Comparative Systems Biology(CSB)

What is CSB?

- Investigate the similarity & difference
- among genes, proteins, genomes, proteomes, metabolomes, organisms, etc.
- @ systems level.

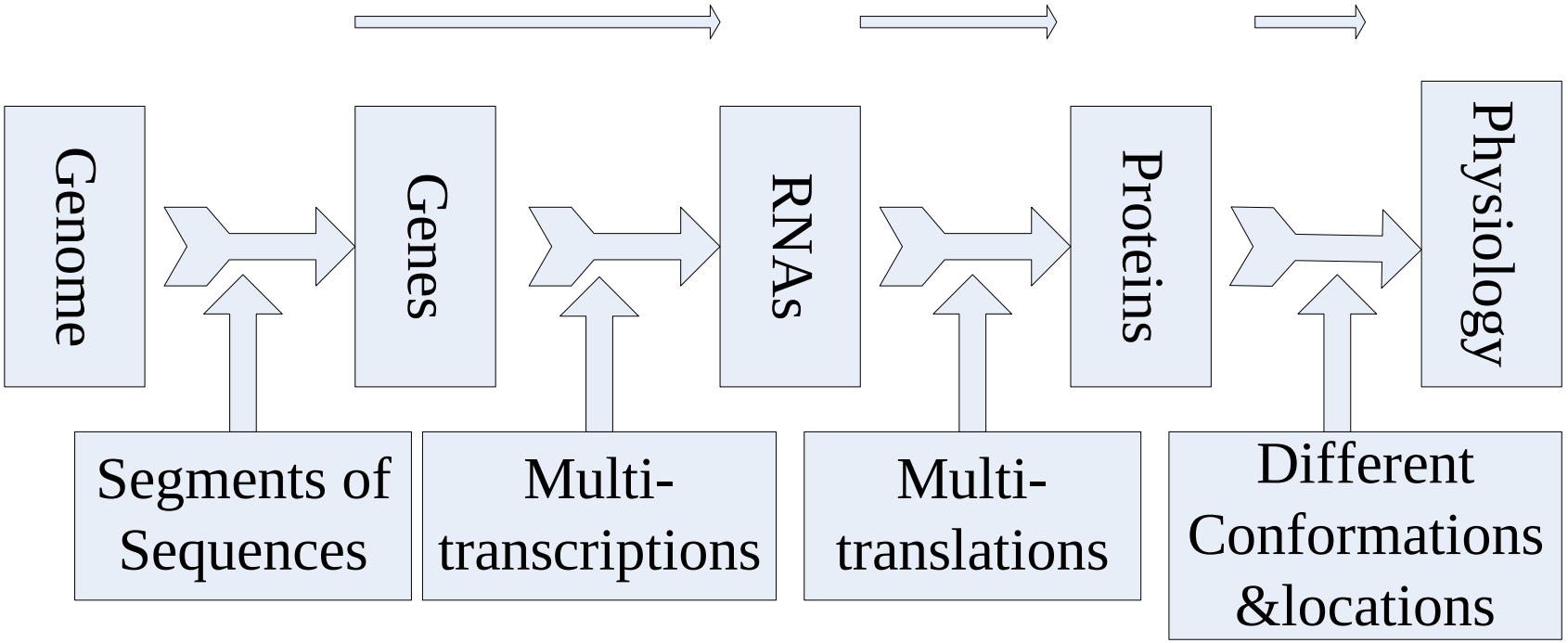
Why we can compare species @ systems level?

- Two main 'laws'-assumptions
- ✓Evolution theory
- ✓One gene to one protein and then one function (the central Dogma)

Evolution Theory

- ✓ Darwin's natural selection (by stochastic mutation)
- ✓ Kimura's Neutral theory of molecular evolution (molecular clock)
- All the species from the same origin
- So, we can find somewhat similar groups among the genes, proteins, genomes, proteomes,etc

The Central Dogma



What can we learn from that?

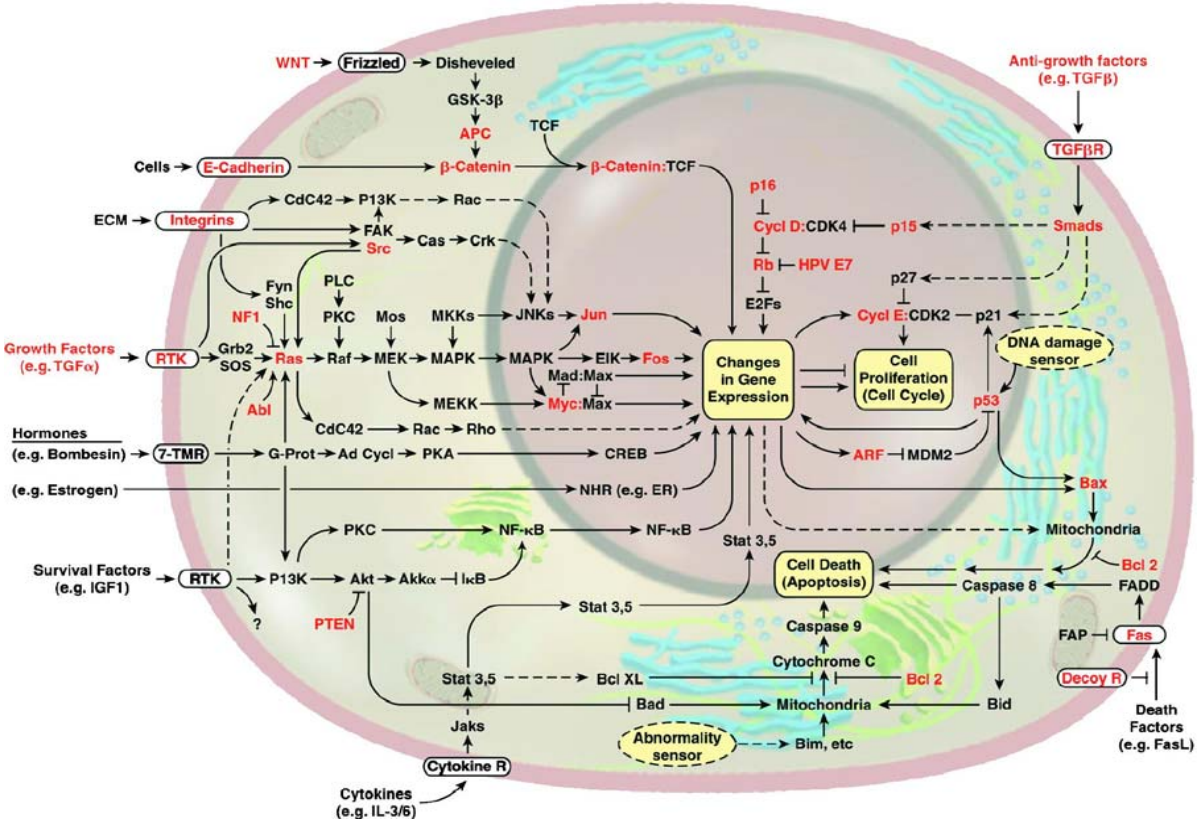
The Central Dogma

- Information Flow

one gene-one protein-one function

- Control Flow (Temporal and Spatial)
- Everything related to genes follows the same pattern above, so, it's comparable

An example of biological network



It exist in mammalian cells, it should have a similar one existed in other cells in different organisms

Versatile Biological Databases

- Each layer and each process is classified into different databases.

http://www.infobiogen.fr/services/dbcat/

http://www3.oup.co.uk/nar/database/c/

- The data in most of database is not complete or only partially known
- All the databases not in fully connected by proper time and space
- So, we can dig out sth among them.

Gene Ontology

- 'The Gene Ontology (GO) project is a collaborative effort to address the need for consistent descriptions of gene products in different databases.'

http://www.geneontology.org/

- It divides all the genes terms (annotations) into three parts: biological processes, cellular components and molecular functions and subdivide into tree-structure

Gene Ontology(cont')

- Each gene product(and its term) has a unique identifier
- The GO database cross-link to many different databases, which provides a uniform querying system

Correlated expression patterns – Similar Functions

- Genes that encode proteins that participate in the same pathway or are part of the same protein complex are often coregulated.

- Clusters of genes with related functions often exhibit expression patterns that are correlated under a large number of diverse conditions in DNA microarray experiments.

Evolutionary conservation

- Evolutionary conservation is a powerful criterion to identify genes that are functionally important from a set of coregulated genes.

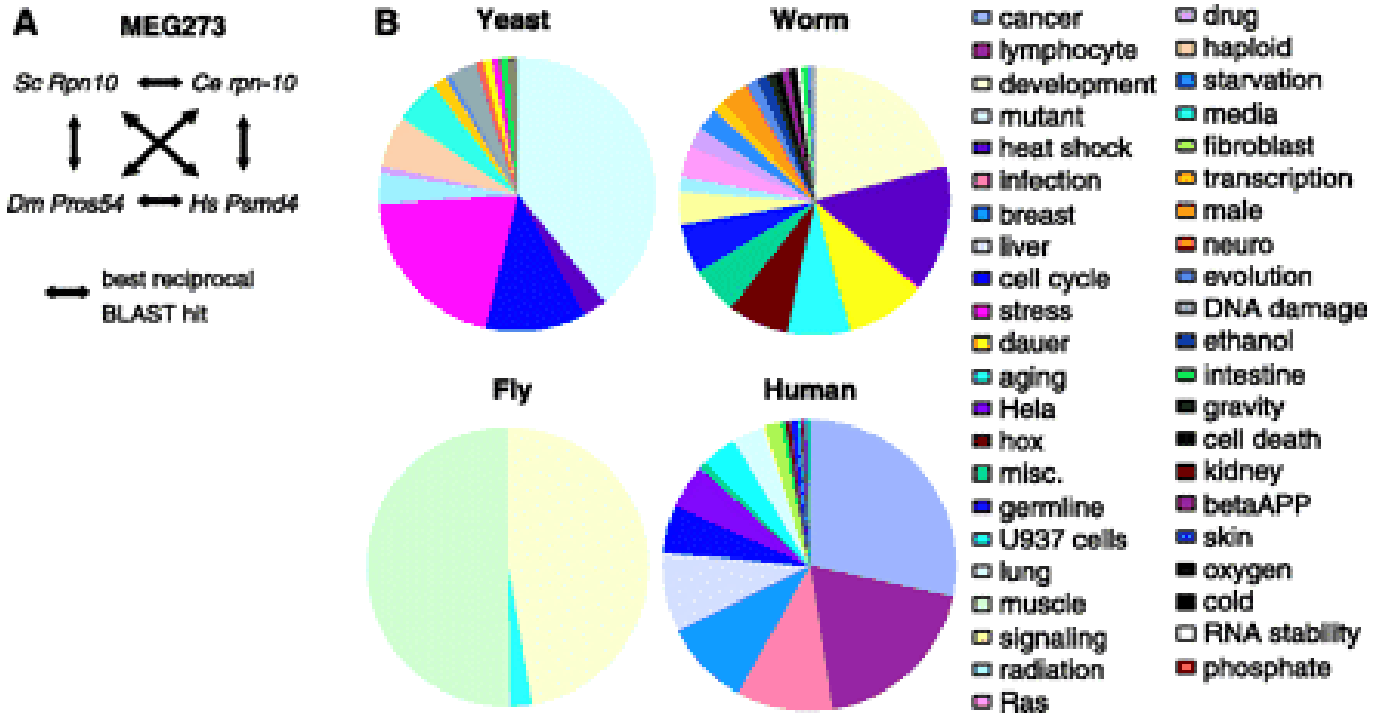
- Coregulation of a pair of genes over large evolutionary distances implies that the coregulation confers a selective advantage, most likely because the genes are functionally related.

Metagene

- **Metagene** as a set of genes across multiple organisms whose protein sequences are one another's best reciprocal BLAST hit

- For example, metagene MEG273 refers to the human gene *Psmd4*, the *C. elegans* gene *rpn-10*, the *D. melanogaster* gene *Pros54*, and the *S. cerevisiae* gene *Rpn10*, all of which encode a non–adenosine triphosphatase subunit of the 19S proteasome cap

Data(1202 DNA microarrays from humans, 979 from worms, 155 from flies, and 643 from yeast)



Building the gene-coexpression network

- Pairs of genes whose expression is significantly correlated in multiple organisms.
- Calculate Pearson correlation of the expression profiles between every pair of genes in the microarray data sets for each organism.
- Rank genes according to their Pearson correlations.

Building the gene-coexpression network..

- Probability method based on order statistics
- Probability of observing a particular configuration of ranks across the different organisms by chance.
- $P<0.05$ is the cutoff to indicate that two metagenes are co-expressed.
- Combined each such link to form a interaction network.

Results & Verification

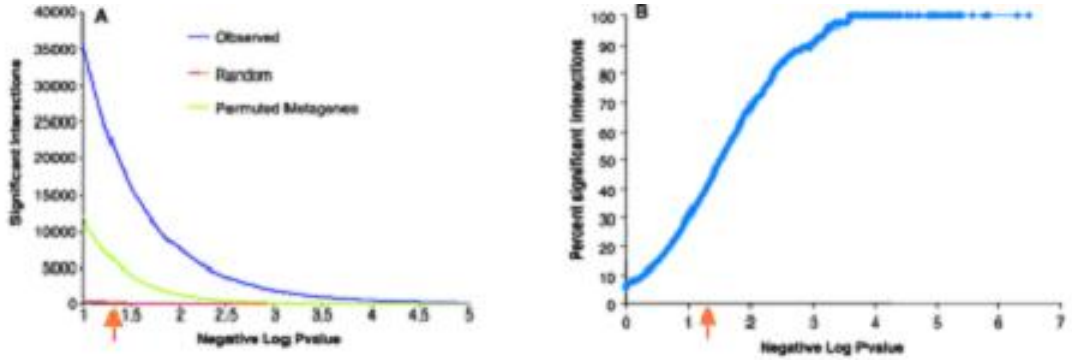
- 3416 Metagenes connected by 22,163 interactions.

- Lot more interactions were observed than the ‘chance’ interactions(236) estimated by the statistical model.

Verification..

- Random pairs of metagenes could have significant co-expression interactions too..
- Metagenes(containing random collection of genes from each organism)
- Built a network and studied the number of siginificant interactions
- Real networks have 3.5 times more interactions than random networks

Verification..



Robustness

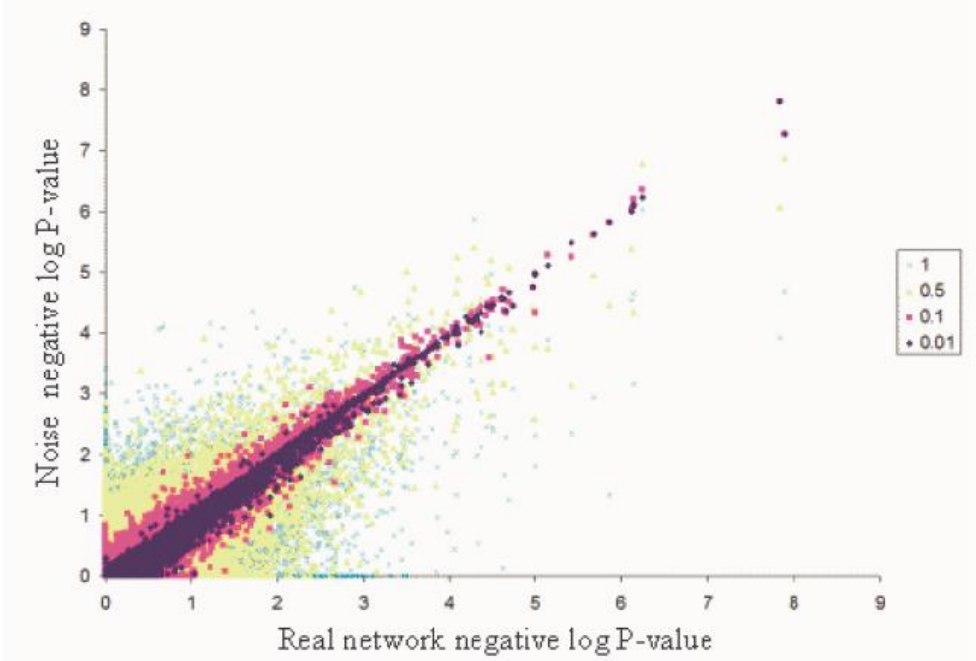
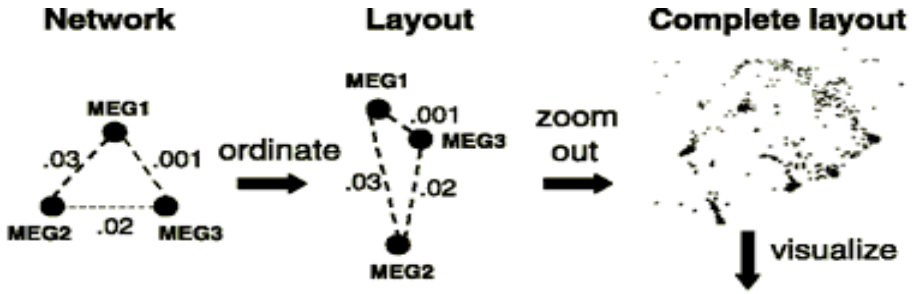
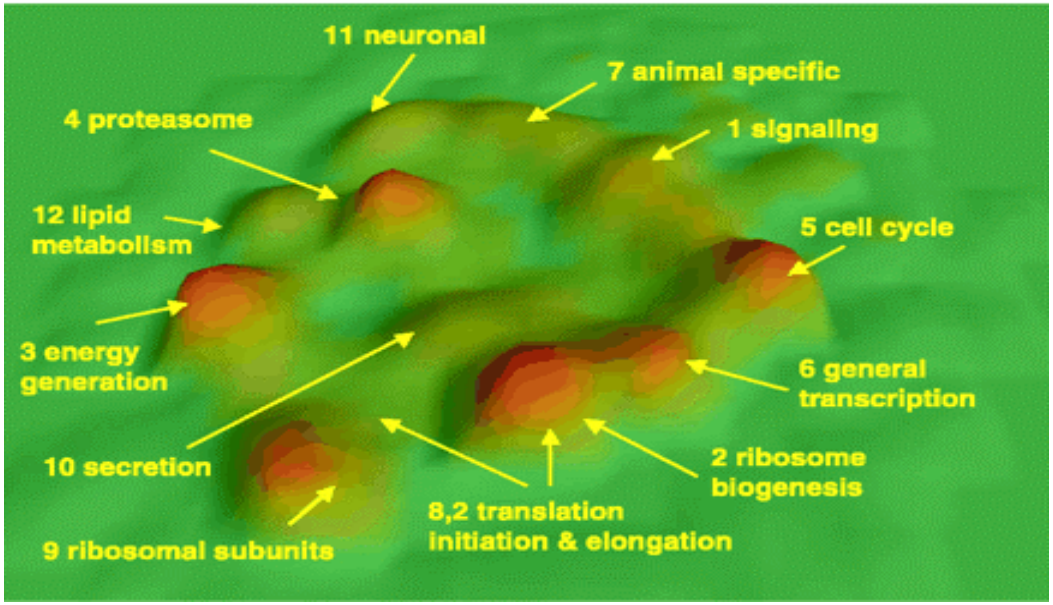
- 

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).

Visualization





Components

- K-means clustering on the x-y co-ordinates
- 12 regions of highly interconnected metagenes (Components)
- Found that each component was enriched with genes involved in similar biological processes

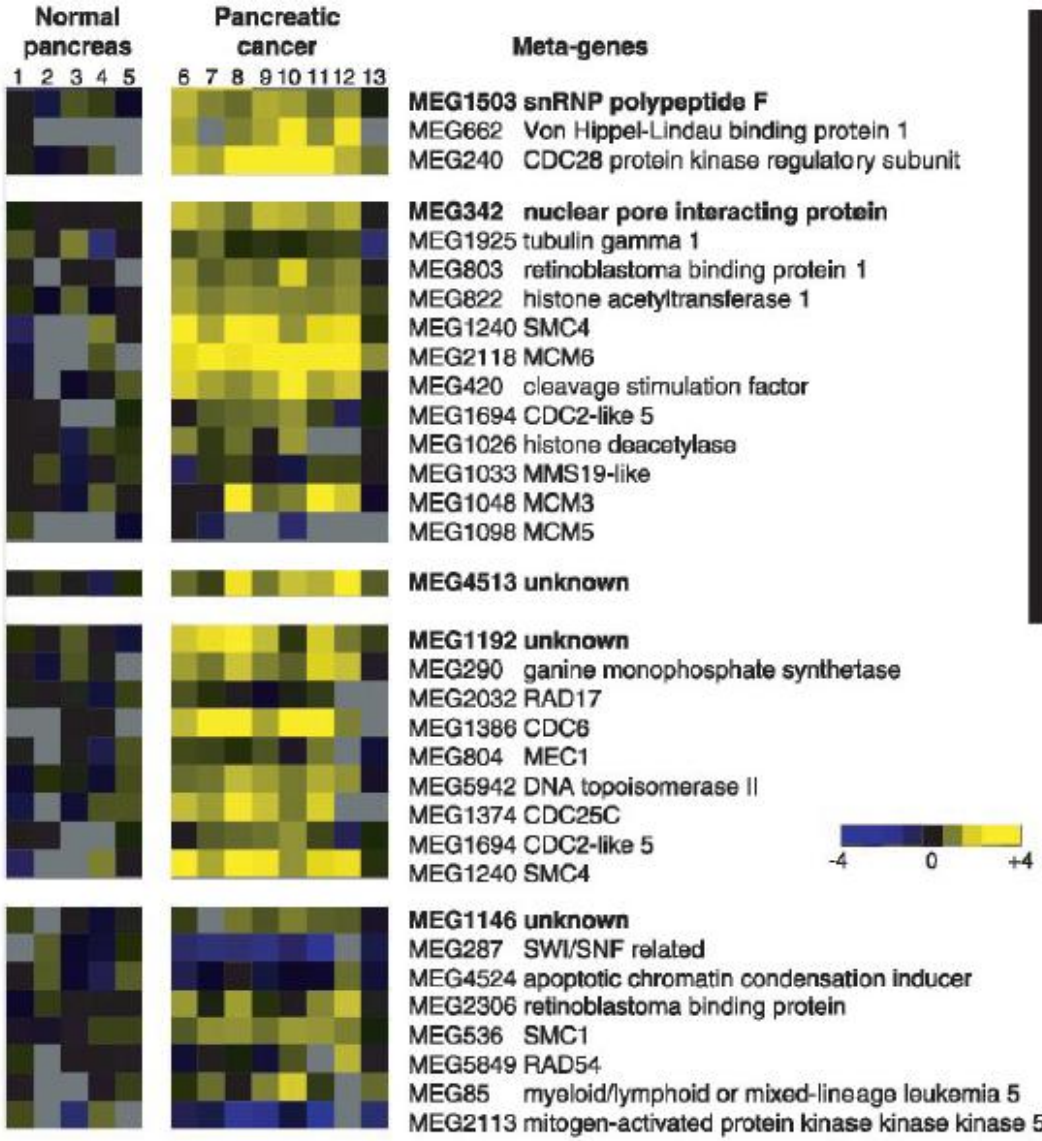
Analysis

- Component 5 – found be strongly enriched with cell cycle metagenes.

- Of 241 metagenes in it, 110 were known to be in cell cycle. The rest 131 could be hypothesized to belong to cell cycle.

Validation

- Meg1503(splicing), Meg342(nucleoporin interacting component),Meg 4513, Meg1192, Meg1146(unknown functions) showed a significant number of links to the cell proliferation metagenes.
- Are these related to cell proliferation ??



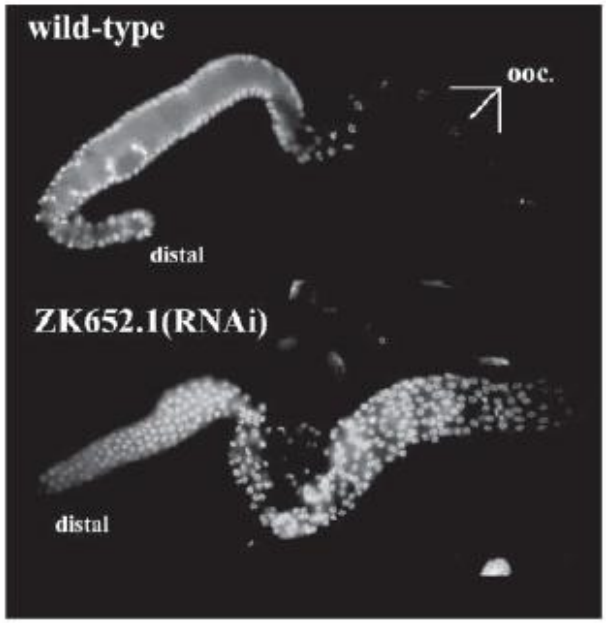
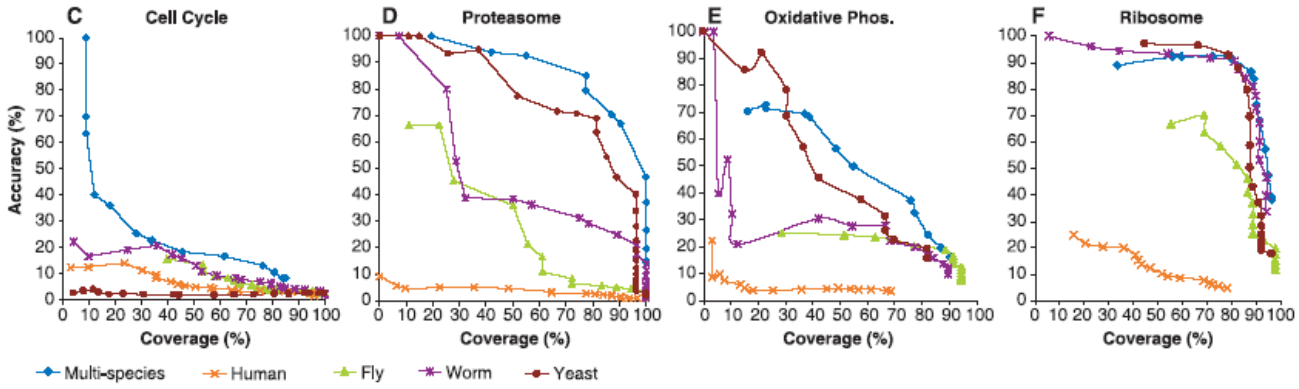
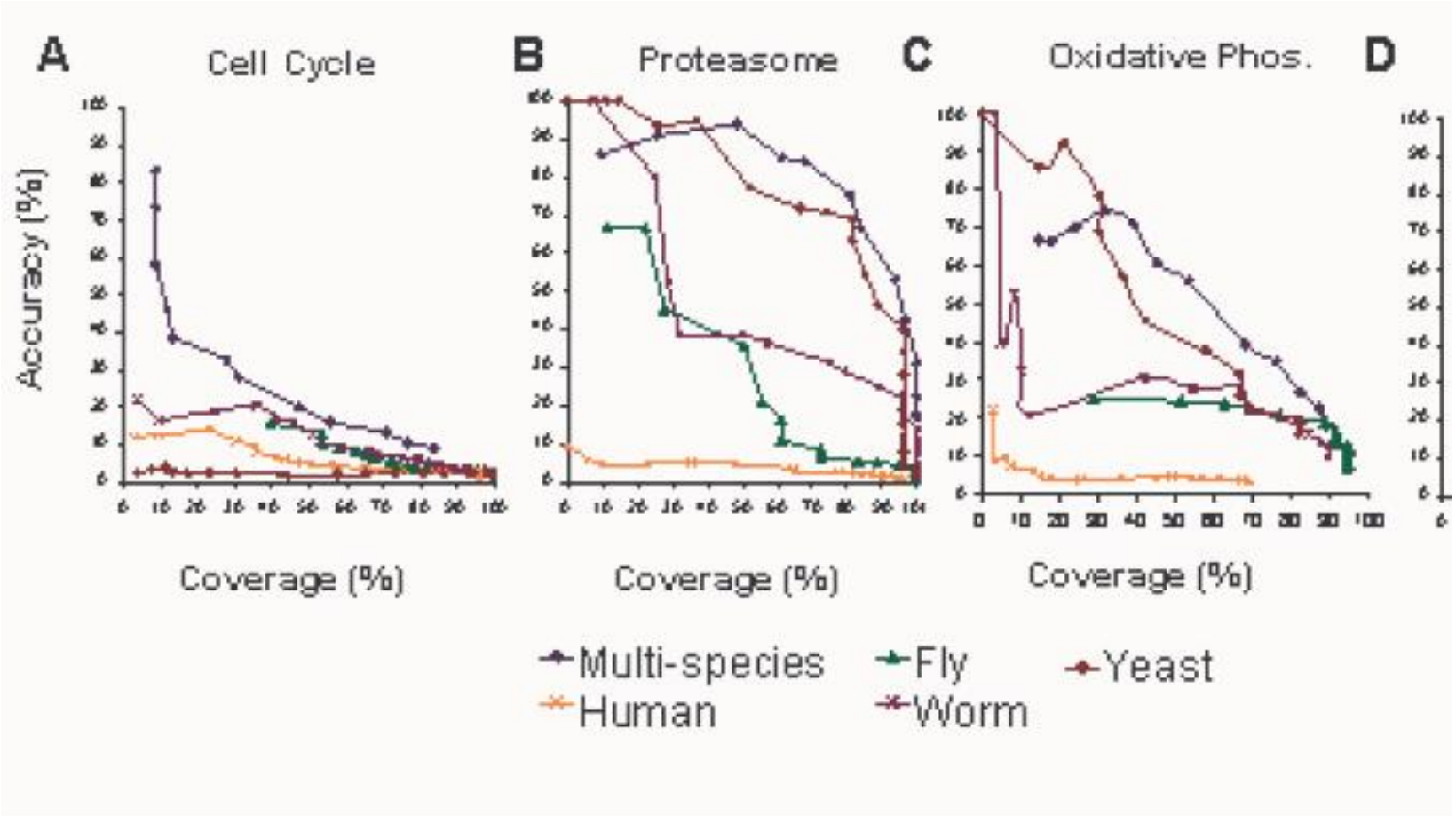


Fig. 4. (A) MEG1503, MEG342, MEG4513, MEG1192, and MEG1146 are overexpressed in pancreatic cancers. We plotted the metagenes with the GeneXPress program (http://genexpress.stanford.edu) using data from (*21*). The first five columns correspond to expression data obtained from normal pancreas specimens (pSF2779N, pSF442N, pSF4N, pSF5NT, and pSF768NT), and the remaining eight columns correspond to expression data obtained from pancreatic cancer specimens [a pancreatic cancer cell line (HS766T), five Hopkins/Goggins pancreatic cancer cell cultures (PL2, PL22, PL21, PL1, and PL8), a poorly differentiated pancreas carcinoma (pSF439T), and a pancreas foamy cell adenocarcinoma specimen (pSF1T)]. Each row corresponds to the expression profile of a single metagene across the 13 pancreatic samples. Bold indicates metagenes with unknown functions that are implicated in cell proliferation by the network. Neighbors of each implicated metagene that were previously known to be involved in cell proliferation or cell cycle are also shown. Scale shows $\log_2$ expression ratio. (B) RNAi-induced phenotype of *ZK652.1*. Shown are wild-type gonads and gonads from worms that were fed bacteria producing *ZK652.1* double-stranded RNA for 2 days (*29*). Gonads were stained with 4′,6′-diamidino-2-phenylindole to show DNA in the nuclei (*30*). *ZK652.1* (RNAi) gonads have more nuclei than the wild type and lack oocytes (ooc.). *rrf-3(pk1426)* worms were used because they are more sensitive to RNAi (*31*).

Quality Evaluation(Multi vs single)

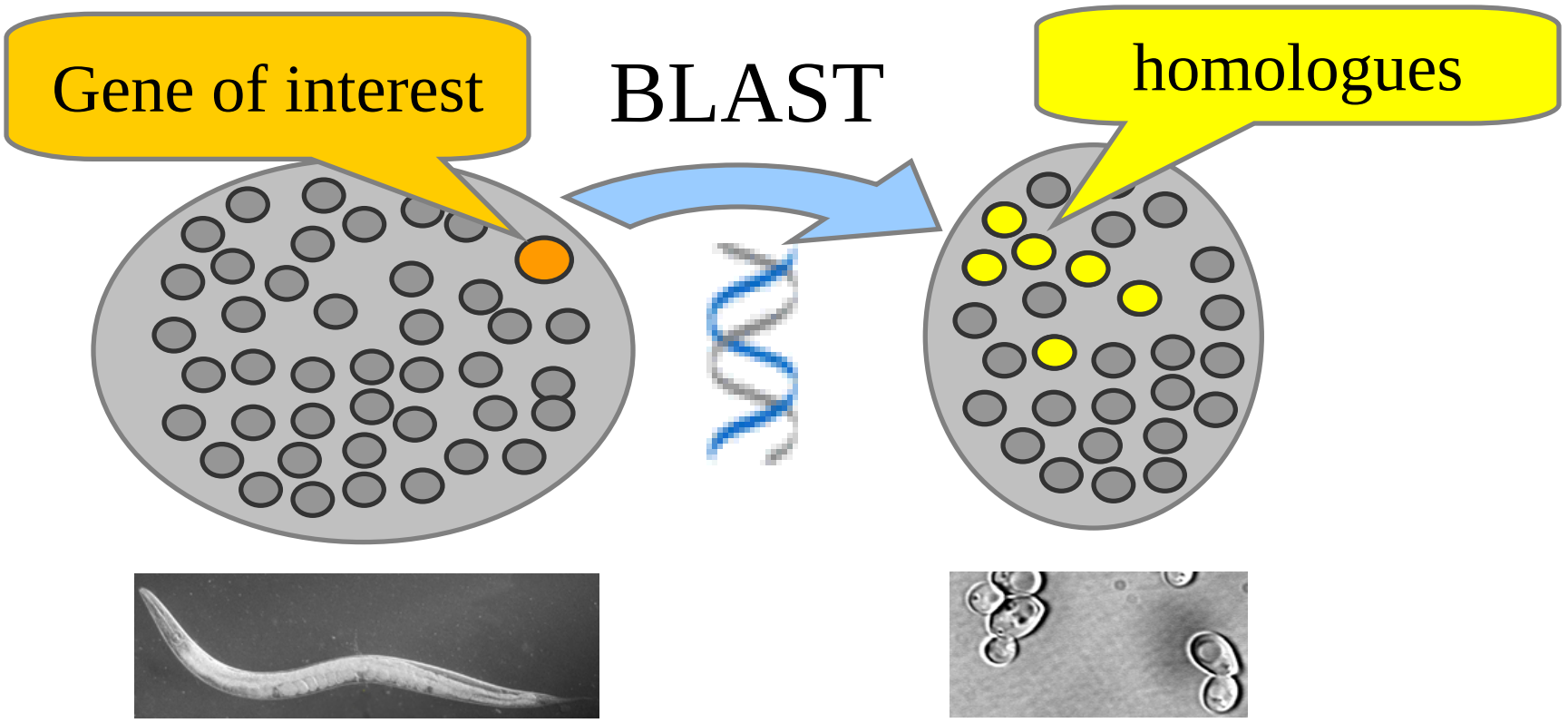
- Accuracy - Percentage of links connecting two members of the category
- Coverage – Percentage of metagenes connected atleast one other metagene in the category





igure S2. Predicting functional categories with a multiple species network built from a naller number of microarrays. **(A-D)** We built a multiple species network from 979 icroarrays instead of the full 3182. We compared its performance predicting KEGG

Using Sequence similarity



Using Sequence Homology(cont..)

- **LIMITATIONS**
  - One genomic sequence may have several close homologues , some of which may be related to different functions.
  - A sequence may have diverged beyond recognition although the gene may have maintained its function

Using Gene Expression Data

- Can be used to provide function links for genes based on the co-expression with known genes

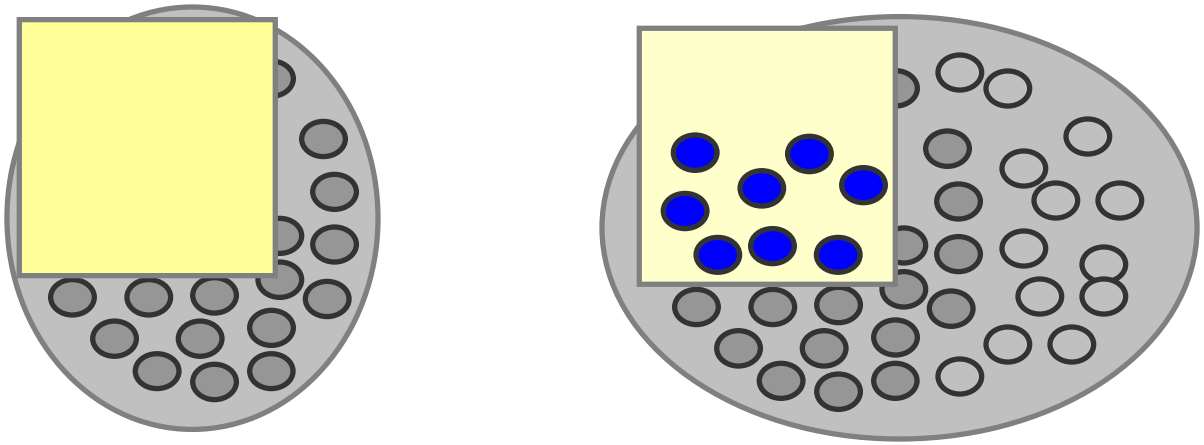
- **Limitations**
  - Can only provide a functional link between genes of the same organism.Difficult for cross-specie comparison
  - Due to the noise in the expression data the inferred co-expression could be accidental and may not necessarily reflect some similar biological function

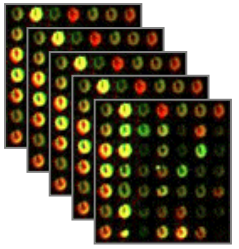
Combining Gene Expression Data and Sequence Data

- The limitations of using either(only sequence or only expression) alone may be reduced
- Homologue genes whose function has been preserved are expected to be co-regulated with genes that have similar function
- This distinguishes from similar homologues whose function has diverged

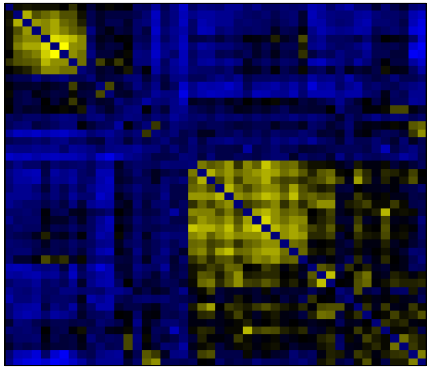
Combining Gene Expression Data and Sequence Data

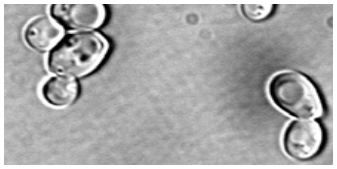
BLAST

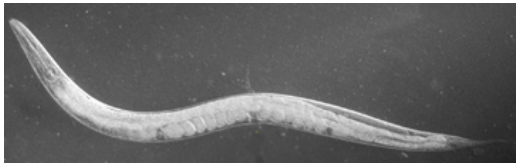


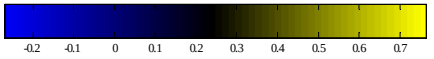


correlation









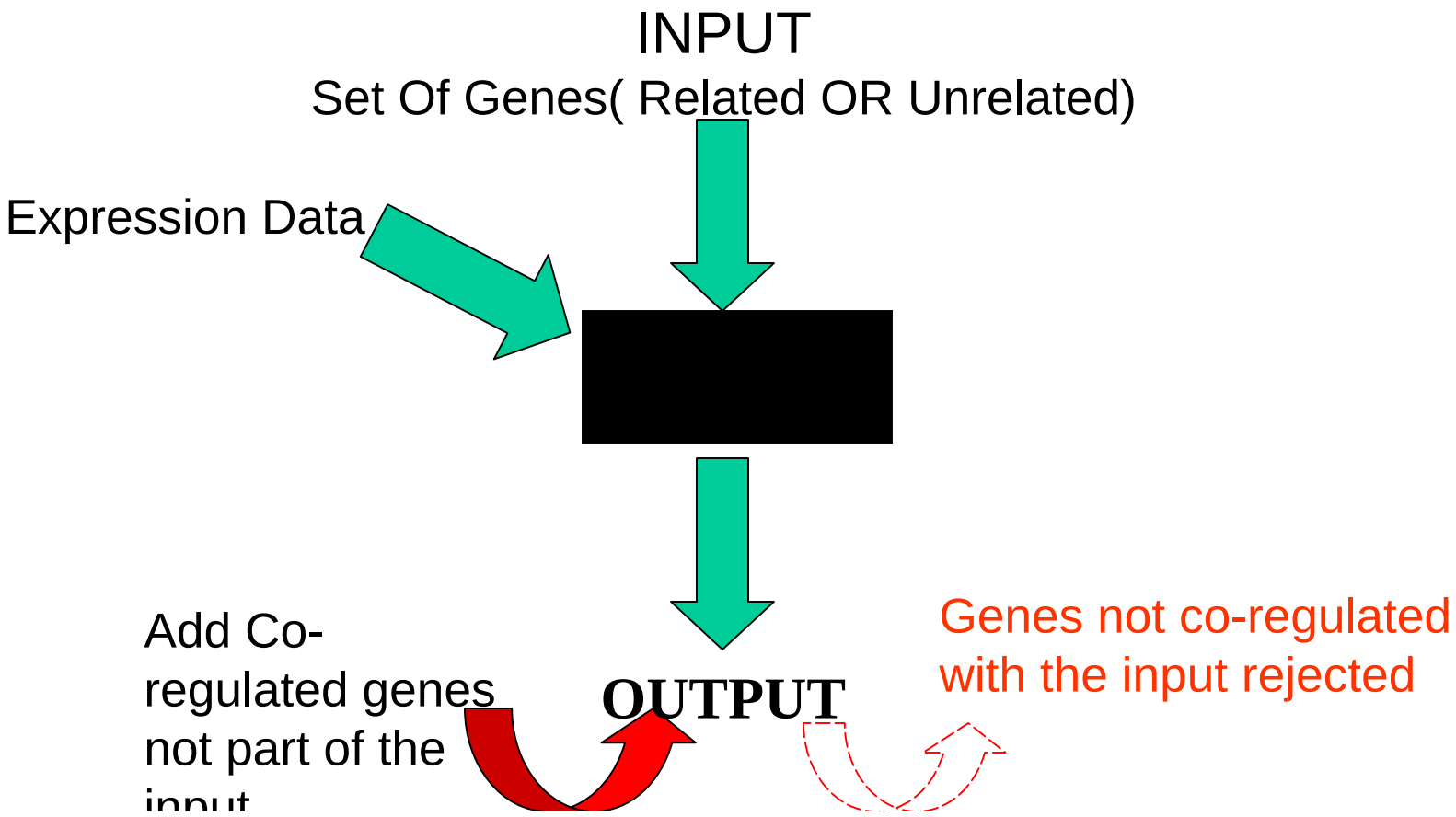
0 1

Standard Cluster Algorithms

- **Limitations**
  - They assign each gene to a single cluster ,whereas in fact genes may participate in several functions and could be in several clusters
  - These algorithms classify genes on the basis of there expression under all experimental conditions, whereas cellular processes are generally affected by a subset of conditions.Most conditions that do not contribute information contribute to the background noise

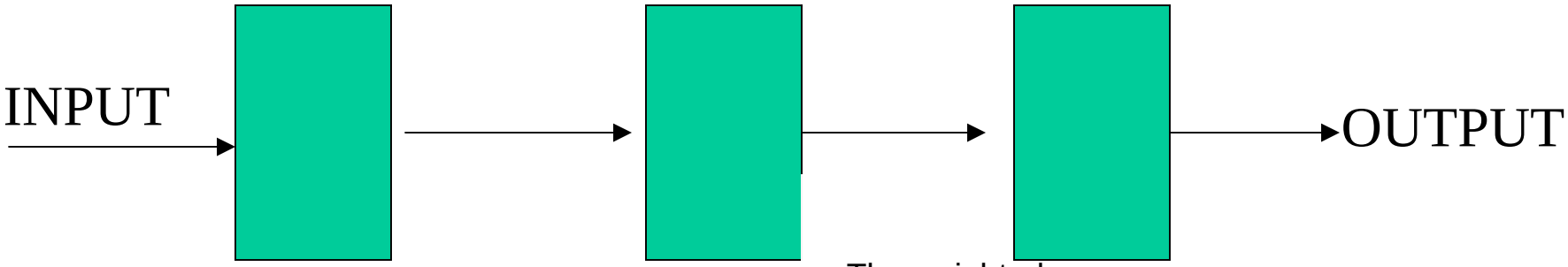
Signature Algorithms

- Takes a set of related or random genes
- Uses expression data and generates an Output



The Algorithm

- Identifies the co-regulated genes and also the experimental conditions under which they are co-expressed
- Algorithm proceeds in two stages:
  - Identifies the experimental conditions under which the genes are co regulated most tightly
  - Selects those genes that show a significant and consistent expression under the conditions selected in the first stage

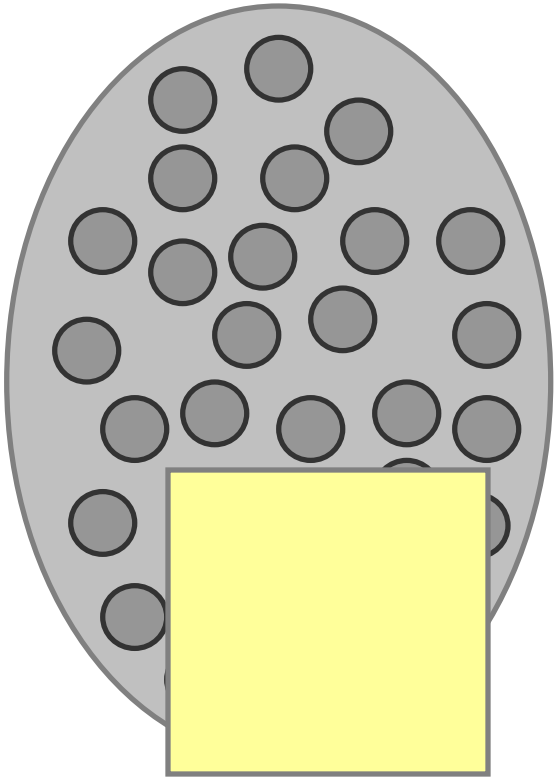


Average change calculated for each condition in the expression of the input and called "Condition score" . Only scores that are large enough are selected

The weighted avg. change in expression for each gene calculated Score referred to as "gene score". High gene score selected

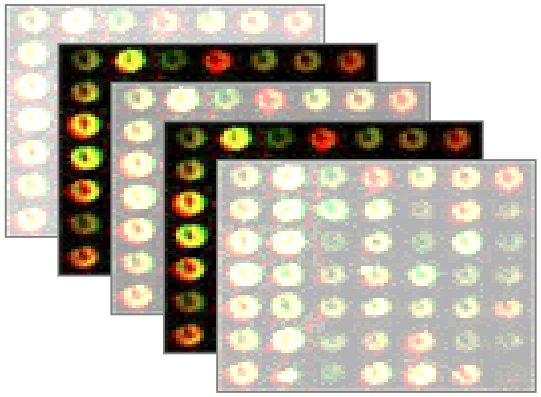
Signature Algorithm Output

Co-regulated genes



+

Co-regulating conditions



Iterative Signature Algorithm

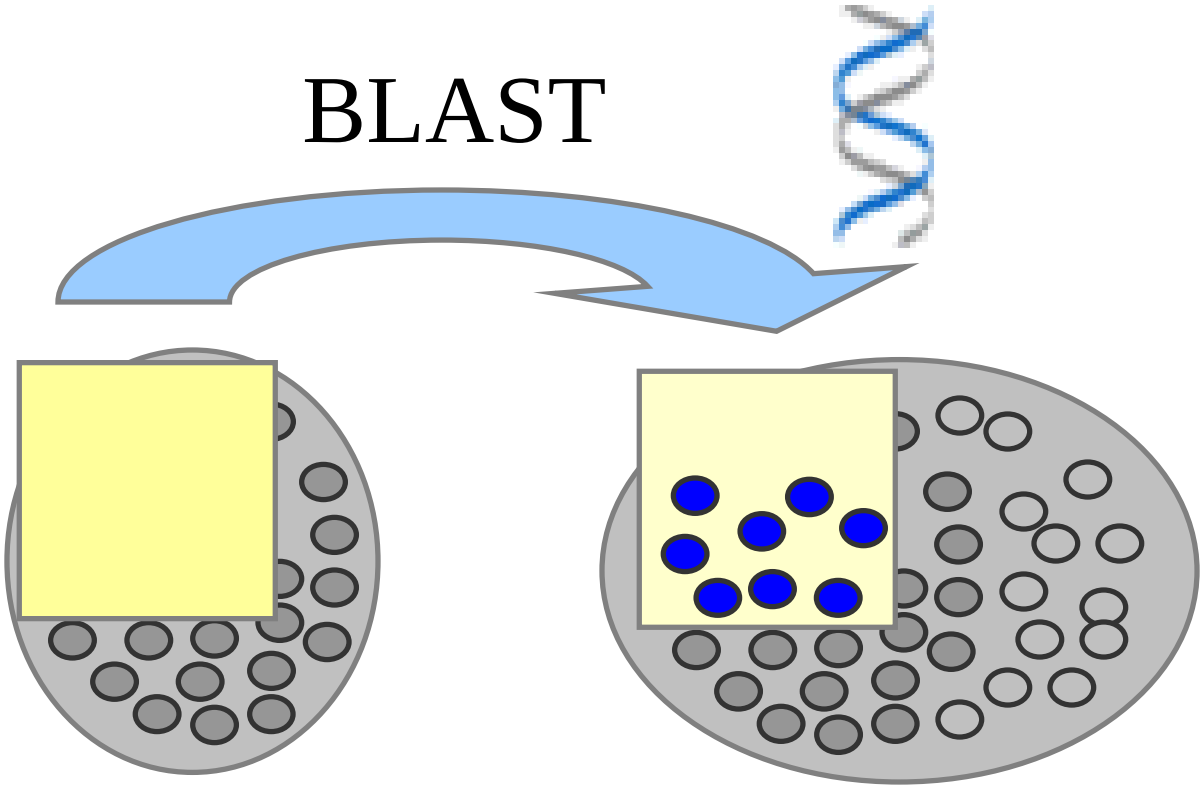
- OUPUT re-used as INPUT, such that further iterations can bring in more co-regulated genes
- Procedure repeated OUTPUT equals INPUT
- Final OUTPUT is called **“transcription module”.**
- Contains set of co-regulated genes and the conditions that induce their co-regulation.
- By definition, all genes outside the module are less co-regulated than the module genes under these conditions

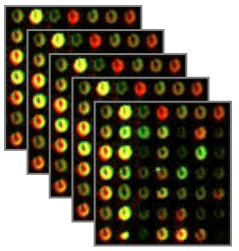
INPUT=OUTPUT

Using Transcription Modules

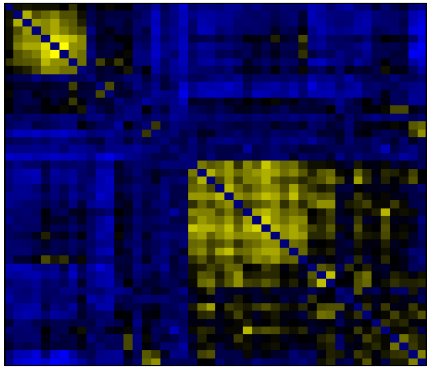
(Combining Expression and sequence info.)

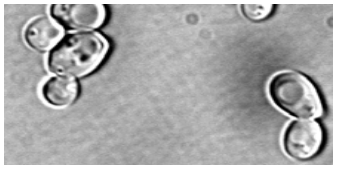
BLAST

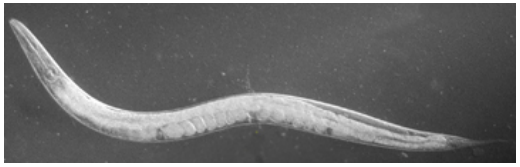


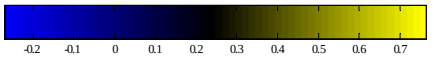


Correlation









0 1

Homologue Modules

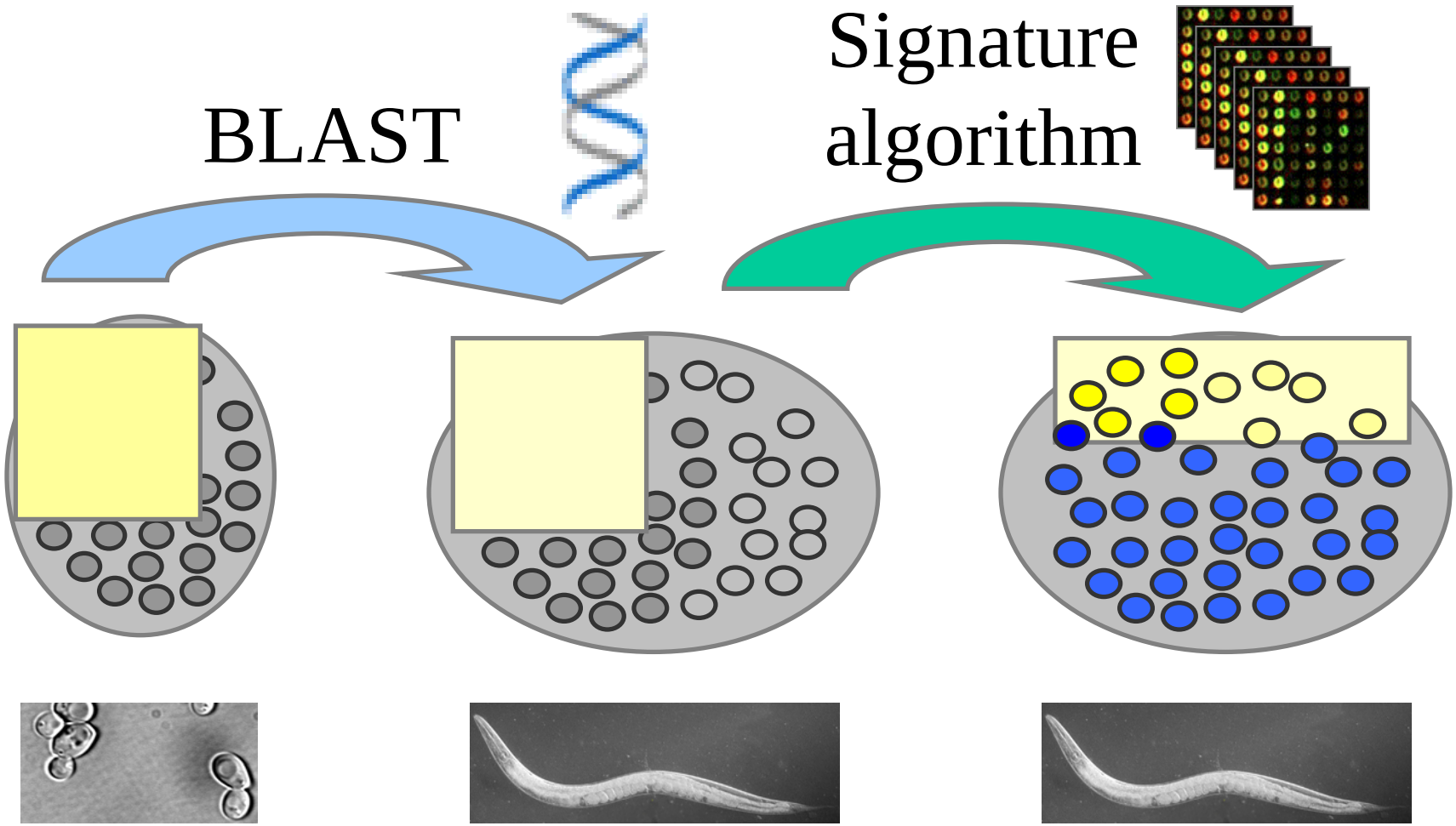
- Using the Signature Algorithm we create a Transcription Module starting out with genes of an organism that are associated with a function …
  - Eg: yeast genes associated with cellular function and end up with a TM.
- Using this TM, find homologues in other organism
- In the paper at hand five organisms (E.coli,A.thaliana,C.elegans,D.melanogaster,H.sapiens,S.cerevisiae) are used and five different homologue modules were created on sequence similarity
- The Assumption is that the co expression of functionally linked genes is often conserved
- The results indicate an average correlation between the genes of the homologue module to be significant

Unrefined Module

- The homologue modules that result here has limitations
- The average correlation is significant, but the pair wise correlation reveals that only some are correlated with each other

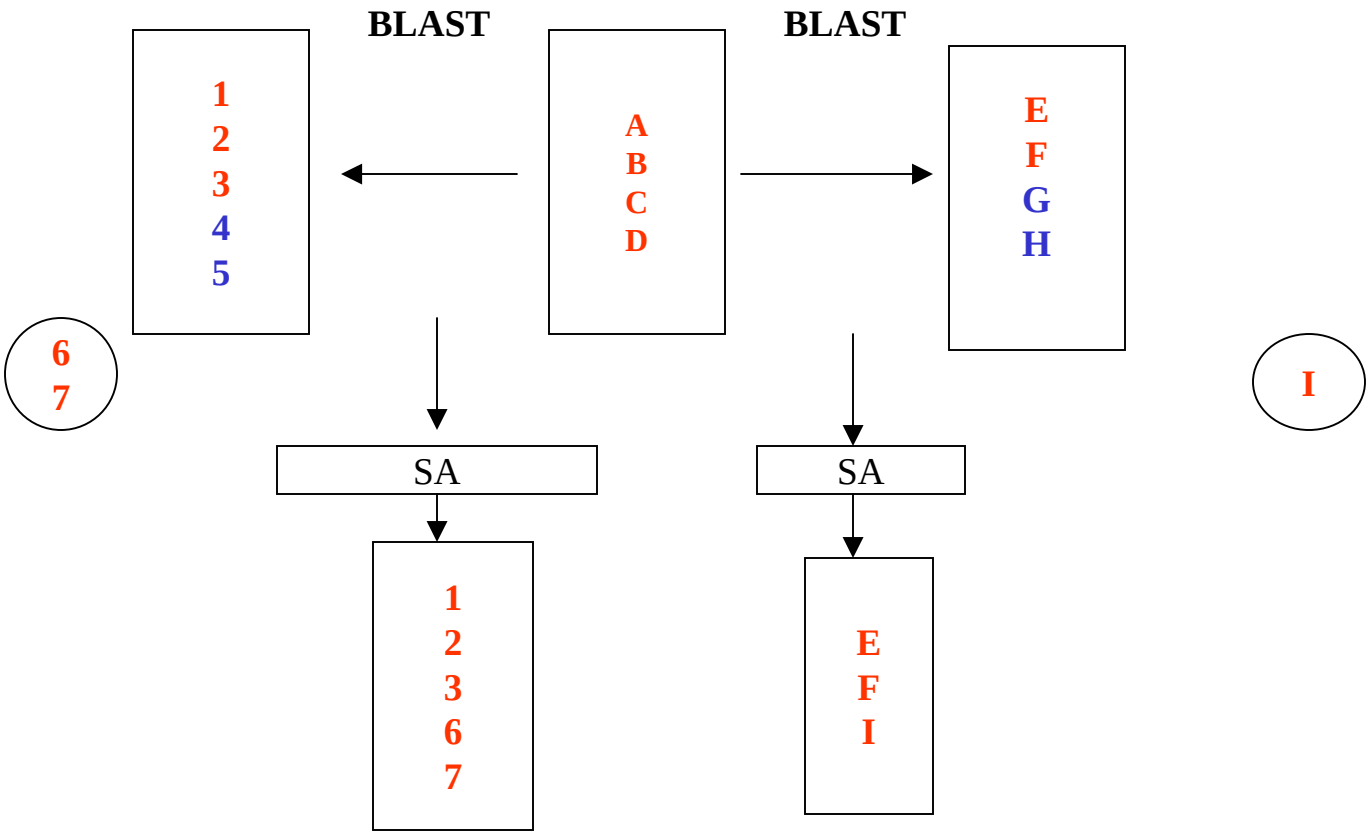
- Inference :
  - Some of the homologues in the homologue module that are not co-expressed may have varied functionally over time
  - The module misses genes whose sequence has changed over time but the function has remained the same

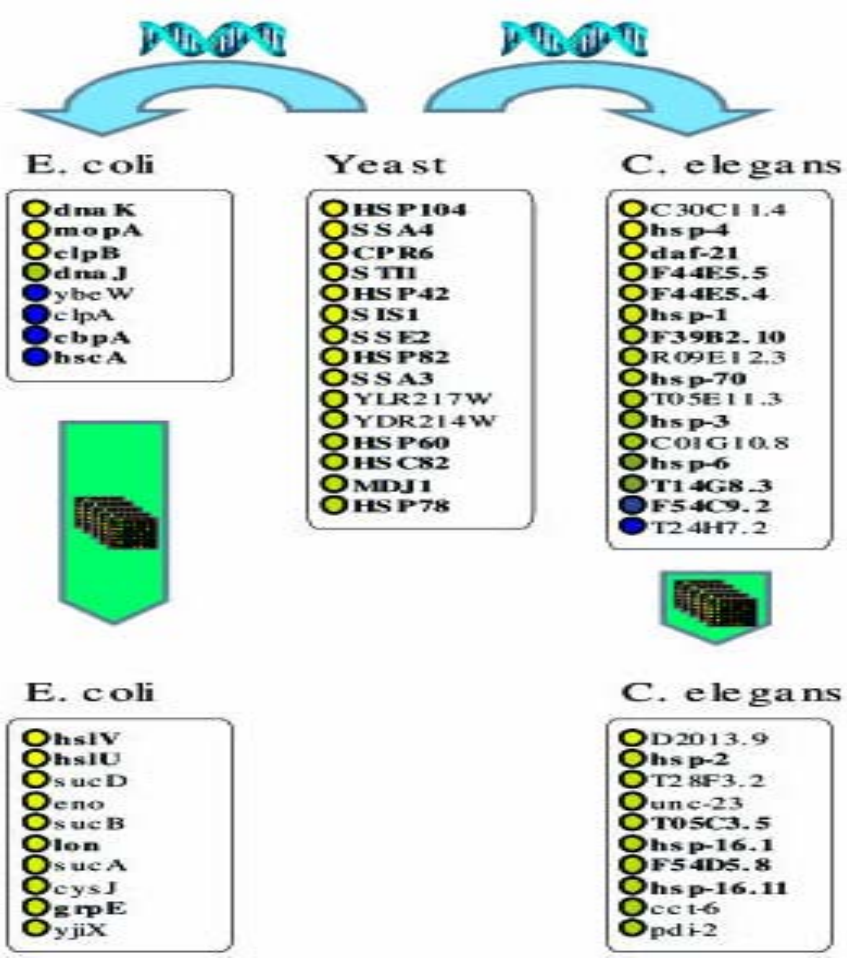
“Gene Refinement”



Signature Algorithm (Ctd)..

- Signature algorithm used to reject genes not co-regulated according to the available expression data
- Co-regulated genes not included by homology added




E. coli
dnaK
mopA
clpB
dnaJ
ybcW
clpA
cbpA
hscA
Yeast
HSP104
SSA4
CPR6
STI1
HSP42
SIS1
SSE2
HSP82
SSA3
YLR217W
YDR214W
HSP60
HSC82
MDJ1
HSP78
C. elegans
C30C11.4
hsp-4
daf-21
F44E5.5
F44E5.4
hsp-1
F39B2.10
R09E12.3
hsp-70
T05E11.3
hsp-3
C01G10.8
hsp-6
T14G8.3
F54C9.2
T24H7.2
E. coli
hslV
hslU
sucD
eno
sucB
lon
sucA
cysJ
grpE
yjiX
C. elegans
D2013.9
hsp-2
T28F3.2
unc-23
T05C3.5
hsp-16.1
F54D5.8
hsp-16.11
cct-6
pdi-2

Within Species

- Still has problems
  - Different chips
  - Different experimental procedures
- Fixes some problems
  - No need for “homologues”
  - Can (sometimes) directly compare expression profiles

Within Species

- Possible uses
  - Cancer vs Non-Cancer
  - Cancer vs Cancer by
    - Type (Lung vs Brain)
    - Degree of progression (metastatic vs not)
    - Grade (high vs low)
- Functional annotation

Analysis vs Meta-Analysis

- Analysis
  - Direct comparison of expression profiles
    - “2-fold increase in cancer”
- Meta-Analysis
  - Comparison of properties of the expression profiles
    - “Well above standard deviation in cancer”

Meta-Signatures

- Group of genes whose differential expression is “most significant”
  - Neoplastic transformation
  - Undifferentiated cancer
- This is NOT the aforementioned “signature algorithm”

Finding Meta-Signatures

- Choose analogous differential expression data sets
- Select direction and significance threshold
- Sort genes by number of signatures in which they appear
- Find intersection
- Calculate the significance of the intersection

Validation

- Leave-one-out voting

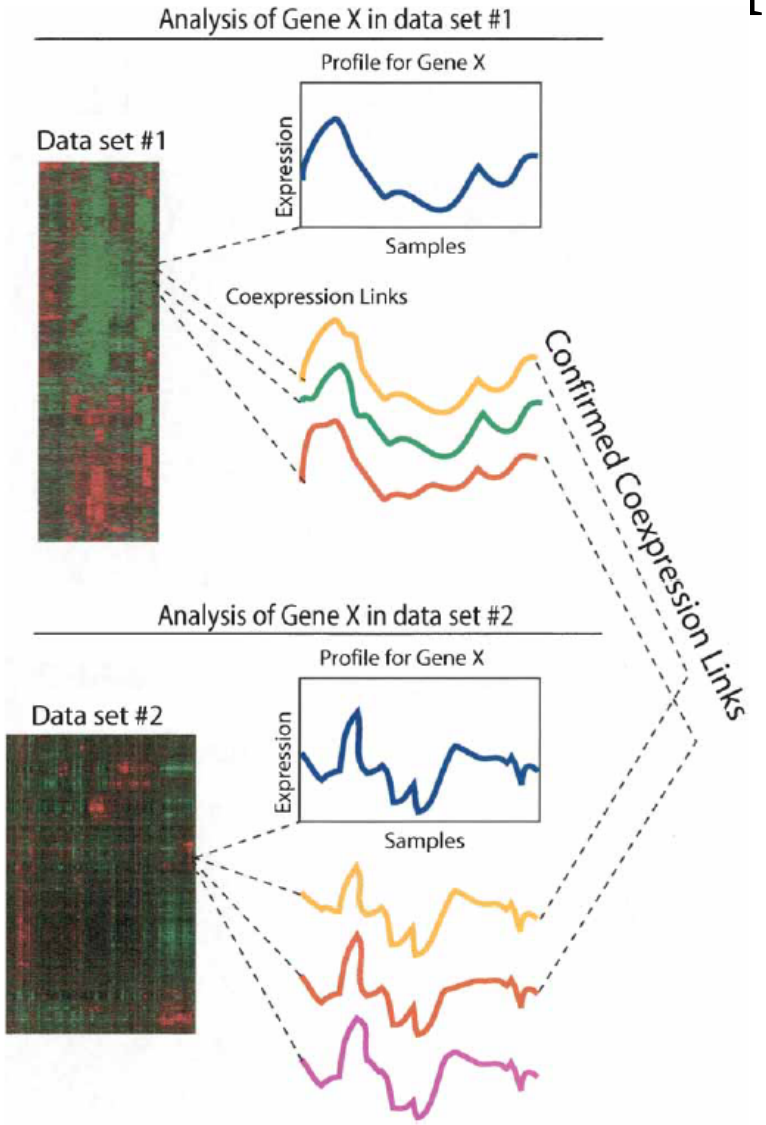
Results

- Neoplastic transformation
  - 36 signatures
  - 183 present in 10/36
  - Contains: cell cycle, transcriptional regulation, protein folding, and the proteasome
- Undifferentiated cancer
  - 7 signatures
  - 69 genes in 4/7

Functional Annotation/ Coexpression Links

- Basic idea: if a pair/set/group of genes are coexpressed in more than one data set, then they are more likely to be coexpressed *in vivo*

Coexpression Links

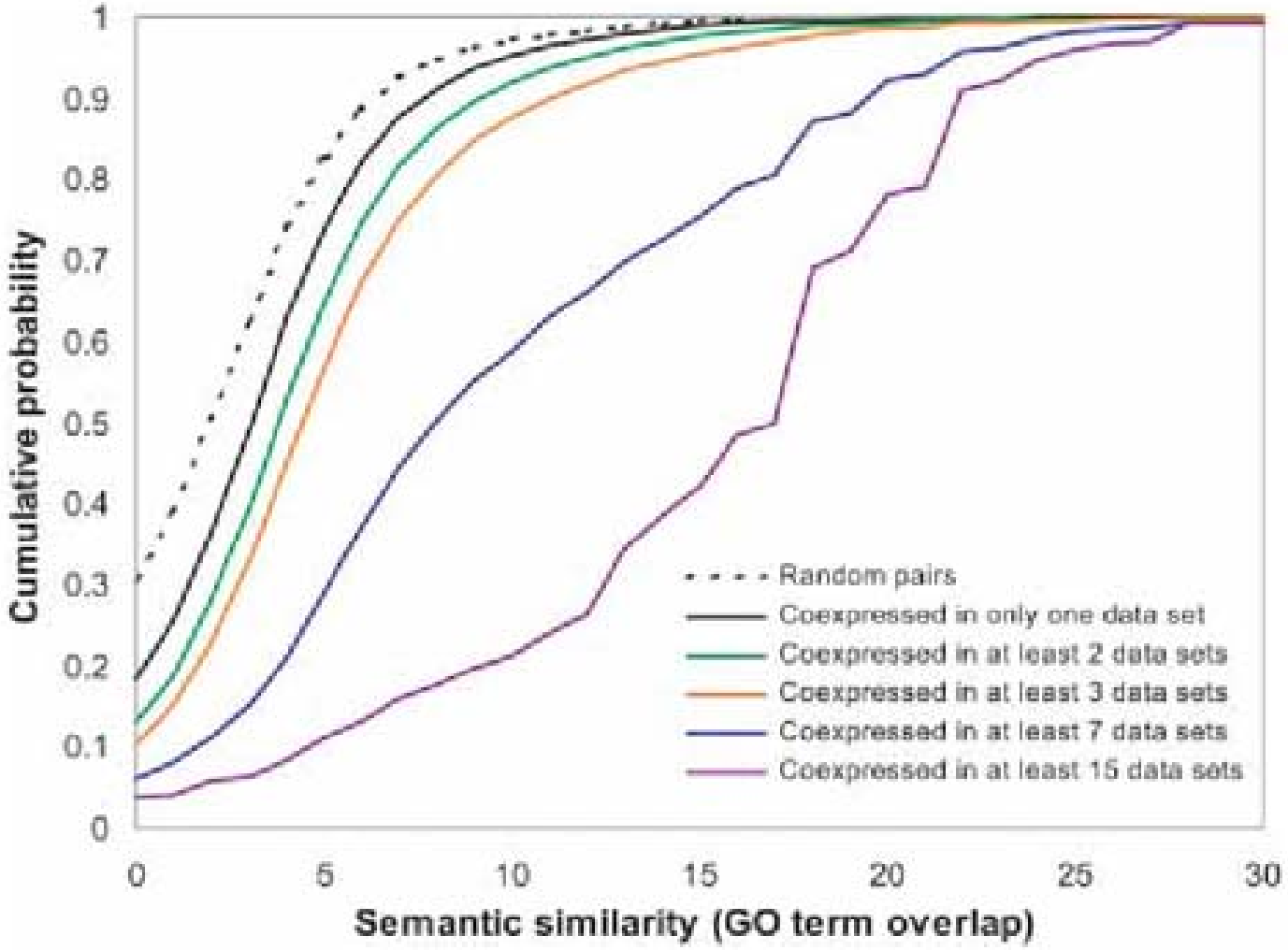


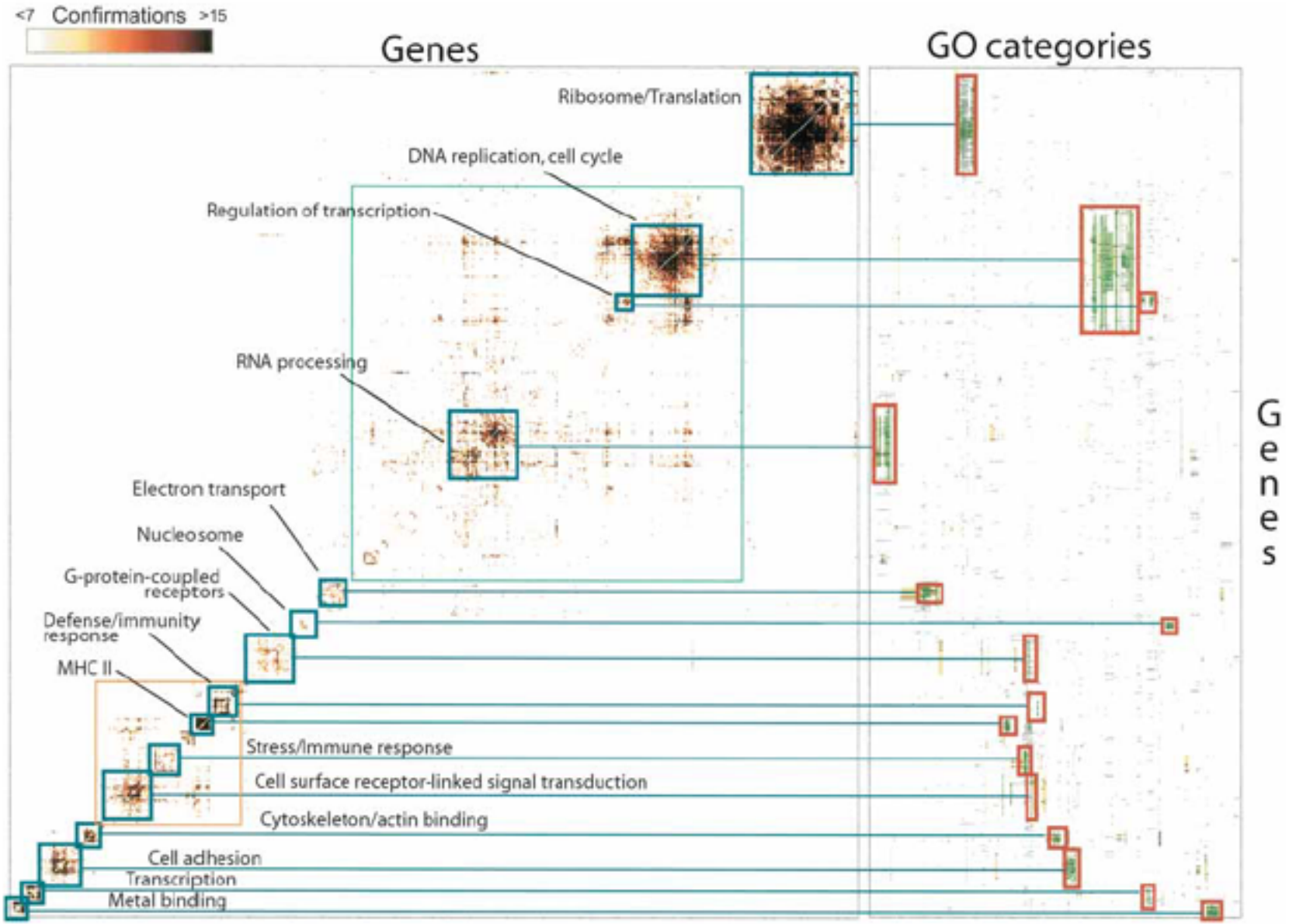
Coexpression Links

- 9.7 million coexpression links in 60 data sets
- 220,649 are 3+ confirmed

Functional Annotation

- GO term overlap




<7 Confirmations >15
Genes
GO categories
Ribosome/Translation
DNA replication, cell cycle
Regulation of transcription
RNA processing
Electron transport
Nucleosome
G-protein-coupled receptors
Defense/immunity response
MHC II
Stress/Immune response
Cell surface receptor-linked signal transduction
Cytoskeleton/actin binding
Cell adhesion
Transcription
Metal binding
Genes