k-k_s}}{\binom{N+(\alpha-1)k_s}{k+(\alpha-1)k_{s_0}}}$$