A Gene Co-expression Network for Global Discovery of Conserved Genetic Modules

Joshua Stuart, Eran Segal, Daphne Koller, Stuart Kim

Science 2003

Presented by

Meghana Kshirsagar

Motivation for cross-species analysis

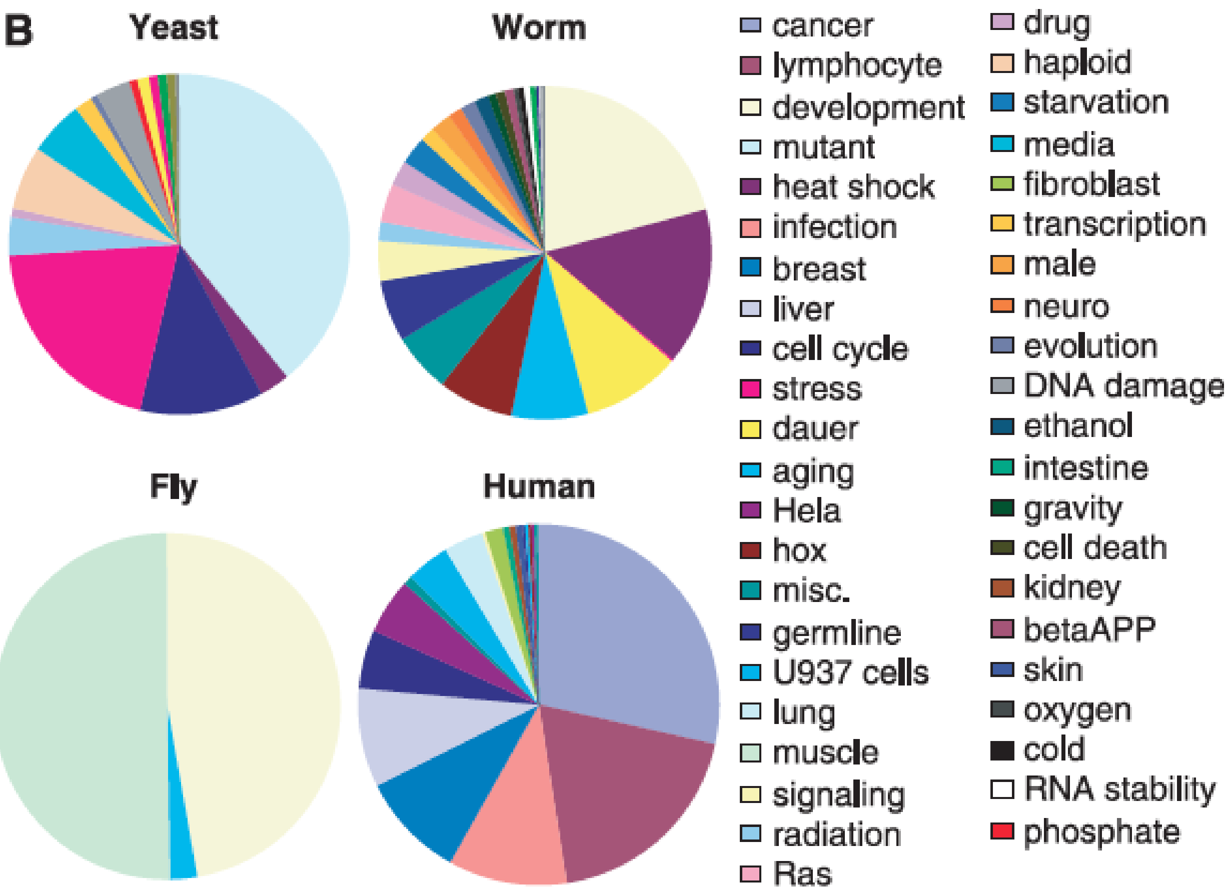
Correlation of genes across experimental conditions ⇨ coregulation of genes

Coregulation of genes across organisms ⇨ functional relationship

Dataset

3182 DNA microarrays from 4 organisms

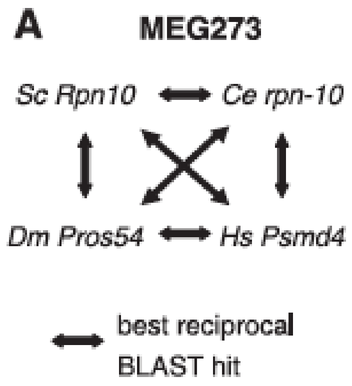
Human	1202
Worm	979
Fly	155
Yeast	643



Technique

- Network of “metagenes”
- Metagenes : orthologous set of genes from different organisms
  - 6307 metagenes

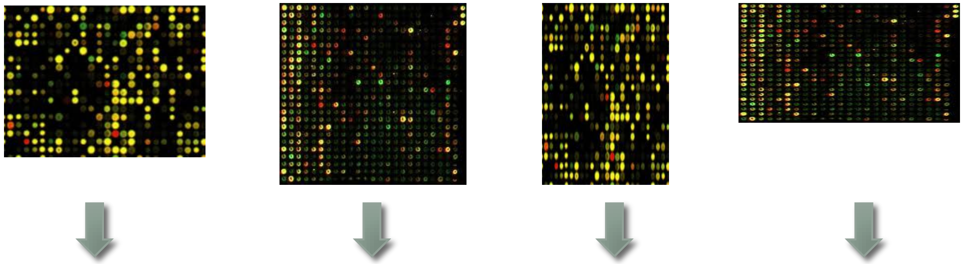
Human	Worm	Fly	Yeast
6591	5180	5802	2434



- Edges between co-expressed metagenes
  - resulted in 3416 metagenes connected by 22163 edges

Defining co-expression edges

Input :



Compute: $\rho\,(g_i\,,\,g_j)$ $\rho\,(g_i\,,\,g_j)$ $\rho\,(g_i\,,\,g_j)$ $\rho\,(g_i\,,\,g_j)$ for all gene pairs

For every directed pair of meta-genes *(m, m')* compute rank-ratios:

m: h_1 y_2 f_3 w_4

m': h'_1 y'_2 f'_3 w'_4

$$r_1 = \frac{\text{rank of } h'_1 \text{ w.r.t } h_1}{\text{\# metagenes with human}}$$

Similarly obtain r_2, r_3, r_4 for other organisms

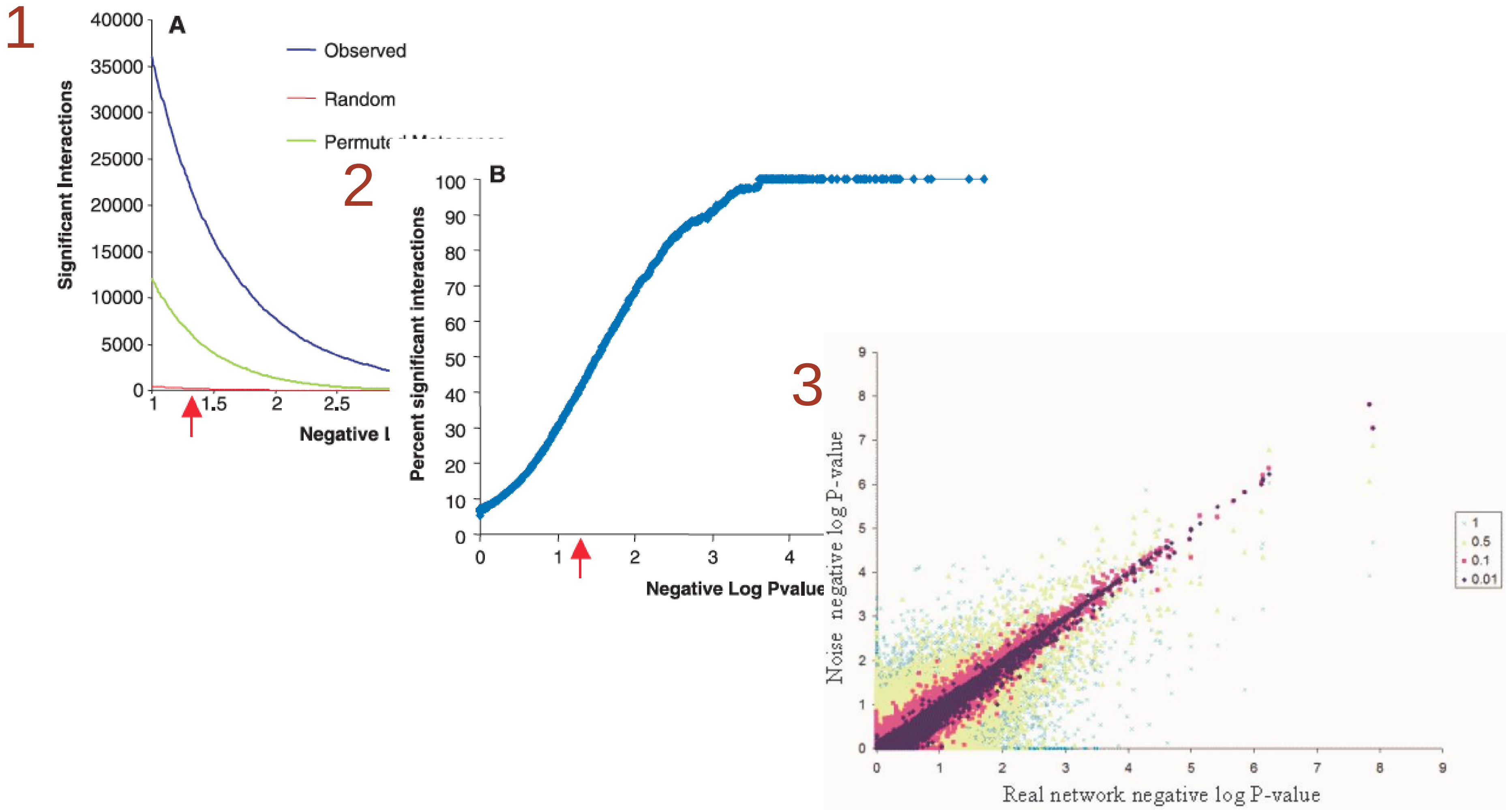
P-value of co-expression edge:

$$P(r_1 r_2, \ldots, r_n) = n! \int_0^{r_1} \int_{s_1}^{r_2} \cdots \int_{s_{n-1}}^{r_n} ds_1 ds_2 \cdots ds_n$$

Threshold on p-value to include edges in the network

Validation tests

Discuss 3 computational tests to validate the network

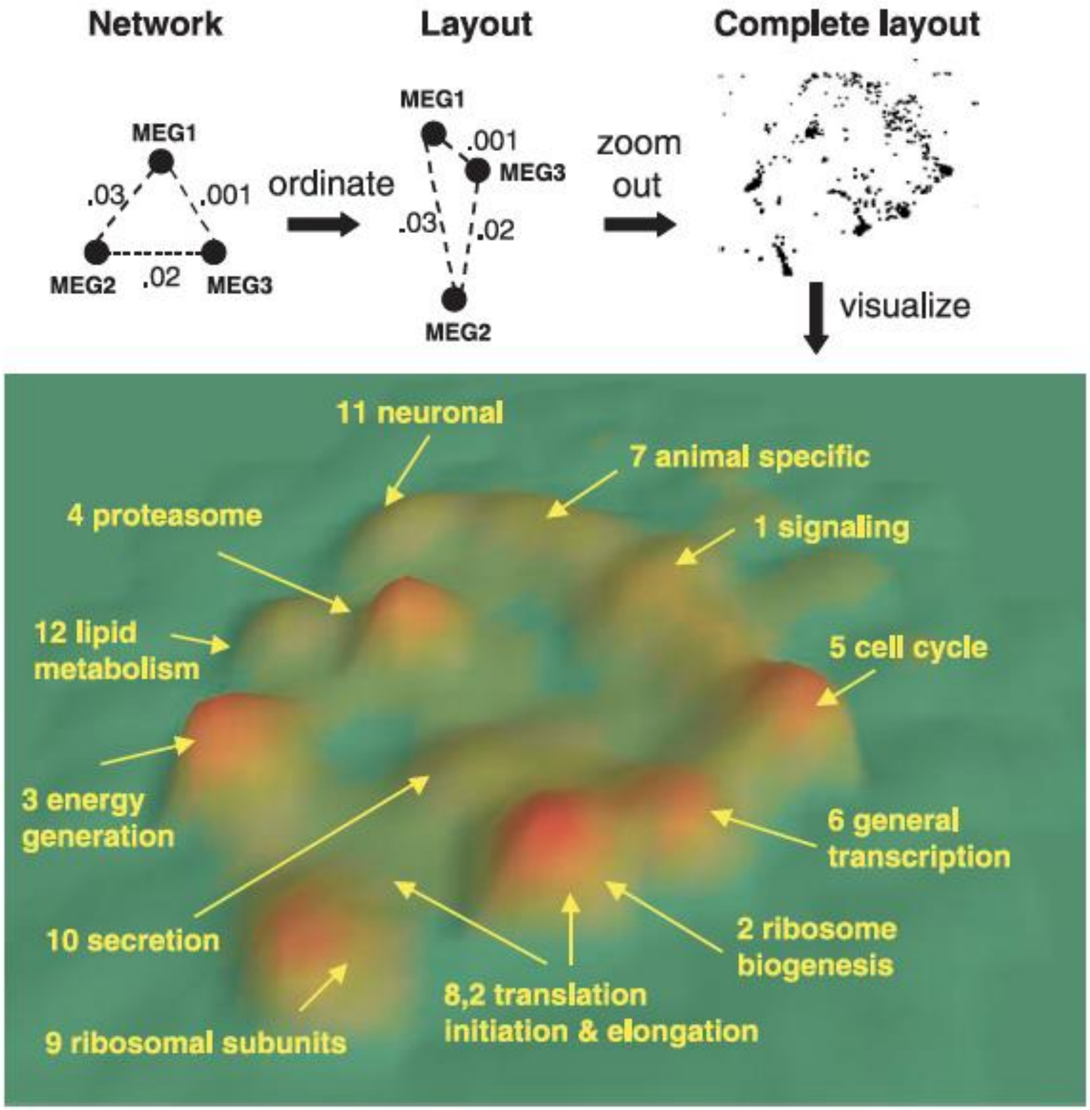


Biological Insights

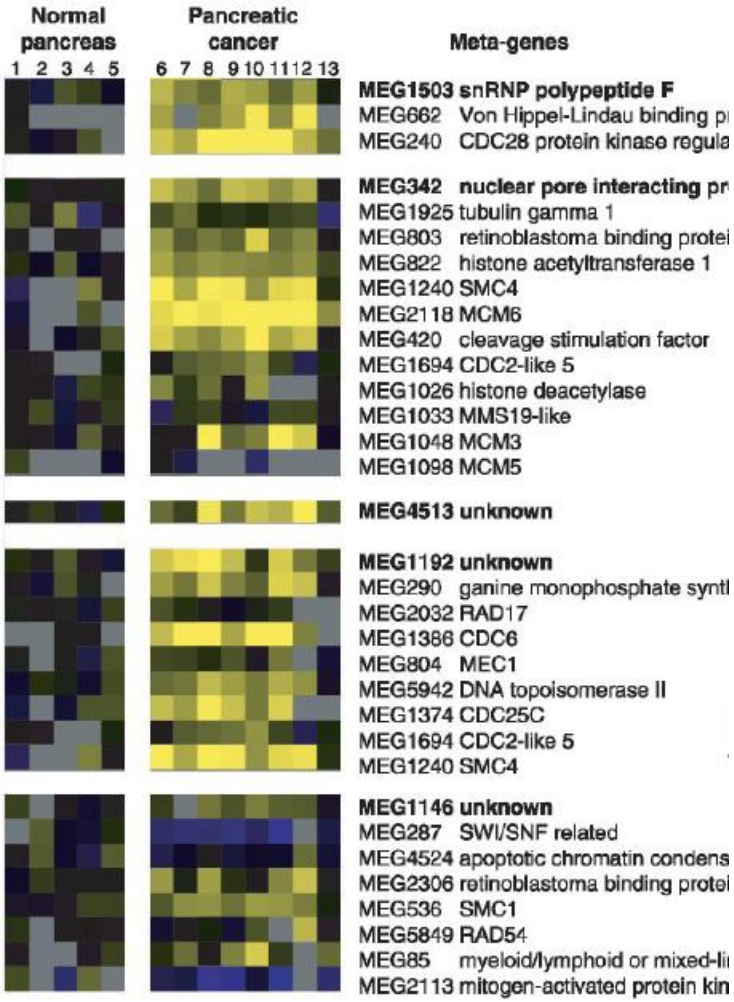
- Plot using VxInsight
- K-means clustering to reveal biological function of components

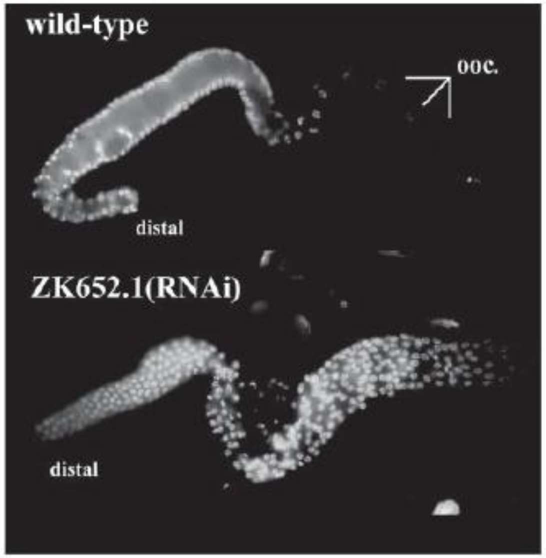
131 genes from component 5 were assigned “cell cycle” functionality due to proximity

5 genes assigned as “cell proliferation” functionality



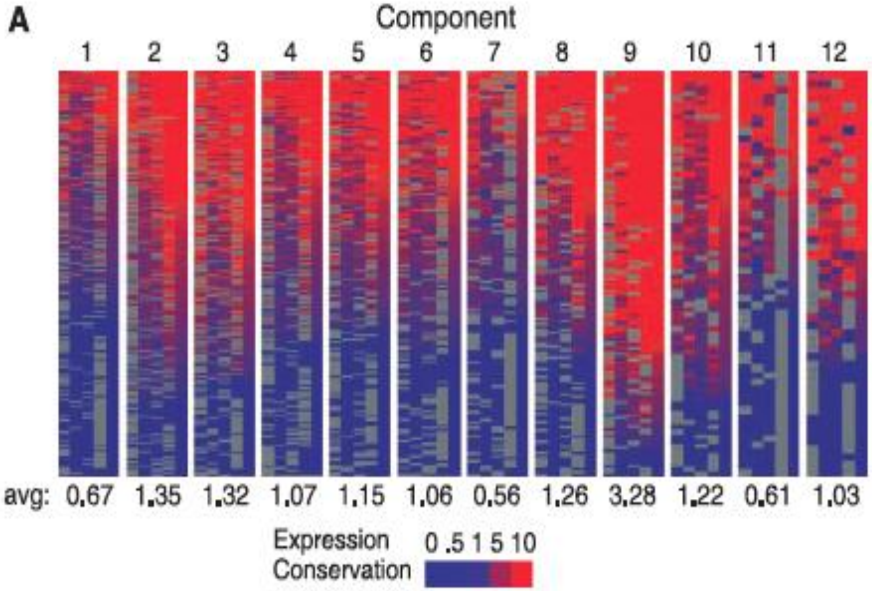
Validation of functional predictions

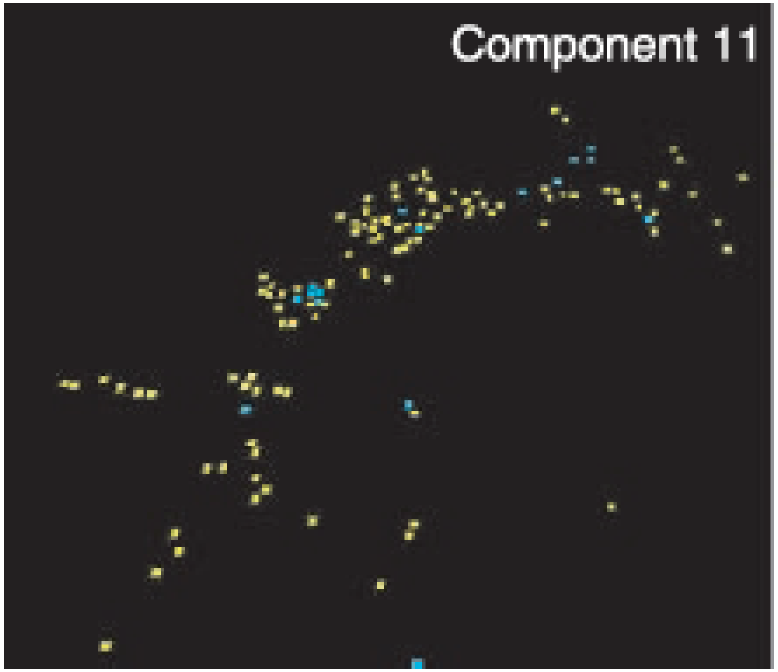


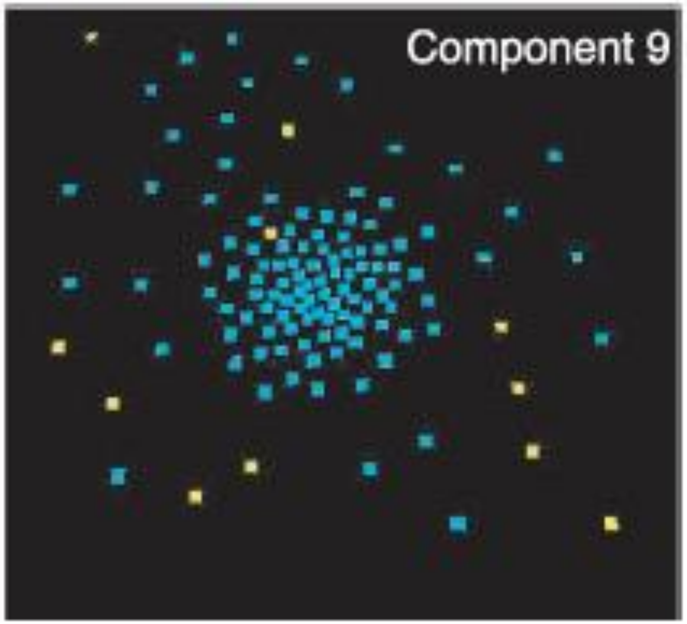


Comparison of multiple-species network with single-species network, using KEGG functional category enrichment

Genetic modules







$$ECI_m = \sum_{s=1}^{n} \frac{I_{gs} \cap I_m}{I_{gs} \cup I_m}$$

3 types of modules:

a) Ancient
b) Evolving
c) Flexible

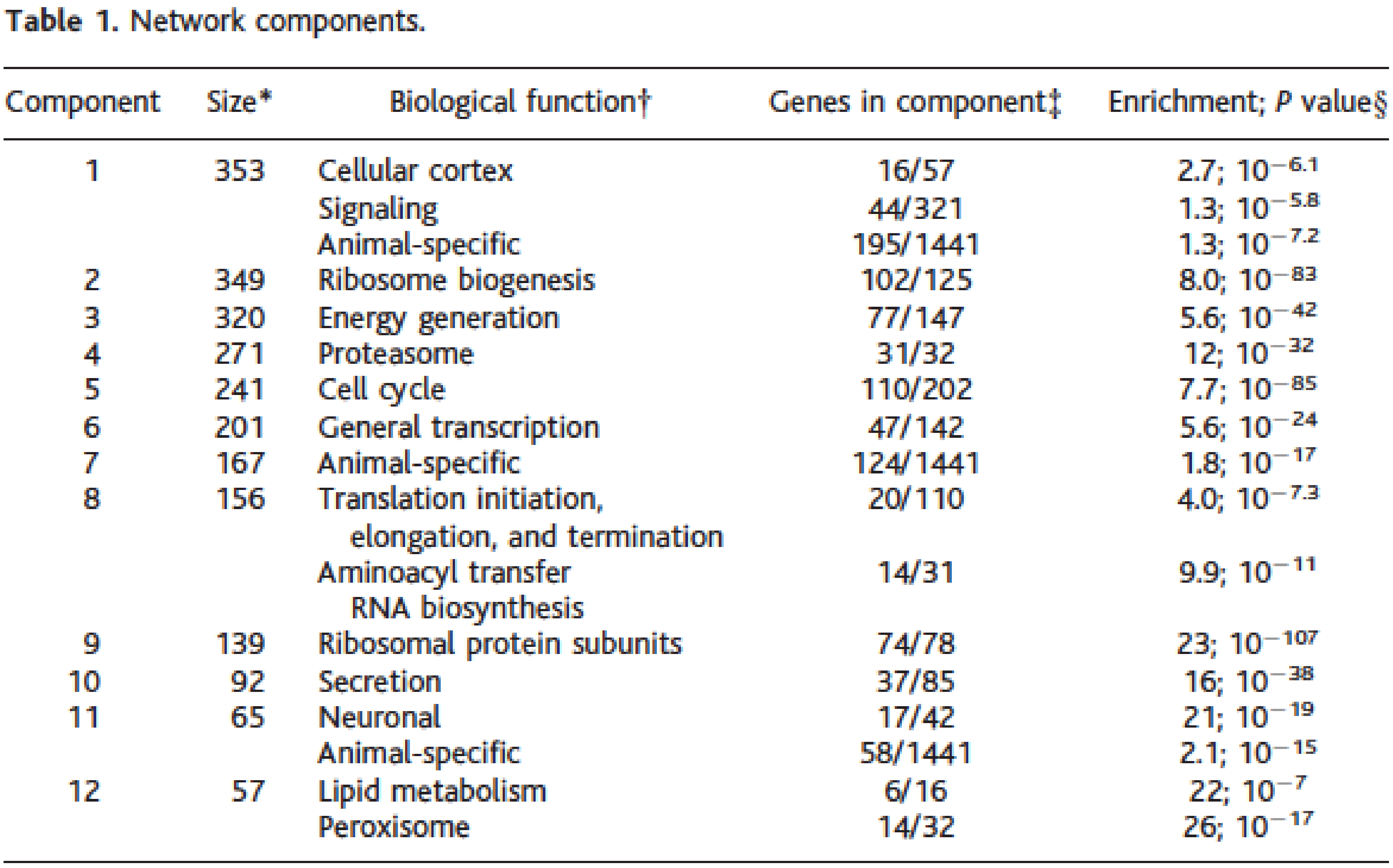
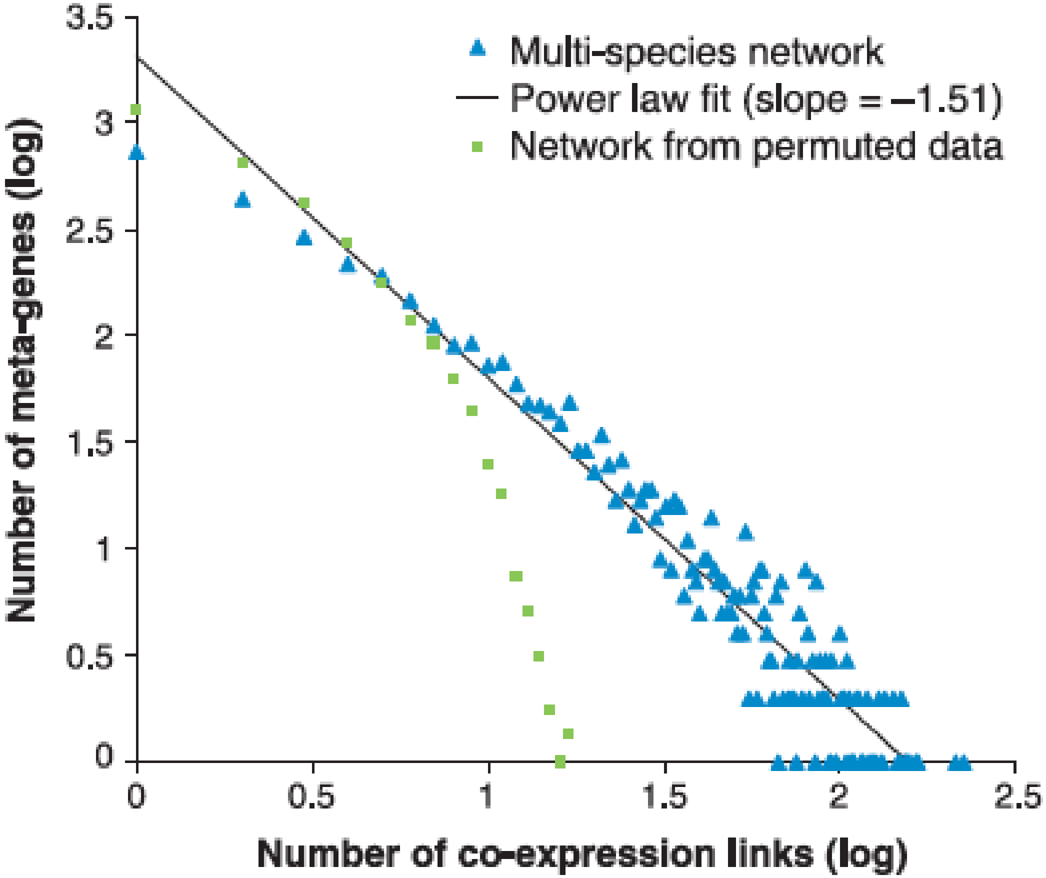


Table 1. Network components.

Component	Size*	Biological function†	Genes in component‡	Enrichment; *P* value§
1	353	Cellular cortex	16/57	2.7; $10^{-6.1}$
		Signaling	44/321	1.3; $10^{-5.8}$
		Animal-specific	195/1441	1.3; $10^{-7.2}$
2	349	Ribosome biogenesis	102/125	8.0; 10^{-83}
3	320	Energy generation	77/147	5.6; 10^{-42}
4	271	Proteasome	31/32	12; 10^{-32}
5	241	Cell cycle	110/202	7.7; 10^{-85}
6	201	General transcription	47/142	5.6; 10^{-24}
7	167	Animal-specific	124/1441	1.8; 10^{-17}
8	156	Translation initiation, elongation, and termination	20/110	4.0; $10^{-7.3}$
		Aminoacyl transfer RNA biosynthesis	14/31	9.9; 10^{-11}
9	139	Ribosomal protein subunits	74/78	23; 10^{-107}
10	92	Secretion	37/85	16; 10^{-38}
11	65	Neuronal	17/42	21; 10^{-19}
		Animal-specific	58/1441	2.1; 10^{-15}
12	57	Lipid metabolism	6/16	22; 10^{-7}
		Peroxisome	14/32	26; 10^{-17}

Co-expression network has power law distribution



Contributions

- Technique to combine thousands of micro array datasets across species.
- Implication of genes involved in pancreatic cancer
- Finding modules conserved across organisms, as well as distinct / animal-specific modules