

Comparative genomics & genetics networks

What is comparative genomics?

“Because all modern genomes have arisen from common ancestral genomes, the relationships between genomes can be studied with this fact in mind. This commonality means that information gained in one organism can have application in other even distantly related organisms. Comparative genomics enables the application of information gained from facile model systems to agricultural and medical problems. The nature and significance of differences between genomes also provides a powerful tool for determining the relationship between genotype and phenotype through comparative genomics and morphological and physiological studies.”

Function prediction with comparative genomics

Why not just use sequence similarity?

- 40% of predicted genes in newly sequenced genomes cannot be assigned function based on sequence similarity
- Assumption: genes are functionally related will be associated across genomes

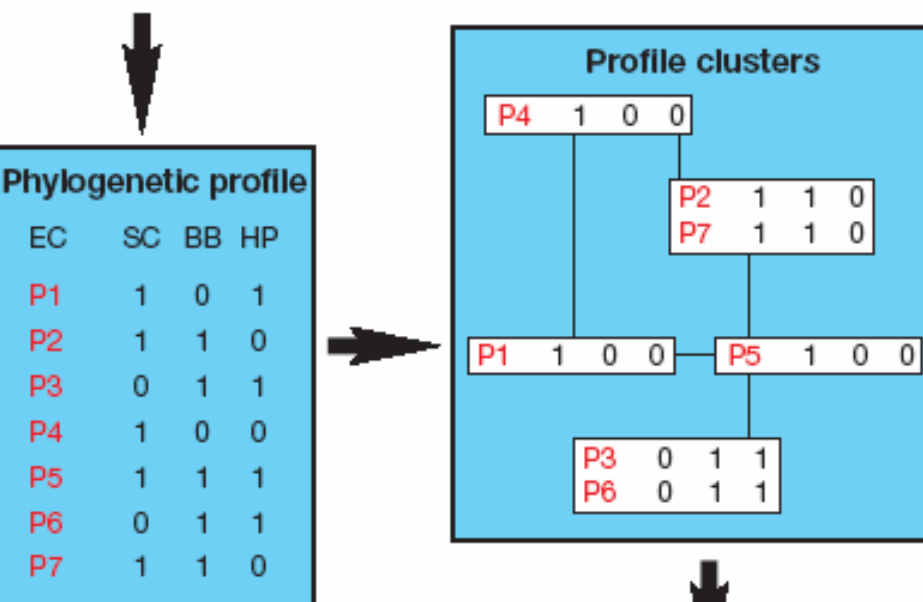
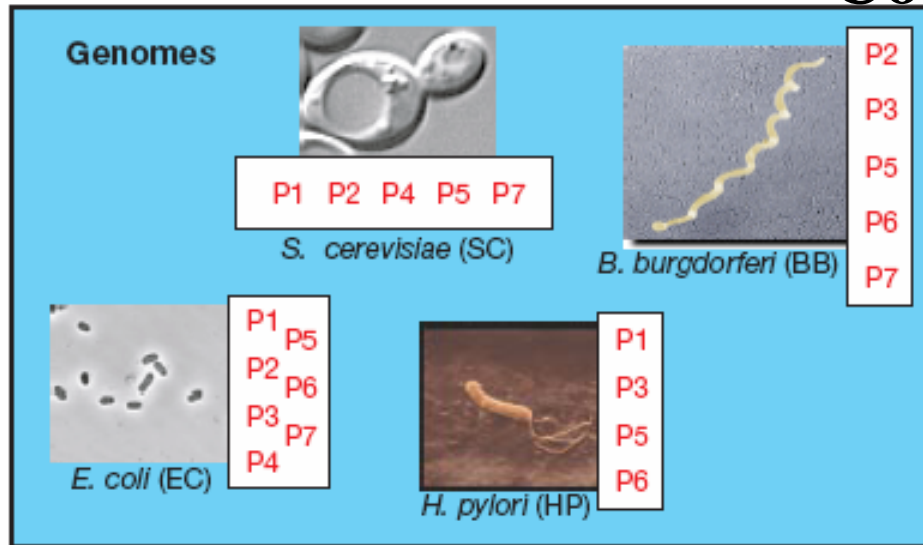
How can comparative genomics aid function prediction?

- Non-sequence-based
 - Co-conservation across genomics
 - Co-localized gene clusters
 - Domain fusion

Co-conservation among genomes

Phylogenetic profiles method

Proteins are considered functionally linked if they share phylogenetic profiles (presence and absence in genomes). Proteins do not have to be homologous by sequence.



Conclusion P2 and P7 are functionally linked, P3 and P6 are functionally linked

Phylogenetic profiles

- Pellegrini et al. compared “phylogenetic profiles” of 4290 *E. coli* proteins against proteins in 16 other fully sequenced genomes
- Using ribosomal protein RL7, can show that proteins with similar profiles are functionally linked. $> \frac{1}{2}$ *E. coli* proteins with phylogenetic profiles similar to the RL7's were ribosomal proteins
- Groups of functionally linked proteins have, on average, more pairs of phylogenetic neighbors than non-functionally-linked proteins

Disadvantages

- Not all functionally linked proteins have similar phylogenetic profiles => false negatives
- Not probabilistic – what does “similar profiles” mean? => false positives
- Optimal “similarity” definition hard to define

Co-localized gene clusters

- Idea: in prokaryotes, groups of functionally related genes are often located close to each other in the genome (exemplified by operons). This functional co-location has been found in some eukaryotes (work & fly, potentially human, but not yeast).

Observed gene locations

Method of correlated gene neighbors



Inferred functional linkage



If two genes (blue and yellow in the figure) are found to be neighbours in several different genomes, a functional linkage may be inferred between the proteins they encode. The method is most robust for microbial genomes but may work to some extent even for human genes where operon-like clusters are observed (see, for example, ref. 26). The gene neighbour method correctly identifies functional links among eight enzymes in the biosynthetic pathway for arginine in *Mycobacterium tuberculosis*.

Gapped alignment method

- Wolf et al. constructed gapped local alignments of conserved gene strings in two genomes
 - Preservation of gene order
 - Mismatches in gene order treated as gaps in scoring
- Gene order is poorly conserved among bacteria & archea:
 - 10% of the genes of each genome belong to conserved gene strings on average
 - Between 5% and 24% (for closely related genomes)
- Thus statistically conserved gene strings can be predicted to form operons

PCBBH & PCH methods

- Overbeek et al. suggested “pairs of close bidirectional best hits” & “pairs of close homologs” methods
- Construct pairs of genes that are closely conserved b/w two species and thus might be functionally related
- Got 343 clusters of “role groups”, paired hundreds of hypothetical proteins with proteins of known function

Disadvantages

- “gene proximity” necessary for functional association hard to define => false positives
- Not all functionally associated genes are close to each other => false negatives
- Application to eukaryotes is limited

Domain fusion

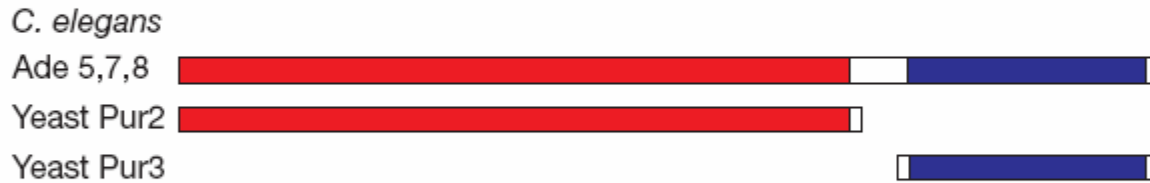
- Certain co-regulated or functionally interacting proteins are fused in another organisms into a single protein (“composite” or “Rosetta Stone” sequence)
- So if a composite protein is similar to two proteins in another species, those two proteins may be interacting and/or functionally related

The Rosetta Stone method

General concept



Top sequence = fused domain that's homologous to two separate seqs from another species



Rosetta stone method results

- Marcotte et al. predict 6809 protein-protein interactions in *E. coli* and 45,502 interactions in yeast
- 68% of pairs shared at least one keyword in their SISS-PROT annotations (vs. 15% of random pairs)
- 5% overlap with pairs identified using phylogenetic profiles method

Disadvantages

- Not all functionally related proteins have been fused in another organism => MANY false negatives
- Potentially not very good for physical interactions, but may have low false positives for functional associations
- Need many diverse related genomes

Comparative genomics for modules discovery

Stuart et al. Science (302) 2003

Why comparative genomics?

- Coregulation doesn't necessarily mean genes are functionally co-regulated
- Can use evolutionary conservation to identify genes that are functionally important from a group of coexpressed genes

Evolutionary conservation

- Coregulation of a pair of functionally related genes may confer selective advantage
- Small & subtle changes in fitness can confer selective advantage during evolution => evolutionary conservation in the wild is a sensitive test for gene function (potentially more sensitive than scoring strong loss-of-function phenotypes in the lab)
- Need to look over large evolutionary distances for this to work

Goal of paper

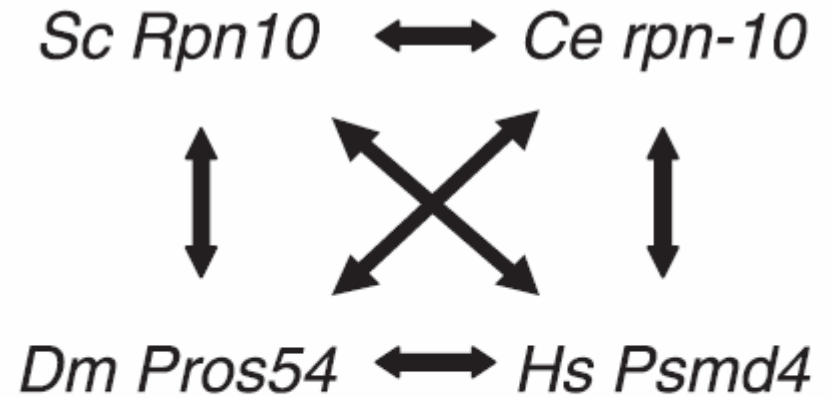
- Analyze coexpression relationships of homologous sets of genes in human, fly, worm, yeast to identify conserved genetic modules
- 3182 microarrays over multiple groups of homologous genes (metagenes)

Metagenes

- Orthologous counterparts in different organisms
- Best reciprocal BLAST hit
- Each gene assigned to at least one metagene

Metagene example: MEG273

- Non-adenosine triphosphatase subunit of the 19S proteasome cap
- Human Psmd4
- Worm rpn-10
- Fly Pros54
- Yeast Rpn10



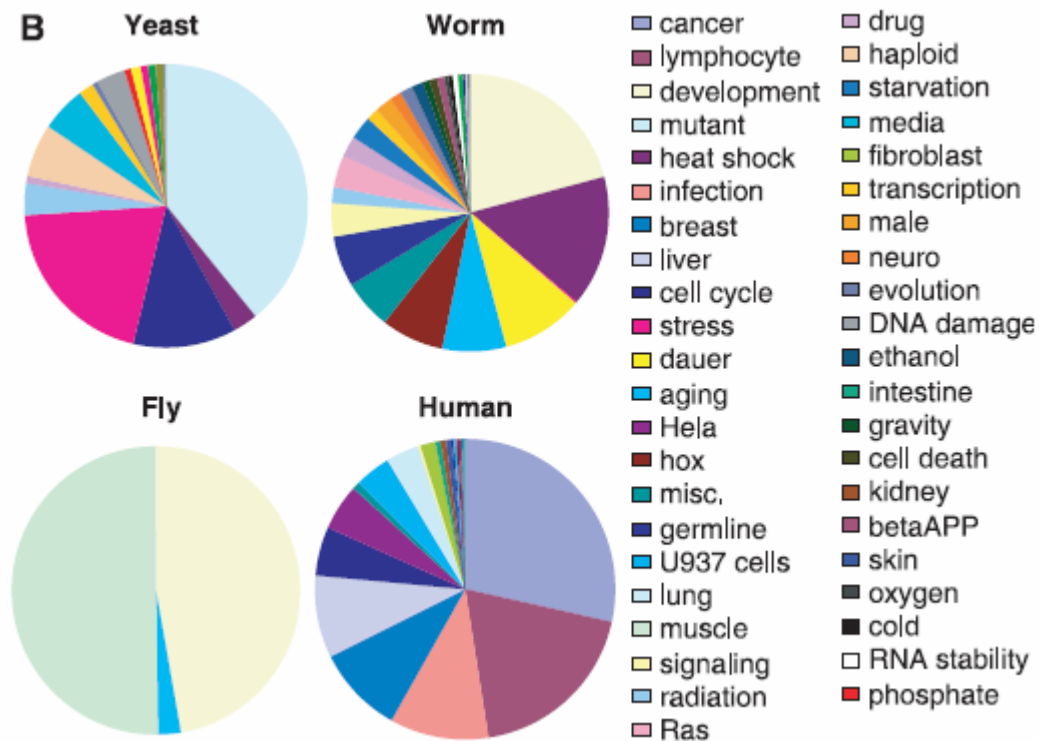
↔ best reciprocal
BLAST hit

Metagenes in this study

- 6307 metages:
 - 6591 human genes
 - 5180 worm genes
 - 5802 fly genes
 - 2434 yeast genes

Coexpression of metages

- Goal: identify sets of metagenes correlated in diverse experiments in multiple organisms
- # of microarrays: 1202 human, 979 worm, 155 fly, 643 yeast



Networks definition

- Network is a combination of links of pairs of co-expressed metagenes
- A link:
 - Pearson correlation b/w metagenes expression profiles
 - Use order statistics to evaluate p-value of observing particular configuration of ranks across different organisms by chance

Evaluating the network

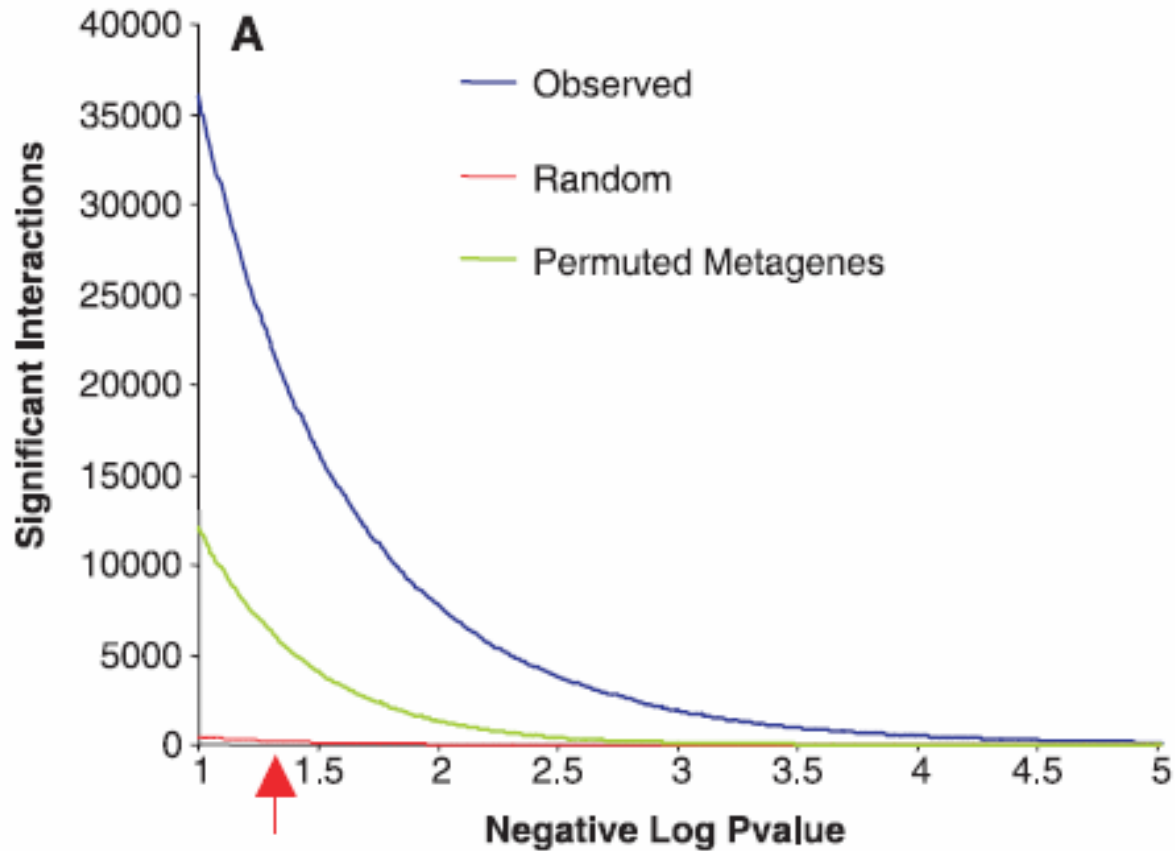
- How does the number of connections compare to a random network?
- Do available microarray experiments represent a good coverage of potential interactions?
- How stable are these conclusions with respect to noise?

Comparison to random network

- Expect only 236 interactions by chance at $p > .05$, but get 22,163 observed interactions
- However, these nonrandom interactions could still be due to the correlation structure inherent in the data => if the set of metagenes exhibit only a few simple types of expression patterns, you are likely to find random coexpression relationships anyway
- What to do to check???

Generate permuted metagenes – a random collection of genes from each organism, and construct a network out of those genes

Comparison to random network



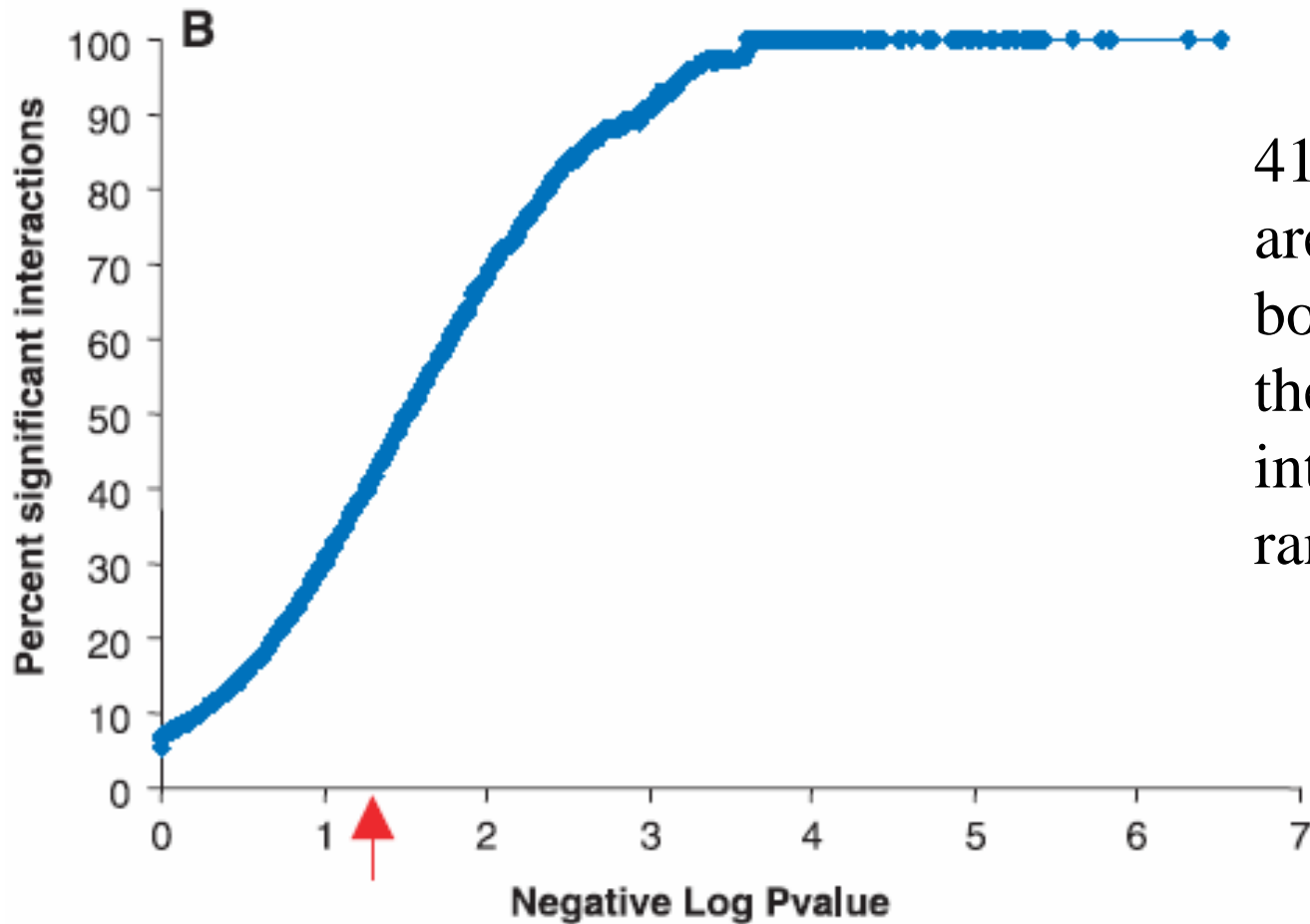
Coverage of the network

- Do microarray experiments cover a reasonable fraction of potential gene interactions?
- How would you check?

If GE data represents only a small fraction of possible gene interactions, particular set of interactions in each organism will be heavily dependent on what subset of microarray experiments was used for analysis.

So should show that a significant fraction of the gene-expression links is present in networks built by only a random half of the data

Coverage



41% of interactions are significant in both networks when the whole net is split into two halves randomly

Stability with respect to noise

- Add increasing levels of Gaussian noise to expression data, and construct networks
- Compare perturbed nets to original one

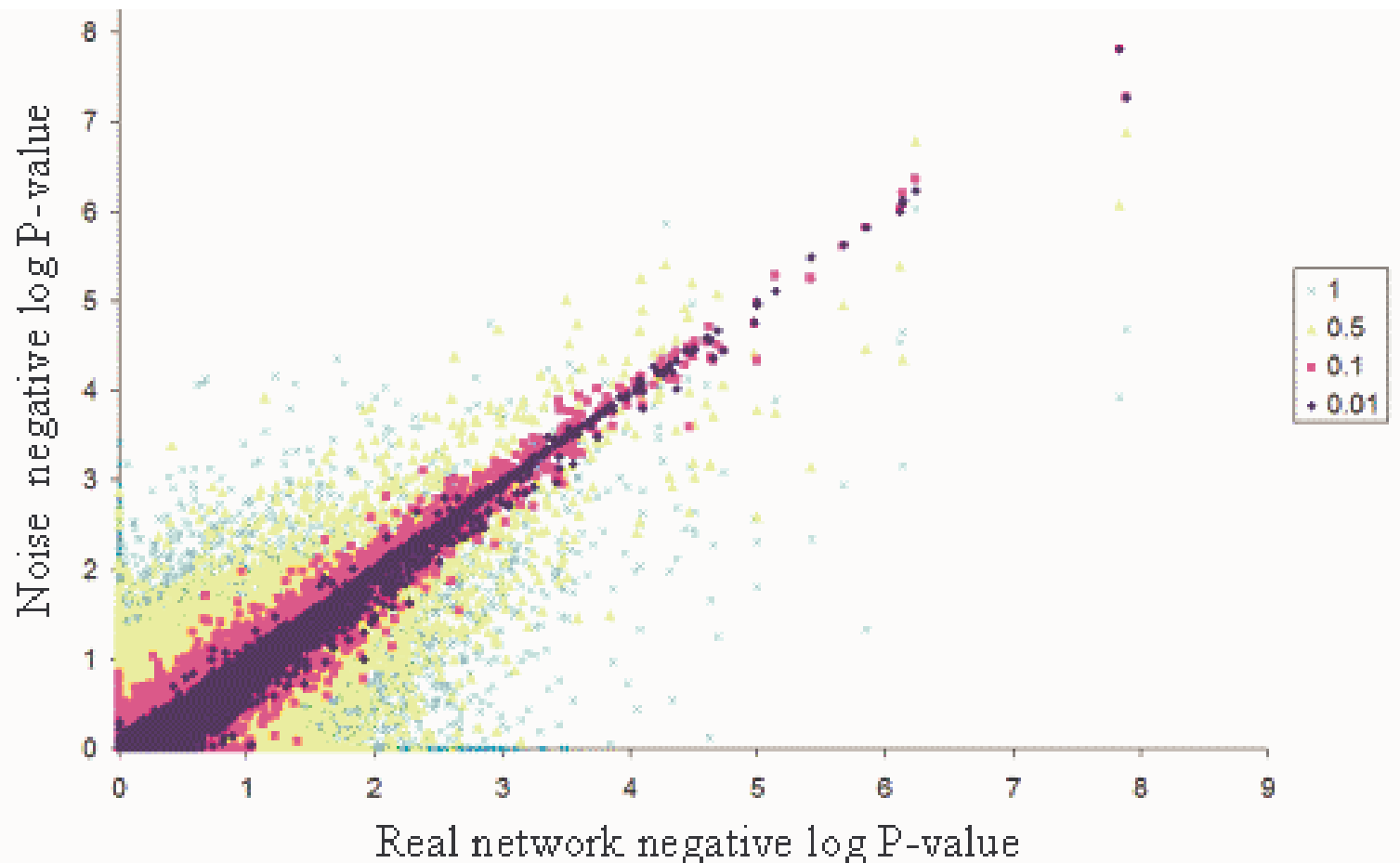
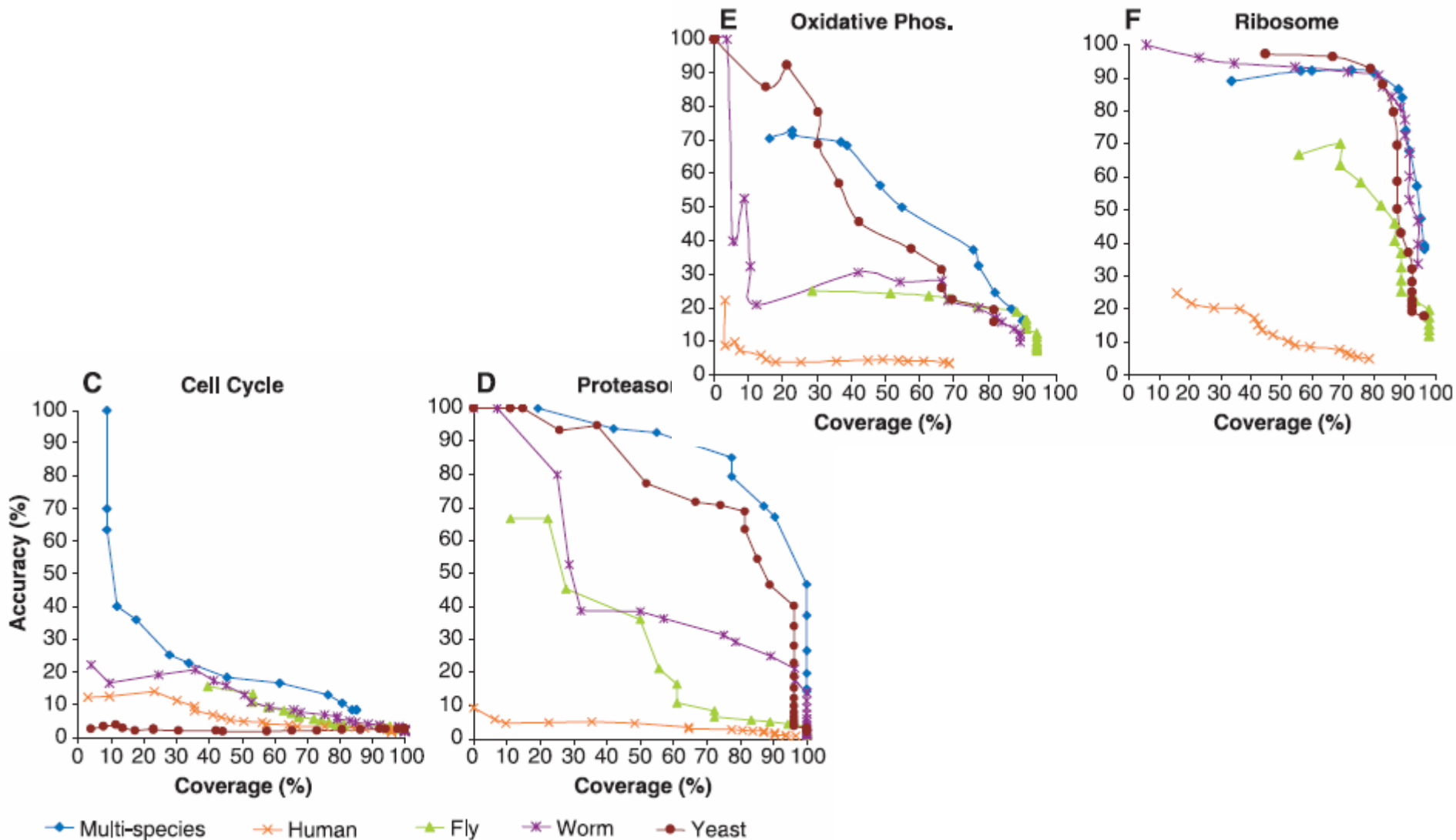


Figure S1. Rubustness to noise analysis. We added increasing levels of Gaussian noise to each organisms' dataset with 0.01σ (blue circles), 0.1σ (pink squares), 0.5σ (yellow triangles), and 1.0σ (light blue crosses). Shown is the negative log P-value of an interaction in the original network (x-axis) plotted against the interaction's P-value in the network constructed from the noise-added data (y-axis).

Multi-species vs. single species coexpression networks

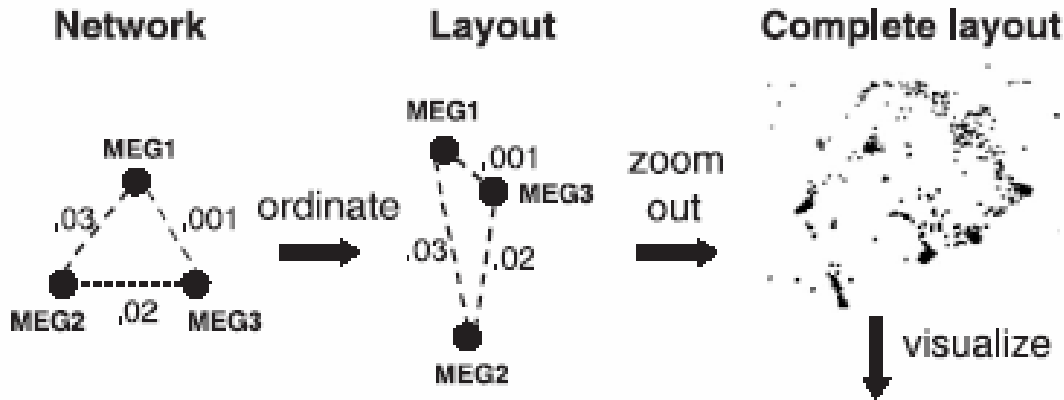
- Multispecies only maps genes that have orthologs => focus on core, conserved biological processes
- Interactions imply evolutionary conservation vs. just correlated expression in single-species

Multi-species vs. single species

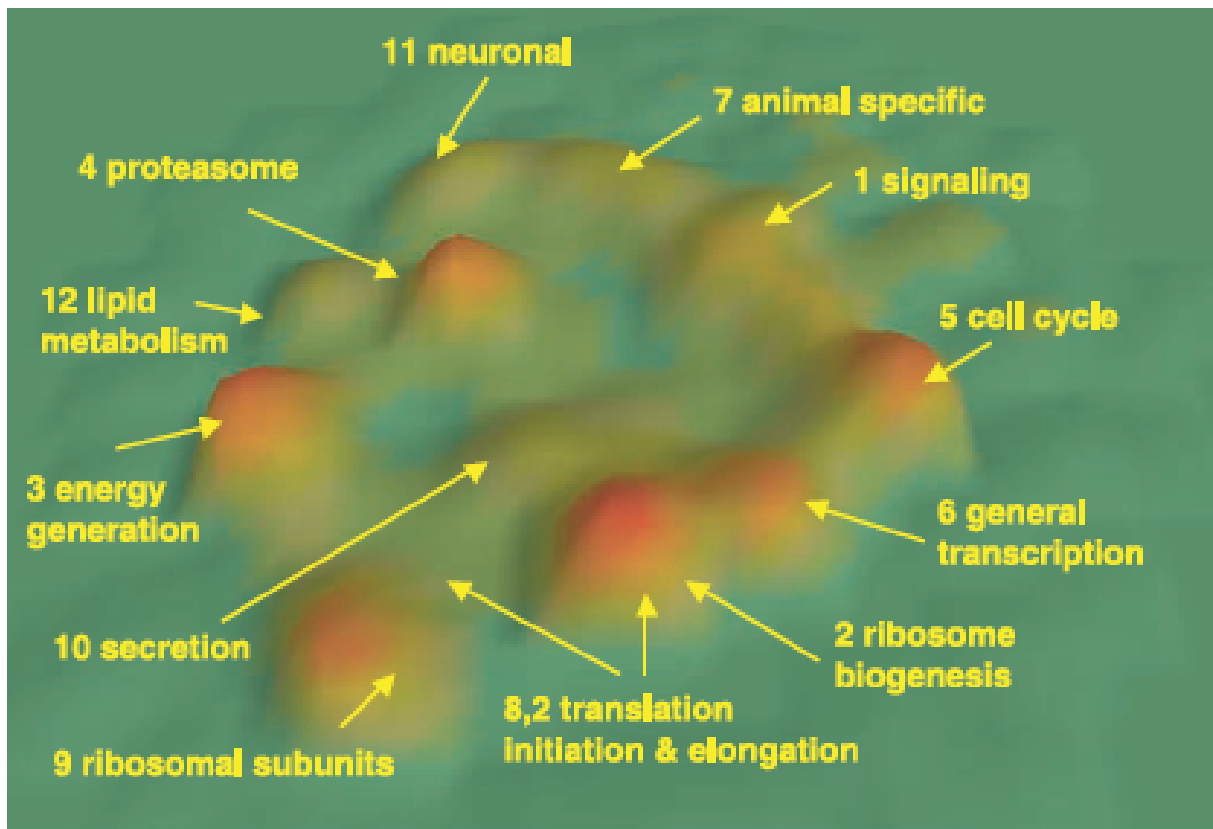


Multi-species vs. single species – is it just more data?

- For many processes, multi-species performs better than any single-species, for others it does as well as the best single species and significantly better than others
- Networks build from fewer multi-species arrays (979, same as all worm arrays) performs almost as well as the complete network



Visualization of the network with VxInsight



- Genes arranged on x-y plane by p-value
- k-means clustering => 12 clusters

Table 1. Network components.

Component	Size*	Biological function†	Genes in component‡	Enrichment; <i>P</i> 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	14/31	9.9; 10^{-11}
		RNA biosynthesis		
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}

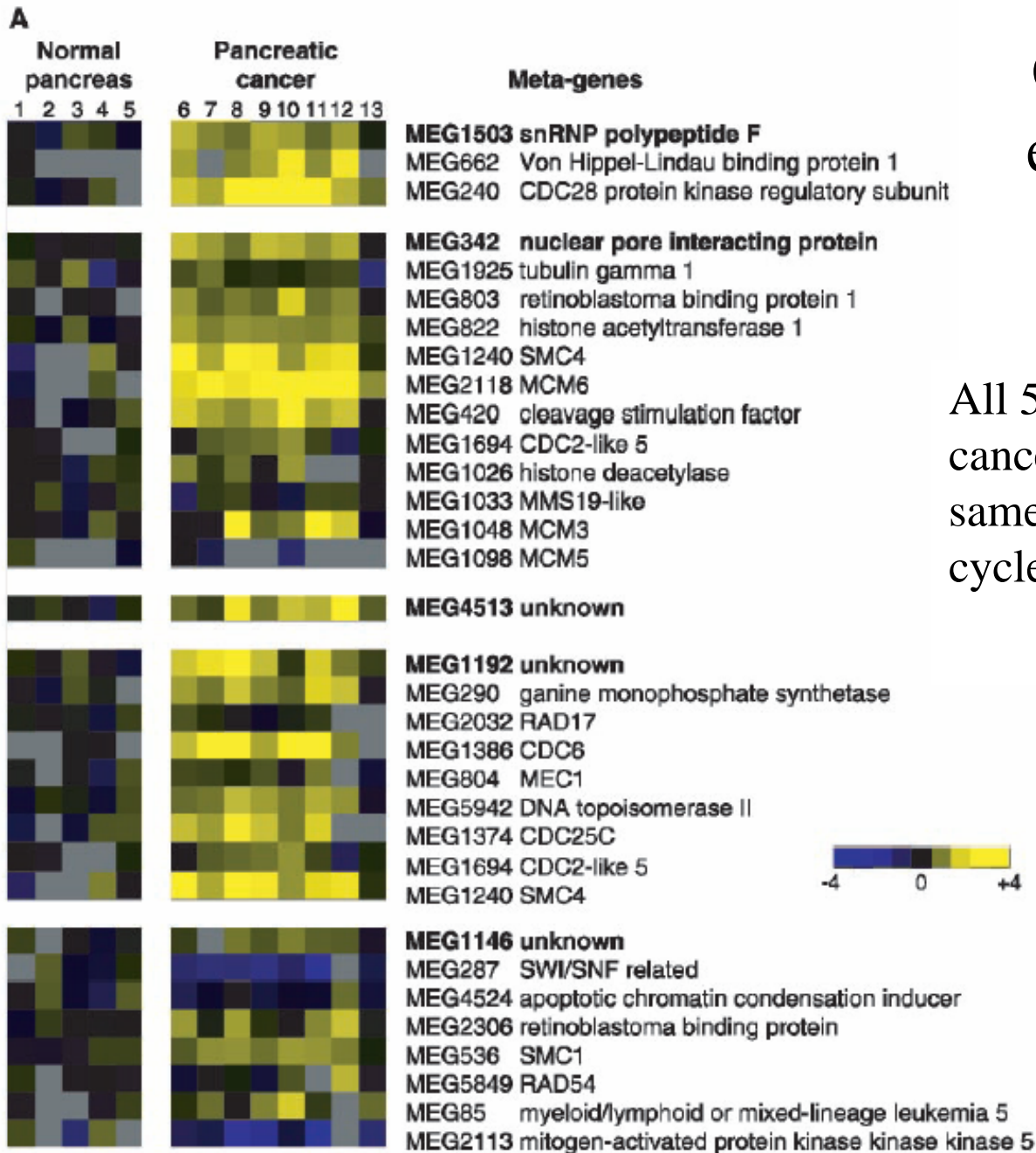
*The total number of metagenes in the component. †Biological functions were based on edited terms from Gene Ontology (15) and the KEGG database (22). ‡The number of metagenes in the biological function group and in the component divided by the total number of metagenes in the biological function group that were also in the network. §The ratio between the number of observed metagenes in a category and the number expected by chance. The *P* value was computed as the probability of obtaining the observed number of overlaps by chance under a hypergeometric distribution.

Clusters eg component 5

- 241 metagenes
- 110 previously known to be cell cycle (p<10⁻⁸⁵ by hypergeometric)
- 30 cell cycle regulation
- 80 terminal cell cycle functions
- **131 genes predicted to be cell cycle**

Function prediction based on comparative genomics

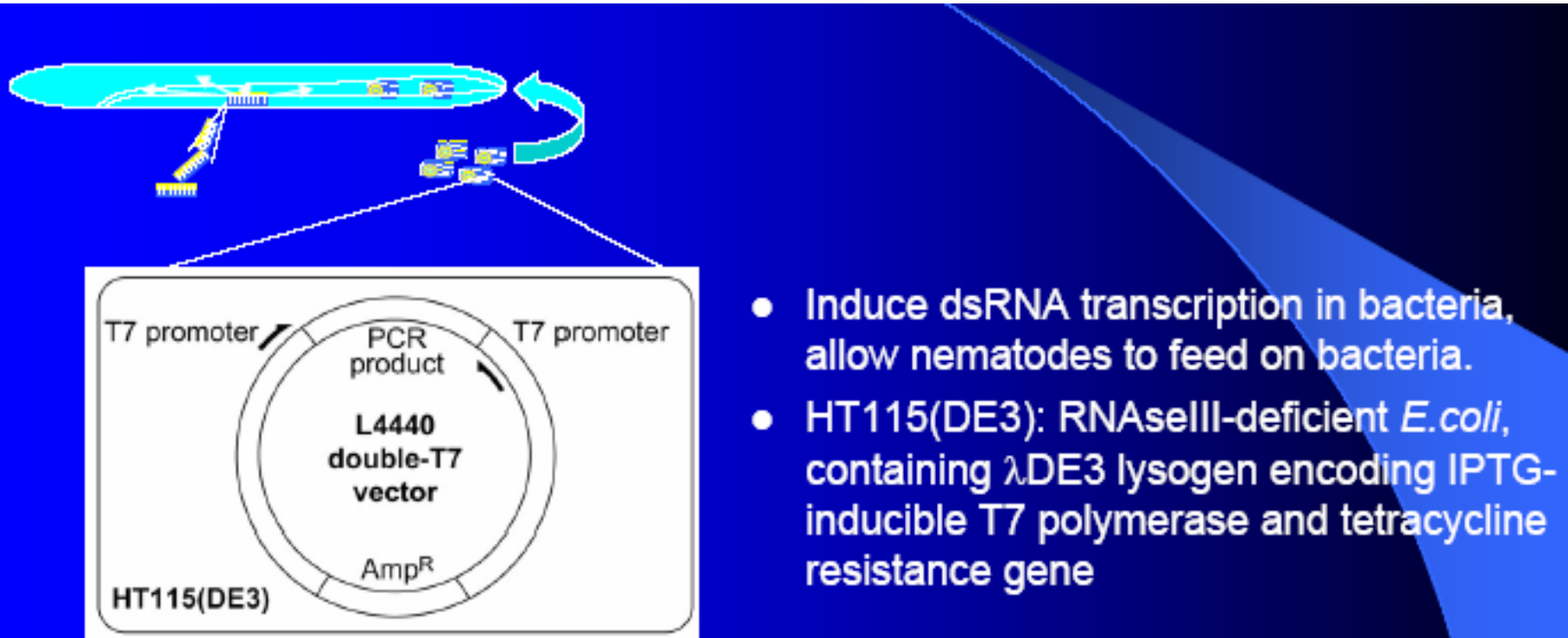
- Unknown genes in functionally enriched clusters => function prediction
- Experimentally & computationally validated 5 metagenes predicted to be involved in proliferation & cell cycle based on statistically enriched functional groups



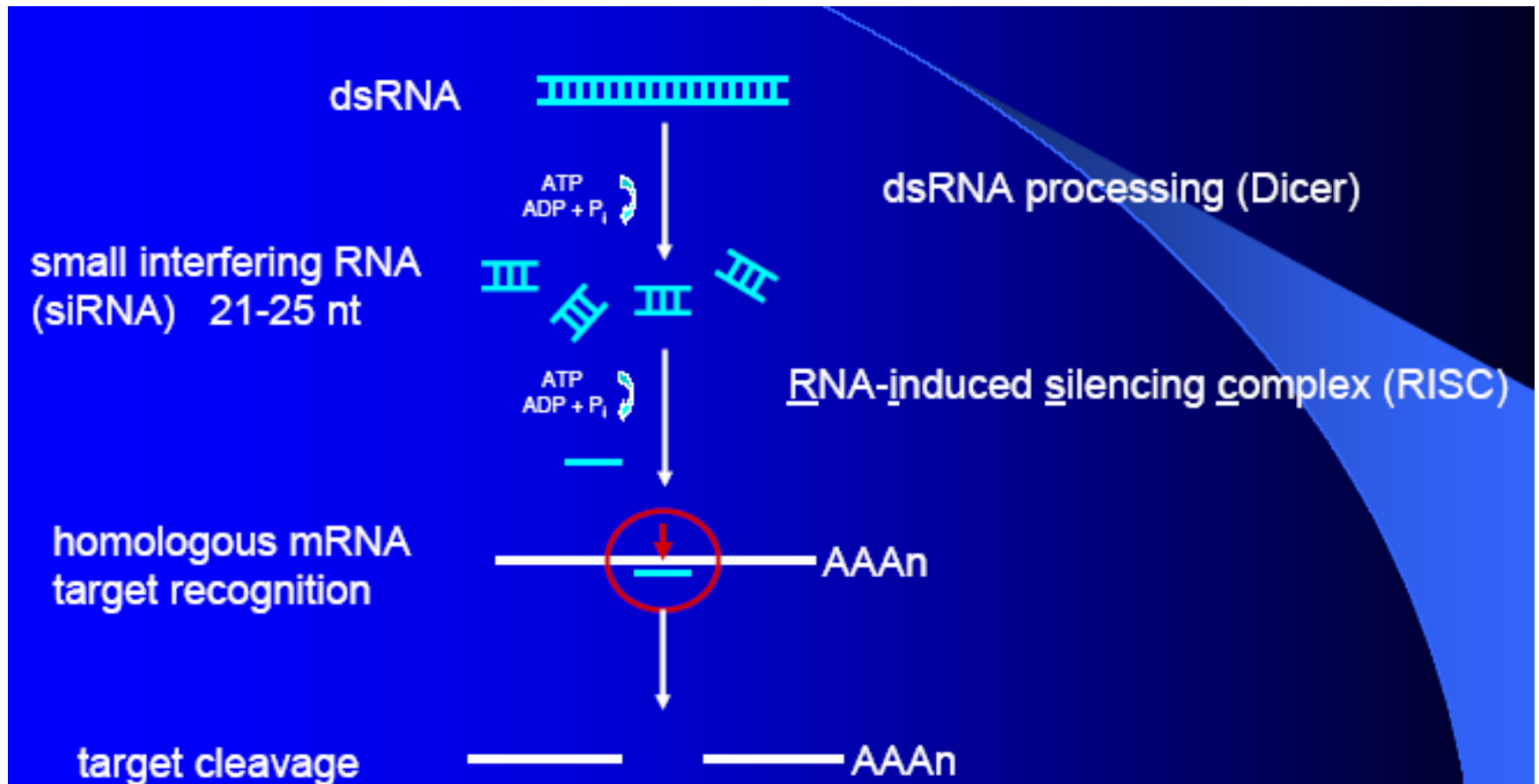
Computational evaluation – an independent dataset

All 5 overexpressed in pancreatic cancer (indep. Dataset) to the same extent as known cell cycle/proliferation genes

Experimental evaluation –RNAi in worms

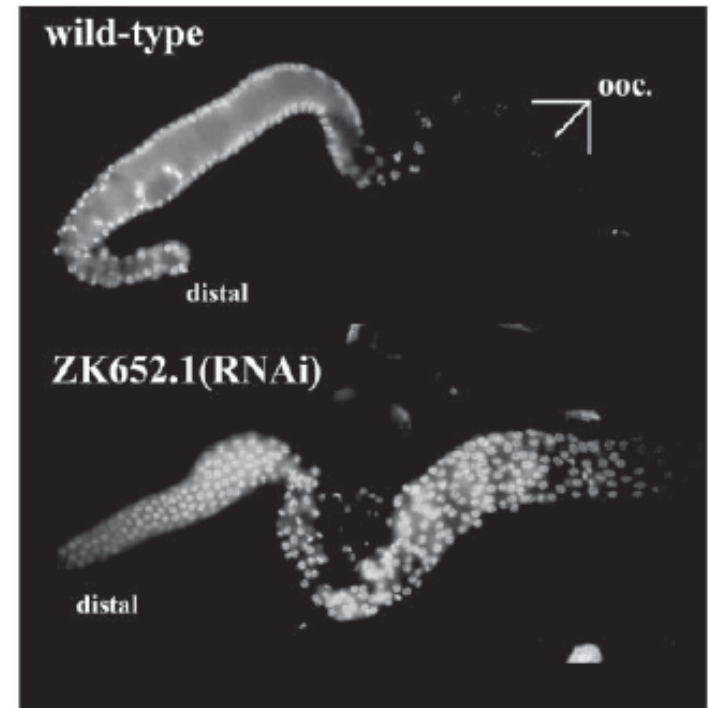


Experimental evaluation- RNAi



RNAi for ZK652

- Gonads stained for DNA, wt worms have less nuclei => wt function of this gene is to suppress germline proliferation



Genetic modules based on conservation – 3 types

- Ancient dedicated modules
- Evolving modules
- Modules with interchangeable parts

Ancient modules

- Responsible for main/core cellular function
- Conserved over long evolutionary distances (yeast to humans)
- Metagenes expected to have highly conserved coding regions
- Metagenes should have conserved gene expression links
- EG: metagenes involved in ribosomal function

Evolving modules

- Show rapid change among species
- Metagenes lack a yeast ortholog
- metagenes show large changes in expression links between invertebrates and human
- EG: neuronal modules

Modules with interchangeable parts

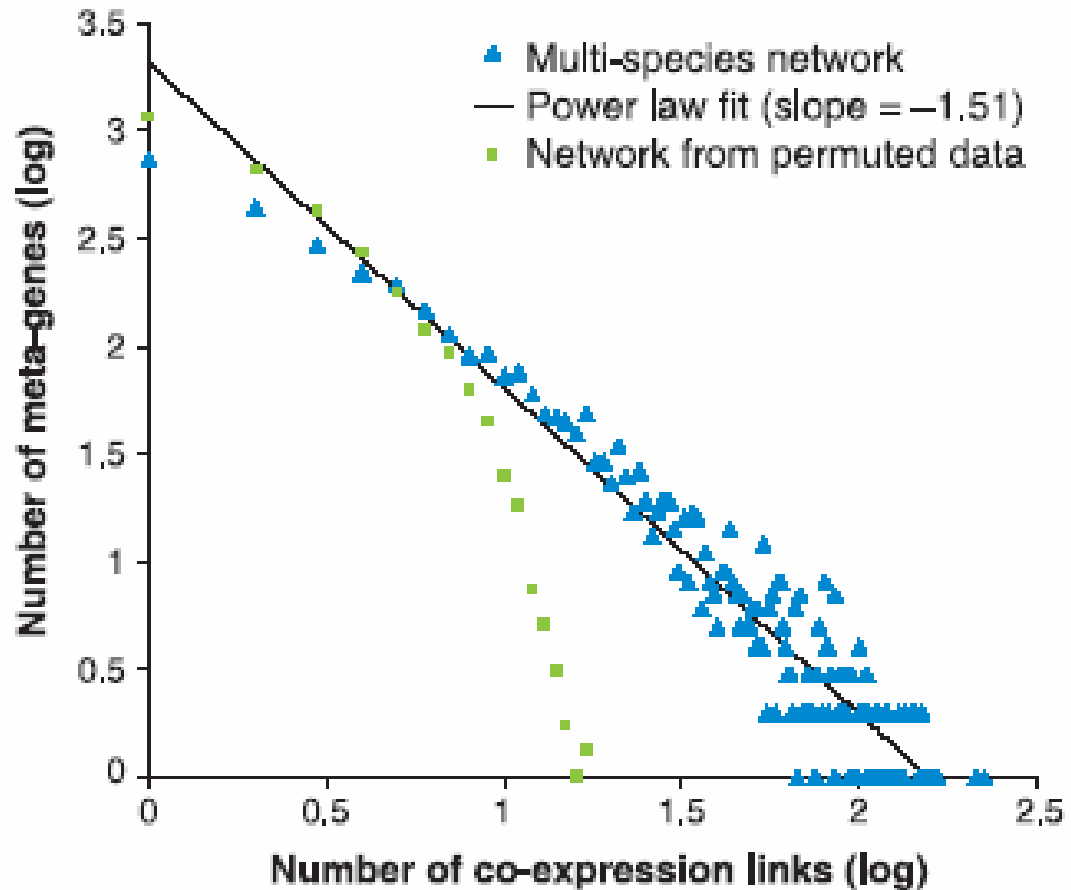
- Composed of metagenes that have different links in different species
- Sequence of metagenes conserved
- EG: *sir-2* –
 - protein involved in regulating chromatin structure and gene expression
 - highly conserved from yeast to human
 - but has different downstream targets in different species
- Also transcription factors, signaling molecules

Connectivity properties of genetic networks

- Count number of neighbors of each metagene, and compare to networks constructed from permuted data
- Highly nonrandom distribution – significantly more metagenes with a larger number of gene expression links than the random networks

Power law again

Fig. 6. Distribution of the number of links for each metagene. Shown is the number of links (x axis, \log_{10} scale) compared with the number of meta-genes that have that number of links (y axis, \log_{10} scale) in the network (blue triangles) and in the networks constructed from permuted data (green squares) (14). The black line (slope of -1.51) depicts the least-squares fit of the data to a linear line in the log-log plot.



Suggests existence of selective force in the overall design of genetic pathways to maintain a highly connected class of genes

Putting it all together – what does
it take to predict a genetic
network

Computational analysis of genetic networks

Q1: What does a gene do?

function prediction – sequence similarity, coregulation, interactions

Q2: How is it controlled?

promoter identification

Q3: What other genes is it co-regulated with?

coregulation? genome position similarity?
evolutionary links?

Q2: What other genes does it interact with?

Q3: What other genes does it control?