

Analysis of Genomes &  
Transcriptomes in terms of the  
Occurrence of Parts and Features:

# Surveys of a Finite Parts List

**Mark Gerstein**

**Molecular Biophysics & Biochemistry and  
Computer Science, Yale University**

*H Hegyi, J Lin, B Stenger, P Harrison, N Echols,  
J Qian, A Drawid, D Greenbaum, R Jansen*

Transcriptome 2000, Paris

8 November 2000

**1995**

Bacteria,  
1.6 Mb,  
~1600 genes  
[*Science* 269: 496]



**1997**

Eukaryote,  
13 Mb,  
~6K genes  
[*Nature* 387: 1]



Genomes  
highlight  
the  
**Finiteness**  
of the  
“Parts” in  
Biology

**1998**

Animal,  
~100 Mb,  
~20K genes  
[*Science* 282:  
1945]



**2000?**

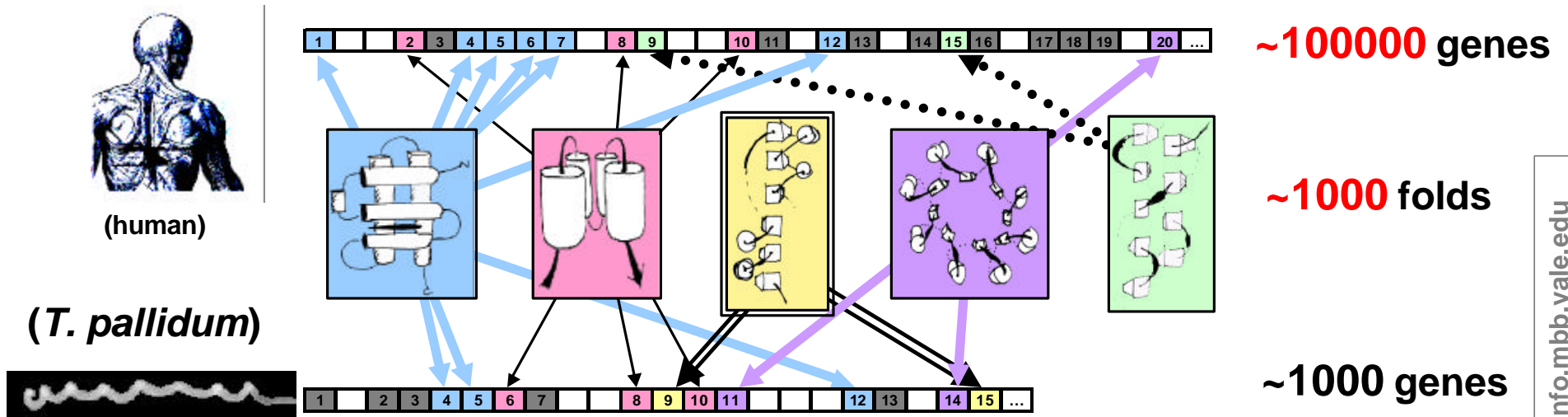
Human,  
~3 Gb,  
~100K  
genes [???



'98 spoof

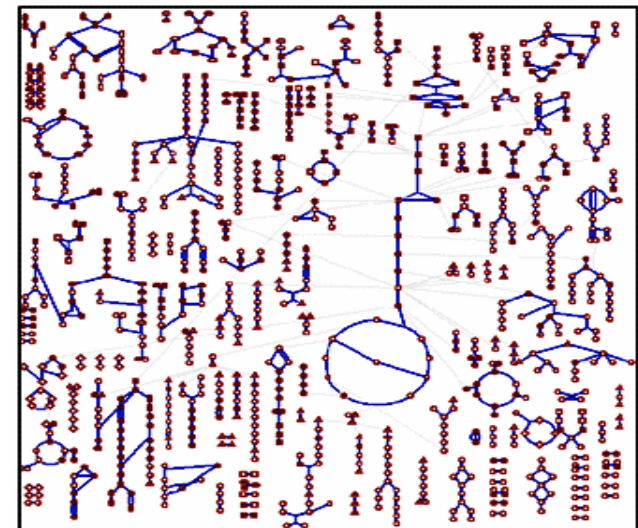
real thing, Apr '00

# Simplifying the Complexity of Genomes: Global Surveys of a Finite Set of Parts from Many Perspectives

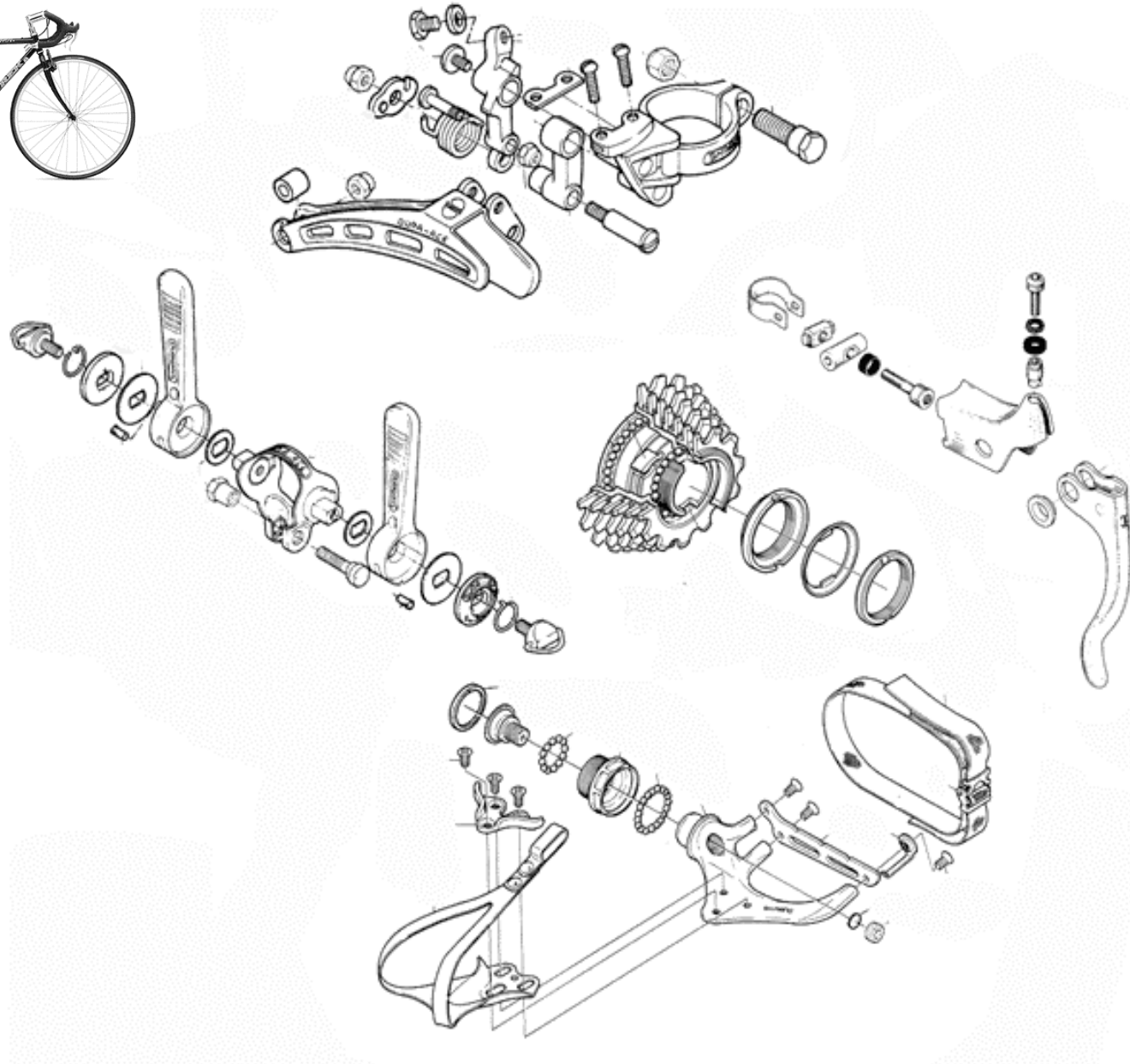


Same logic for sequence families, blocks, orthologs, motifs, pathways, functions....

Functions picture from [www.fruitfly.org/~suzi](http://www.fruitfly.org/~suzi) (Ashburner); Pathways picture from, [ecocyc.pangeasystems.com/ecocyc](http://ecocyc.pangeasystems.com/ecocyc) (Karp, Riley). Related resources: COGS, ProDom, Pfam, Blocks, Domo, WIT, CATH, Scop....

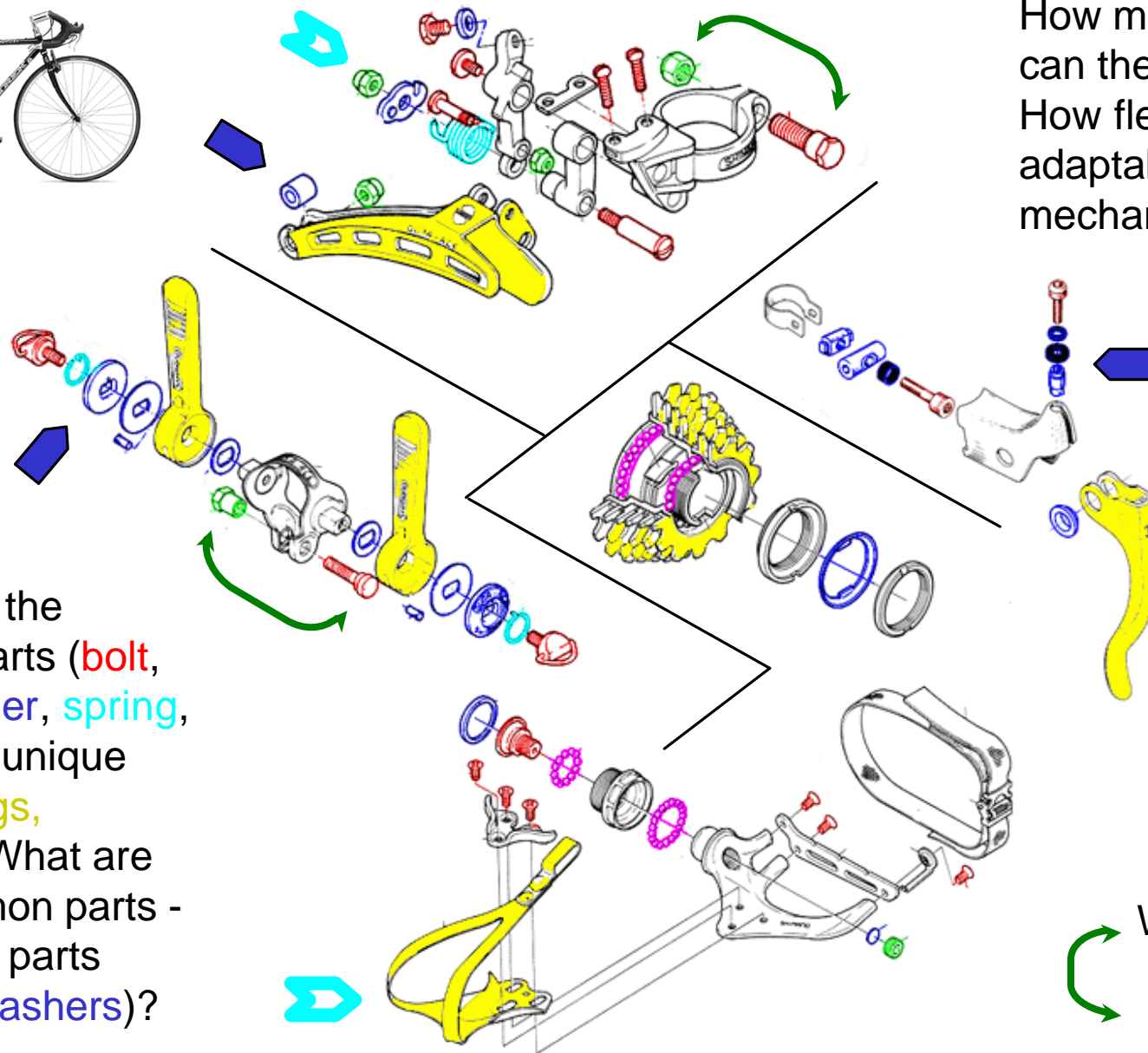


# A Parts List Approach to Bike Maintenance








# A Parts List Approach to Bike Maintenance



What are the shared parts (**bolt**, **nut**, **washer**, **spring**, **bearing**), unique parts (**cogs**, **levers**)? What are the common parts - types of parts (**nuts** & **washers**)?

How many roles can these play?   
How flexible and adaptable are they mechanically? 

Where are the parts located?  
Which parts interact? 

# Analysis of Genomes & Transcriptomes in terms of the Occurrence of Parts & Features

## 1 Using Parts to Interpret Genomes.

Shared and/or unique parts. Venn Diagrams, Fold tree with all- $\beta$  diff. Ortholog tree. Top-10 folds.

## 2 Using Parts to Interpret Pseudogenomes.

In worm, top  $\Psi$ -folds (DNase, hydrolase) v top-folds (lg). chr. IV enriched, dead and dying families (90YG v 1G)

## 3 Using Parts to Interpret Transcriptomes: Expression & Structure.

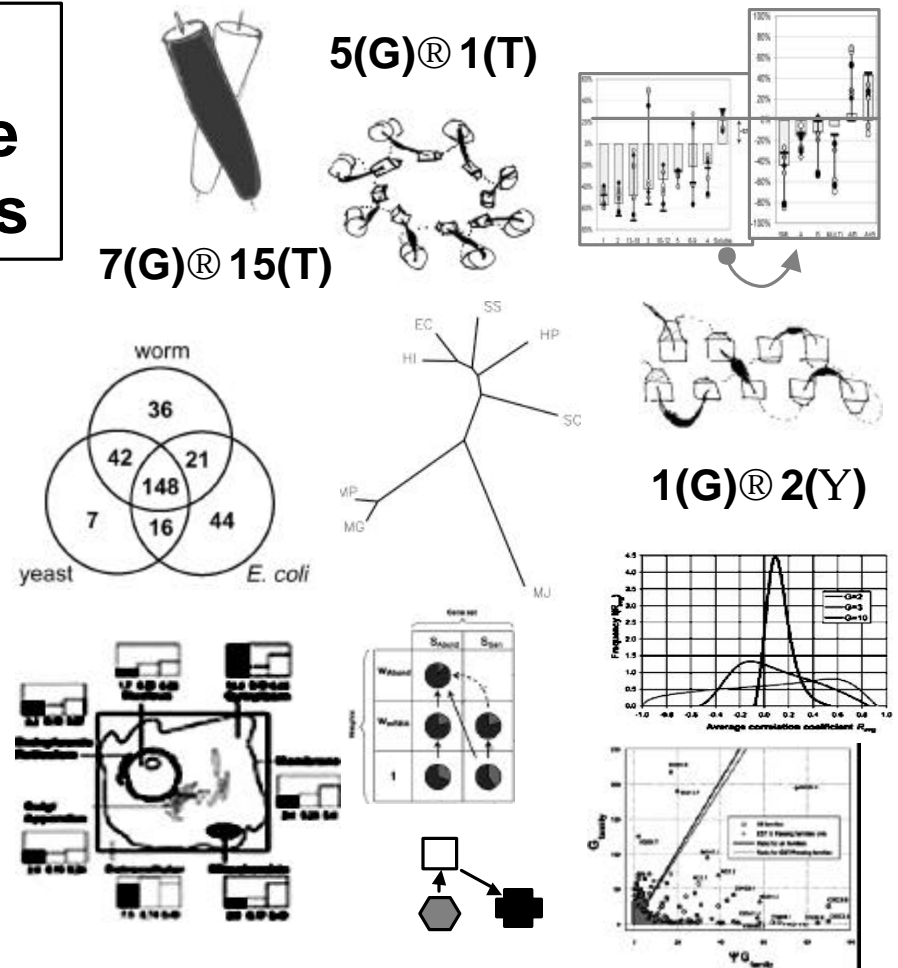
Top-10 parts in mRNA. Enriched in transcriptome:  $\alpha\beta$  folds, energy, synthesis, TIM fold, VGA. Depleted: TMs, transport, transcription, Leu-zip, NS. Compare with prot. abundance.

## 4 Expression & Localization.

Enriched : Cytoplasmic. Depleted: Nuclear. Bayesian localizer

## 5 Expression & Function.

Expression relates to structure & localization but to function, globally? P-value formalism. Weak relation to protein-protein interactions.



[bioinfo.mbb.yale.edu](http://bioinfo.mbb.yale.edu)

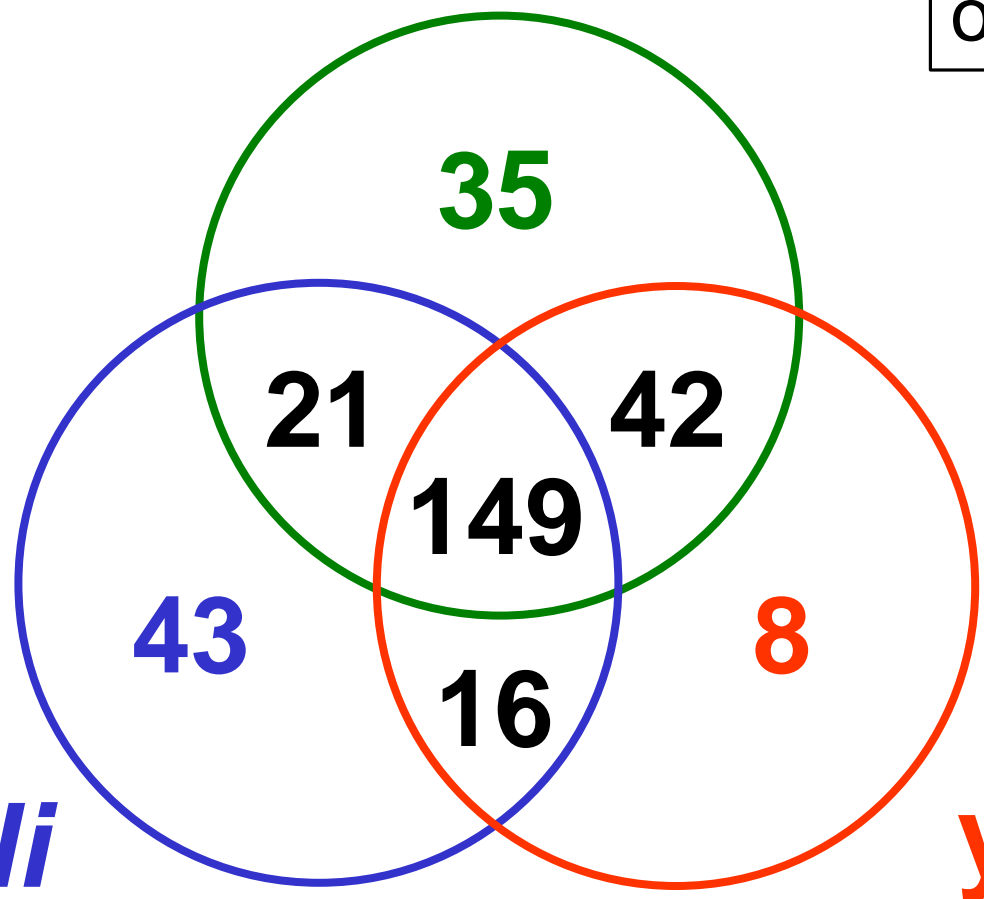
*H Hegyi, J Lin, B Stenger, P Harrison, N Echols, R Jansen, A Drawid, J Qian, D Greenbaum, M Snyder*



Shared  
Folds

worm

of 339

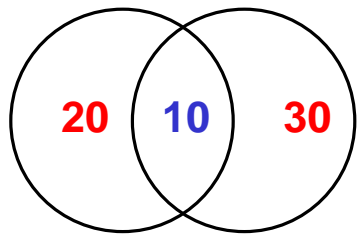
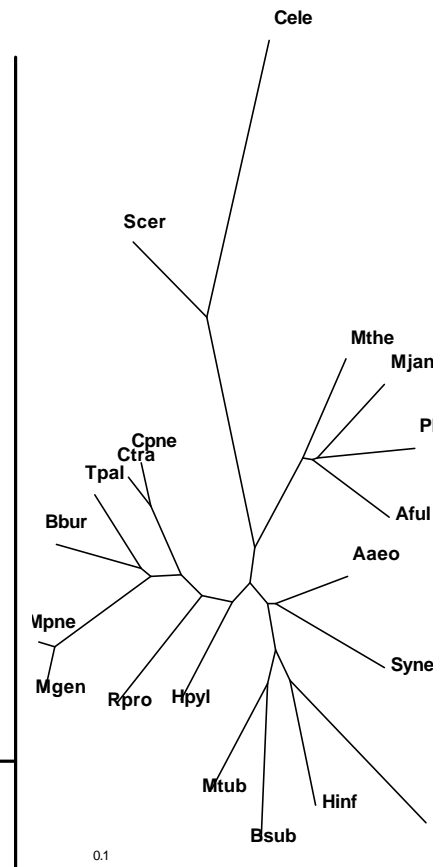
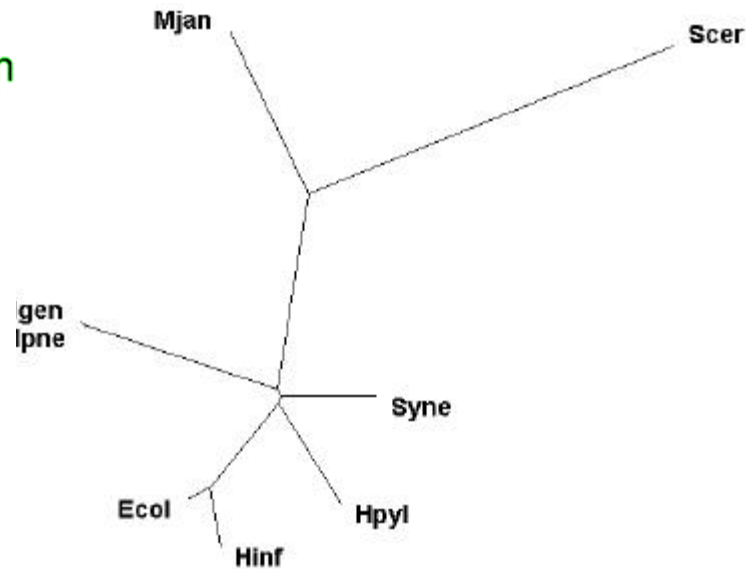
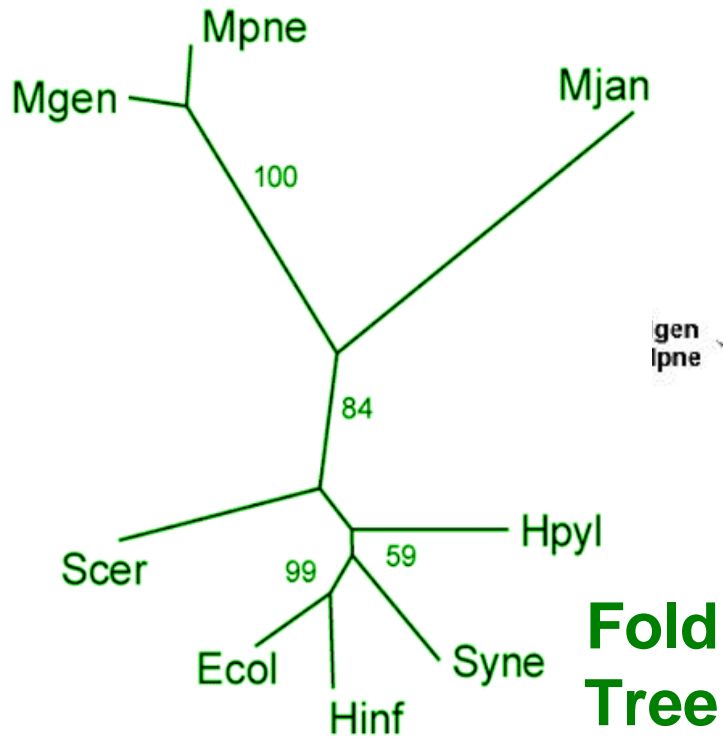


*E. coli*

yeast



# Cluster Trees Grouping Initial Genomes on Basis of Shared Folds



$$D = 10 / (20 + 10 + 30)$$

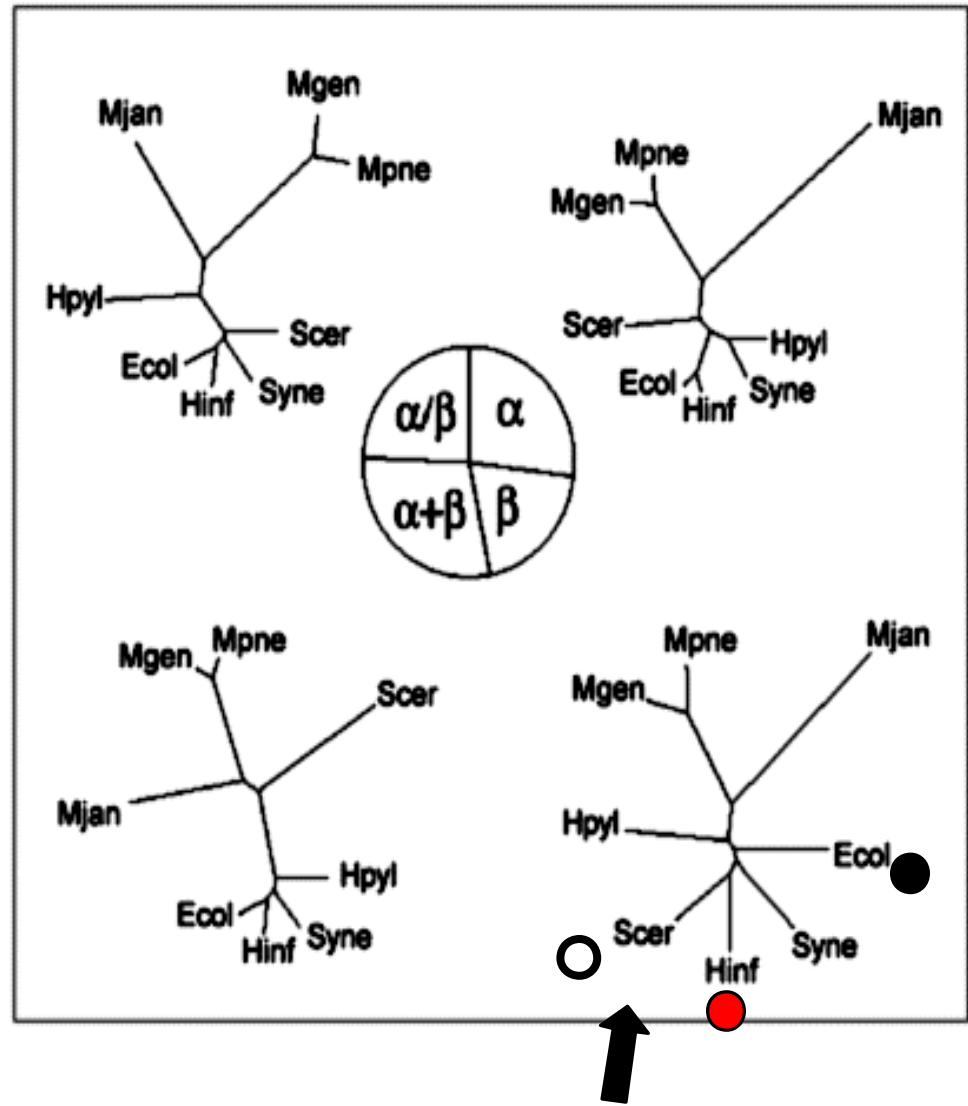
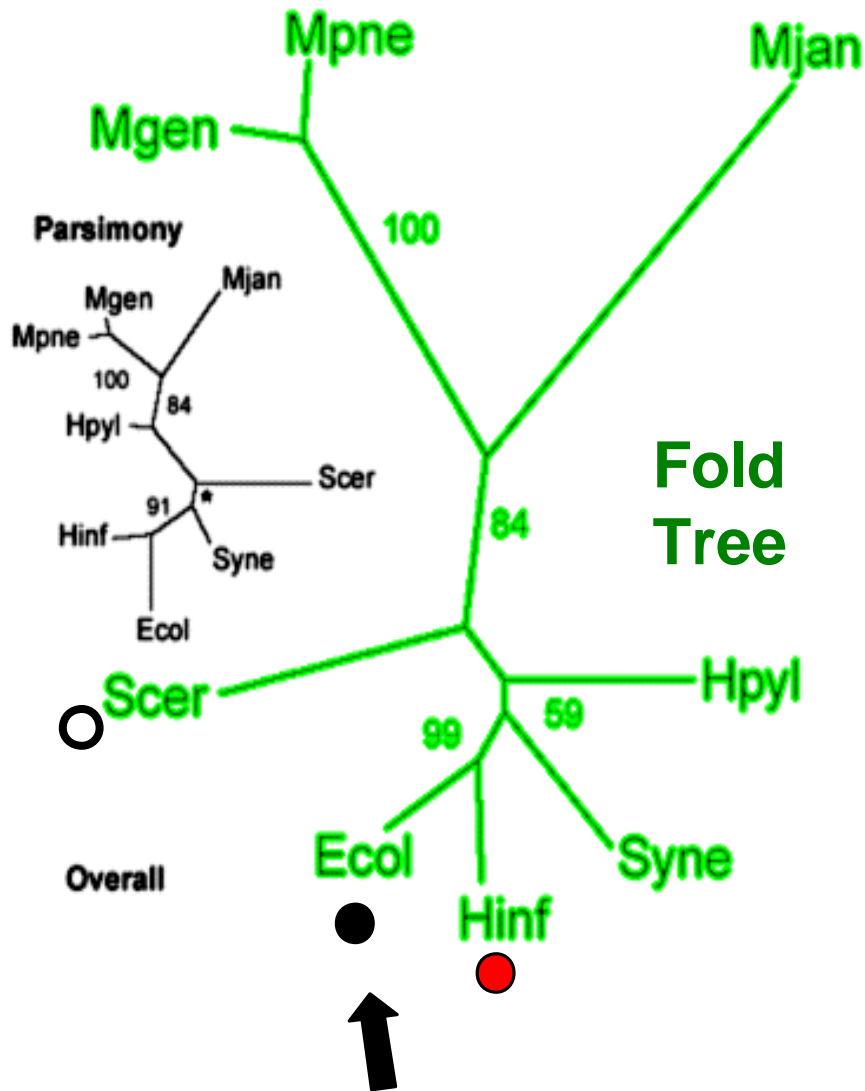
$$D = S / T$$

**S** = # shared folds

**D** = shared fold dist. betw. 2 genomes

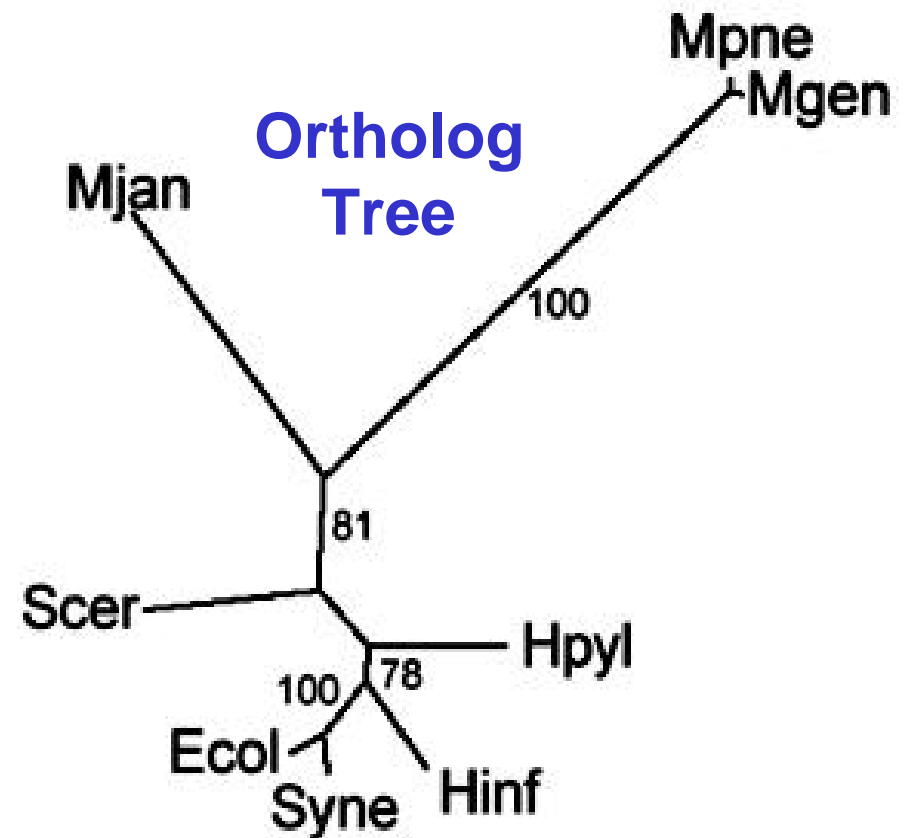
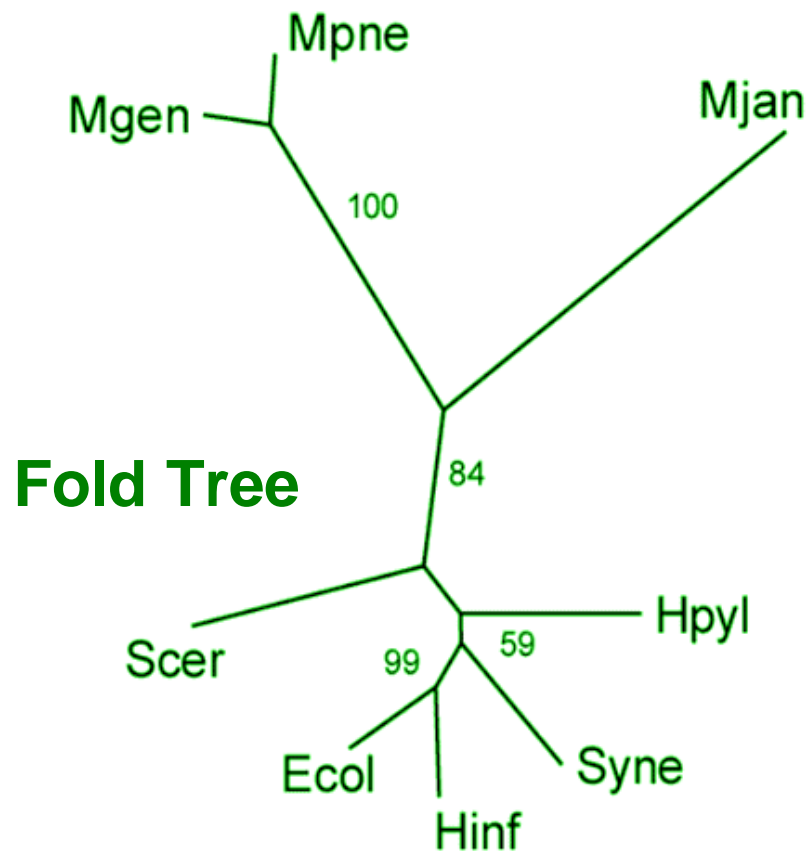
**T** = total # folds in both

# Distribution of Folds in Various Classes



Unusual distribution of all-beta folds

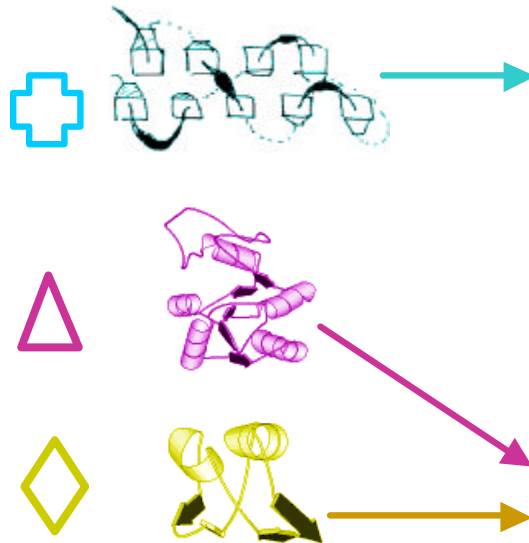
# Compare with Ortholog Occurrence Trees (Part = ortholog v fold)



(based on COGs scheme of Koonin & Lipman, similar approaches by Dujon, Bork, &c.)

# Common Folds in Genome, Varies Betw. Genomes

Depends on comparison method, DB, sfams v folds, &c (new top superfamilies via  $\psi$ -Blast, Intersection of top-10 to get shared and common)



Top-10 Worm Folds		class	num. matches in worm genome (N)	frac. all worm dom. (F)	in EC?	in SC?
Ig	B	830	1.7%			
Knottins	SML	565	1.1%			
Protein kinases (cat. core)	MULT	472	0.9%			
C-type lectin-like	A+B	322	0.6%			
corticoid recep. (DNA-bind dom.)	SML	276	0.5%			
Ligand-bind dom. nuc. receptor	A	257	0.5%			
alpha-alpha superhelix	A	247	0.5%			
C2H2 Zn finger	SML	239	0.5%			
P-loop NTP Hydrolase	A/B	235	0.5%			
Ferredoxin	A+B	207	0.4%			

Rank	<i>M. genitalium</i> Superfamily #	<i>B. subtilis</i> Superfamily #	<i>E. coli</i> Superfamily #
1	P-loop hydrolase 60	P-loop hydrolyase 173	P-loop hydrolase 191
2	SAM methyl-transferase 16	Rossmann domain 165	Rossmann domain 158
3	Rossmann domain 13	Phosphate-binding barrel 79	Phosphate-binding barrel 64
4	Class I synthetase 12	PLP-transferase 44	PLP-transferase 38
5	Class II synthetase 11	CheY-like domain 36	CheY-like domain 36
6	Nucleic acid binding dom. 11	SAM methyl-transferase 30	Ferredoxins 35
Total ORFs	479	4268	4268
with Common Superfamilies	105 (22%)	465 (11%)	458 (11%)

Eubacteria

Rank	<i>M. thermoautotrophicum</i> Superfamily #	<i>A. fulgidus</i> Superfamily #
1	P-loop hydrolyase 93	P-loop hydrolyase 118
2	Phosphate-binding barrel 54	Rossmann domain 104
3	Rossmann domains 53	Phosphate-binding barrel 56
4	Ferredoxins 48	Ferredoxins 49
5	SAM methyl-transferase 17	SAM methyl-transferase 24
6	PLP-transferases 15	PLP-transferases 18
Total ORFs	1869	2409
with Common Superfamilies	252 (14%)	309 (13%)

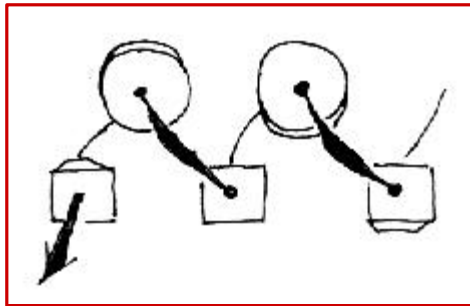
Archaea

Rank	<i>S. cerevisiae</i> Superfamily #
1	P-loop hydrolyase 249
2	Protein kinase 123
3	Rossmann domain 90
4	RNA-binding domain 75
5	SAM methyl-transferase 63
6	Ribonuclease H-like 57
Total ORFs	6218
with Common Superfamilies	560 (9%)

Yeast



# Common, Shared Folds: $\beta\alpha\beta$ structure

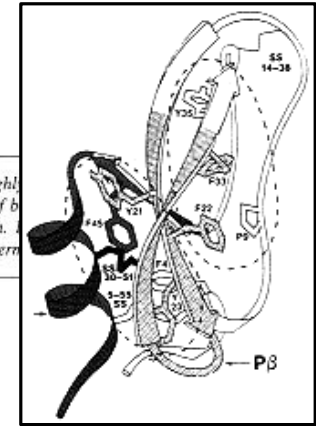


## A peptide model of a protein folding intermediate

Terrence G. Oas & Peter S. Kim

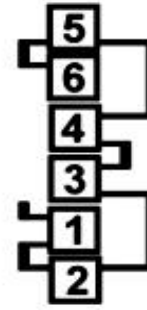
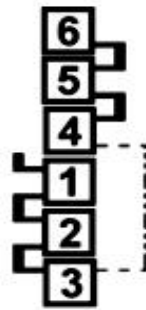
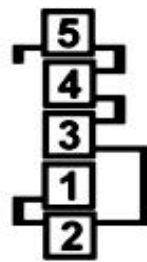
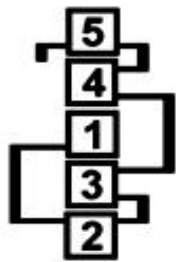
Whitehead Institute for Biomedical Research, Nine Cambridge Center, Cambridge, Massachusetts 02142, USA  
Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA

*It is difficult to determine the structures of protein folding intermediates because folding is a highly disulphide-bonded peptide pair, designed to mimic the first crucial intermediate in the folding of b inhibitor, contains secondary and tertiary structure similar to that found in the native protein. I circumvent the problem of cooperativity and permit characterization of structures of folding inter*



336: 42

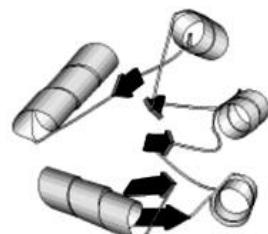
All share  $\alpha/\beta$  structure with repeated R.H.  $\beta\alpha\beta$  units connecting adjacent strands or nearly so (18+4+2 of 24)



HI, MJ, SC  
vs scop  
1.32



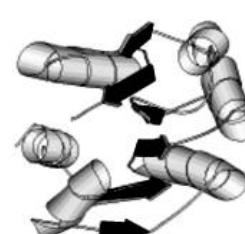
P-loop  
hydrolase



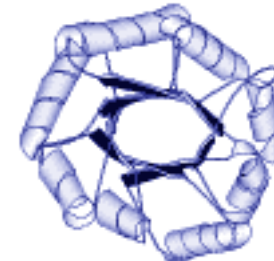
Flavodoxin  
like



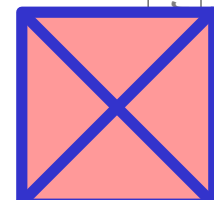
Rossmann  
Fold



Thiamin  
Binding



TIM-  
barrel



# Analysis of Genomes & Transcriptomes in terms of the Occurrence of Parts & Features

## 1 Using Parts to Interpret Genomes.

Shared and/or unique parts. Venn Diagrams, Fold tree with all- $\beta$  diff. Ortholog tree. Top-10 folds.

## 2 Using Parts to Interpret Pseudogenomes.

In worm, top  $\Psi$ -folds (DNase, hydrolase) v top-folds (lg). chr. IV enriched, dead and dying families (90YG v 1G)

## 3 Using Parts to Interpret Transcriptomes: Expression & Structure.

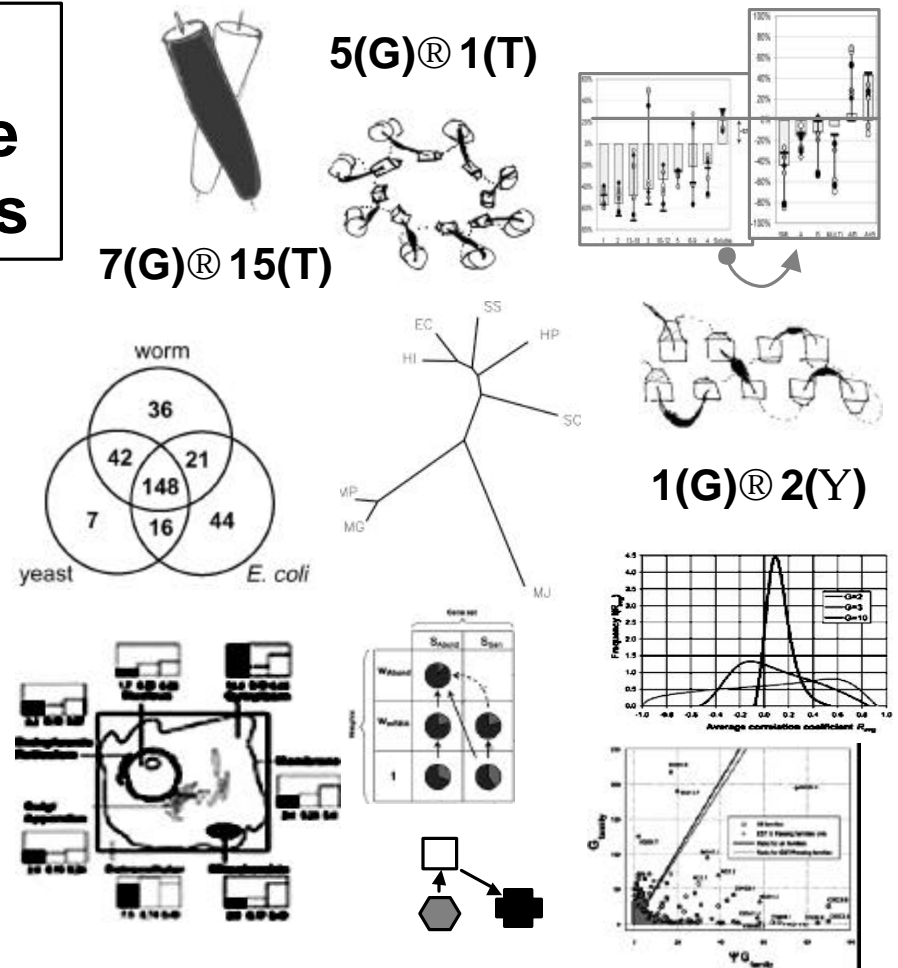
Top-10 parts in mRNA. Enriched in transcriptome:  $\alpha\beta$  folds, energy, synthesis, TIM fold, VGA. Depleted: TMs, transport, transcription, Leu-zip, NS. Compare with prot. abundance.

## 4 Expression & Localization.

Enriched : Cytoplasmic. Depleted: Nuclear. Bayesian localizer

## 5 Expression & Function.

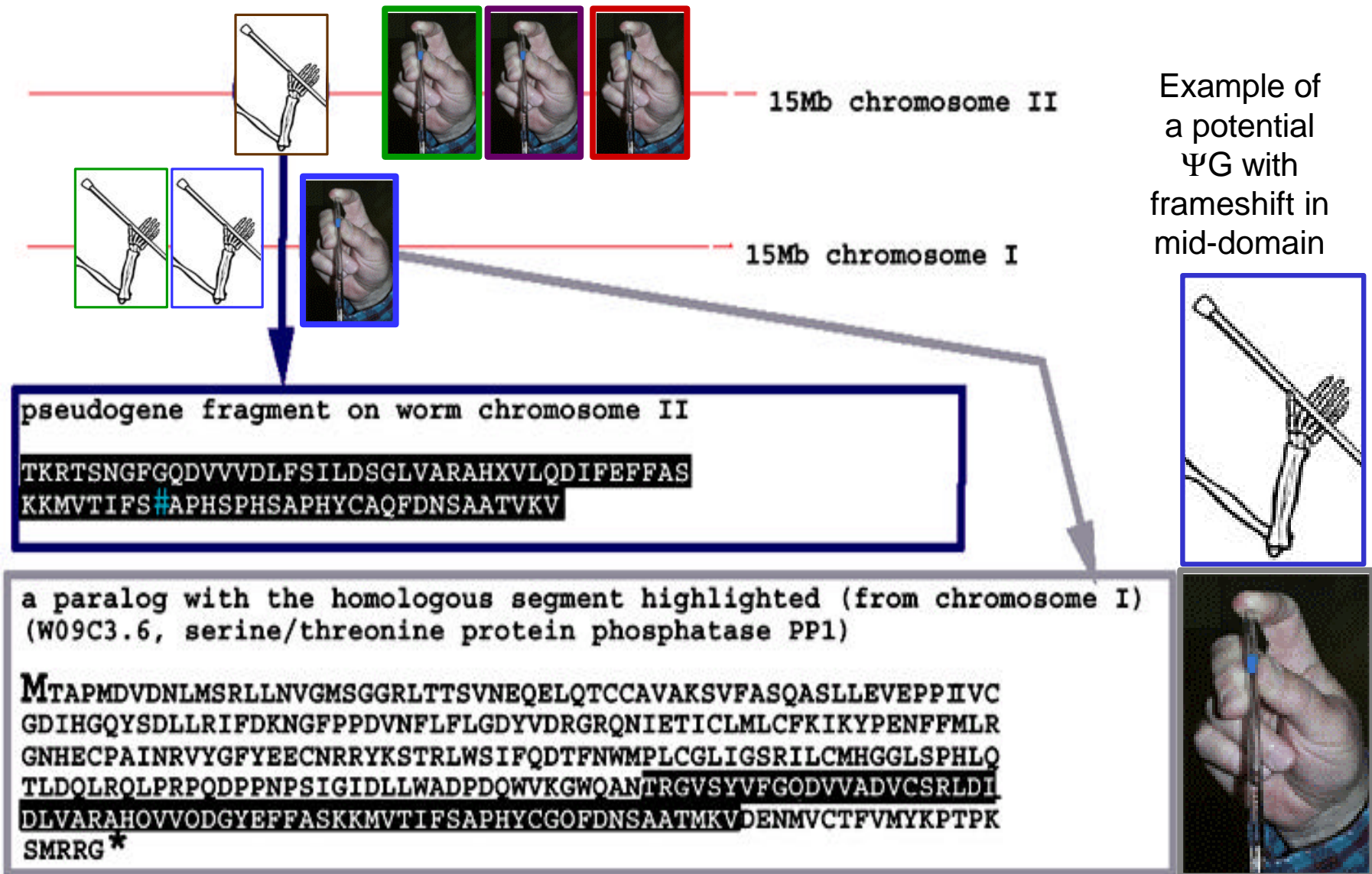
Expression relates to structure & localization but to function, globally? P-value formalism. Weak relation to protein-protein interactions.



[bioinfo.mbb.yale.edu](http://bioinfo.mbb.yale.edu)

*H Hegyi, J Lin, B Stenger,  
P Harrison, N Echols,  
R Jansen, A Drawid, J Qian,  
D Greenbaum, M Snyder*

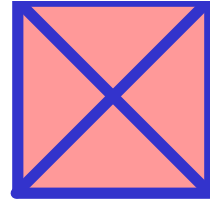
# Pseudogenomics: Surveying “Dead” Parts



(Our def'n: ΨG = obvious homolog to known protein with frameshift or stop in mid-domain)

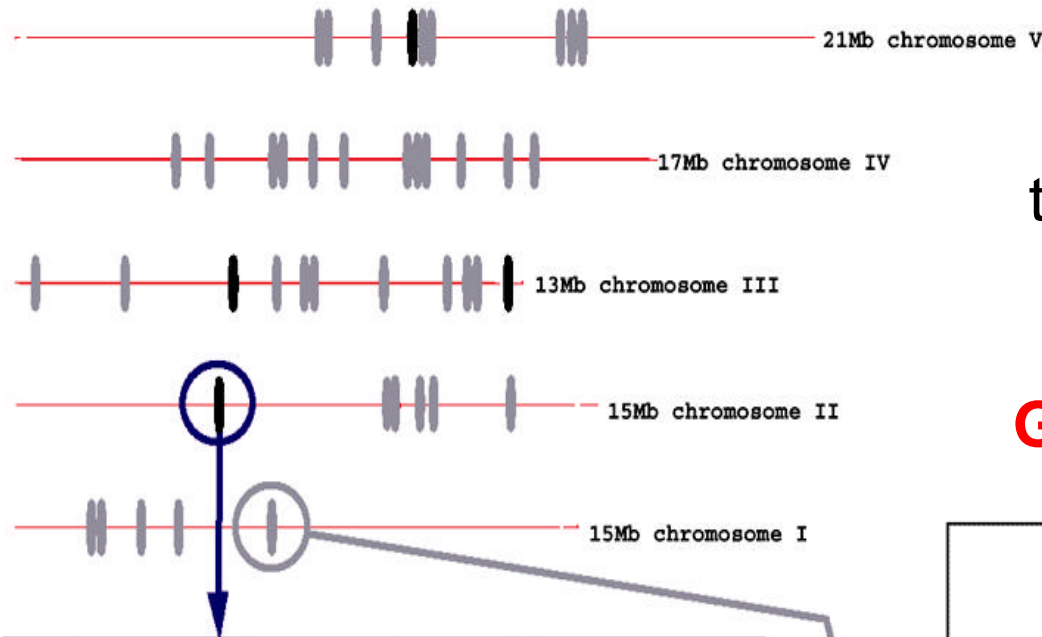


# Folds in Pseudogenes



## YG identification pipeline to Summary of Pseudogenes in worm

**G=19K** **G<sub>E</sub>=8K** **YG=4K (2K)**



pseudogene fragment on worm chromosome II

```
TKRRTSNGFGQDVVVDLFSILDSGLVARAHXVLQDIFEFFAS
KKMVTIFS#APHSPHSAPHYCAQFDNSAATVKV
```

a paralog with the homologous segment highlighted (from chromosome I) (W09C3.6, serine/threonine protein phosphatase PP1)

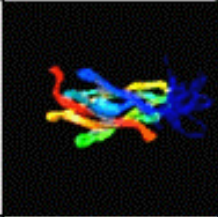
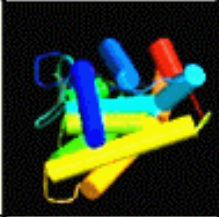
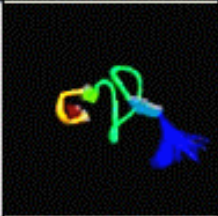
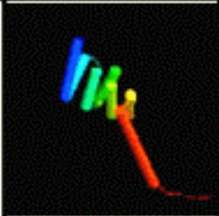
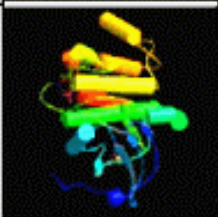
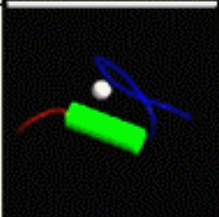
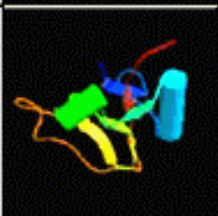
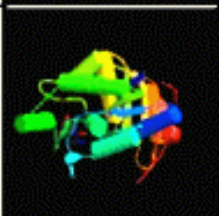
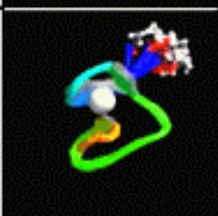

```
MTAPMDVDNLMRSLLNVGMSGGRLLTTSVNEQELQTCCAVAKSVFASQASLLEVEPPPIVC
GDIHQYSDLLRIFDKNGFPDVFNLFGLGDYVDRGRQNIETICLMLCFKIKYPENFFMLR
GNHECPAINRVYGFYEENRRYKSTRLWSIFQDTFNWMPCLGLIGSRILCMHGGLSPHLO
TLDLROLRPRPODPPNPSIGIDLLWADPDOWVKGWOANTRCVSYVFGODVVADVCSRLDI
DLVARAHOVVODGYEFFASKKMVTIFSAPHYCGOFDNSAATMKVDENMVCTFVMYKPTPK
SMRRG*
```

Example of a potential  $\Psi$ G with frameshift in mid-domain


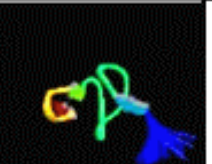


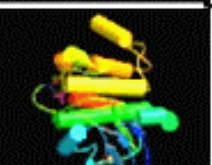
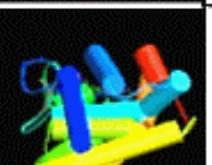
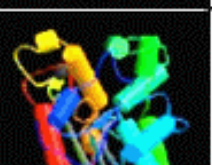
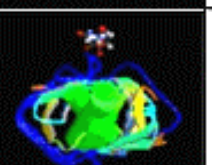
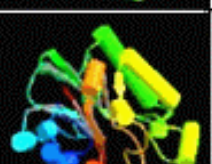
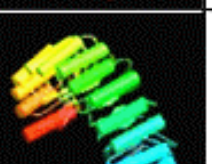
	Category	Total number	Number for genes with EST match	Genes with EST match as percentage of Category	Number for genes in paralog families with EST match	Genes in paralog families with EST match as percentage of Category
Genes	Total	18,576 (G)	7,829 (G <sub>E</sub> )	42%	13,417 (G <sub>P</sub> )	72%
	Singletons	5,913	2,788	47%	---	---
Pseudogenes and pseudogene fragments	Total	3,814 ( $\Psi$ G)	997 ( $\Psi$ G <sub>E</sub> )	26%	2,729 ( $\Psi$ G <sub>P</sub> )	72%
	Singletons	637 (17% of $\Psi$ G)	233	36%	---	---
	Intronic pseudogenes *	1,155 (30% of $\Psi$ G)	351	30%	704	61%

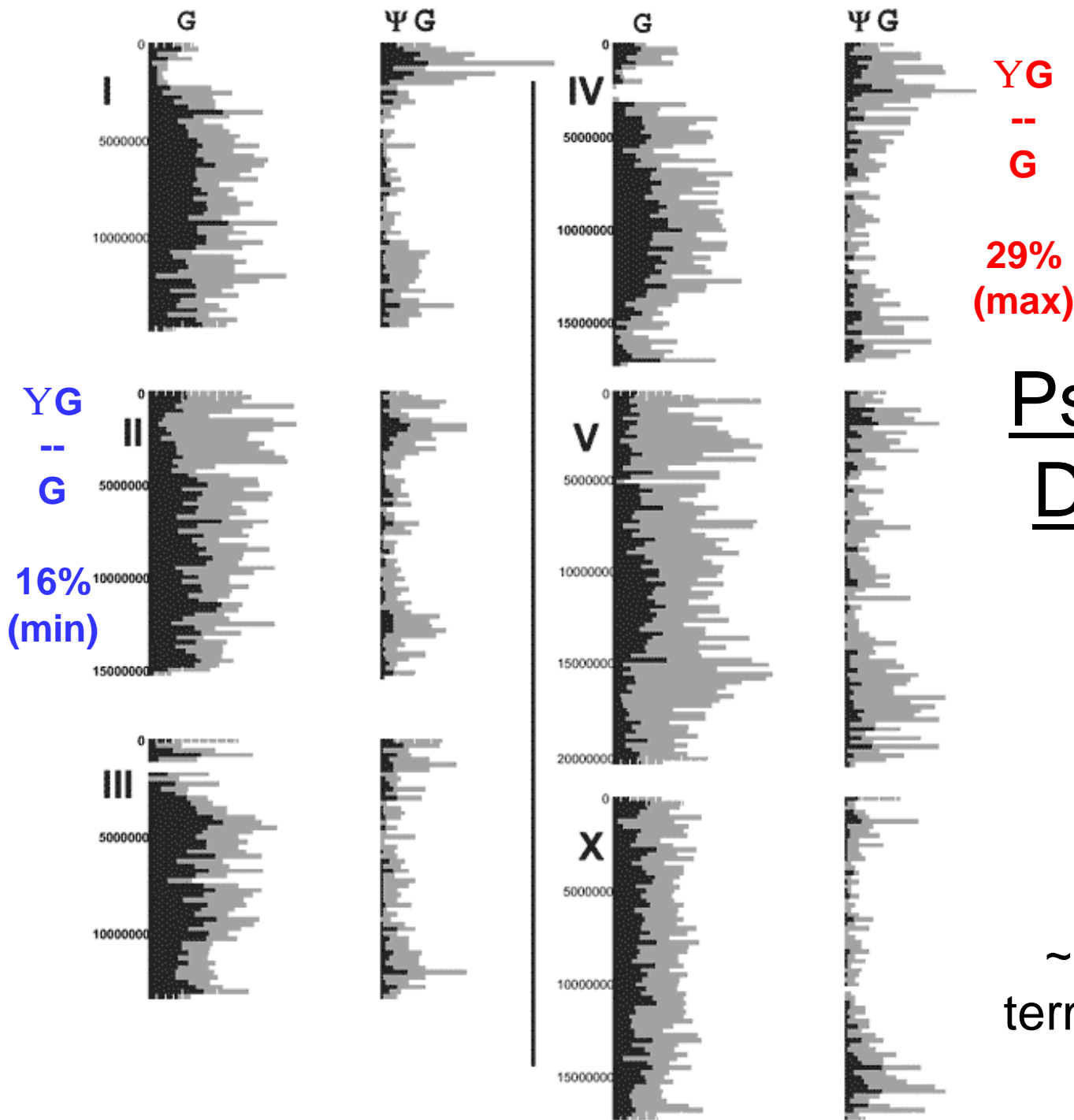


# Most Common Worm “Pseudofolds” #1

G Rank (Number matches)	ΨG Rank	Fold	Representative Domain, SCOP 1.39 Number, Description	G Rank (Number matches)	ΨG Rank	Fold	Representative Domain, SCOP 1.39 Number, Description
<b>1</b> (769)	<b>2</b>		d1ajw__ 2.1 Immuno- globulin	<b>6</b> (246)	<b>8</b>		d21bd__ 1.95 Nuc. receptor ligand-binding domain
<b>2</b> (555)	<b>6</b>		d1dec__ 7.3 Knottin	<b>7</b> (243)	34		d1a17__ 1.91 Alpha/alpha superhelix
<b>3</b> (434)	<b>3</b>		d3lck__ 5.1 Protein kinase	<b>8</b> (227)	17		d1sp2__ 7.31 Classic zinc finger
<b>4</b> (302)	<b>1</b>		d1tsg__ 4.105 C-type lectin	<b>9</b> (215)	20		d1dai__ 3.29 P-loop NTP hydrolase
<b>5</b> (274)	<b>7</b>		d1zfo__ 7.33 Glucocorticoid receptor DNA- binding dom.	<b>10</b> (197)	13		d2aw0__ 4.34 Ferredoxin

# Most Common Worm “Pseudofolds” #2

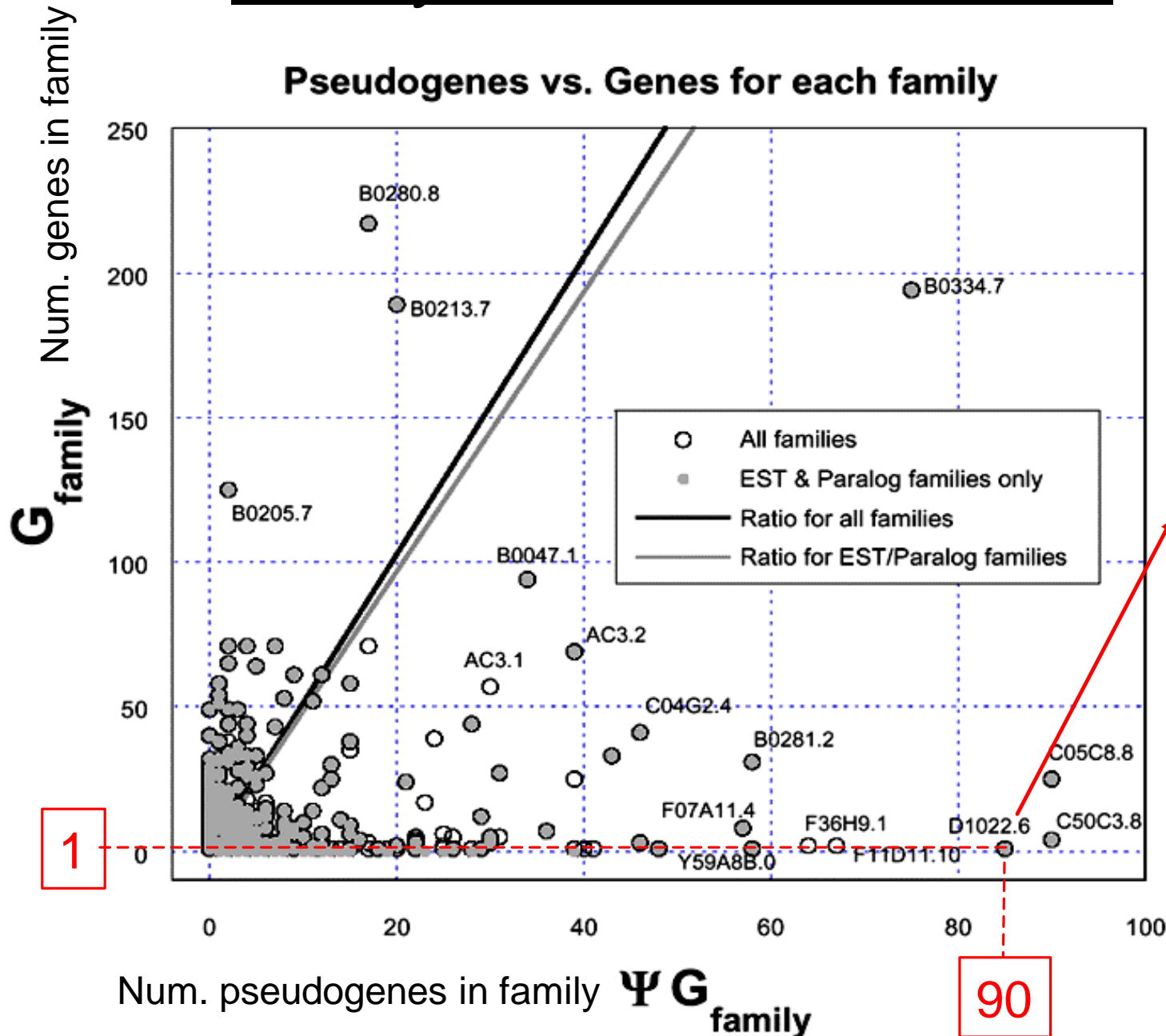
$\Psi$ G Rank (Number matches)	G Rank	Fold	Representative Domain, SCOP 1.39 Number, Description	$\Psi$ G Rank (Number matches)	G Rank	Fold	Representative Domain, SCOP 1.39 Number, Description
<b>1</b> (39)	<b>4</b>		d1tsg__ 4.105 C-type lectin	<b>6</b> (18)	<b>2</b>		d1dec__ 7.3 Knottin
<b>2</b> (32)	<b>1</b>		d1ajw__ 2.1 Immunoglobulin	<b>7</b> (17)	<b>5</b>		d1zfo__ 7.33 Glucocorticoid receptor DNA-binding dom.
<b>3</b> (27)	<b>3</b>		d3lck__ 5.1 Protein kinase	<b>8</b> (15)	<b>6</b>		d21bd__ 1.95 Nuc. receptor ligand-binding domain
<b>4</b> (25)	11		d1cvl__ 3.56 Alpha/beta-hydrolase	<b>9</b> (13)	58		d1bus__ 7.14 Ovomucoid PCI inhibitor fold
<b>5</b> (23)	63		d1ako__ 4.93 DNase-I fold	<b>9</b> (13)	19		d2bnh__ 3.7 Leu-rich, right-handed beta/alpha superhelix



# Pseudogene Distribution on Chomo- somes

~50% ΨG in  
terminal 3Mb vs  
~30% G

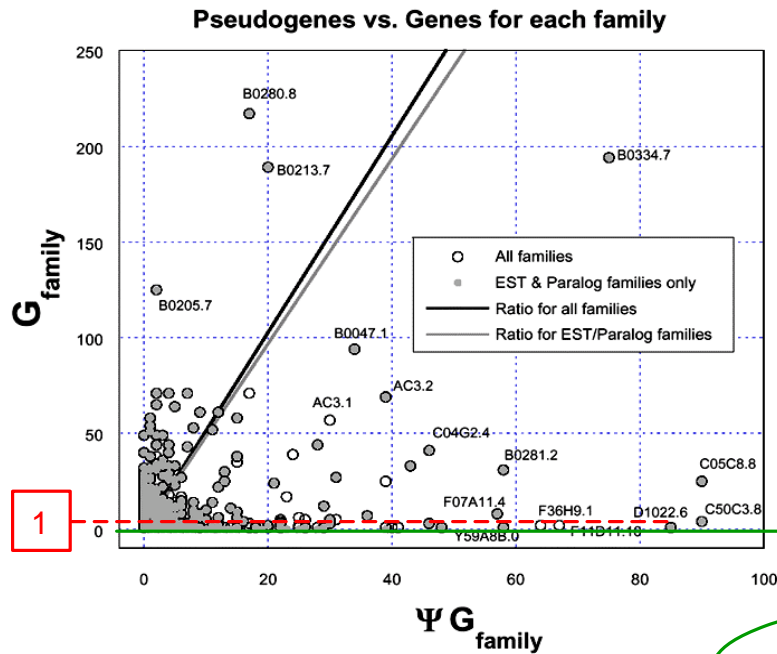
# Decayed Lines of Genes?



**D1022.6** has 90 dead fragments of itself – a disused line of chemo-receptors?

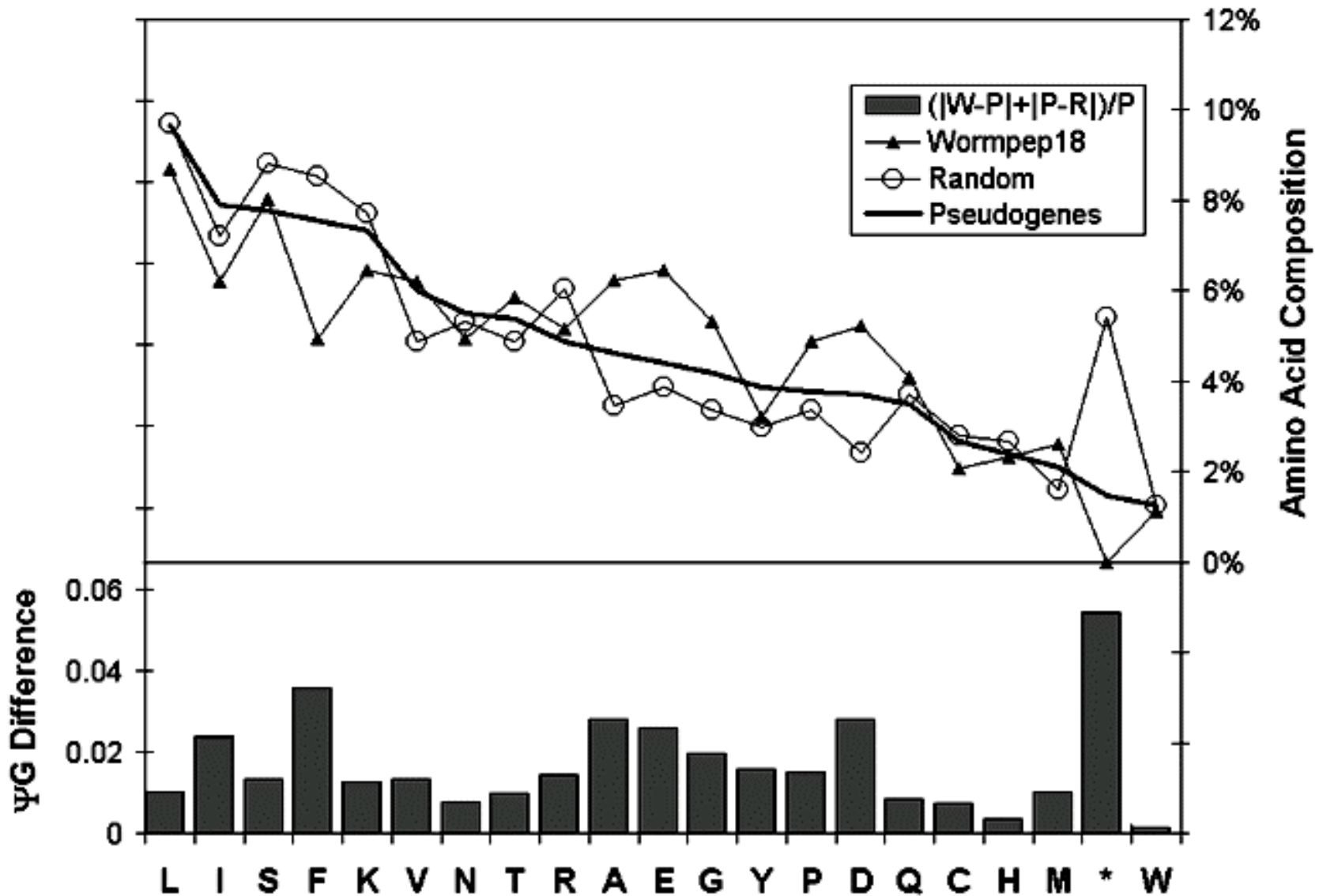


# Completely Dead Families



Rank	Number matches	Organism of closest match*	PROTOMAP family representative	Notes on representative
#1	7 *****	Yeast	YJA7_YEAST	Hypothetical protein in yeast
#2 =	5 *****	Human	XPD_MOUSE	Xeroderma pigmentosum group D complementing protein
#2 =	5 *****	Cow	CPSA_BOVIN	Cleavage and polyadenylation specificity factor
#4 =	4 ****	Frog	THB_RANCA	Thyroid hormone receptor beta
#4 =	4 ****	Human	SEX_HUMAN	SEX gene
#4 =	4 ****	Fly	MDR1_RAT	Multidrug resistance protein 1
#7 =	3 ***	Vaccinia virus	YVFB_VACCC	Hypothetical vaccinia virus protein
#7 =	3 ***	Fly	VHRP_VACCC	Host range protein from vaccinia
#7 =	3 ***	Human	IF4V_TOBAC	Eukaryotic initiation factor 4A
#7 =	3 ***	<i>E. coli</i>	ACRR_ECOLI	Acrab operon repressor

# Amino Acid Composition of Pseudogenes is Midway between Proteins and Random



# Analysis of Genomes & Transcriptomes in terms of the Occurrence of Parts & Features

## 1 Using Parts to Interpret Genomes.

Shared and/or unique parts. Venn Diagrams, Fold tree with all- $\beta$  diff. Ortholog tree. Top-10 folds.

## 2 Using Parts to Interpret Pseudogenomes.

In worm, top  $\Psi$ -folds (DNase, hydrolase) v top-folds (lg). chr. IV enriched, dead and dying families (90YG v 1G)

## 3 Using Parts to Interpret Transcriptomes:

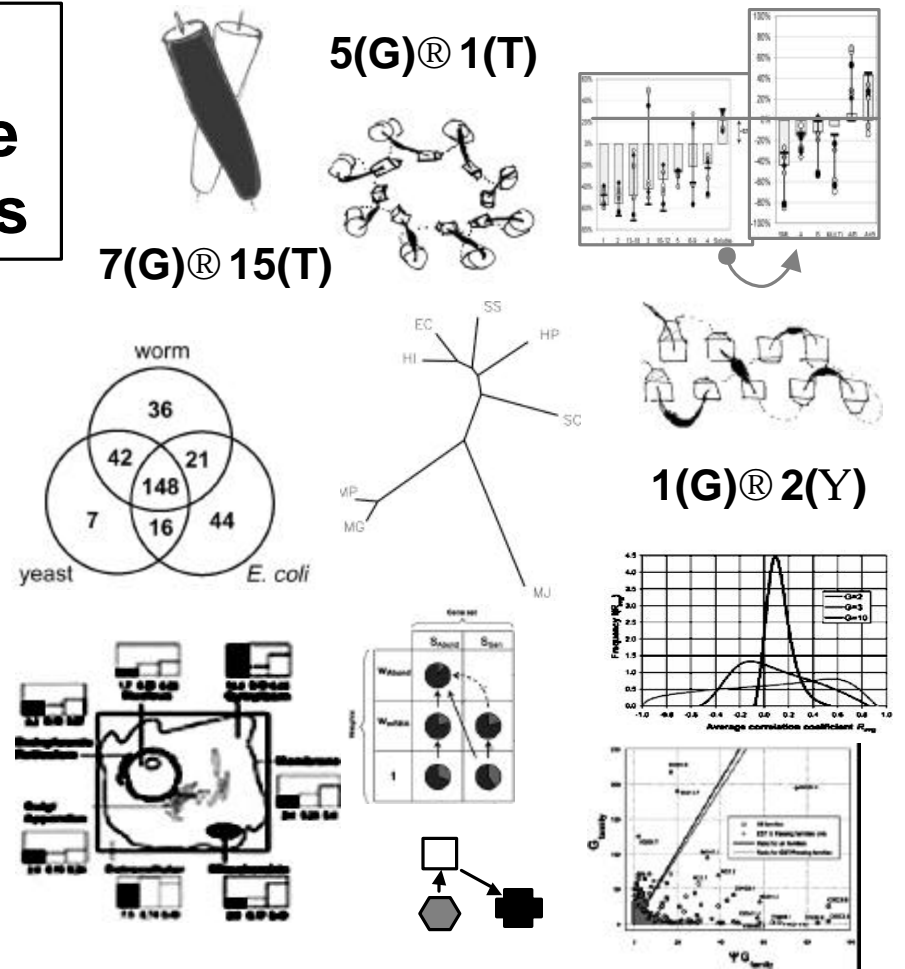
**Expression & Structure.** Top-10 parts in mRNA. Enriched in transcriptome:  $\alpha\beta$  folds, energy, synthesis, TIM fold, VGA. Depleted: TMs, transport, transcription, Leu-zip, NS. Compare with prot. abundance.

## 4 Expression & Localization.

Enriched : Cytoplasmic. Depleted: Nuclear. Bayesian localizer

## 5 Expression & Function.

Expression relates to structure & localization but to function, globally? P-value formalism. Weak relation to protein-protein interactions.

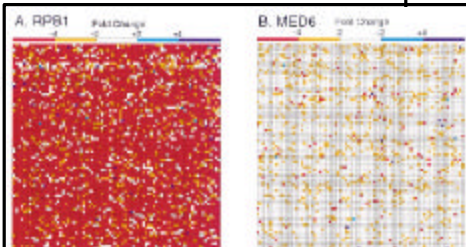


[bioinfo.mbb.yale.edu](http://bioinfo.mbb.yale.edu)

**H Hegyi, J Lin, B Stenger,**  
**P Harrison, N Echols,**  
**R Jansen, A Drawid, J Qian,**  
**D Greenbaum, M Snyder**

### Dissecting the Regulatory Circuitry of a Eukaryotic Genome

Frank C. P. Holsteg,<sup>1</sup> Ezra G. Jennings,<sup>1</sup> John J. Wyrick,<sup>1</sup> Tong Ihn Lee,<sup>1</sup> Christoph J. Hangartner,<sup>1</sup> Michael R. Green,<sup>1</sup> Todd R. Golub,<sup>2</sup> Eric S. Lander,<sup>1</sup> and Richard A. Young<sup>1</sup>  
<sup>1</sup>Whitehead Institute for Biomedical Research, Cambridge, Massachusetts 02142  
<sup>2</sup>Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139  
<sup>3</sup>Howard Hughes Medical Institute Program in Molecular Medicine, University of Massachusetts Medical Center



# Young, Church... Affymetrix GeneChips Abs. Exp.

regulation which is superimposed on that due to specific transcription factors, a novel mechanism for the regulation of specific sets of genes

Figure 2. Genome-Wide Expression Data for Selected Components of the RNA Polymerase II Holoenzyme  
Genes in which levels were increased in response to the inorganic salt ligand (as indicated by a grid format) in left and right square regions. In the left square region, the left-most gene is chromosome 1, and the square to its right contains a sequence of genes in tandem through chromosome 1, from 1, 2, 3, etc., etc., with the last gene on the right side of chromosome 1X in shaded gray. The results are shown for (A) RPB1, (B) MED6, (C) RPB1, and (D) RPB1.

**The Brown Lab**  
Stanford University Department of Biochemistry

**The MGuide**  
The Complete Guide to MicroArrays  
Build your own arrayer and scanner!

The transcriptional program in the response of human fibroblasts to serum

use it with that obtained by its inactivation. Comparison of the two data sets reveals that expression decay kinetics in yeast is more like that in mammalian cells. Proteolytic cleavage of the web site for Med6, indicated by its disruption of Med6 when an could be made, the mRNA of 196 (with similar kinetics in the Med6 and 1). Thus, the expression of 10% of yeast genes on Med6 as they are on Rpb1 are most likely to have a direct regulatory function. The genes whose transcript the Rpb1 kinetics could have a direct effect for Med6 function, or the effects a gene are a secondary consequence of some other gene's altered mRNA levels. The 500 genes we have identified that require Med6 function to the same extent as Rpb1 function are those which protein-associated transcriptional regulators are most likely to function through interactions with Med6. Srb5 is a component of the Srbmediator complex whose function is also not known (Thompson et al., 1992; Kim et al., 1994; Kozlov and Young, 1996; Hangartner et al., 1998; Myers et al., 1998). To determine the genome-wide dependence of gene expression on Srb5, a strain lacking an SRB5 gene and its end-origo counterpart, were compared to the web site for detailed information. The results indicate that 16% of all genes require Srb5 function for their expression. With the SRB5 deletion strain and other constitutive mutants

# Brown, marrays, Rel. Exp. over Timecourse

# Also: SAGE (mRNA); 2D gels for Protein Abundance (Aebersold, Futcher)

# Gene Expression Datasets: the Yeast Transcriptome

## Yeast Expression Data: 6000 levels! Integrated Gene Expression Analysis System: X-ref. Parts and Features against expression data...

Proc. Natl. Acad. Sci. USA  
Vol. 94, pp. 190-195, January 1997  
Genetics

### A multipurpose transposon system for analyzing protein production, localization, and function in *Saccharomyces cerevisiae*

PETRA ROSS-MACDONALD, AMY SHEEHAN, G. SEIBLIEEN ROEDER, AND MICHAEL SNYDER\*

Department of Biology, Yale University, P.O. Box 2080, New Haven, Connecticut 06510

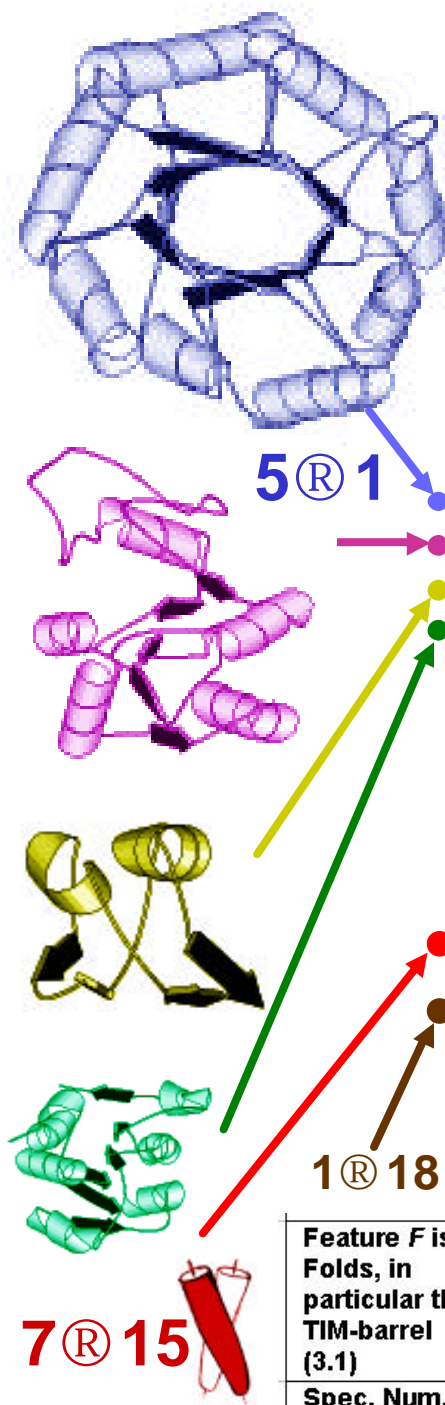
**ABSTRACT** Analysis of the function of a particular product typically involves determining the expression pattern of the gene, the subcellular location of the protein, and phenotype of a null strain lacking the protein. Conditional alleles of the gene are often created as an additional tool to have developed a multifunctional, transposon-based system that simultaneously generates constructs for all the analyses and is suitable for mutagenesis of any given *Saccharomyces cerevisiae* gene. Depending on the transposon used, yeast gene is fused to a coding region for  $\beta$ -galactosidase, green fluorescent protein. Gene expression can thereby monitored by chemical or fluorescence assays. The transposons create insertion mutations in the target gene, also phenotypic analysis. The transposon can be reduced by site-specific recombination to a smaller element that leaves a unique tag inserted in the encoded protein. In addition, utility for a variety of immediate purposes, the system

**INTRODUCTION** The ability to analyze the function of a particular product typically involves determining the expression pattern of the gene, the subcellular location of the protein, and phenotype of a null strain lacking the protein. Conditional alleles of the gene are often created as an additional tool to have developed a multifunctional, transposon-based system that simultaneously generates constructs for all the analyses and is suitable for mutagenesis of any given *Saccharomyces cerevisiae* gene. Depending on the transposon used, yeast gene is fused to a coding region for  $\beta$ -galactosidase, green fluorescent protein. Gene expression can thereby monitored by chemical or fluorescence assays. The transposons create insertion mutations in the target gene, also phenotypic analysis. The transposon can be reduced by site-specific recombination to a smaller element that leaves a unique tag inserted in the encoded protein. In addition, utility for a variety of immediate purposes, the system

**RESULTS** The system was used to analyze the function of a particular product typically involves determining the expression pattern of the gene, the subcellular location of the protein, and phenotype of a null strain lacking the protein. Conditional alleles of the gene are often created as an additional tool to have developed a multifunctional, transposon-based system that simultaneously generates constructs for all the analyses and is suitable for mutagenesis of any given *Saccharomyces cerevisiae* gene. Depending on the transposon used, yeast gene is fused to a coding region for  $\beta$ -galactosidase, green fluorescent protein. Gene expression can thereby monitored by chemical or fluorescence assays. The transposons create insertion mutations in the target gene, also phenotypic analysis. The transposon can be reduced by site-specific recombination to a smaller element that leaves a unique tag inserted in the encoded protein. In addition, utility for a variety of immediate purposes, the system



# Common Parts: the Transcriptome

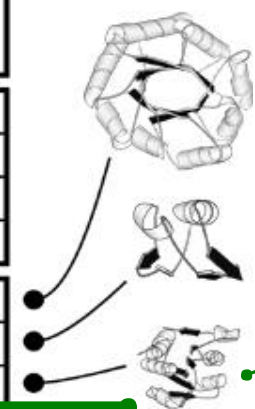


Fold	Fold Class	Rep. PDB	Composition			Rank										
			Genome [%]	Transcriptome [%]	Rel. Diff. [%]	Genome	Young	Samson	Church-a	Church-alpha	Church-gal	Church-heat	SAGE-GM	SAGE-L	SAGE-S	
TIM barrel	$\alpha \beta$	1byb	4.2	8.3	+98	5	1	1	1	1	1	1	1	1	1	1
P-loop NTP hydrolases	$\alpha \beta$	1gky	5.8	5.2	-11	3	2	2	4	4	4	5	5	6	7	
Ferredoxin like	$\alpha\beta$	1fxd	3.9	3.4	-14	6	3	7	11	9	8	10	4	10	11	
Rossmann fold	$\alpha \beta$	1xel	3.3	3.3	0	8	4	3	3	3	2	2	19	15	9	
7-bladed beta-propeller	$\beta$	1mda*	6.4	2.9	-55	2	5	4	5	6	6	7	9	9	16	
alpha-alpha superhelix	$\alpha$	2bct	4.4	2.7	-37	4	6	11	15	16	12	12	8	5	8	
Thioredoxin fold	$\alpha \beta$	2trx	1.7	2.7	+63	14	7	6	8	2	5	4	11	10	6	
G3P dehydrogenase-like	$\alpha\beta$	1drwt	0.2	2.7	+1316	81	8	12	2	5	3	3	35	19	30	
beta grasp	$\alpha\beta$	1igd	0.6	2.6	+348	36	9	10	21	9	18	21	82	122	120	
HSP70 C-term. fragment	multi	1dky	0.8	2.6	+231	31	10	16	17	11	16	12	48	25	56	
Leu-zipper	$\alpha$	1zta	3.8	2.1	-46	7	15	8	14	21	15	19	21	20	33	
Protein kinases (cat. core)	multi	1hcl	6.8	1.6	-77	1	18	19	9	16	11	15	13	16	17	
alpha/beta hydrolases	$\alpha \beta$	2ace	2.2	0.9	-62	10	32	31	25	26	21	23	26	26	26	
Zn2/C6 DNA-bind. dom.	sml	1aw6	2.6	0.3	-89	9	75	94	27	50	32	40	48	39	50	

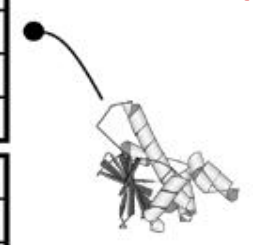
Feature F is Folds, in particular the TIM-barrel (3.1)	Number of TIM-barrel fold matches in yeast genome	Number of matches with all folds in yeast genome	Genome composition of TIM-barrel fold matches	Number of TIM-barrel fold matches weighted by expression	Number of matches with all folds weighted by expression	Transcriptome composition of TIM-barrel fold matches	Relative enrichment of TIM-barrel matches in transcriptome
Spec. Num.	65	1560	4.2%	389	4709	8.3%	97.8%

Fold of	Freq.		Change					Rep. PDB
	Genome	Transcriptome	CDC28	CDC15	Diauxic Shift	Sporulation	E. coli heat shock	
Protein kinases (cat. core)	1	18	94	98	139	60	100	1p38
β-propeller	2	5	160	108	109	82	-	1mda
P-loop NTP hydrolases	3	2	100	88	91	57	39	1gky
α-α superhelix	4	6	136	90	110	44	55	2bct
TIM-barrel	5	1	58	57	39	24	91	1byb
Ferredoxin-like	6	3	135	61	63	70	144	1fxd
Rossmann fold	8	4	55	99	43	56	92	1xel
Ribonucleotide reductase (R1)	100	143	1	-	-	-	35	1rlr
ATPase dom. of HSP90	100	91	2	4	72	73	2	1ah6
Homing endonuclease-like	130	164	3	136	85	175	41	1af5
Aminoacid dehydrogenases; dim. dom.	-	-	4	169	121	3	51	1hup
DNA topo I (N-term)	-	-	175	1	148	126	-	1ois
DNA clamp	130	115	8	2	87	11	60	2pol
Metallothionein	100	14	89	3	33	12	-	1mhu
Phosphoenolpyruvate carboxykinase	130	190	51	26	1	98	169	1ayl
Citrate synthase	81	120	14	8	2	28	51	1csh
N-carbamoylsarcosine amidohydrolase	130	112	9	-	3	138	118	1nba
TBP-like	81	91	46	38	4	75	100	1bv1
5'-3' exonuclease	67	150	32	125	162	1	157	1tfr
α/α toroid	62	132	169	145	114	2	100	1gai
Cyclin-like	20	61	20	15	129	4	-	1vin
ATPase domain of GroEL	36	34	183	143	61	151	1	1aon
Head domain of GrpE	130	135	196	31	165	165	3	1dkg
HSP70 (C-term)	31	10	16	11	58	117	4	1dkz

Common Folds



Folds that change a lot in frequency

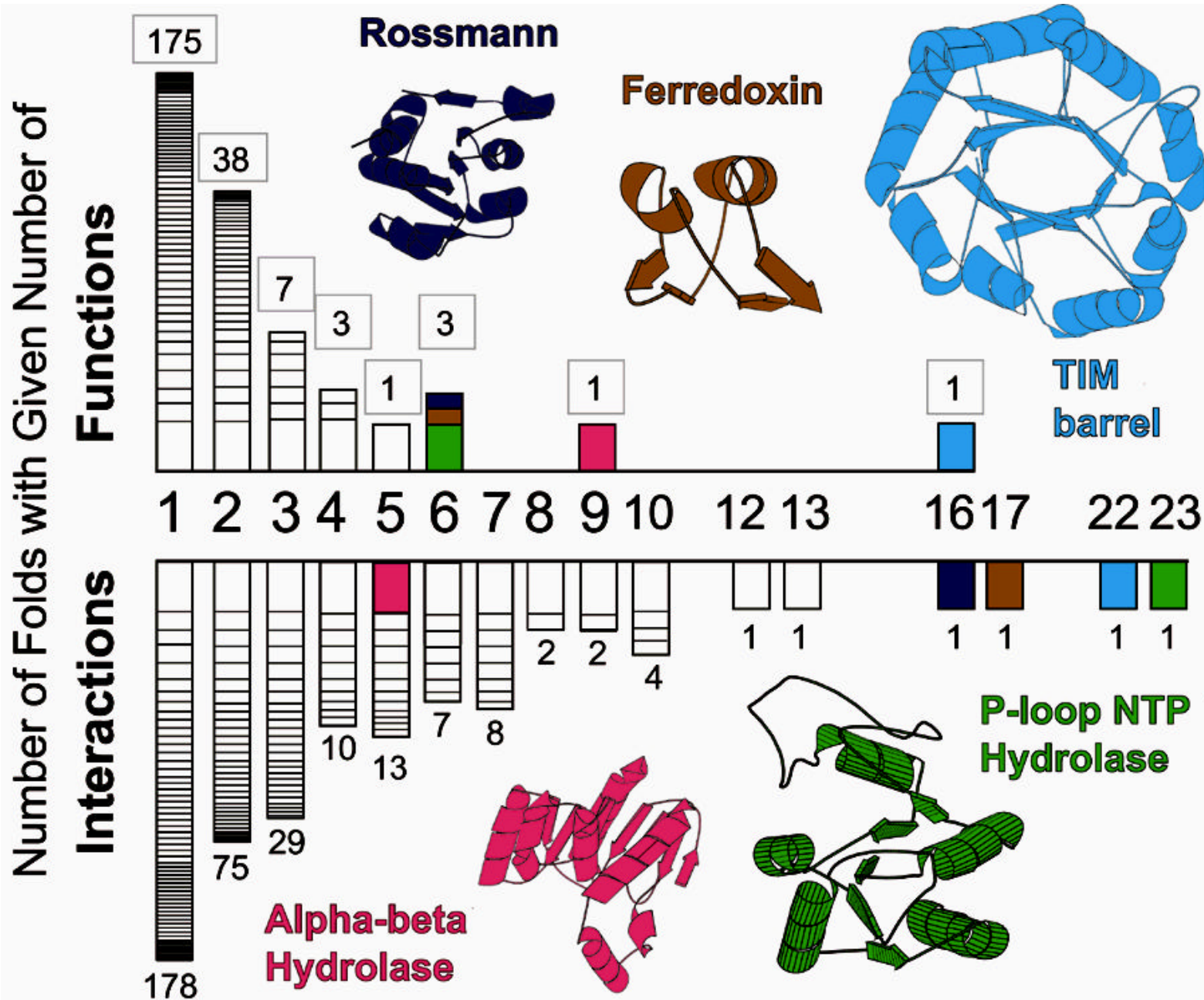


are not common



Changing Folds

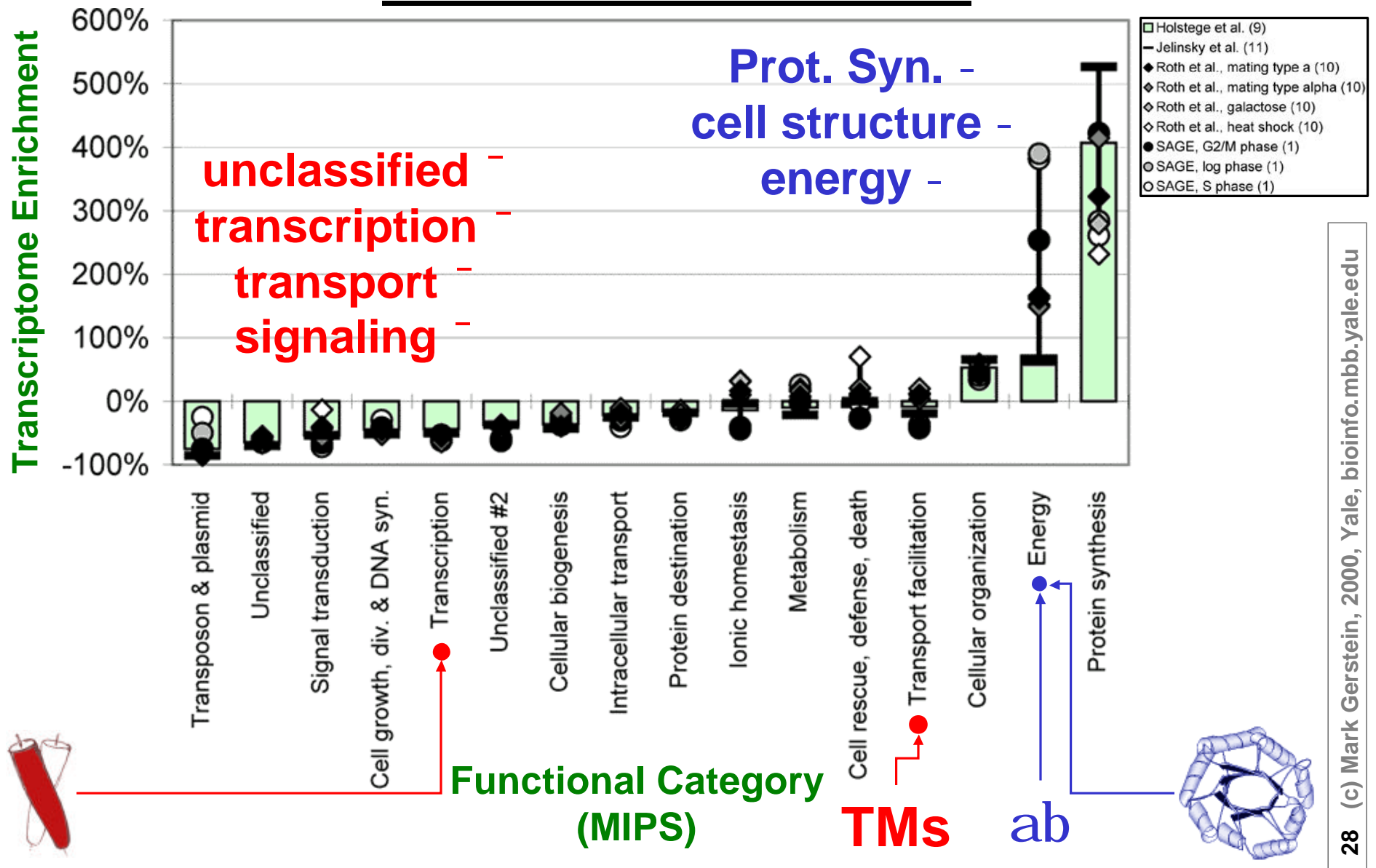
# Most Versatile Folds – Relation to Interactions



Similar results  
Martin et al.  
(1998)

The number of interactions for each fold = the number of other folds it is found to contact in the PDB

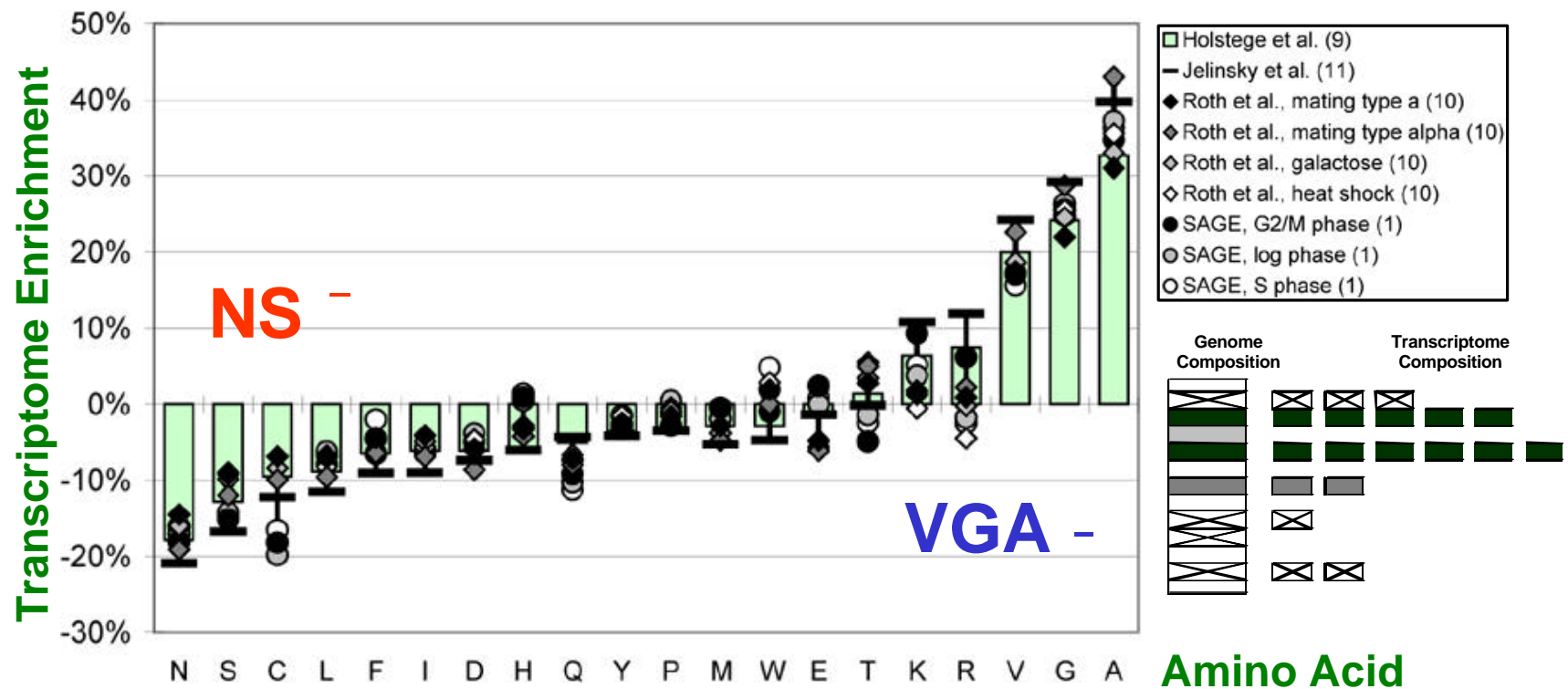
# Composition of Transcriptome in terms of Functional Classes





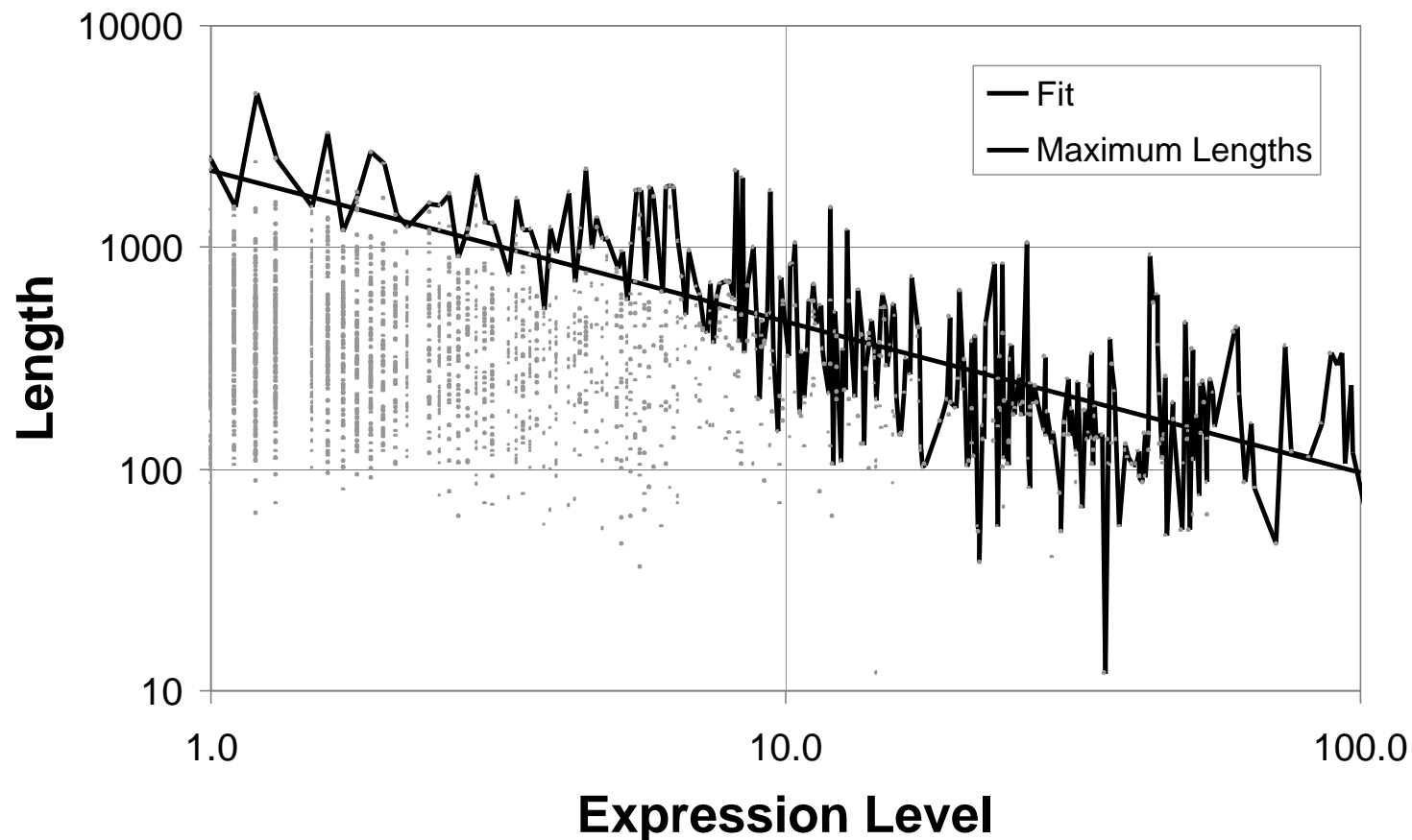
# Composition of Genome vs. Transcriptome

	$\sum_{\text{orf } i} n_i(F)$	$\sum_F \sum_{\text{orf } i} n_i(F)$	$G(F)$	$\sum_{\text{orf } i} e_i n_i(F)$	$\sum_F \sum_{\text{orf } i} e_i n_i(F)$	$T(F)$	$D(F)$
<b>Feature F is Amino acids, in particular Ala</b>	Number of Ala in yeast	Number of amino acids in yeast	Genome composition of Ala in yeast	Number of Ala weighted by expression	Number of amino acids weighted by expression	Transcriptome composition of Ala in yeast	Relative enrichment of Ala in transcriptome
<b>Spec. Num.</b>	<b>141890</b>	<b>2574876</b>	<b>5.5%</b>	<b>347807</b>	<b>4758441</b>	<b>7.3%</b>	<b>32.7%</b>
<b>Feature F is Folds, in particular the TIM-barrel (3.1)</b>	Number of TIM-barrel fold matches in yeast genome	Number of matches with all folds in yeast genome	Genome composition of TIM-barrel fold matches	Number of TIM-barrel fold matches weighted by expression	Number of matches with all folds weighted by expression	Transcriptome composition of TIM-barrel fold matches	Relative enrichment of TIM-barrel matches in transcriptome
<b>Spec. Num.</b>	<b>65</b>	<b>1560</b>	<b>4.2%</b>	<b>389</b>	<b>4709</b>	<b>8.3%</b>	<b>97.8%</b>

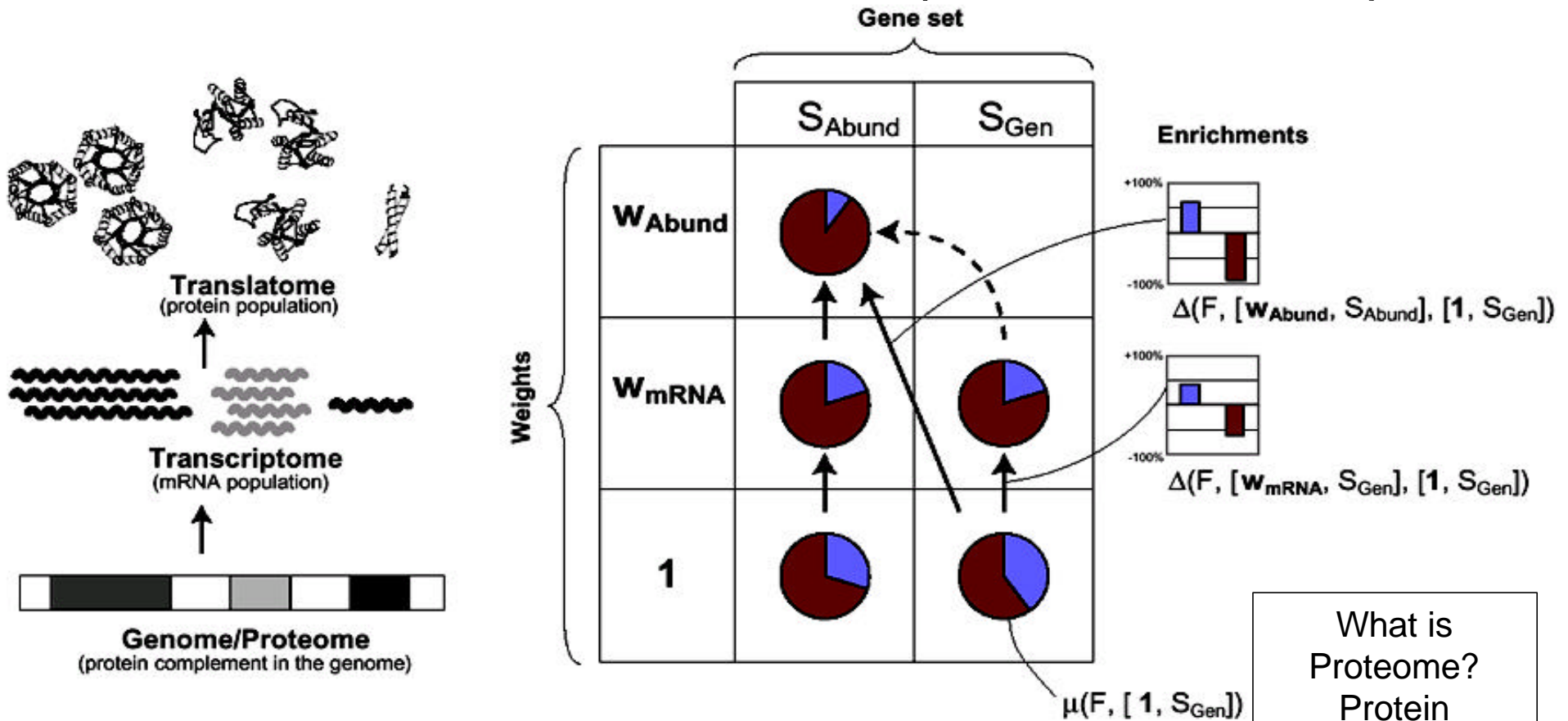


# Relation between Length & Expression

Max Expression (e.g. transcripts/cell)  $\sim$  (Length)<sup>-2/3</sup>  
Shorter proteins can be more highly expressed



# Relating the Transcriptome to Cellular Protein Abundance (Translatome)

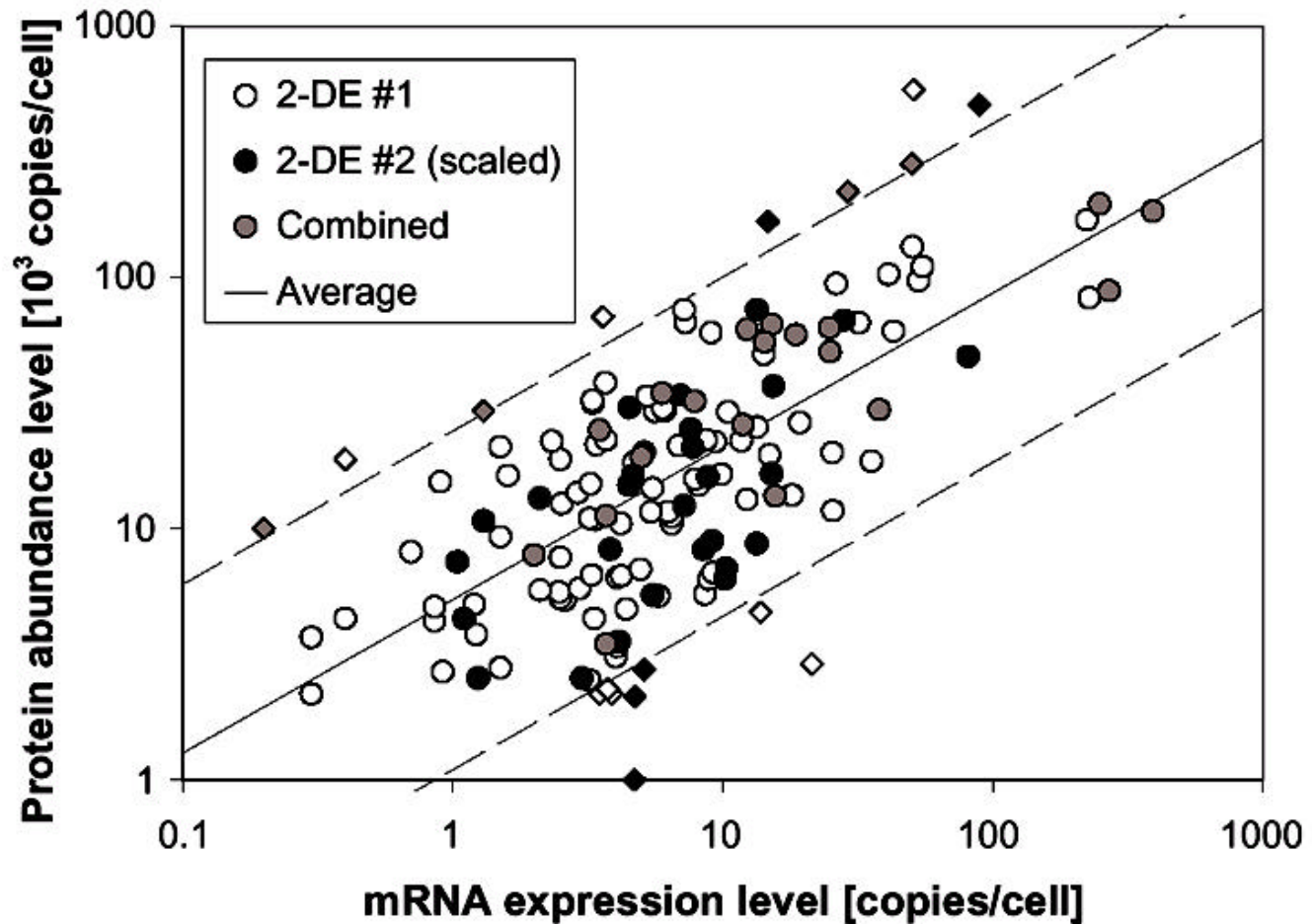


2D-gel electrophoresis Data sets: Futcher (71), Aebersold (156), scaled set with 171 proteins  
 New effect is dealing with gene selection bias

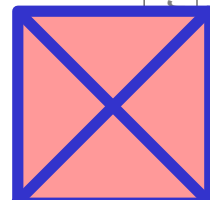
What is Proteome?  
 Protein complement in genome or cellular protein population?

# mRNA and protein abundance related, roughly

~150 protein abundance values from merging results of 2D gel expts. of Aebersold & Futcher

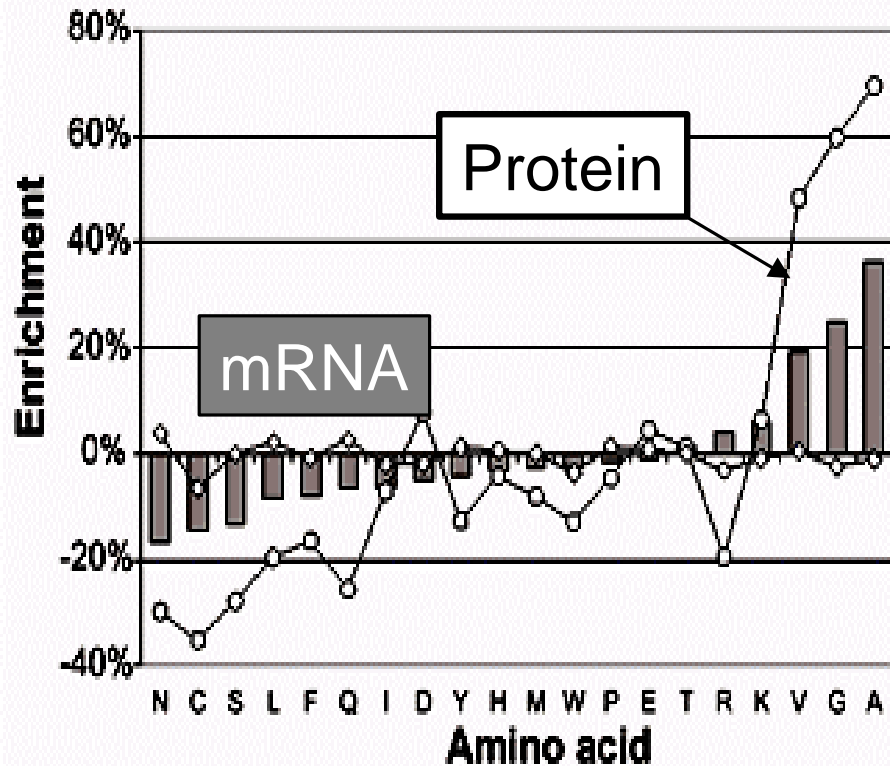


mRNA values for same 150 genes from merging and scaling 6 yeast expressions

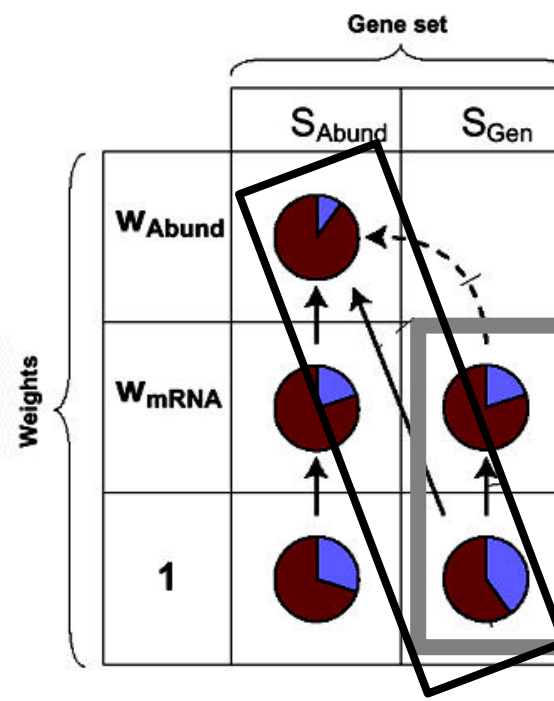




# Amino Acid Enrichment

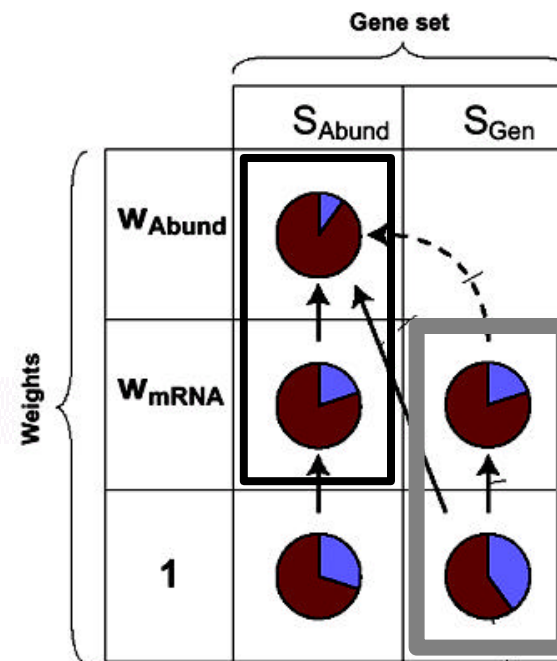
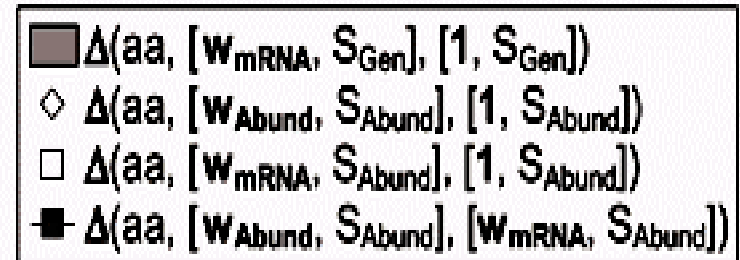
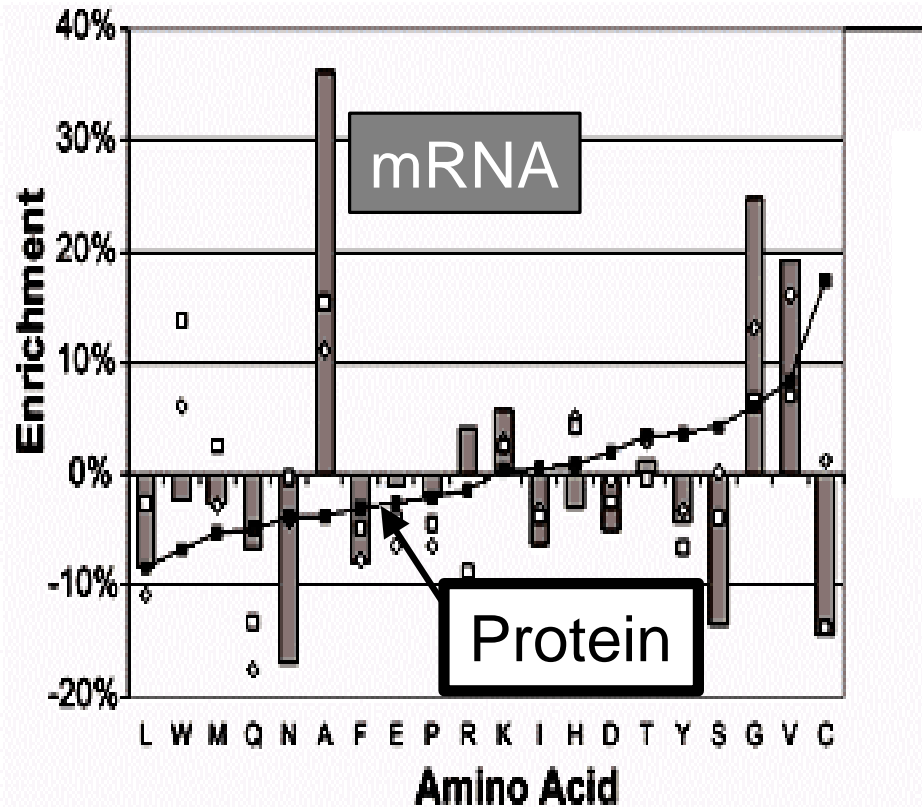


$$\begin{aligned} & \blacksquare \Delta(aa, [W_{mRNA}, S_{Gen}], [1, S_{Gen}]) \\ & \circ \Delta(aa, [W_{Abund}, S_{Gen}], [1, S_{Gen}]) \\ & \diamond \Delta(aa, [W_{\beta-Gal}, S_{\beta-Gal}], [1, S_{Gen}]) \end{aligned}$$



Simple story is transcriptome is enriched in same way as proteome

# Amino Acid Enrichment – Complexities



# Analysis of Genomes & Transcriptomes in terms of the Occurrence of Parts & Features

## 1 Using Parts to Interpret Genomes.

Shared and/or unique parts. Venn Diagrams, Fold tree with all- $\beta$  diff. Ortholog tree. Top-10 folds.

## 2 Using Parts to Interpret Pseudogenomes.

In worm, top  $\Psi$ -folds (DNase, hydrolase) v top-folds (lg). chr. IV enriched, dead and dying families (90YG v 1G)

## 3 Using Parts to Interpret Transcriptomes: Expression & Structure.

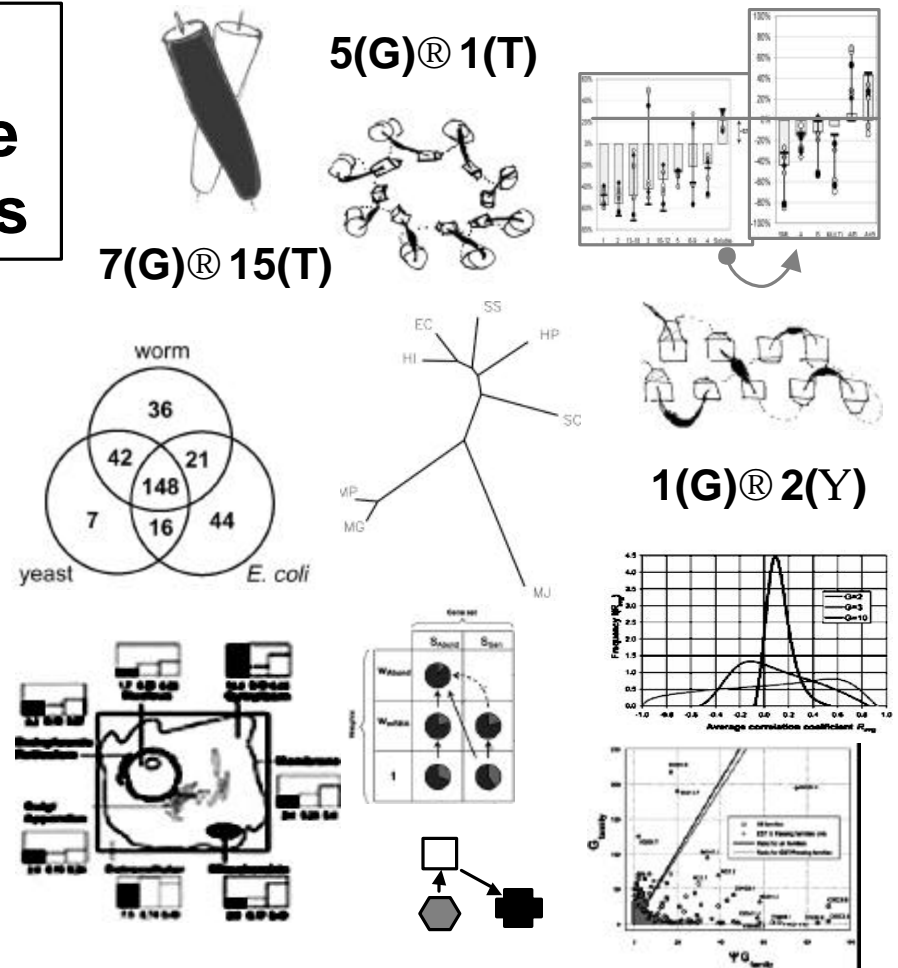
Top-10 parts in mRNA. Enriched in transcriptome:  $\alpha\beta$  folds, energy, synthesis, TIM fold, VGA. Depleted: TMs, transport, transcription, Leu-zip, NS. Compare with prot. abundance.

## 4 Expression & Localization.

Enriched : Cytoplasmic. Depleted: Nuclear. Bayesian localizer

## 5 Expression & Function.

Expression relates to structure & localization but to function, globally? P-value formalism. Weak relation to protein-protein interactions.



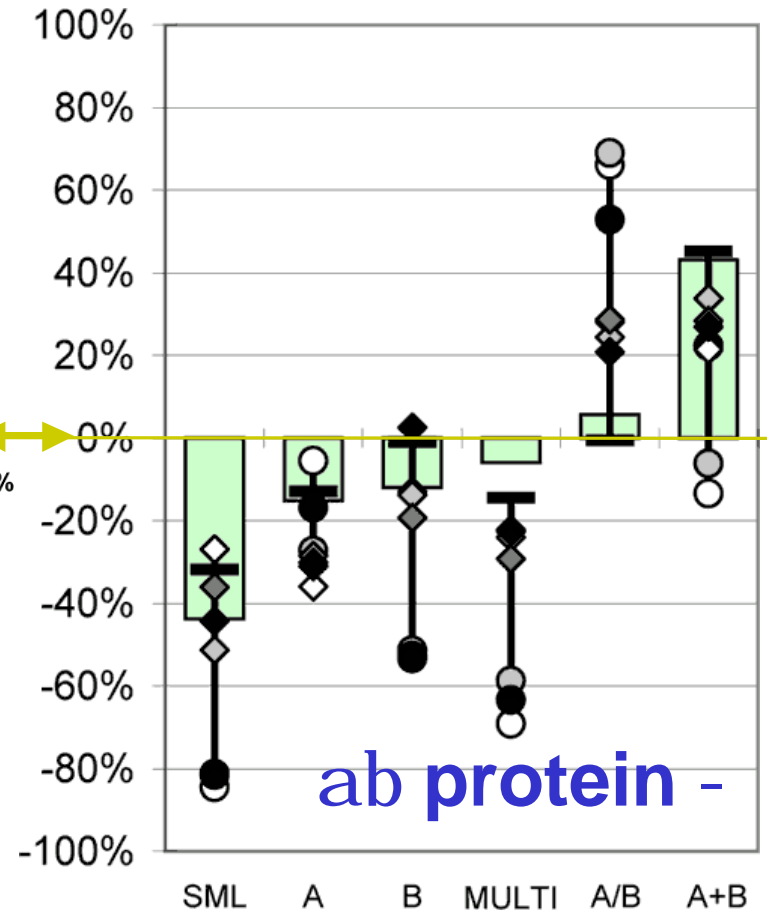
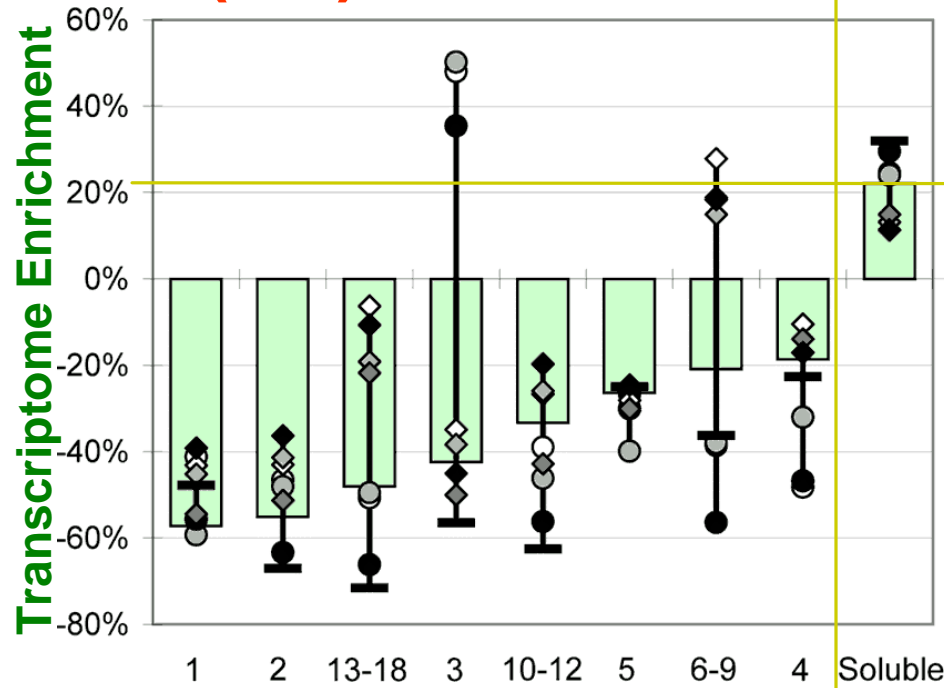
[bioinfo.mbb.yale.edu](http://bioinfo.mbb.yale.edu)

*H Hegyi, J Lin, B Stenger,  
P Harrison, N Echols,  
R Jansen, A Drawid, J Qian,  
D Greenbaum, M Snyder*

# Composition of Transcriptome in terms of Broad Structural Classes

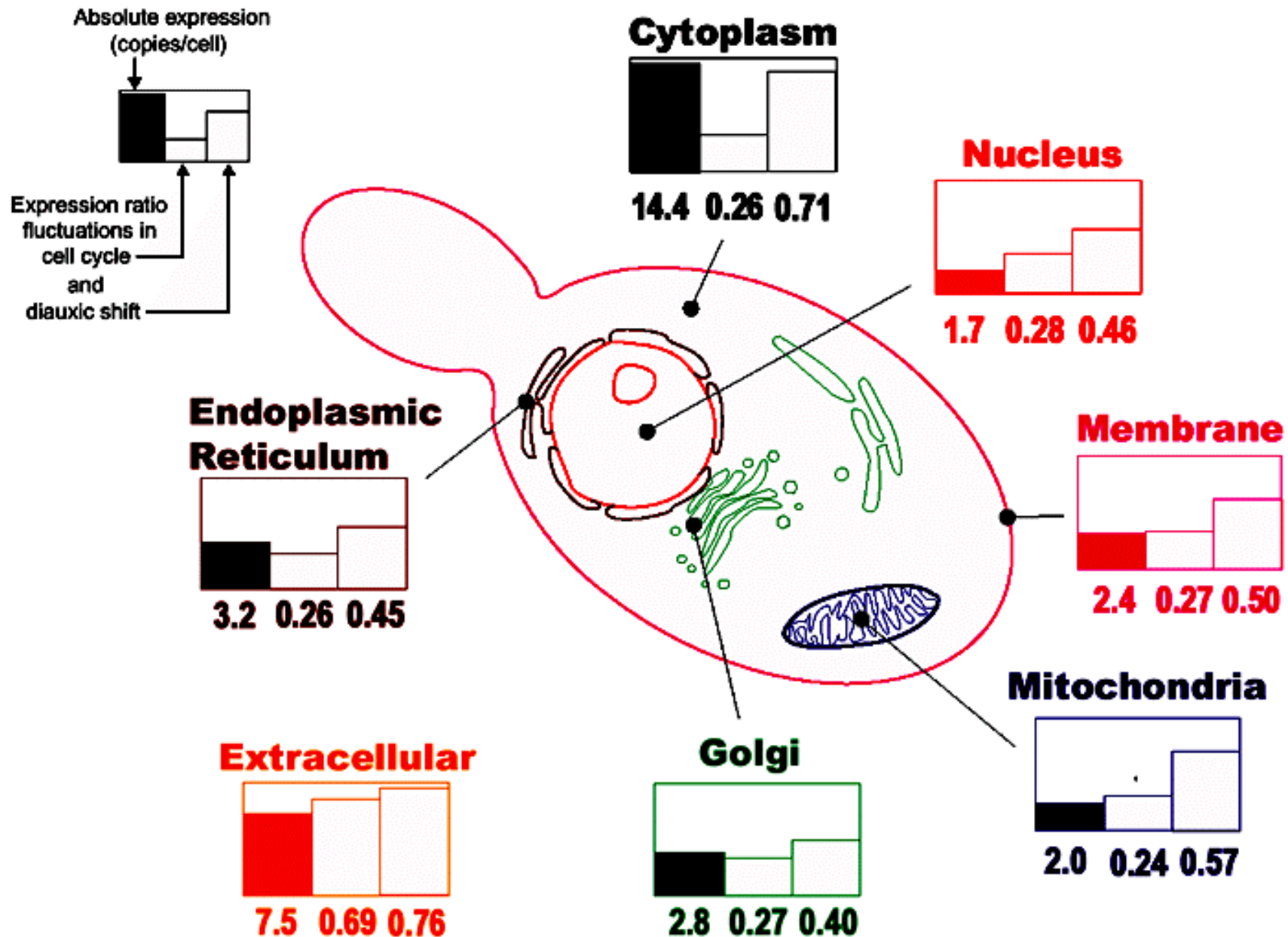
- Holstege et al. (9)
- Jelinsky et al. (11)
- ◆ Roth et al., mating type a (10)
- ◇ Roth et al., mating type alpha (10)
- ◇ Roth et al., galactose (10)
- ◇ Roth et al., heat shock (10)
- SAGE, G2/M phase (1)
- SAGE, log phase (1)
- SAGE, S phase (1)

**Membrane (TM) Protein -**

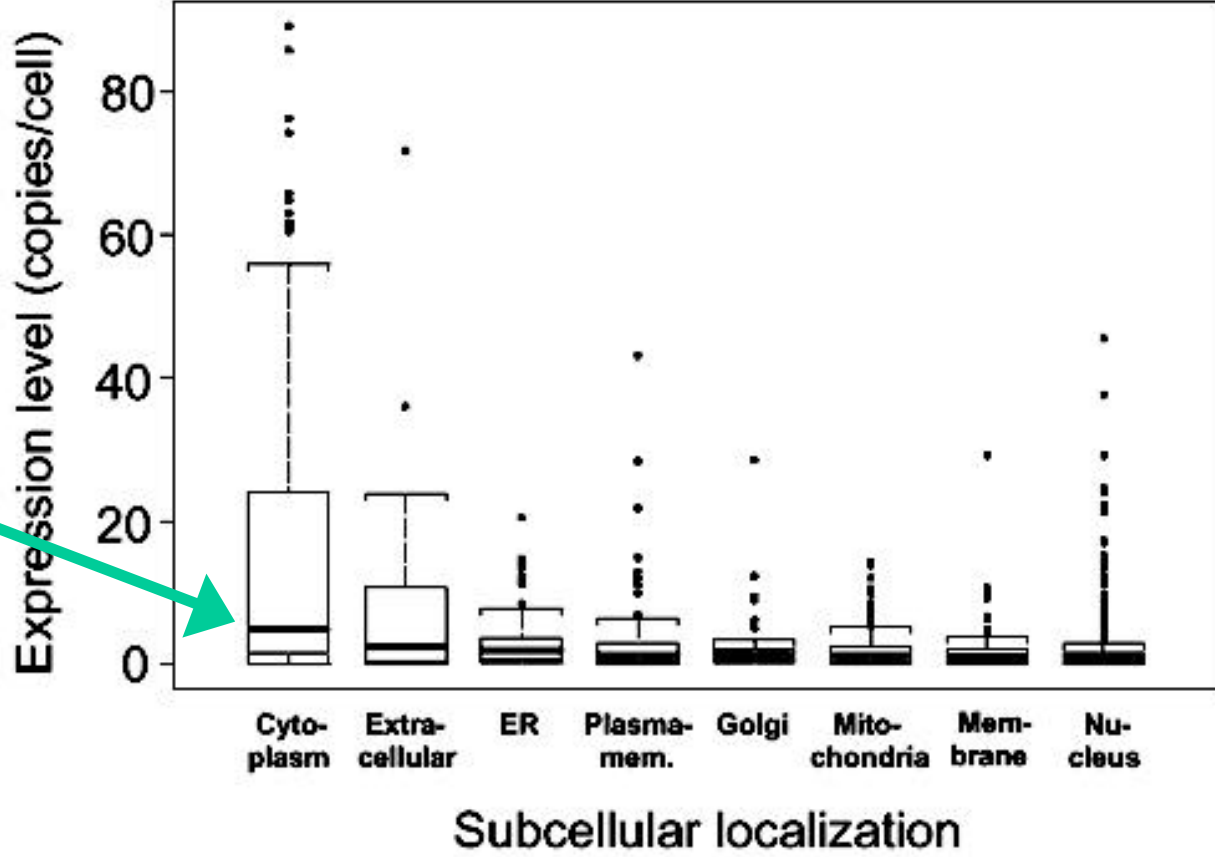
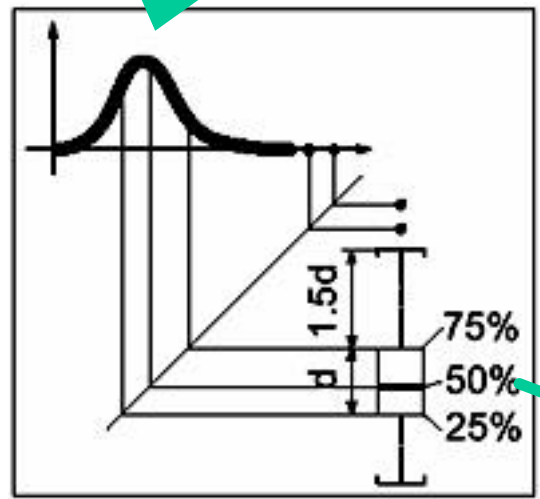
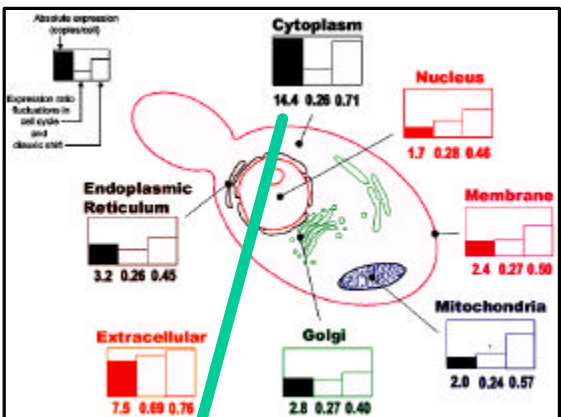




# Expression Level is Related to Localization

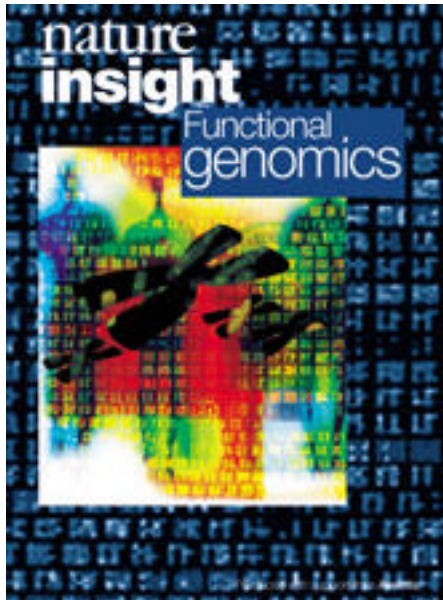


# Distributions of Expression Levels



~6000 yeast genes  
with expression levels

but only ~2000 with localization....



insight review articles

## Genomics, gene expression and DNA arrays

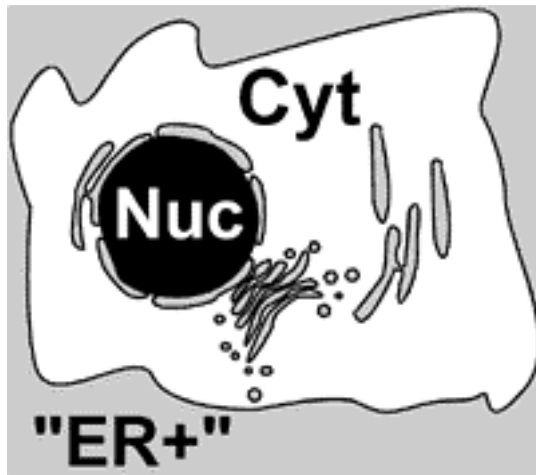
David J. Lockhart & Elizabeth A. Winzeler

*Genomics Institute of the Novartis Research Foundation, 3115 Merryfield Row, San Diego, California 92121, USA*

Experimental genomics in combination with the growing body of sequence information promise to revolutionize the way cells and cellular processes are studied. Information on genomic sequence can be used experimentally with high-density DNA arrays that allow complex mixtures of RNA and DNA to be interrogated in a parallel and quantitative fashion. DNA arrays can be used for many different purposes, most prominently to measure levels of gene expression (messenger RNA abundance) for tens of thousands of genes simultaneously. Measurements of gene expression and other applications of arrays embody much of what is implied by the term (genomics); they are broad in scope, large in scale, and take advantage of all available sequence information for experimental design and data interpretation in pursuit of biological understanding.

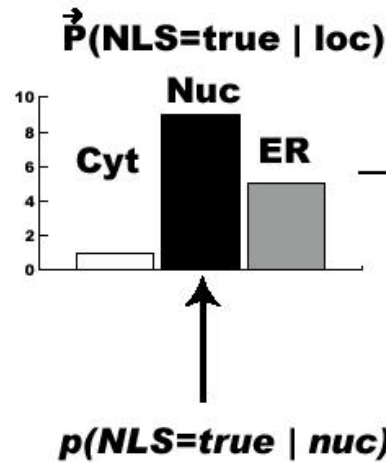
# Bayesian System for Localizing Proteins

**loc=**

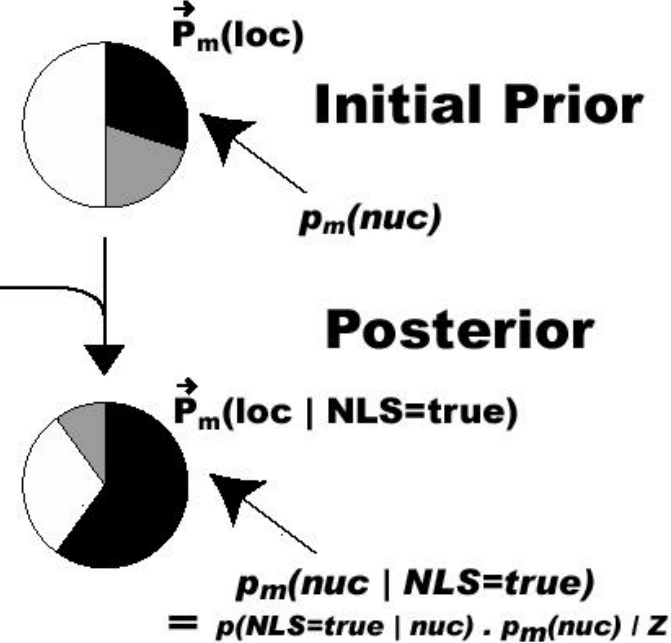


Represent localization of each protein by the state vector  $\vec{P}(\text{loc})$  and each feature by the feature vector  $\vec{P}(\text{feature}|\text{loc})$ . Use Bayes rule to update.

**Feature Vectors**  
 $\vec{P}(\text{feature}|\text{loc})$



**State Vectors**

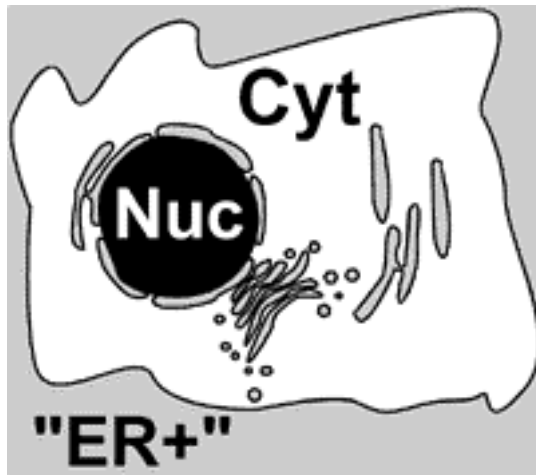


18 Features: Expression Level (absolute and fluctuations), signal seq., KDEL, NLS, Essential?, aa composition



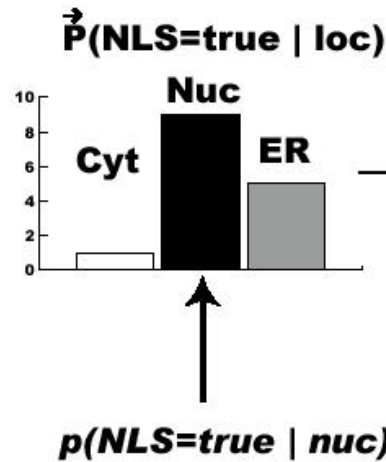
# Bayesian System for Localizing Proteins

loc=

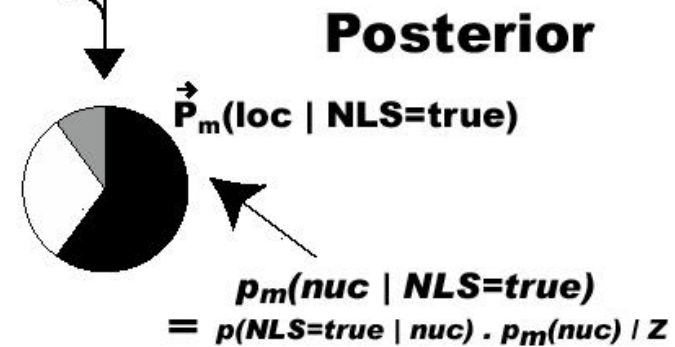
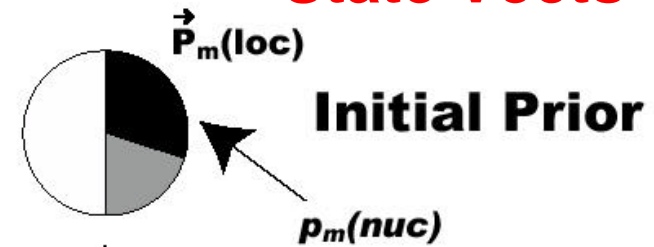


Represent localization of each protein by the state vector  $\vec{P}(\text{loc})$  and each feature by the feature vector  $\vec{P}(\text{feature}|\text{loc})$ . Use Bayes rule to update.

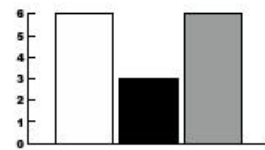
**Feature Vectors**  
 $\vec{P}(\text{feature}|\text{loc})$



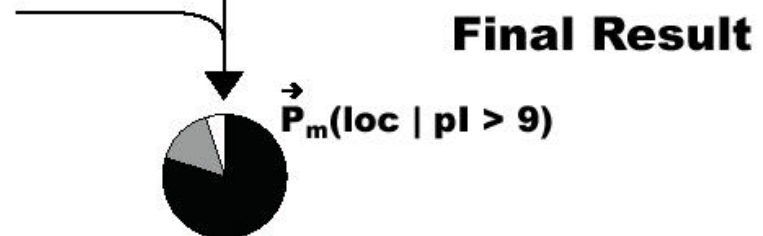
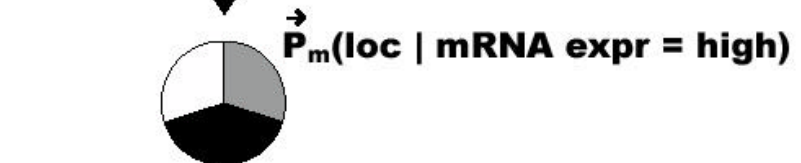
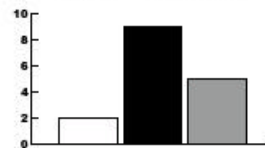
**State Vectors**



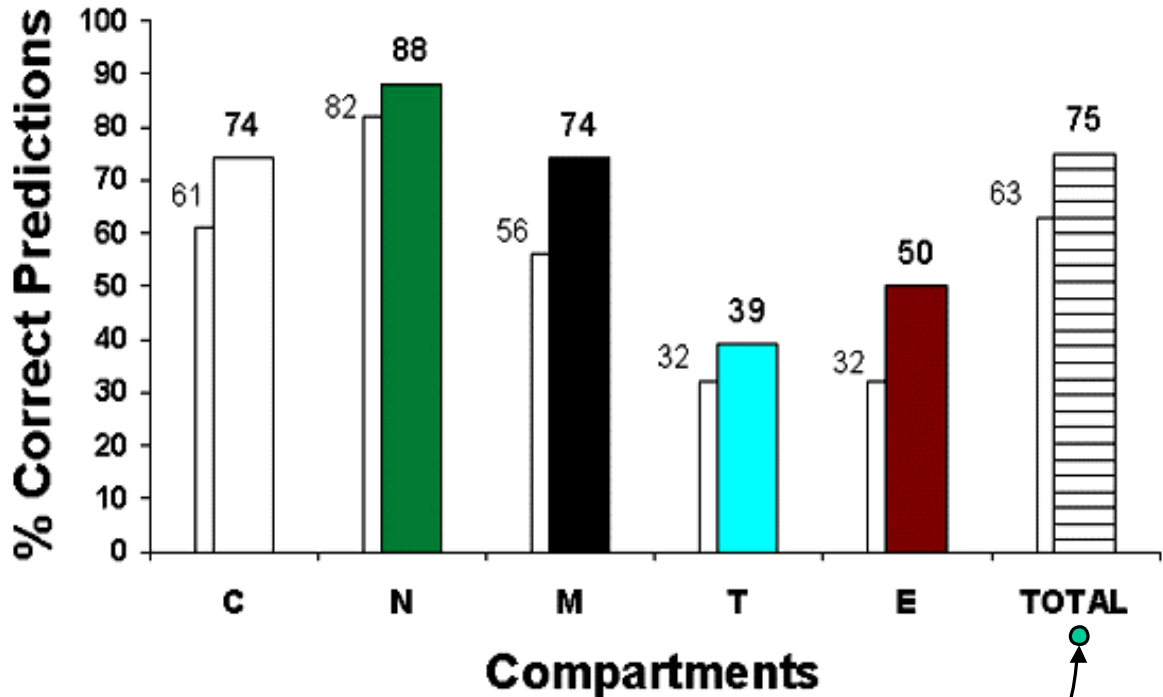
**Feature Vectors**  
 $\vec{P}(\text{mRNA expr=high} | \text{loc})$



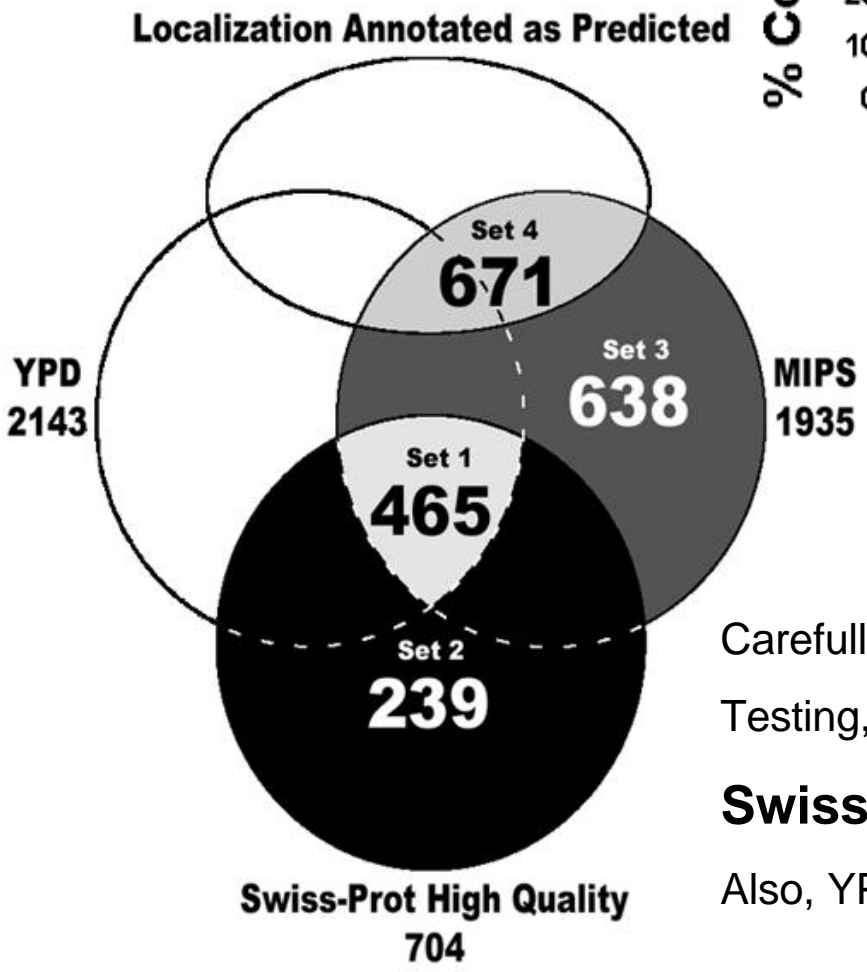
**Feature Vectors**  
 $\vec{P}(\text{pI} > 9 | \text{loc})$



# Results on Testing Data



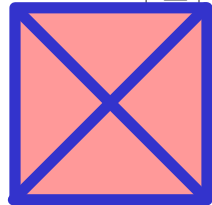
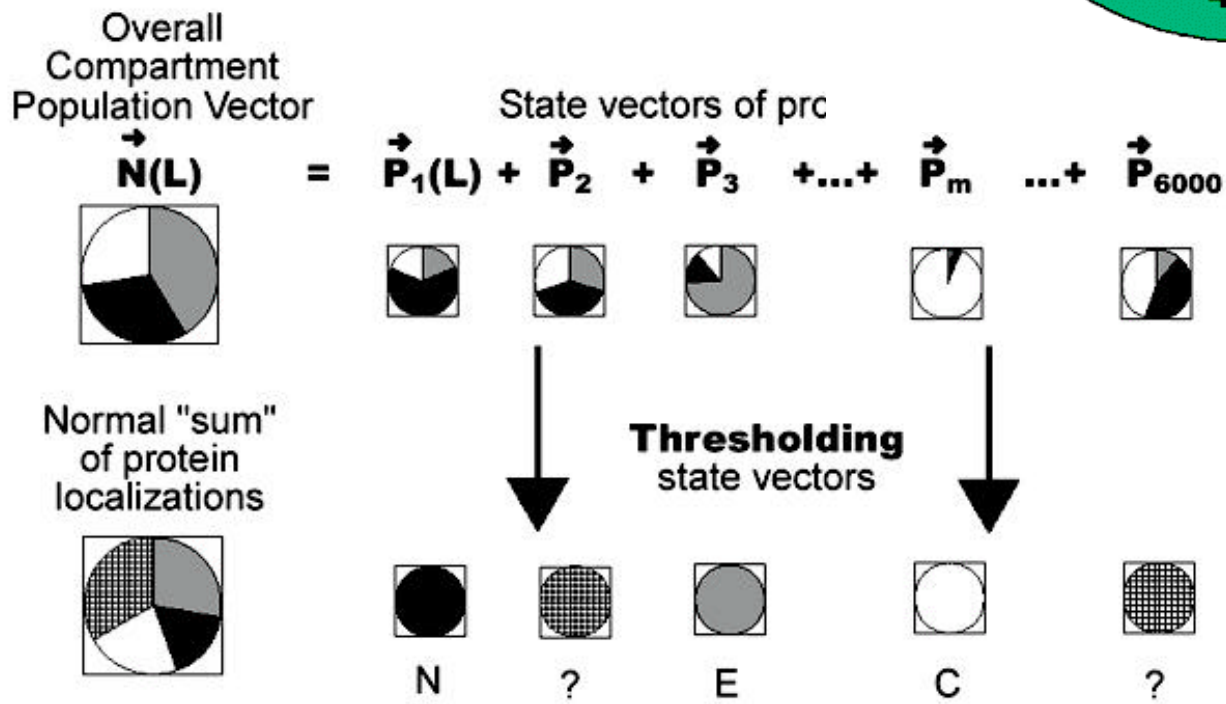
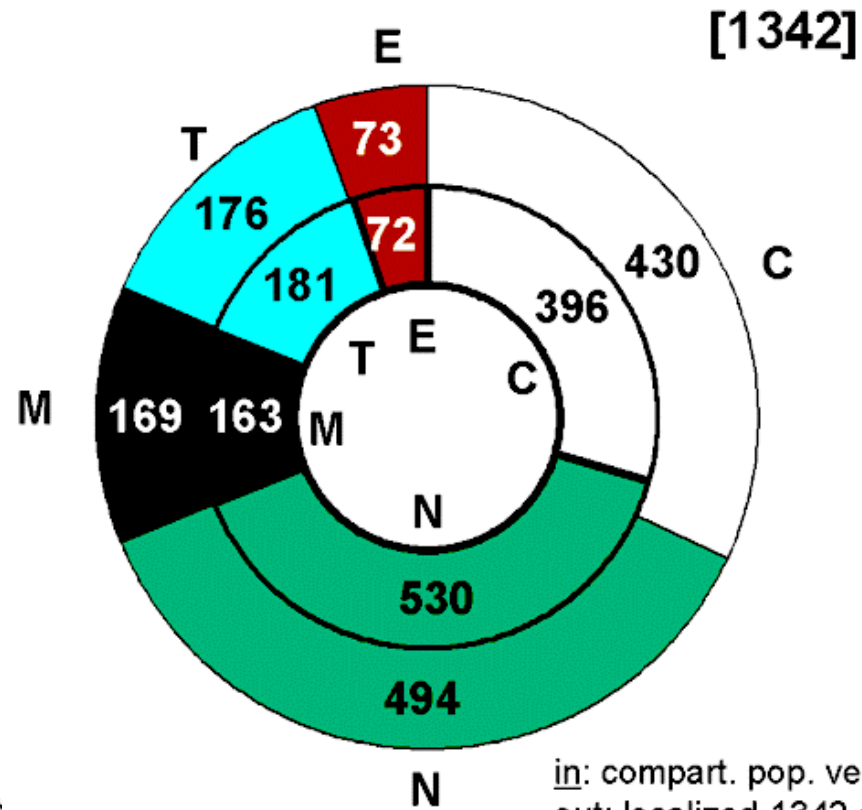
**Individual proteins: 75% with cross-validation**



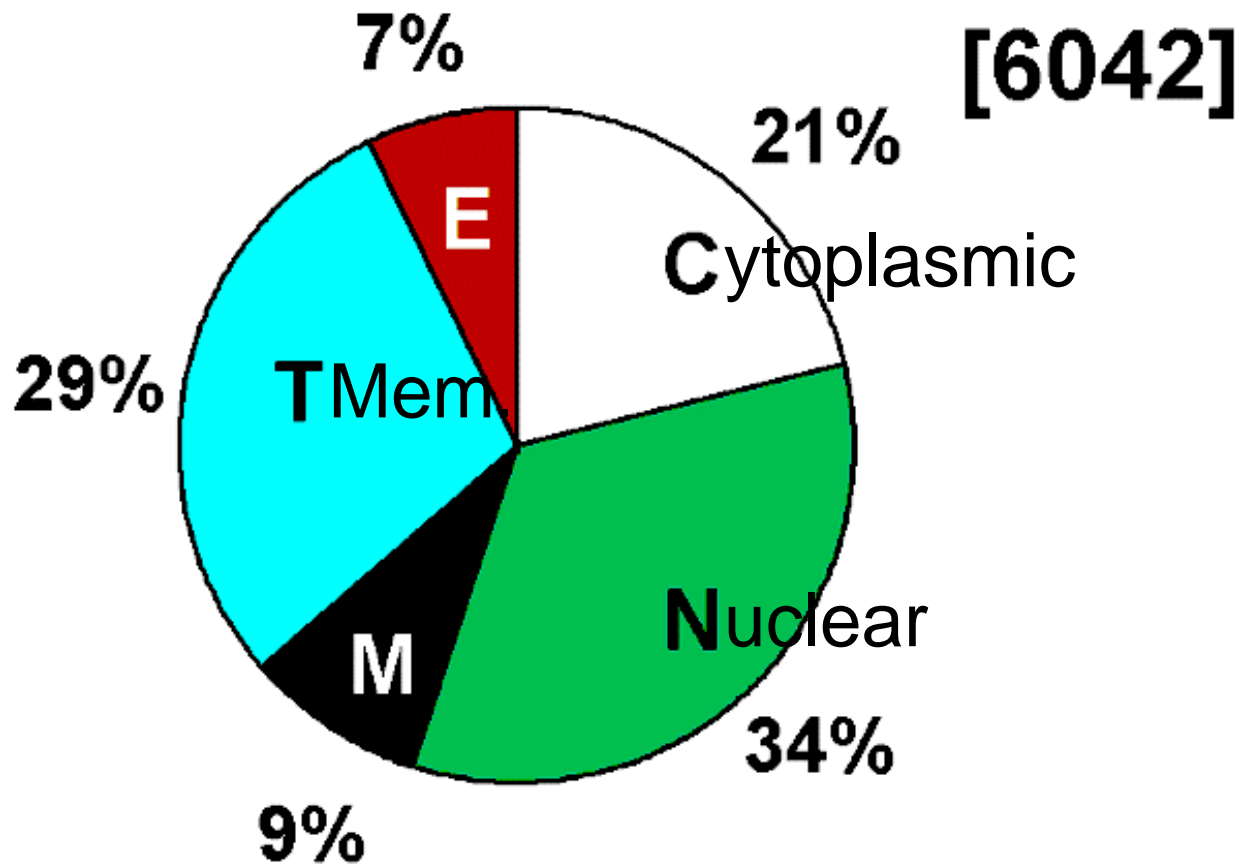
Carefully clean training dataset to **avoid circular logic**  
 Testing, training data, Priors: ~2000 proteins from  
**Swiss-Prot Master List**  
 Also, YPD, MIPS, Snyder Lab

# Results on Testing Data #2

Compartment Populations. Like QM, directly sum state vectors to get population. Gives **96%** pop. similarity.



Extrapolation to Compartment  
Populations of Whole Yeast Genome:  
~4000 predicted + ~2000 known





# Analysis of Genomes & Transcriptomes in terms of the Occurrence of Parts & Features

## 1 Using Parts to Interpret Genomes.

Shared and/or unique parts. Venn Diagrams, Fold tree with all- $\beta$  diff. Ortholog tree. Top-10 folds.

## 2 Using Parts to Interpret Pseudogenomes.

In worm, top  $\Psi$ -folds (DNase, hydrolase) v top-folds (lg). chr. IV enriched, dead and dying families (90YG v 1G)

## 3 Using Parts to Interpret Transcriptomes: Expression & Structure.

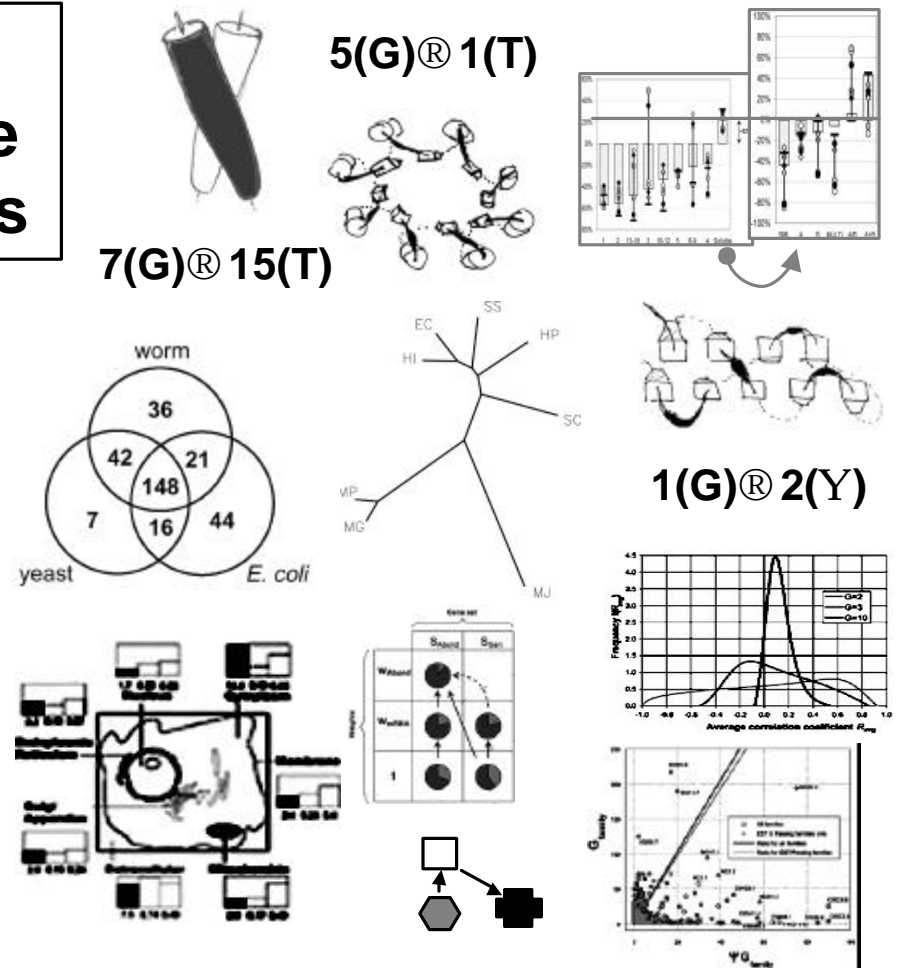
Top-10 parts in mRNA. Enriched in transcriptome:  $\alpha\beta$  folds, energy, synthesis, TIM fold, VGA. Depleted: TMs, transport, transcription, Leu-zip, NS. Compare with prot. abundance.

## 4 Expression & Localization.

Enriched : Cytoplasmic. Depleted: Nuclear. Bayesian localizer

## 5 Expression & Function.

Expression relates to structure & localization but to function, globally? P-value formalism. Weak relation to protein-protein interactions.

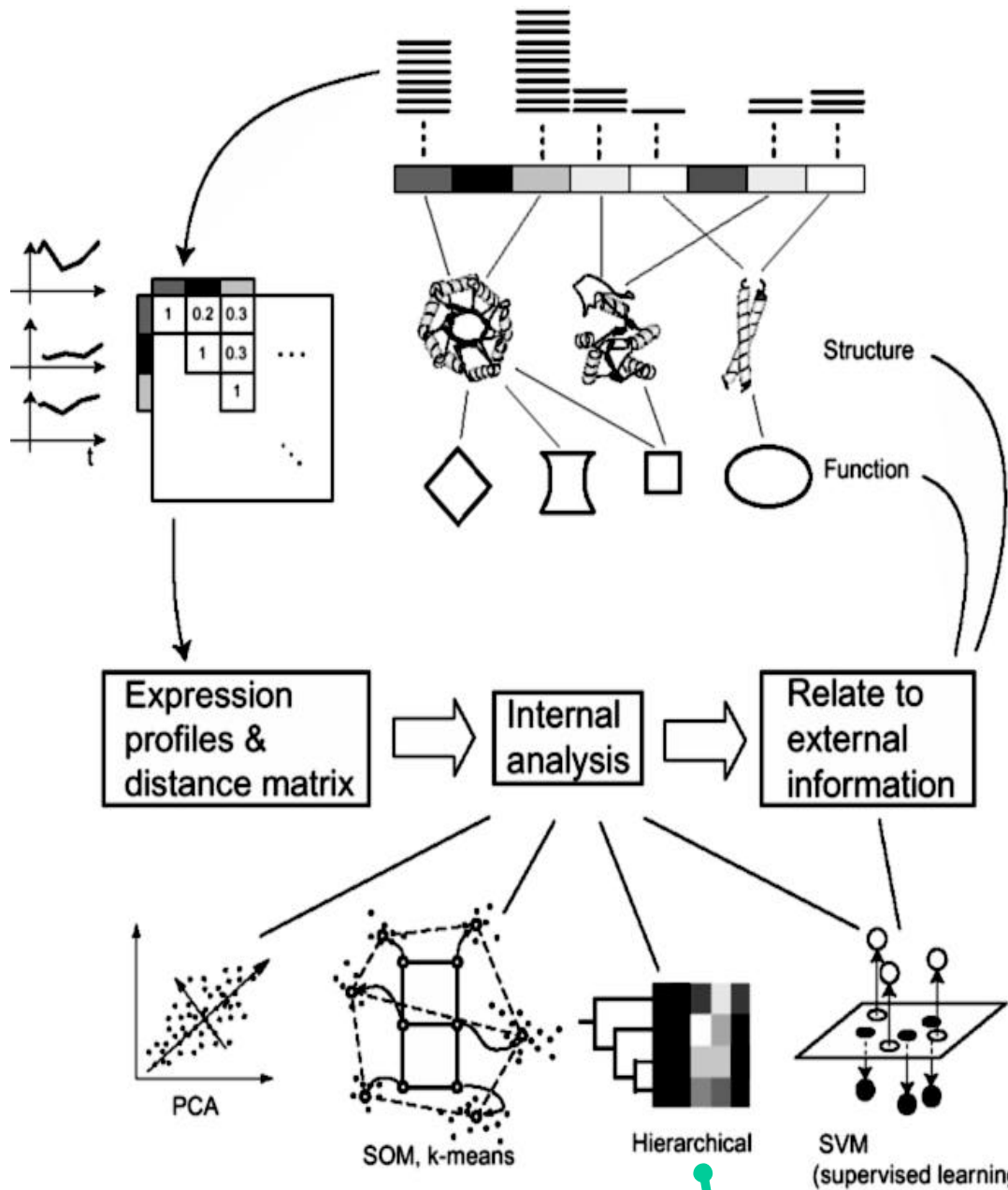


[bioinfo.mbb.yale.edu](http://bioinfo.mbb.yale.edu)

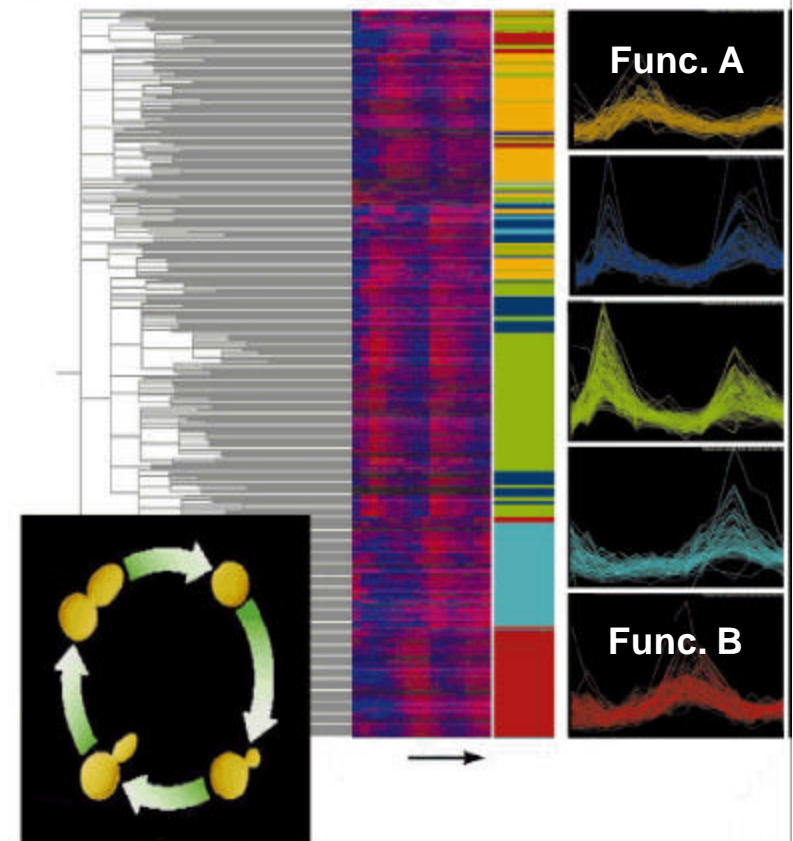
*H Hegyi, J Lin, B Stenger,  
P Harrison, N Echols,  
R Jansen, A Drawid, J Qian,  
D Greenbaum, M Snyder*

# Do Expression Clusters Relate to Protein Function?

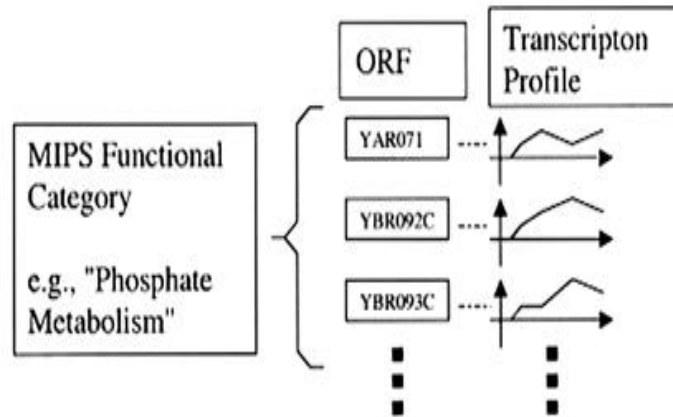
Can they predict functions?



- Clustering of expression profiles
- Grouping functionally related genes together (?)
- **Botstein (Eisen)**, Lander, Haussler, and Church groups, Eisenberg



# Distributions of Gene Expression Correlations, for All Possible Gene Groupings



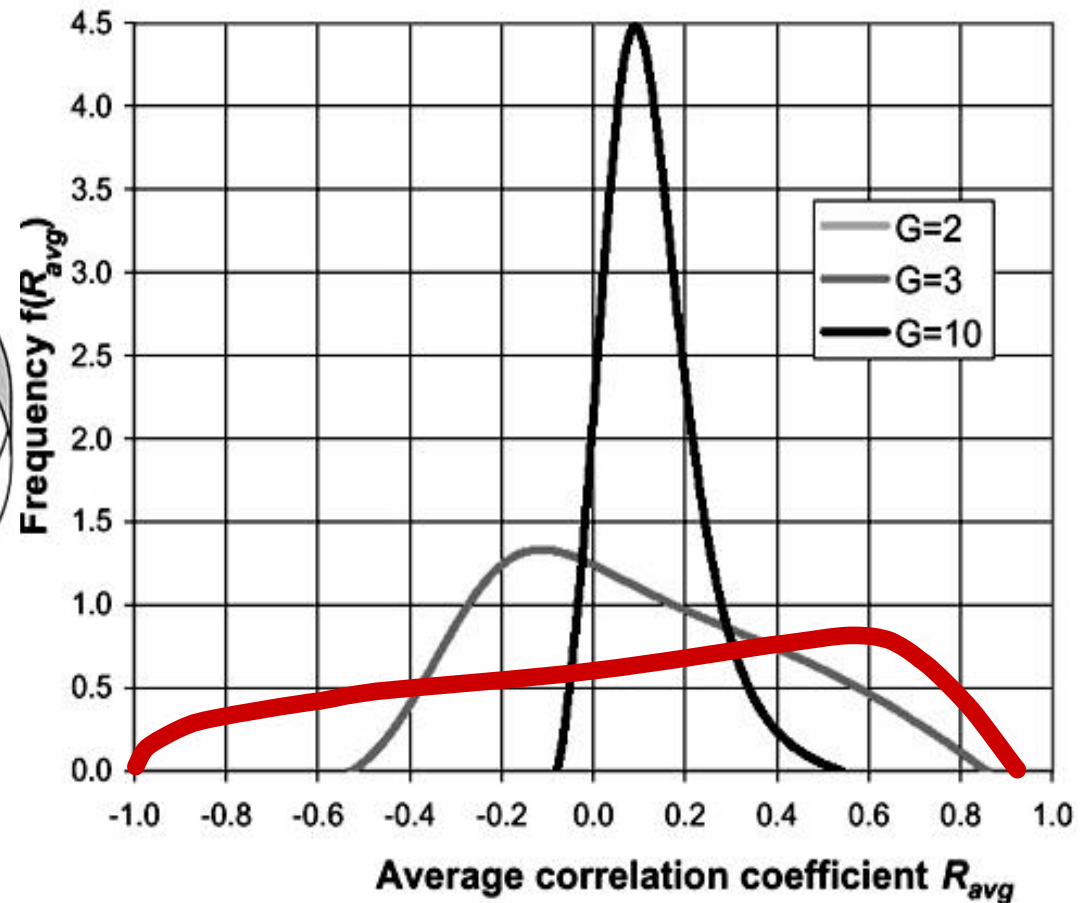
	YAR071W	YBR092C	YBR093C	...
YAR071	1.	0.2	0.3	
YBR092C	0.2	1.	0.4	
YBR093C	0.3	0.4	1.	
...				...

Correlation Coefficient Matrix (Pearson Coefficient)

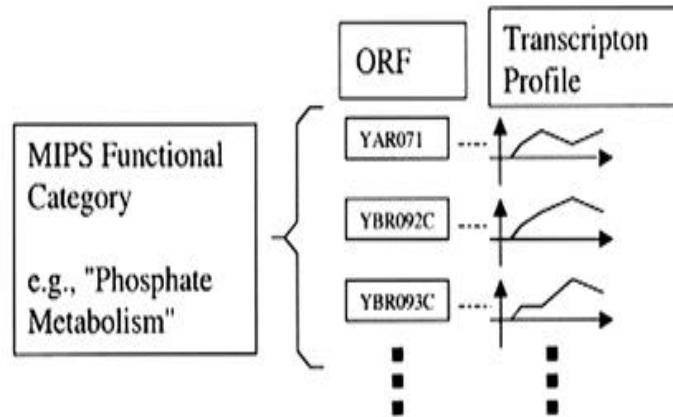
## Average Correlation Coefficient for Group of Genes

Sample for Diauxic shift Expt. (Brown),

$$\text{Ex. } R_{\text{avg}, G=3} = \frac{[ R(\text{gene-1}, \text{gene-3}) + R(\text{gene-1}, \text{gene-4}) + R(\text{gene-5}, \text{gene-7}) ]}{3}$$



# Distributions of Gene Expression Correlations, for All Possible Gene Groupings 2



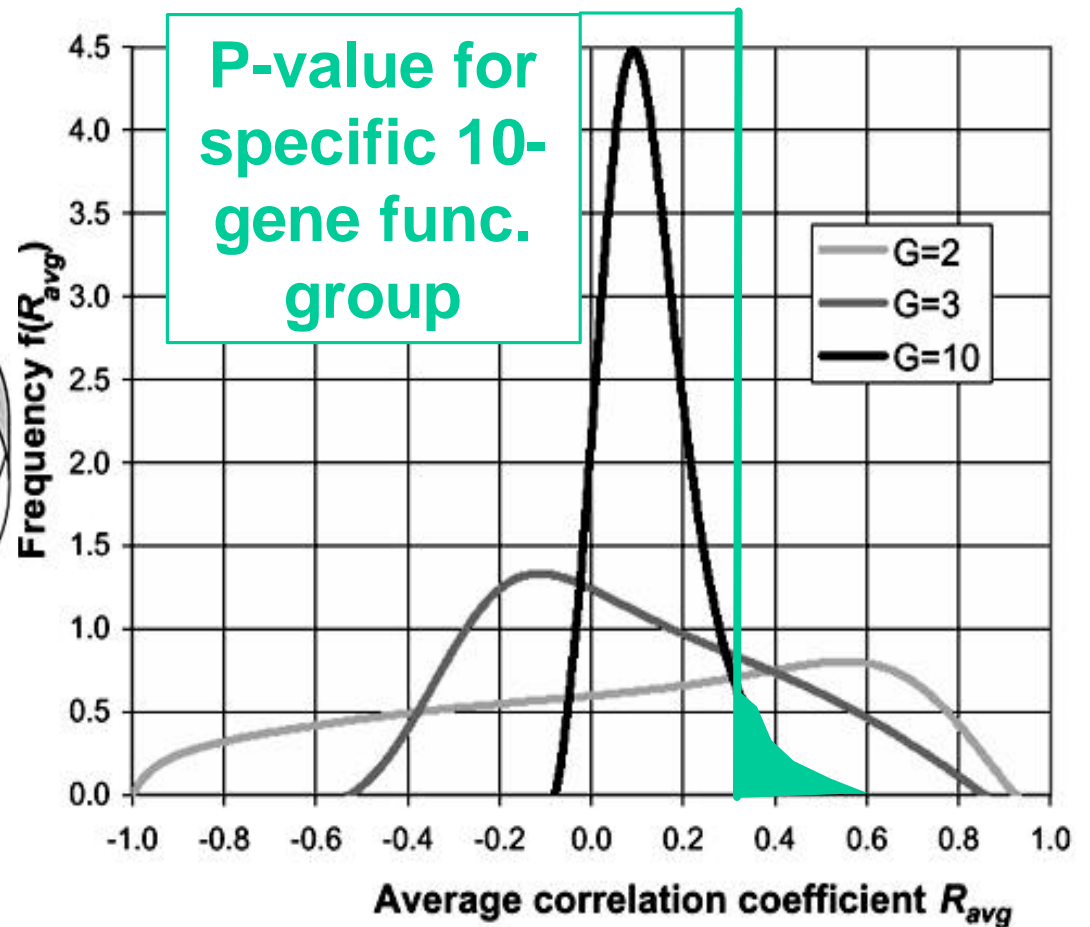
	YAR071	YBR092C	YBR093C	...
YAR071	1.	0.2	0.3	
YBR092C	0.2	1.	0.4	
YBR093C	0.3	0.4	1.	
...				...

Correlation Coefficient Matrix (Pearson Coefficient)

## Average Correlation Coefficient for Group of Genes

Sample for Diauxic shift Expt. (Brown),

$$\text{Ex. } R_{\text{avg}, G=3} = \frac{[ R(\text{gene-1}, \text{gene-3}) + R(\text{gene-1}, \text{gene-4}) + R(\text{gene-5}, \text{gene-7}) ]}{3}$$





	Experiment			
	Cell Cycle (CDC28)	Cell cycle (CDC15)	Diauxic shift	Sporulation
Cell growth, division & DNA syn.	>4	>4	>4	>4
Protein synthesis	>4	>4	>4	>4
Transcription	>4	>4	>4	1.6
Cellular organization	>4	>4	0.3	0.3
Energy	>4	>4	0.1	0.9
Cell rescue, defense, death	>4	>4	0	0
Intracellular transport	>4	>4	0	0
Ionic homeostasis	>4	>4	0	0.8
Metabolism	>4	>4	0	0
Transport facilitation	>4	>4	0	0
Signal transduction	2.5	1.6	0.1	0.6
Unclassified	2.3	>4	0	0
Cellular biogenesis	2.0	>4	0.4	0.2
Protein destination	0.3	>4	0.2	0.6
Retrotransposon & plasmid	0	2.8	1.9	1.0

	Experiment			
	Cell Cycle (CDC28)	Cell cycle (CDC15)	Diauxic shift	Sporulation
Respiration	>4	>4	>4	3.4
TCA pathway	>4	>4	>4	0.6
Glycogen, trehalose metabolism	>4	>4	1.2	0.7
Glycolysis	>4	>4	0.9	2.1
Gluconeogenesis	3.7	>4	0.1	1.7
Glyoxylate cycle	1.6	0.7	3.0	2.3
Pentose-phosphate pathway	1.5	0.8	0	0.6
Fermentation	1.3	>4	0	2.2
Other energy generation activities	0.7	0.1	0.1	0.2
Beta-oxidation of fatty acids	0.5	0.4	0.4	0.2

Correlation:

**Always Significant**

**Sometimes Significant (depends on expt.)**

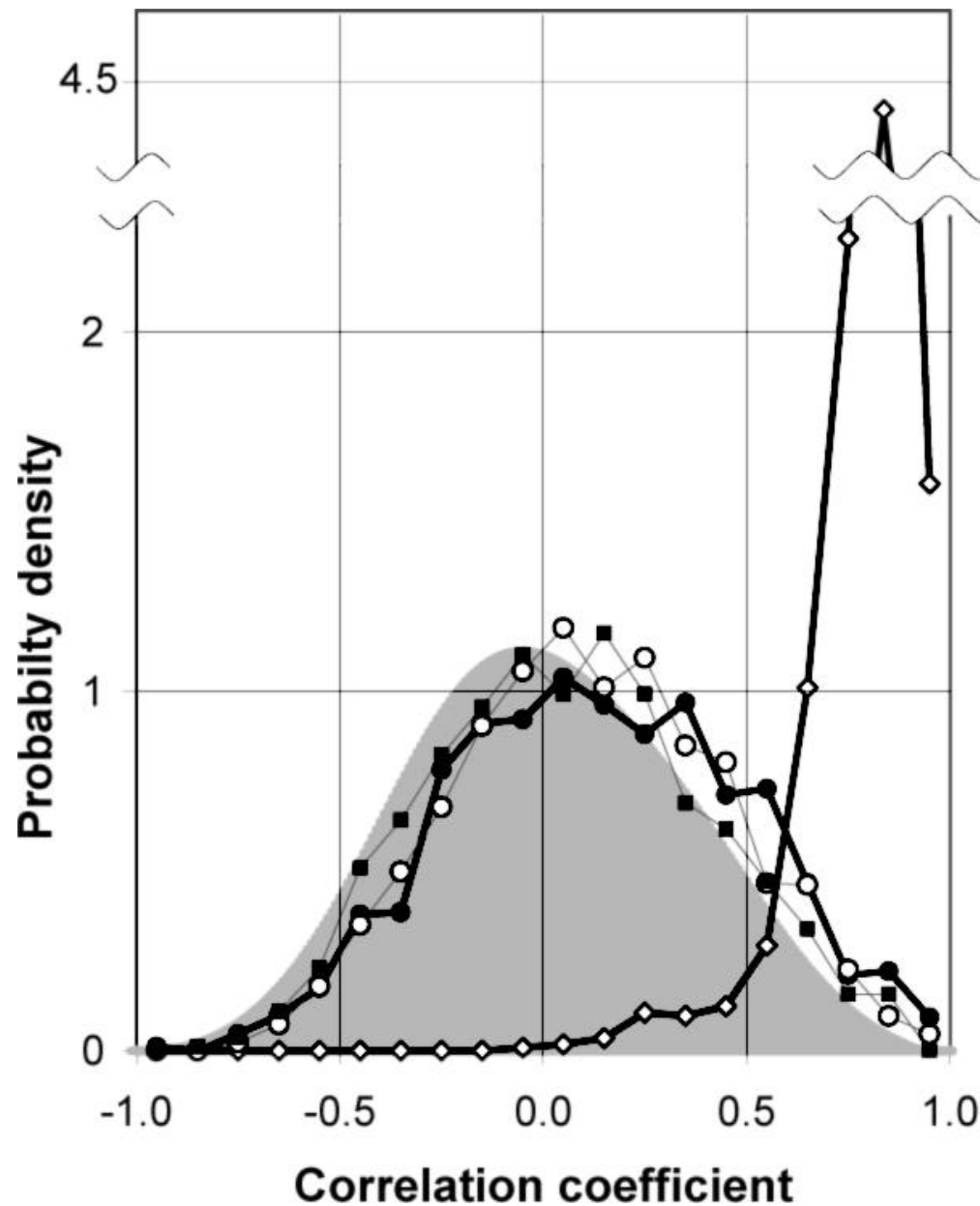
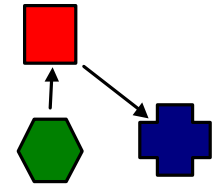
**Never Significant**

Based on Distributions,  
Correlation of  
Established Functional  
Categories, Computer  
Clusterings

	Fraction of significant groups				Total # groups
	CDC28	CDC15	Diauxic Shift	Sporulation	
MIPS 1	63%	81%	19%	13%	16
MIPS 2	50%	63%	17%	13%	102
MIPS 3	23%	33%	5%	4%	73
"Energy" (2 <sup>nd</sup> level)	40%	60%	20%	0%	10
SOM	93%	-	-	-	30
hierarch. Clustering	80%				25

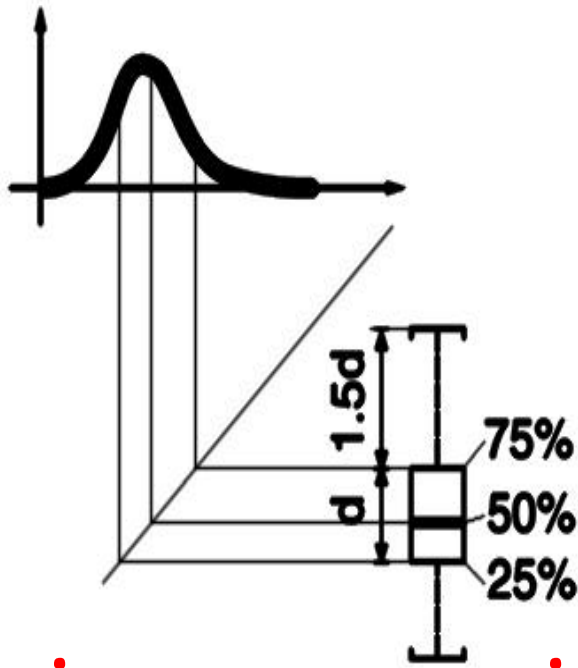
# Protein-Protein Interactions & Expression

Use same formalism to assess how closely related expression timecourses to sets of known p-p interactions

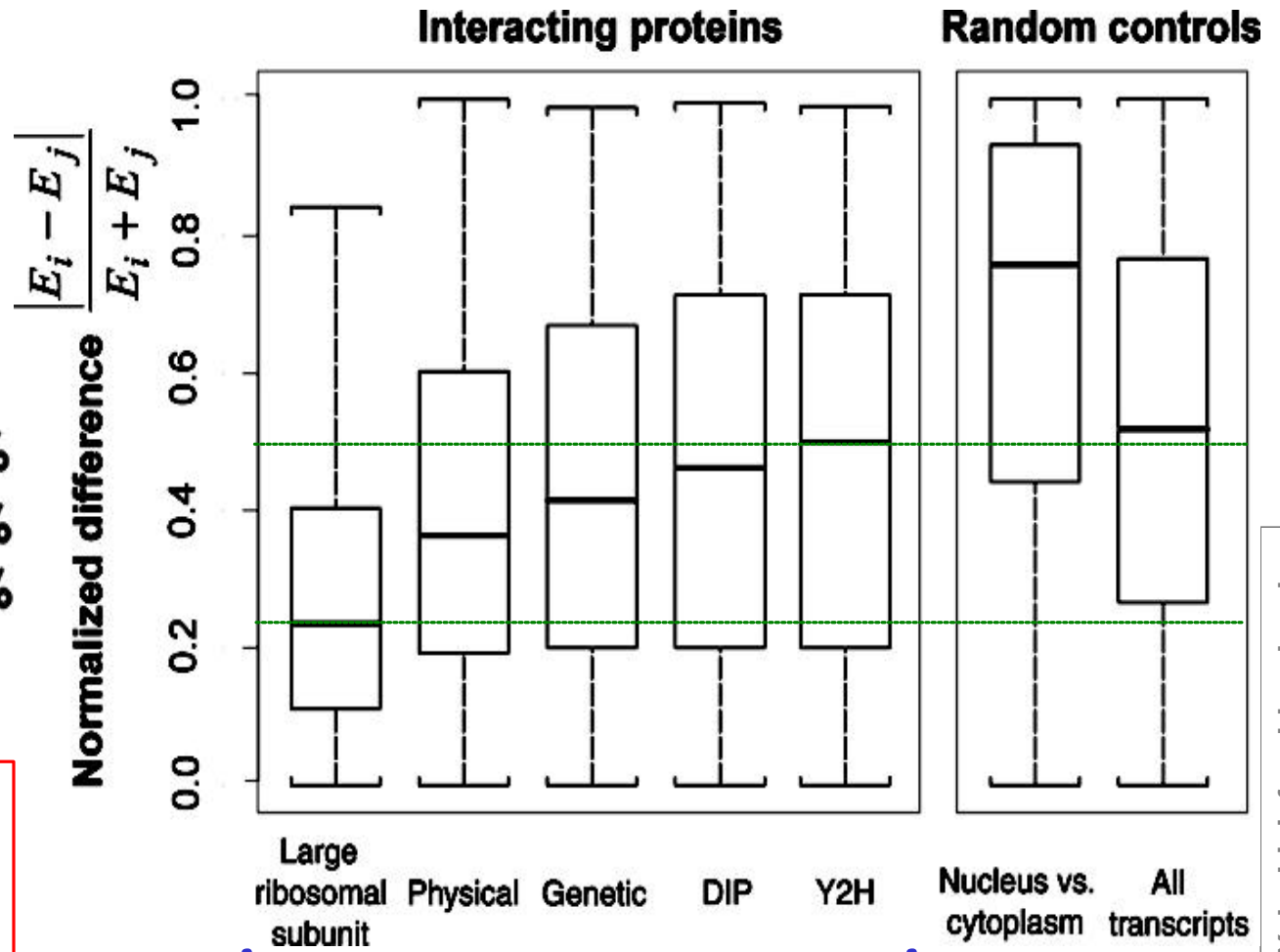


between selected expression timecourses in CDC28 expt. (Davis)

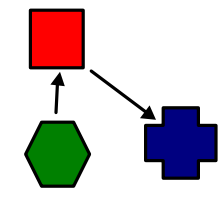
Sets of interactions	
■	Random (cell cycle CDC28) (control)
○	physical (from MIPS)
●	genetic (Uetz et al.)
■	Y2H (Uetz et al.)
◇	Large ribosomal subunit (strong interaction, clearly diff.)



**Distribution of Normalized Expression Levels**



**Sets of Interacting Proteins**



for

Relation of P-P Interactions to Abs. Expression Level

# Can we define FUNCTION well enough to relate to expression?

**Fold, Localization, Interactions & Regulation** are attributes of proteins that are much more clearly defined

Problems defining function:

**Multi-functionality:** 2 functions/protein (also 2 proteins/function)

**Conflating of Roles:** molecular action, cellular role, phenotypic manifestation.

**Non-systematic Terminology:**

'suppressor-of-white-apricot' & 'darkener-of-apricot'

## Functional Classification

**COGs**  
(cross-org., just conserved, NCBI Koonin/Lipman)

**GenProtEC**  
(*E. coli*, Riley)

**ENZYME**  
(SwissProt Bairoch/ Apweiler, just enzymes, cross-org.)

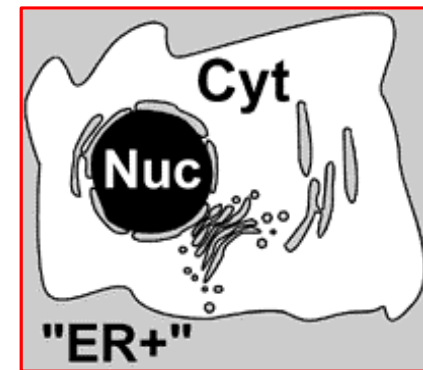
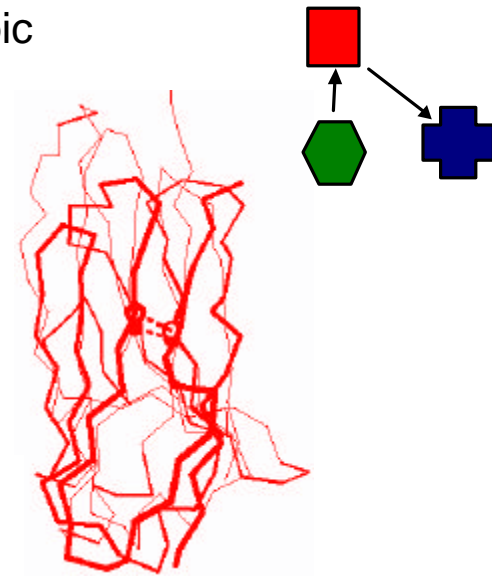
**"Fly"**  
(fly, Ashburner) now extended to **GO** (cross-org.)

**MIPS/PEDANT**  
(yeast, Mewes)

Also:  
Other SwissProt Annotation  
WIT, KEGG (just pathways)  
TIGR EGAD (human ESTs)

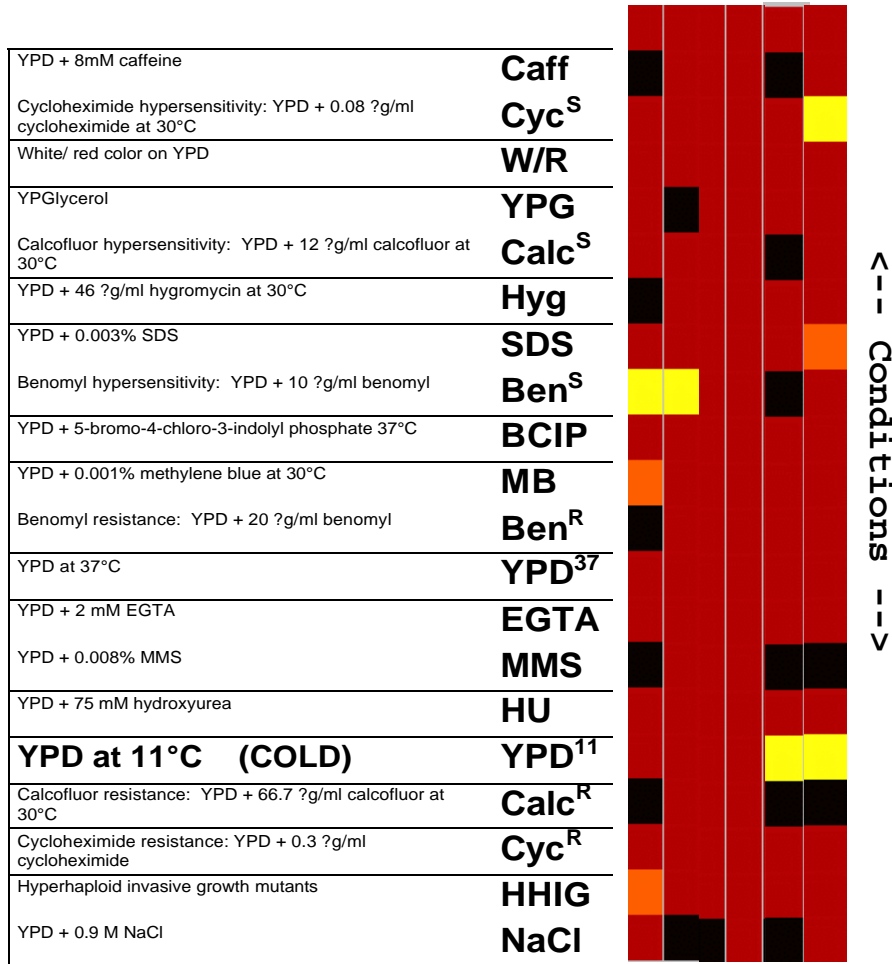
24 (c) Mark Gerstein, 2000, Yale, bioinfo.mbb.yale.edu

**VS.**



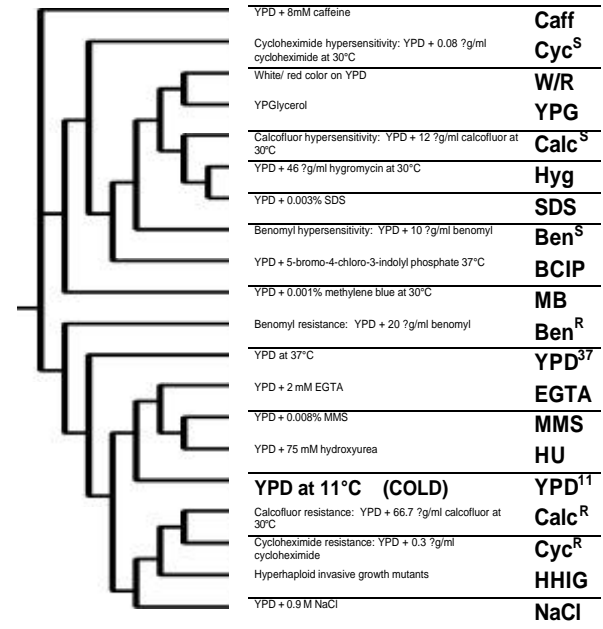
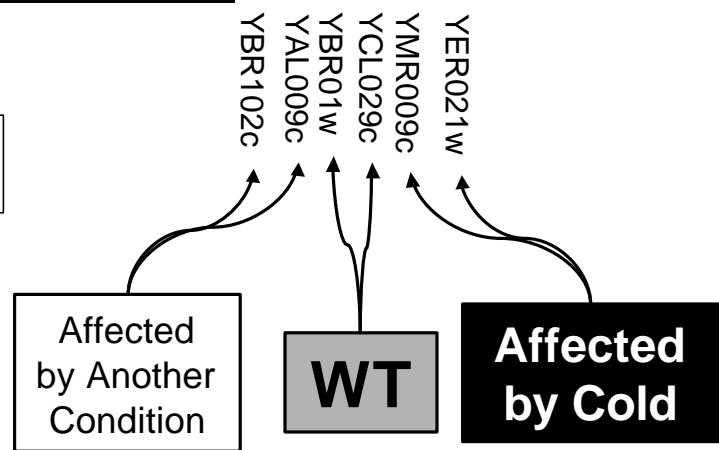
# Whole Genome Phenotype Profiles

Transposon insertions into (almost) each yeast gene to see how yeast is affected in 20 conditions. Generates a phenotype pattern vector, which can be treated **similarly to expression data**



←-- Conditions -->

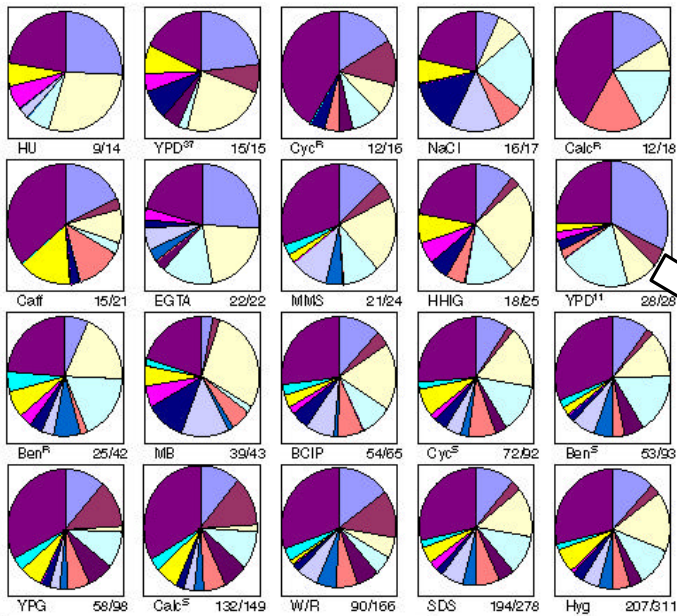
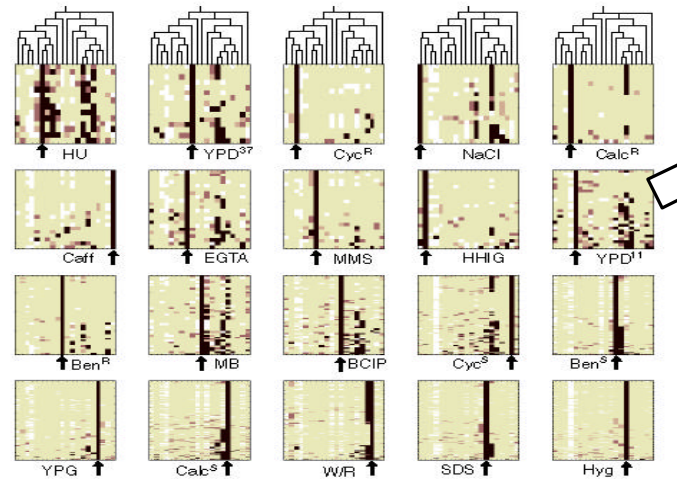
M Snyder



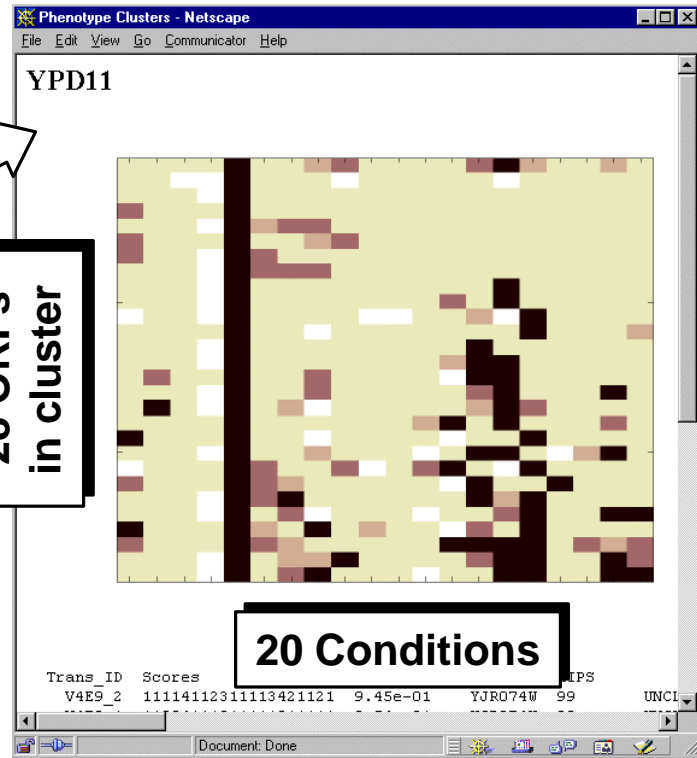
Clustering Conditions



# Phenotype ORF Clusters from Transposon Expt.

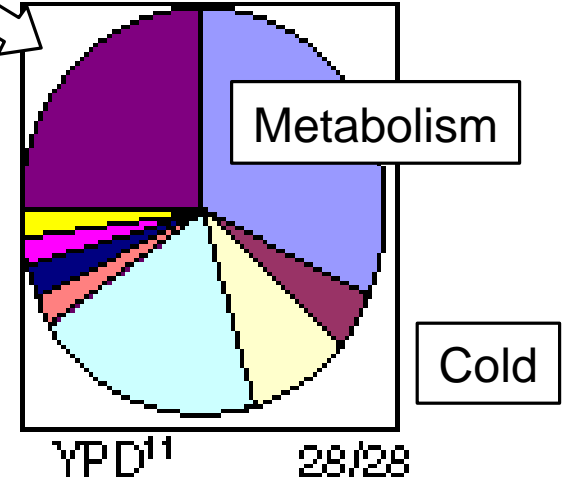


- METABOLISM
- CELL GROWTH, DIVISION AND DNA SYNTHESIS
- PROTEIN SYNTHESIS
- TRANSPORT FACILITATION
- CELLULAR BIOGENESIS
- CELL RESCUE, DEFENSE, CELL DEATH AND AGEING
- CELLULAR ORGANIZATION
- ENERGY
- TRANSCRIPTION
- PROTEIN DESTINATION
- INTRACELLULAR TRANSPORT
- SIGNAL TRANSDUCTION
- IONIC HOMEOSTASIS



Transposon insertions into (almost) each yeast gene to see how yeast is affected in 20 conditions. Generates a phenotype pattern vector, which can be treated **similarly to expression data**

k-means clustering of ORFs based on "phenotype patterns," cross-ref. to MIPs Functional Classes



Cluster showing cold phenotype (containing genes most necessary in cold) is enriched in metabolic functions

M Snyder, A Kumar, et al....

# Analysis of Genomes & Transcriptomes in terms of the Occurrence of Parts & Features

## 1 Using Parts to Interpret Genomes.

Shared and/or unique parts. Venn Diagrams, Fold tree with all- $\beta$  diff. Ortholog tree. Top-10 folds.

## 2 Using Parts to Interpret Pseudogenomes.

In worm, top  $\Psi$ -folds (DNase, hydrolase) v top-folds (lg). chr. IV enriched, dead and dying families (90YG v 1G)

## 3 Using Parts to Interpret Transcriptomes: Expression & Structure.

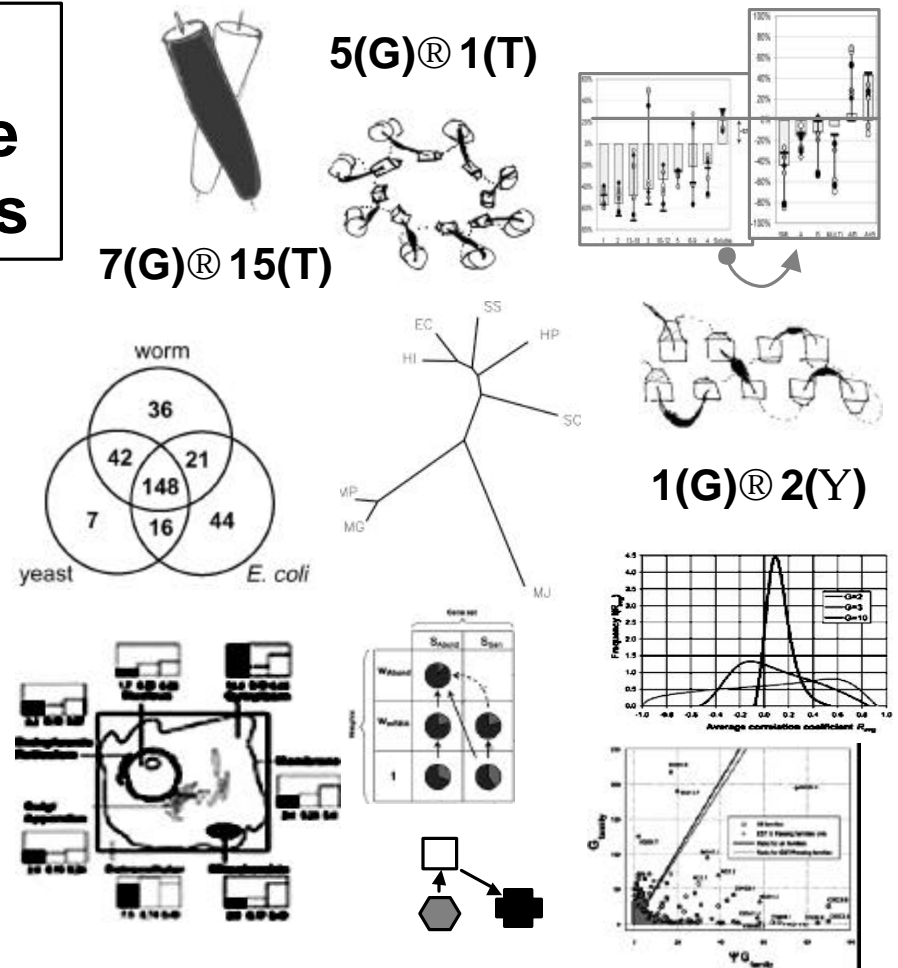
Top-10 parts in mRNA. Enriched in transcriptome:  $\alpha\beta$  folds, energy, synthesis, TIM fold, VGA. Depleted: TMs, transport, transcription, Leu-zip, NS. Compare with prot. abundance.

## 4 Expression & Localization.

Enriched : Cytoplasmic. Depleted: Nuclear. Bayesian localizer

## 5 Expression & Function.

Expression relates to structure & localization but to function, globally? P-value formalism. Weak relation to protein-protein interactions.

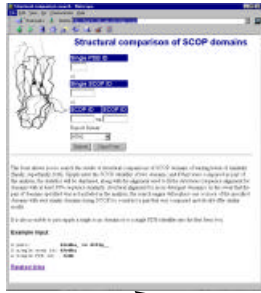


[bioinfo.mbb.yale.edu](http://bioinfo.mbb.yale.edu)

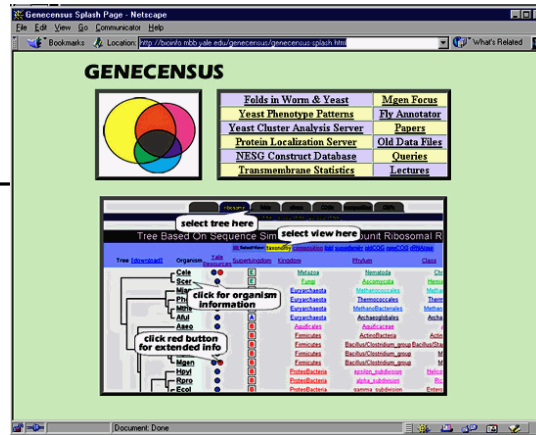
*H Hegyi, J Lin, B Stenger,  
P Harrison, N Echols,  
R Jansen, A Drawid, J Qian,  
D Greenbaum, M Snyder*

# GeneCensus

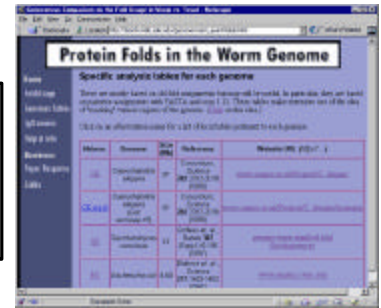
[bioinfo.mbb.yale.edu](http://bioinfo.mbb.yale.edu)



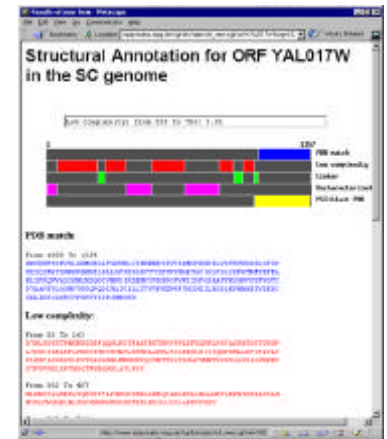
Alignment Database  
Alignment Server



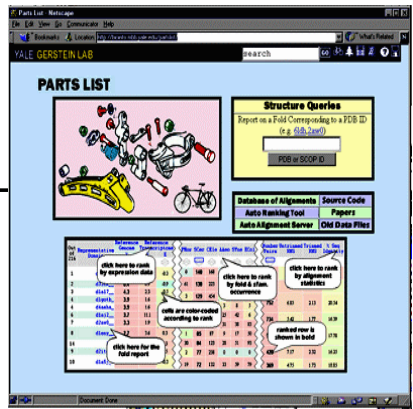
Detailed Tables



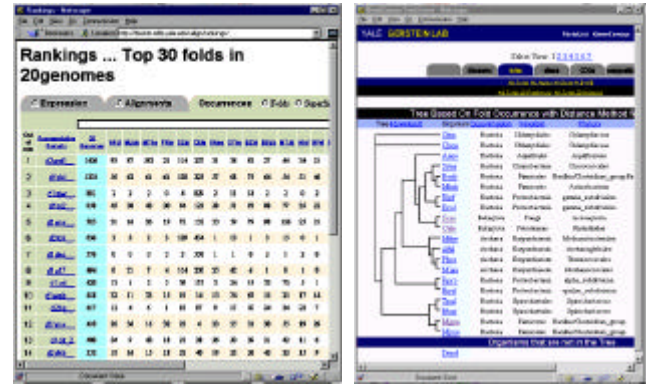
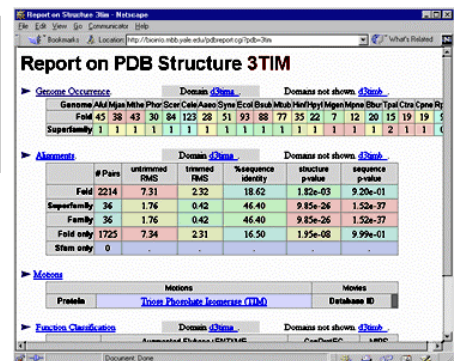
ORF Query



Ranks  
Trees

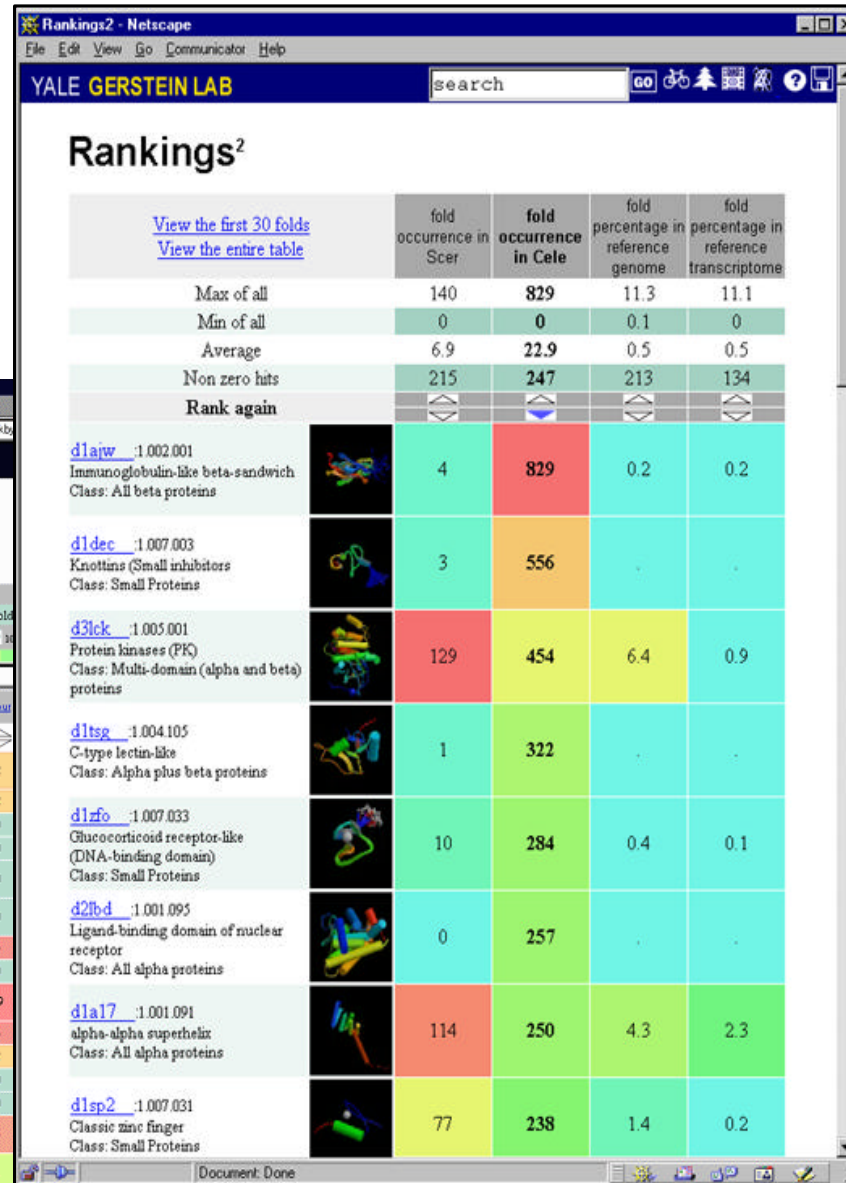
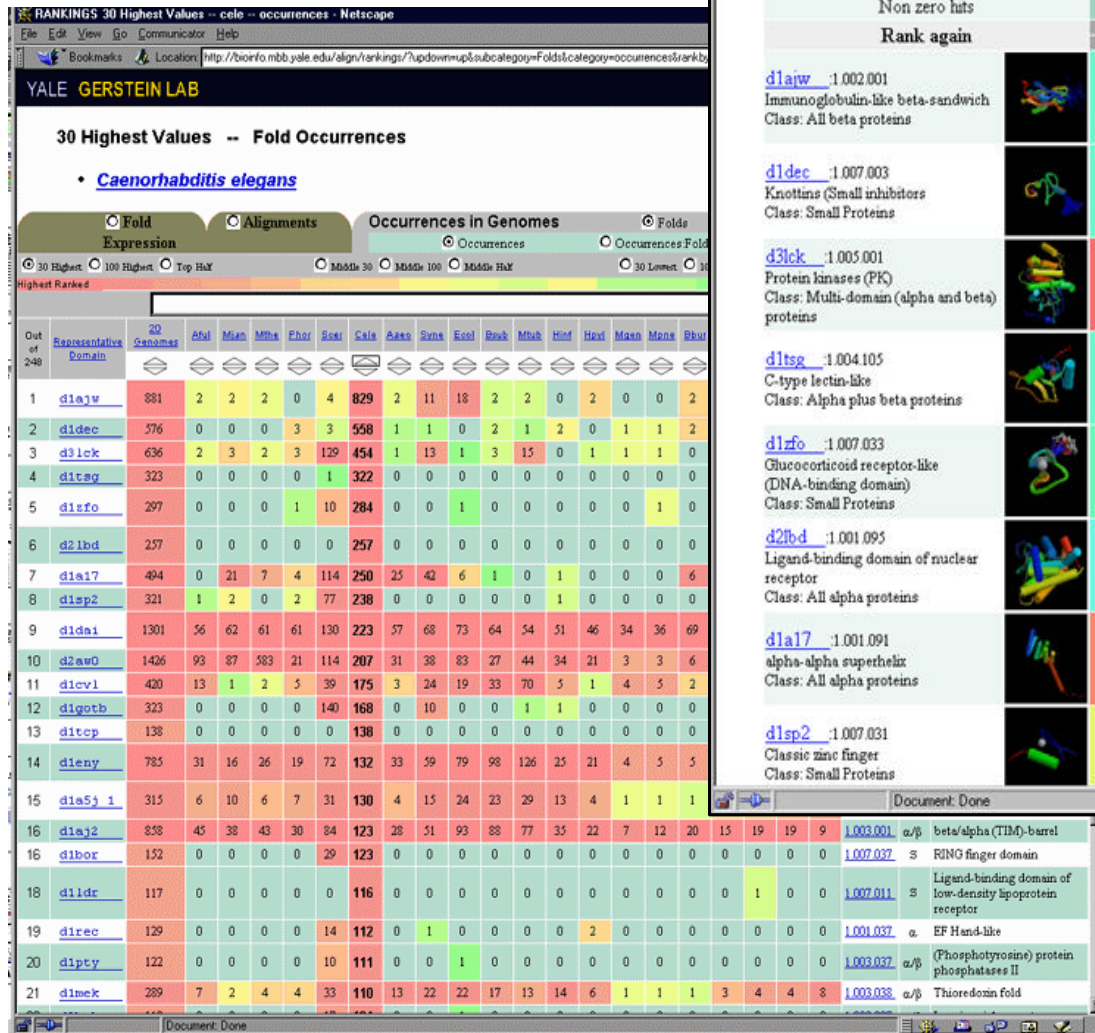


PDB Query





# PartsList Ranking Viewers

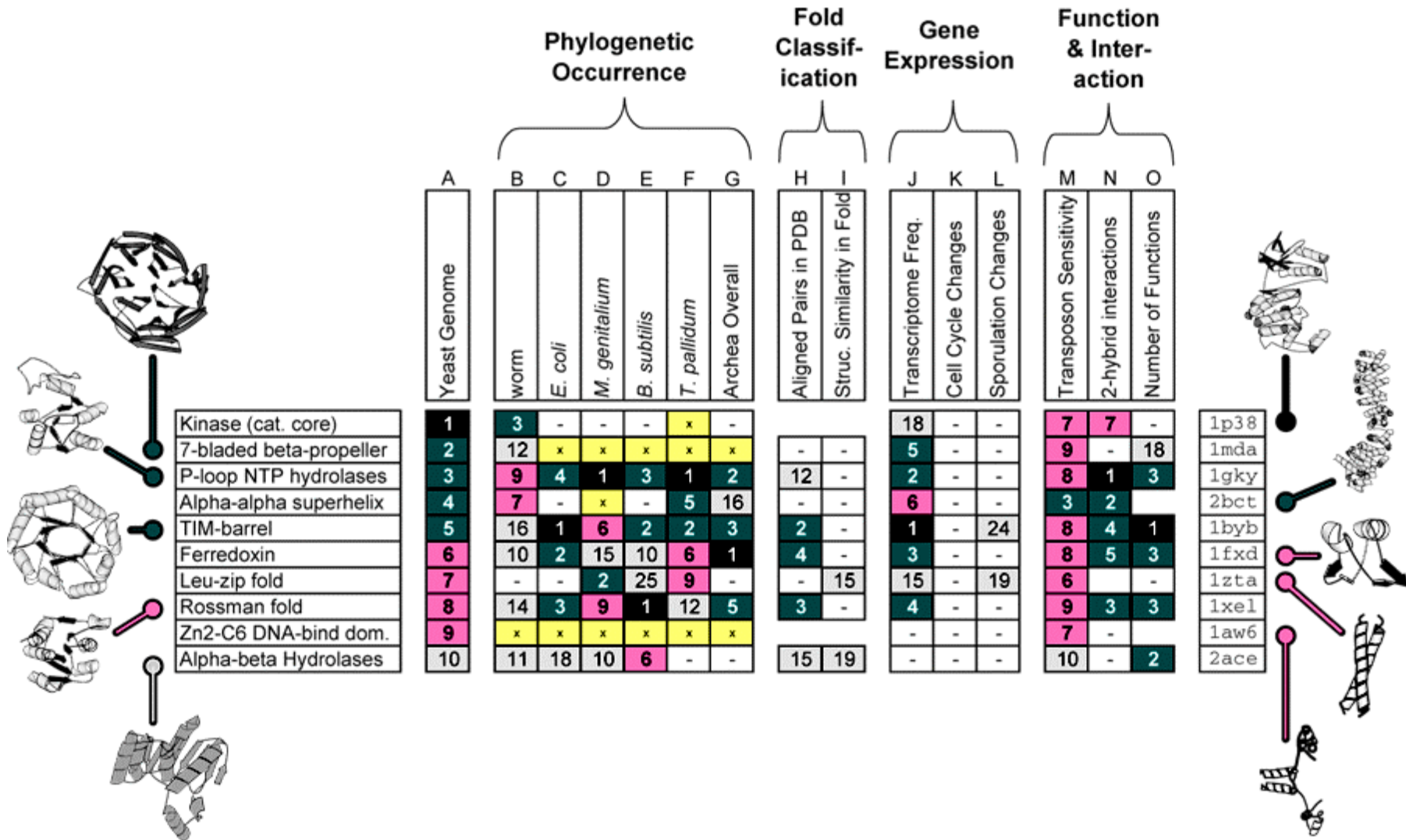


J Qian,  
B Stenger,  
J Lin....



Rank Folds by Genome  
Occurrence, Expression, Fold  
Clustering, Length, &c

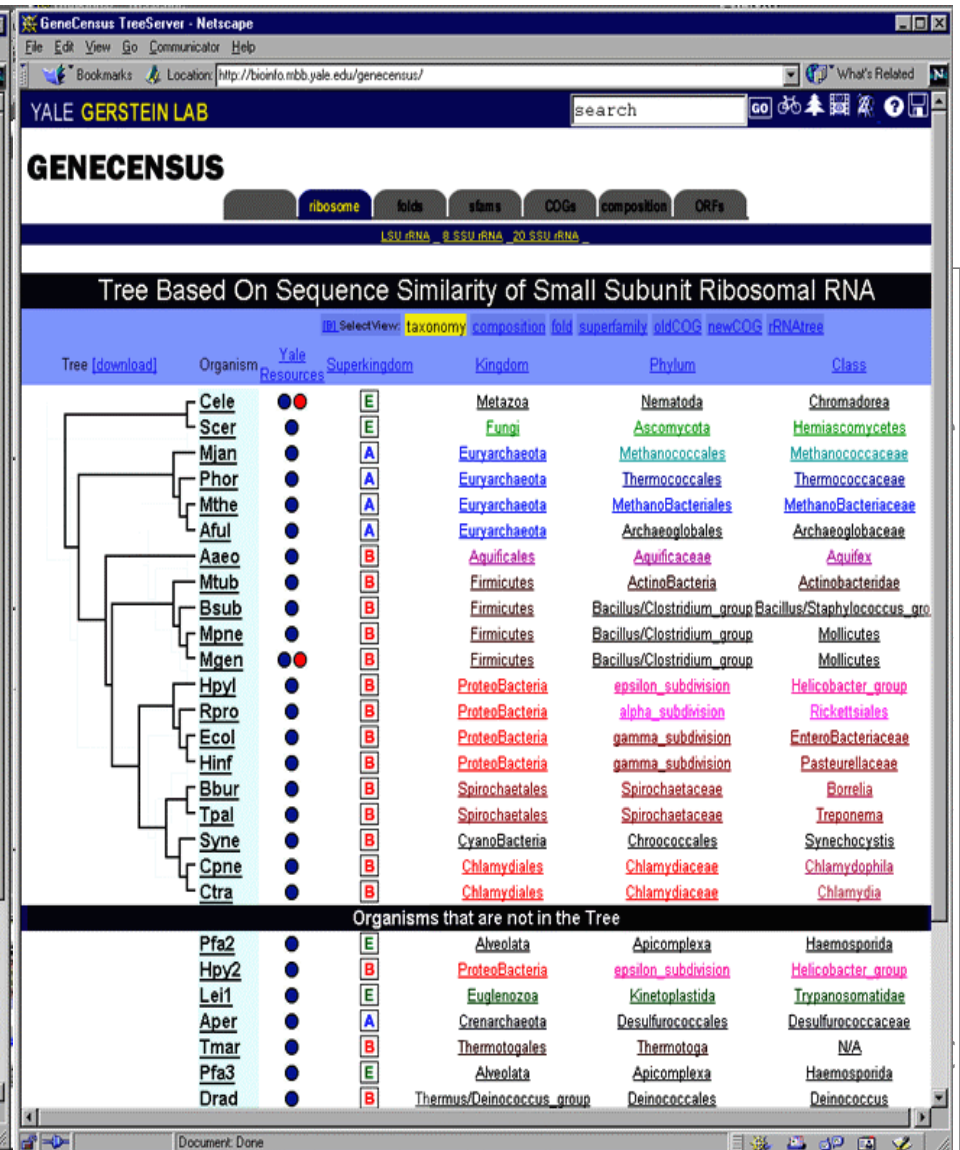
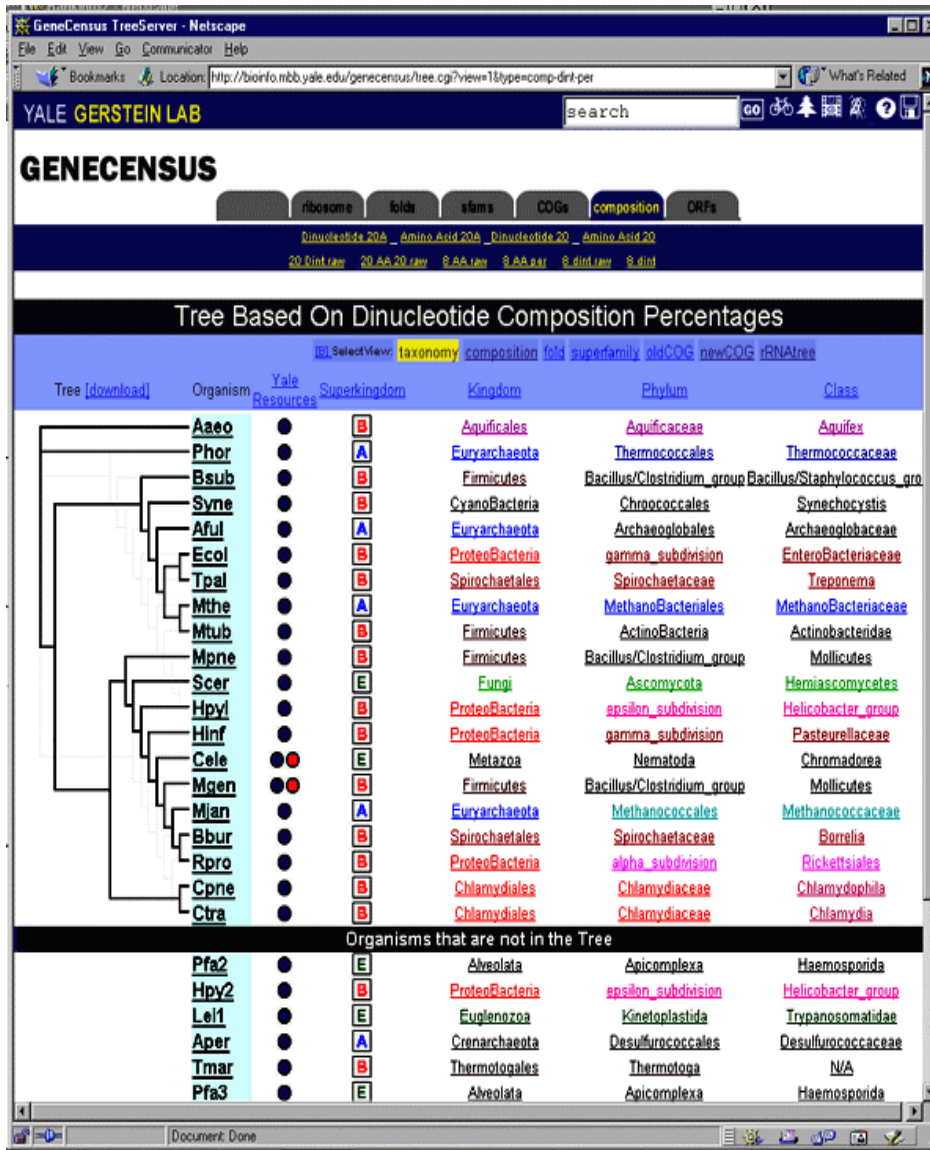
# Surveying a Finite PartsList from Many Perspectives





# GeneCensus Dynamic Tree Viewers

Recluster organisms based on folds, composition, &c and compare to traditional taxonomy



# Analysis of Genomes & Transcriptomes in terms of the Occurrence of Parts & Features

## 1 Using Parts to Interpret Genomes.

Shared and/or unique parts. Venn Diagrams, Fold tree with all- $\beta$  diff. Ortholog tree. Top-10 folds.

## 2 Using Parts to Interpret Pseudogenomes.

In worm, top  $\Psi$ -folds (DNase, hydrolase) v top-folds (lg). chr. IV enriched, dead and dying families (90YG v 1G)

## 3 Using Parts to Interpret Transcriptomes:

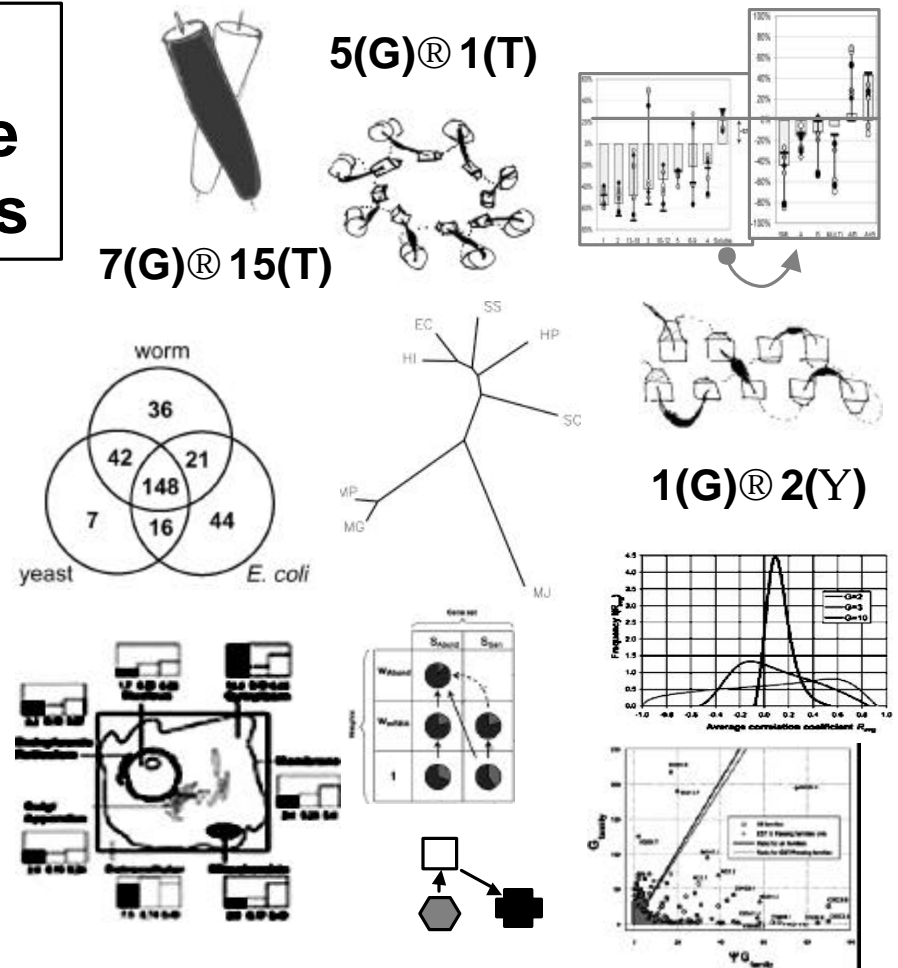
**Expression & Structure.** Top-10 parts in mRNA. Enriched in transcriptome:  $\alpha\beta$  folds, energy, synthesis, TIM fold, VGA. Depleted: TMs, transport, transcription, Leu-zip, NS. Compare with prot. abundance.

## 4 Expression & Localization.

Enriched : Cytoplasmic. Depleted: Nuclear. Bayesian localizer

## 5 Expression & Function.

Expression relates to structure & localization but to function, globally? P-value formalism. Weak relation to protein-protein interactions.



[bioinfo.mbb.yale.edu](http://bioinfo.mbb.yale.edu)

*H Hegyi, J Lin, B Stenger,  
P Harrison, N Echols,  
R Jansen, A Drawid, J Qian,  
D Greenbaum, M Snyder*