Hidden in plain sight: The eukaryotically conserved unstudied proteins and a framework for their classification and characterisation.

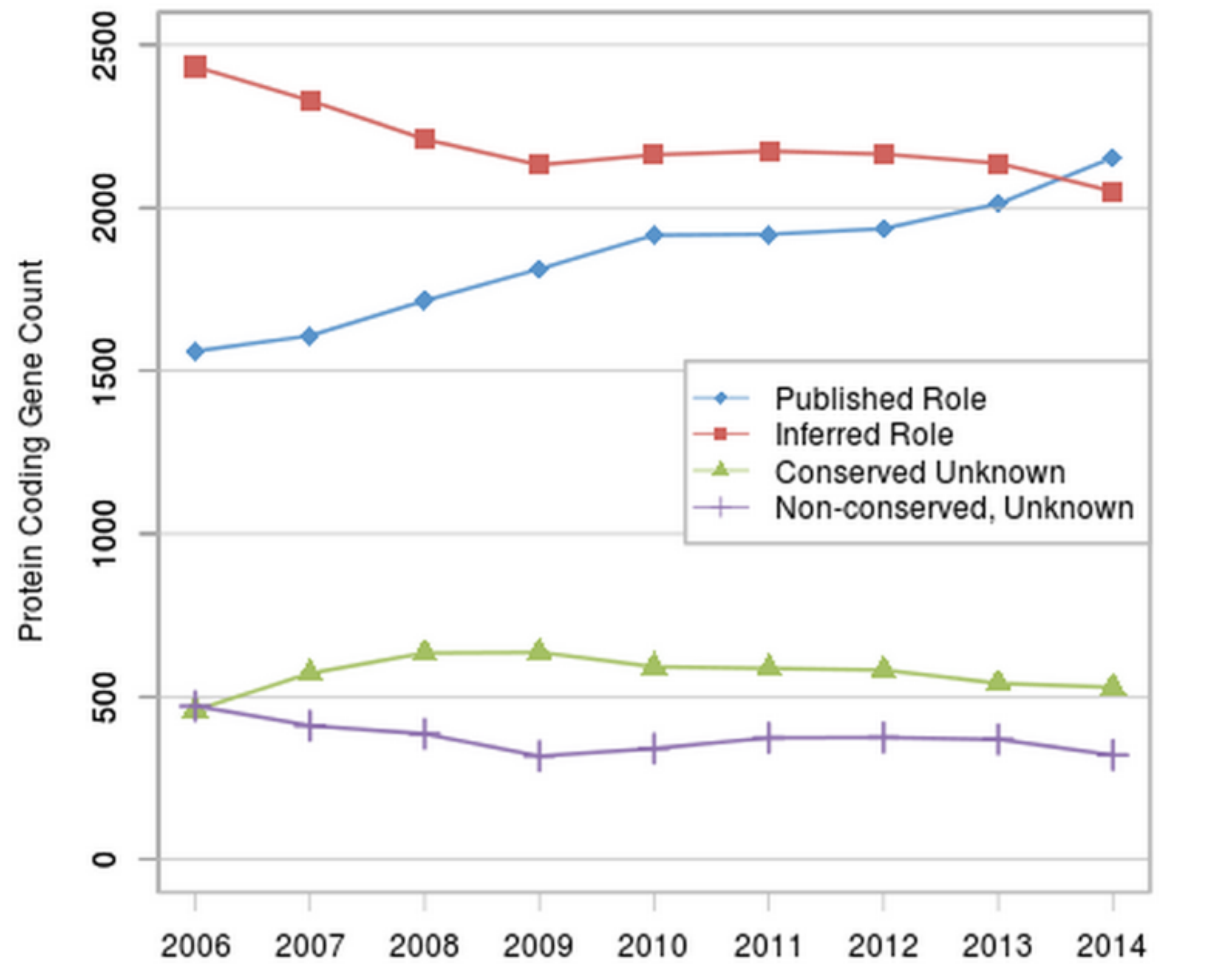
Wood V [1], Bahler J [2], Harris M [1], Lock A [2], Oliver SG [1]

ABSTRACT

Proteins conserved widely among eukaryotes play fundamentally important roles in the shared, basic mechanisms of life. The roles of many broadly conserved proteins remain unknown, however, despite almost a century of gene- and gene product-specific genetic and biochemical investigation. Even the recent emergence of genome-wide experimental techniques and the availability of near-complete protein inventories for many intensively studied eukaryotic model species have shed light on the functions of few previously uncharacterised conserved proteins. Because the success of many endeavours in basic and translational research, including drug discovery, metabolomics, and systems biology, depends critically on comprehensive representation of conserved functions, a more complete understanding of protein components conserved throughout eukaryotes would have far-reaching benefits for biological research in many species and on a wide range of scales.

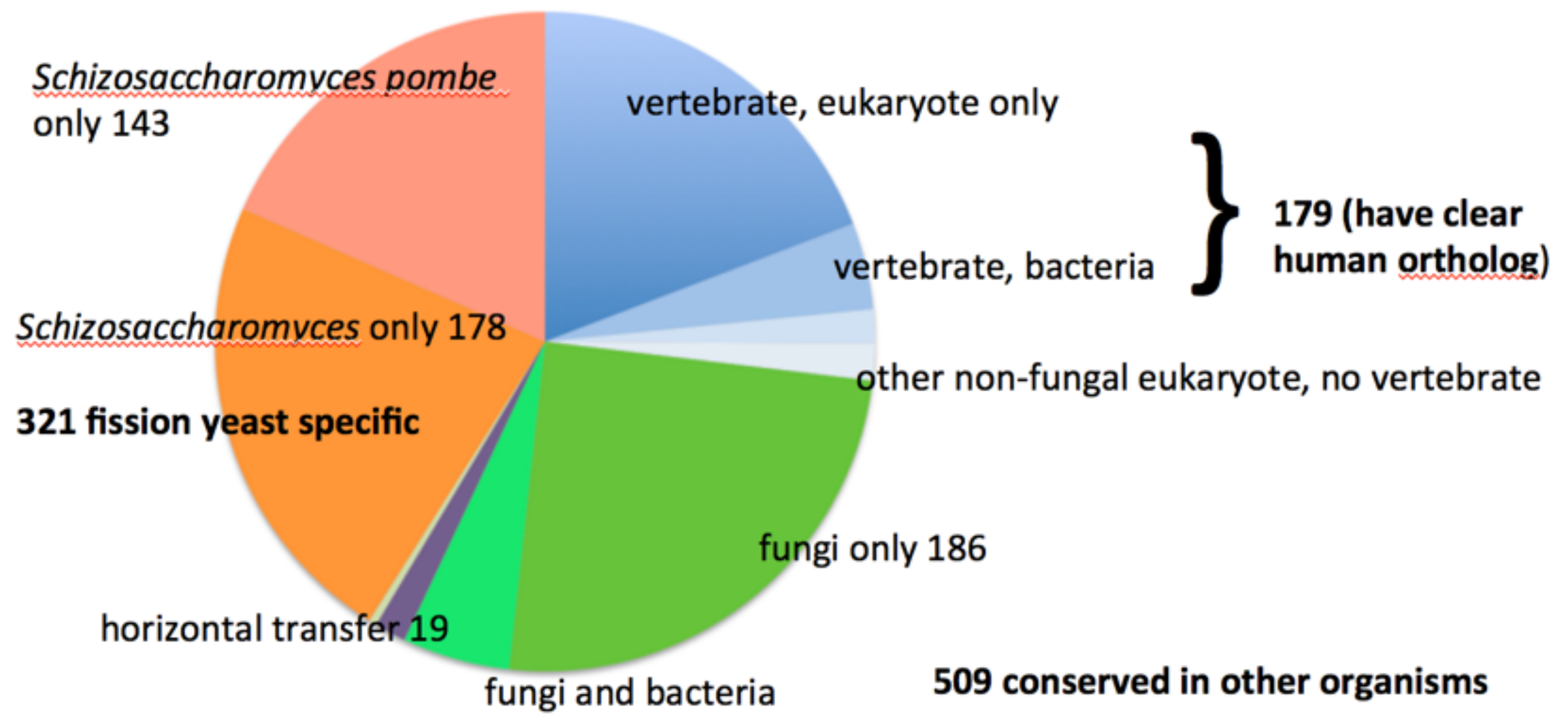
To identify priority targets for experimental investigation, PomBase provides an inventory of fission yeast proteins that are conserved among eukaryotes but whose broad biological roles remain unknown. A broad functional classification of the known proteome using a selection of Gene Ontology biological process categories ("GO Slim") has revealed correlations with features such as subcellular localization and morphological phenotype. Combining available data from genome-wide phenotype and localization experiments with insights from the functional classification of known proteins facilitates prediction of biological roles, and thereby guides specific experimental characterisation of unknown proteins.

We tend to study what we know

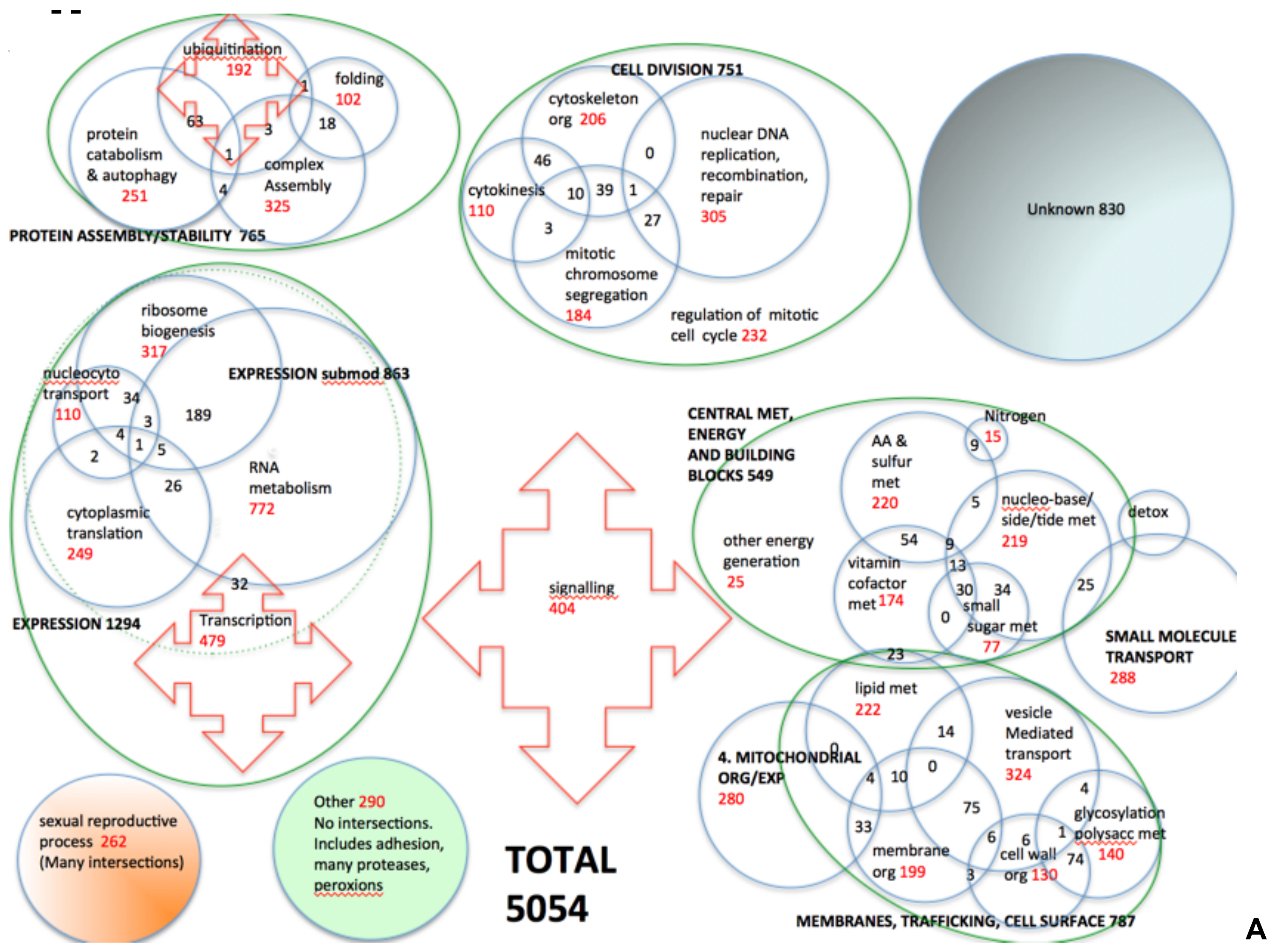


lack of change in the conserved unknown inventory over the pcast decade

Taxonomic distribution of Unknowns :

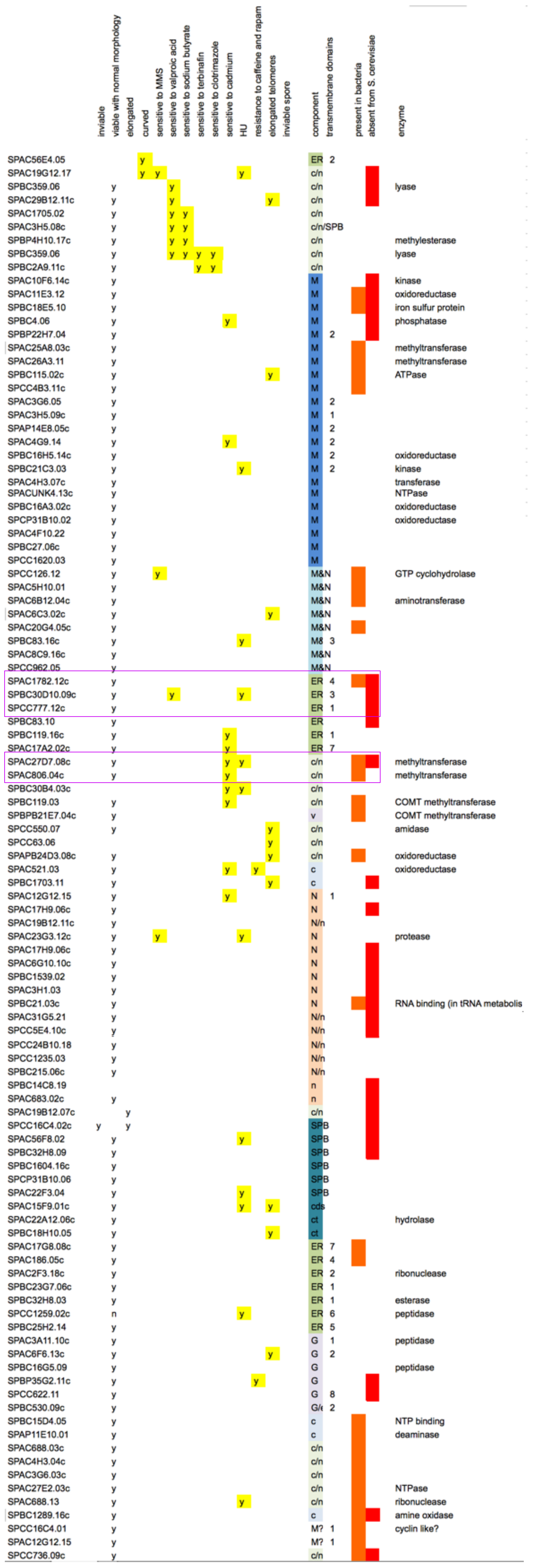


Classifying Knowns: A Visual GO slim:


ubiquitination 192
folding 102
protein catabolism & autophagy 251
complex Assembly 325
63
1
3
18
1
4
PROTEIN ASSEMBLY/STABILITY 765
CELL DIVISION 751
cytoskeleton org 206
nuclear DNA replication, recombination, repair 305
cytokinesis 110
46
0
10
39
1
3
27
mitotic chromosome segregation 184
regulation of mitotic cell cycle 232
Unknown 830
ribosome biogenesis 317
nucleocyto transport 110
EXPRESSION submod 863
34
3
189
4
1
5
2
26
RNA metabolism 772
cytoplasmic translation 249
32
Transcription 479
EXPRESSION 1294
signalling 404
CENTRAL MET, ENERGY AND BUILDING BLOCKS 549
AA & sulfur met 220
Nitrogen 15
9
5
nucleo-base/ side/tide met 219
detox
other energy generation 25
54
9
13
vitamin cofactor met 174
30
34
small sugar met 77
0
25
SMALL MOLECULE TRANSPORT 288
23
lipid met 222
14
vesicle Mediated transport 324
4. MITOCHONDRIAL ORG/EXP 280
0
4
10
0
75
4
33
6
6
1
glycosylation polysacc met 140
membrane org 199
cell wall org 130
3
74
sexual reproductive process 262 (Many intersections)
Other 290 No intersections. Includes adhesion, many proteases, peroxions
TOTAL 5054
MEMBRANES, TRAFFICKING, CELL SURFACE 787

First step: Aim to assign unknowns to these broad classifiers

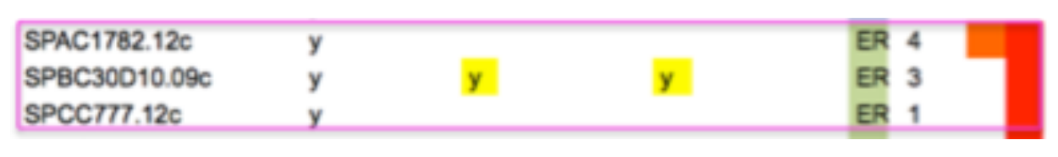
Classifying Unknowns



	inviable	viable with normal morphology	elongated	curved	sensitive to MMS	sensitive to valproic acid	sensitive to sodium butyrate	sensitive to terbinafin	sensitive to clotrimazole	sensitive to cadmium	HU	resistance to caffeine and rapam	elongated telomeres	inviable spore	component	transmembrane domains	present in bacteria	absent from S. cerevisiae	enzyme
SPAC56E4.05				y											ER	2			
SPAC19G12.17				y	y						y				c/n			🟥	
SPBC359.06		y				y									c/n			🟥	lyase
SPAC29B12.11c		y				y							y		c/n			🟥	
SPAC1705.02		y				y	y								c/n				
SPAC3H5.08c		y				y	y								c/n/SPB				
SPBP4H10.17c		y				y	y								c/n				methylesterase
SPBC359.06		y				y	y	y	y						c/n				lyase
SPBC2A9.11c		y						y	y						c/n				
SPAC10F6.14c		y													M			🟥	kinase
SPAC11E3.12		y													M		🟧	🟥	oxidoreductase
SPBC18E5.10		y													M		🟧	🟥	iron sulfur protein
SPBC4.06		y								y					M			🟥	phosphatase
SPBP22H7.04		y													M	2		🟥	
SPAC25A8.03c		y													M		🟧		methyltransferase
SPAC26A3.11		y													M		🟧		methyltransferase
SPBC115.02c		y											y		M		🟧		ATPase
SPCC4B3.11c		y													M		🟧		
SPAC3G6.05		y													M	2			
SPAC3H5.09c		y													M	1			
SPAP14E8.05c		y													M	2			
SPAC4G9.14		y								y					M	2			
SPBC16H5.14c		y													M	2			oxidoreductase
SPBC21C3.03		y									y				M	2			kinase
SPAC4H3.07c		y													M				transferase
SPACUNK4.13c		y													M				NTPase
SPBC16A3.02c		y													M				oxidoreductase
SPCP31B10.02		y													M				oxidoreductase
SPAC4F10.22		y													M				
SPBC27.06c		y													M				
SPCC1620.03		y													M				
SPCC126.12		y			y										M&N		🟧		GTP cyclohydrolase
SPAC5H10.01		y													M&N		🟧		
SPAC6B12.04c		y													M&N		🟧		aminotransferase
SPAC6C3.02c		y											y		M&N				
SPAC20G4.05c		y													M&N		🟧		
SPBC83.16c		y									y				M8	3			
SPAC8C9.16c		y													M&N				
SPCC962.05		y													M&N				
SPAC1782.12c		y													ER	4	🟧	🟥	
SPBC30D10.09c		y				y					y				ER	3		🟥	
SPCC777.12c		y													ER	1		🟥	
SPBC83.10		y													ER			🟥	
SPBC119.16c		y								y					ER	1			
SPAC17A2.02c		y								y					ER	7			
SPAC27D7.08c		y								y	y				c/n		🟧	🟥	methyltransferase
SPAC806.04c		y								y					c/n		🟧		methyltransferase
SPBC30B4.03c										y	y				c/n				
SPBC119.03		y								y					c/n		🟧		COMT methyltransferase
SPBPB21E7.04c		y													v		🟧		COMT methyltransferase
SPCC550.07		y											y		c/n				amidase
SPCC63.06													y		c/n				
SPAPB24D3.08c		y											y		c/n		🟧		oxidoreductase
SPAC521.03		y								y		y			c				oxidoreductase
SPBC1703.11		y											y		c			🟥	
SPAC12G12.15		y								y					N	1			
SPAC17H9.06c		y													N			🟥	
SPAC19B12.11c		y													N/n				
SPAC23G3.12c		y			y						y				N				protease
SPAC17H9.06c		y													N			🟥	
SPAC6G10.10c		y													N			🟥	
SPBC1539.02		y													N			🟥	
SPAC3H1.03		y													N			🟥	
SPBC21.03c		y													N		🟧	🟥	RNA binding (in tRNA metabolis
SPAC31G5.21		y													N/n			🟥	
SPCC5E4.10c		y													N/n			🟥	
SPCC24B10.18		y													N/n				
SPCC1235.03		y													N/n				
SPBC215.06c		y													N/n				
SPBC14C8.19															n			🟥	
SPAC683.02c		y													n			🟥	
SPAC19B12.07c			y												c/n			🟥	
SPCC16C4.02c	y		y												SPB			🟥	
SPAC56F8.02		y									y				SPB			🟥	
SPBC32H8.09		y													SPB			🟥	
SPBC1604.16c		y													SPB				
SPCP31B10.06		y													SPB				
SPAC22F3.04		y									y				SPB				
SPAC15F9.01c		y									y		y		cds				
SPAC22A12.06c		y													ct				hydrolase
SPBC18H10.05		y											y		ct				
SPAC17G8.08c		y													ER	7	🟧		
SPAC186.05c		y													ER	4	🟧		
SPAC2F3.18c		y													ER	2			ribonuclease
SPBC23G7.06c		y													ER	1			
SPBC32H8.03		y													ER	1			esterase
SPCC1259.02c		n									y				ER	6			peptidase
SPBC25H2.14		y													ER	5			
SPAC3A11.10c		y													G	1			peptidase
SPAC6F6.13c		y											y		G	2			
SPBC16G5.09		y													G				peptidase
SPBP35G2.11c		y										y			G			🟥	
SPCC622.11		y													G	8		🟥	
SPBC530.09c		y													G/c	2			
SPBC15D4.05		y													c		🟧		NTP binding
SPAP11E10.01		y													c		🟧		deaminase
SPAC688.03c		y													c/n		🟧		
SPAC4H3.04c		y													c/n		🟧		
SPAC3G6.03c		y													c/n		🟧		
SPAC27E2.03c		y													c/n		🟧		NTPase
SPAC688.13		y									y				c/n		🟧		ribonuclease
SPBC1289.16c		y													c		🟧	🟥	amine oxidase
SPCC16C4.01		y													M?	1	🟧		cyclin like?
SPAC12G12.15		y													M?	1	🟧		
SPCC736.09c		y													c/n		🟧	🟥	

Classification unknowns conserved to human

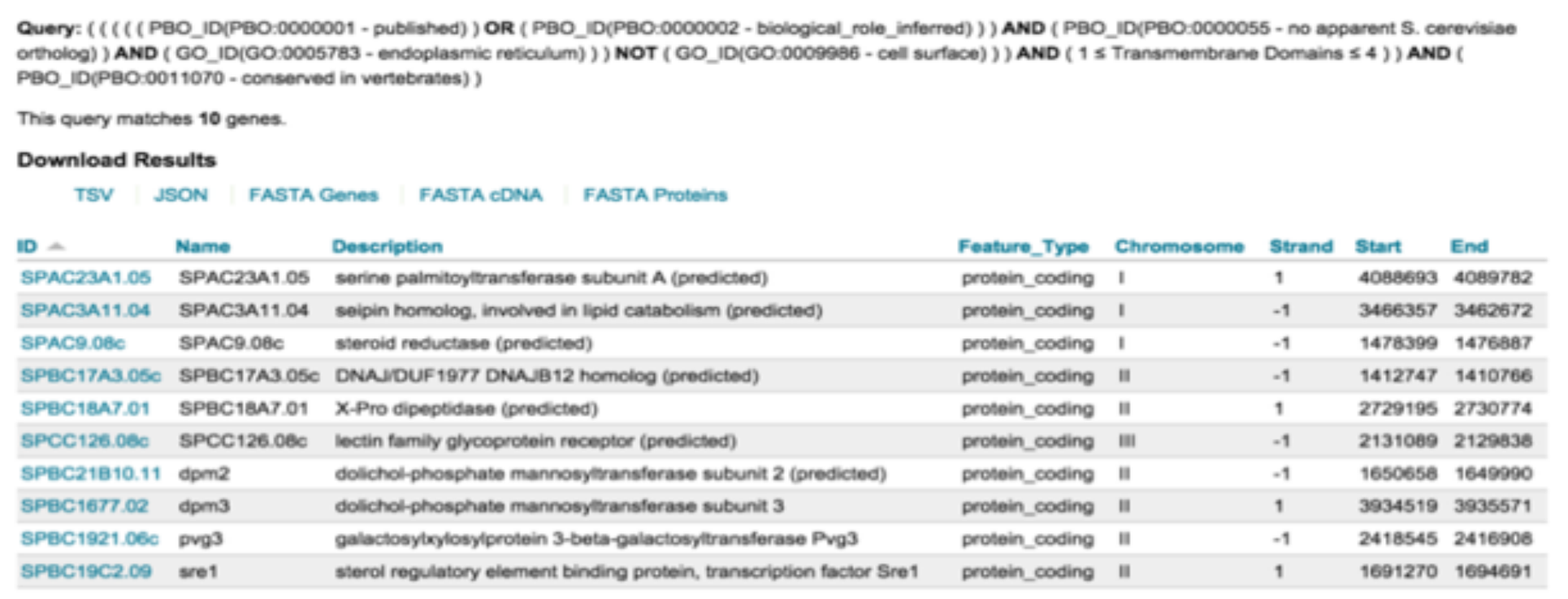
Predicting processes E.g.1



SPAC1782.12c	y			ER	4	🟧	🟥
SPBC30D10.09c	y	y	y	ER	3		🟥
SPCC777.12c	y			ER	1		🟥

1. ER localization
2. >1 <4 TM domain
3. Absent from *S. cerevisiae*
4. Conserved in vertebrates

“which known genes best match these profiles?”



Query: (((((PBO_ID(PBO:0000001 - published)) OR (PBO_ID(PBO:0000002 - biological_role_inferred))) AND (PBO_ID(PBO:0000055 - no apparent S. cerevisiae ortholog)) AND (GO_ID(GO:0005783 - endoplasmic reticulum))) NOT (GO_ID(GO:0009986 - cell surface))) AND (1 ≤ Transmembrane Domains ≤ 4)) AND (PBO_ID(PBO:0011070 - conserved in vertebrates))

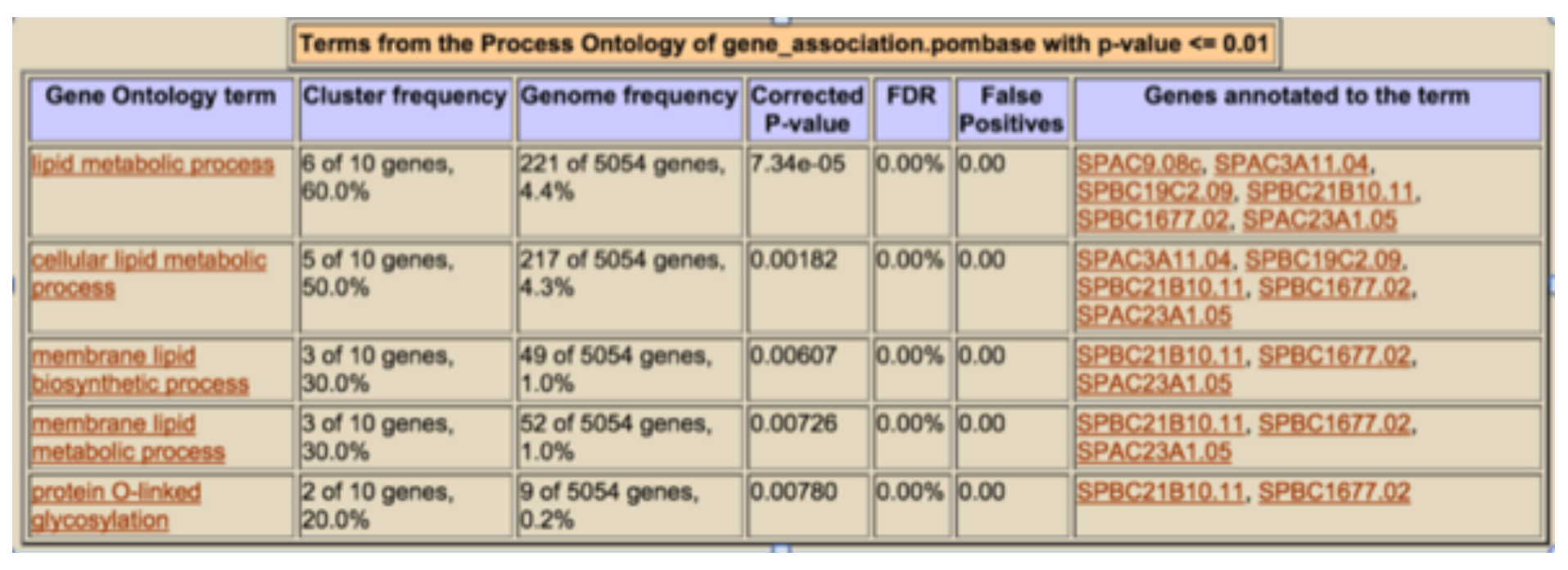
This query matches **10** genes.

Download Results

TSV | JSON | FASTA Genes | FASTA cDNA | FASTA Proteins

ID	Name	Description	Feature_Type	Chromosome	Strand	Start	End
SPAC23A1.05	SPAC23A1.05	serine palmitoyltransferase subunit A (predicted)	protein_coding	I	1	4088693	4089782
SPAC3A11.04	SPAC3A11.04	seipin homolog, involved in lipid catabolism (predicted)	protein_coding	I	-1	3466357	3462672
SPAC9.08c	SPAC9.08c	steroid reductase (predicted)	protein_coding	I	-1	1478399	1476887
SPBC17A3.05c	SPBC17A3.05c	DNAJ/DUF1977 DNAJB12 homolog (predicted)	protein_coding	II	-1	1412747	1410766
SPBC18A7.01	SPBC18A7.01	X-Pro dipeptidase (predicted)	protein_coding	II	1	2729195	2730774
SPCC126.08c	SPCC126.08c	lectin family glycoprotein receptor (predicted)	protein_coding	III	-1	2131089	2129838
SPBC21B10.11	dpm2	dolichol-phosphate mannosyltransferase subunit 2 (predicted)	protein_coding	II	-1	1650658	1649990
SPBC1677.02	dpm3	dolichol-phosphate mannosyltransferase subunit 3	protein_coding	II	1	3934519	3935571
SPBC1921.06c	pvg3	galactosylxylosylprotein 3-beta-galactosyltransferase Pvg3	protein_coding	II	-1	2418545	2416908
SPBC19C2.09	sre1	sterol regulatory element binding protein, transcription factor Sre1	protein_coding	II	1	1691270	1694691

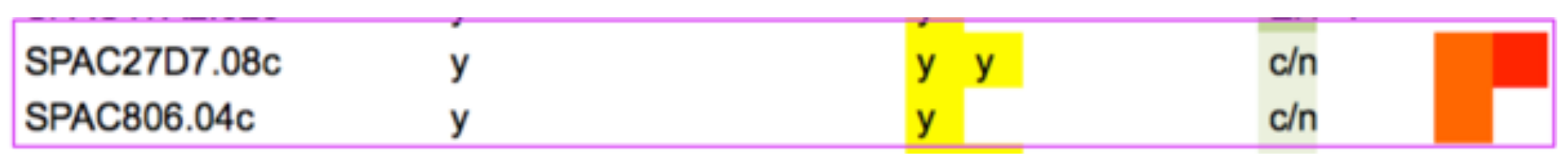
What are these genes enriched for?



Terms from the Process Ontology of gene_association.pombase with p-value <= 0.01

Gene Ontology term	Cluster frequency	Genome frequency	Corrected P-value	FDR	False Positives	Genes annotated to the term
lipid metabolic process	6 of 10 genes, 60.0%	221 of 5054 genes, 4.4%	7.34e-05	0.00%	0.00	SPAC9.08c, SPAC3A11.04, SPBC19C2.09, SPBC21B10.11, SPBC1677.02, SPAC23A1.05
cellular lipid metabolic process	5 of 10 genes, 50.0%	217 of 5054 genes, 4.3%	0.00182	0.00%	0.00	SPAC3A11.04, SPBC19C2.09, SPBC21B10.11, SPBC1677.02, SPAC23A1.05
membrane lipid biosynthetic process	3 of 10 genes, 30.0%	49 of 5054 genes, 1.0%	0.00607	0.00%	0.00	SPBC21B10.11, SPBC1677.02, SPAC23A1.05
membrane lipid metabolic process	3 of 10 genes, 30.0%	52 of 5054 genes, 1.0%	0.00726	0.00%	0.00	SPBC21B10.11, SPBC1677.02, SPAC23A1.05
protein O-linked glycosylation	2 of 10 genes, 20.0%	9 of 5054 genes, 0.2%	0.00780	0.00%	0.00	SPBC21B10.11, SPBC1677.02

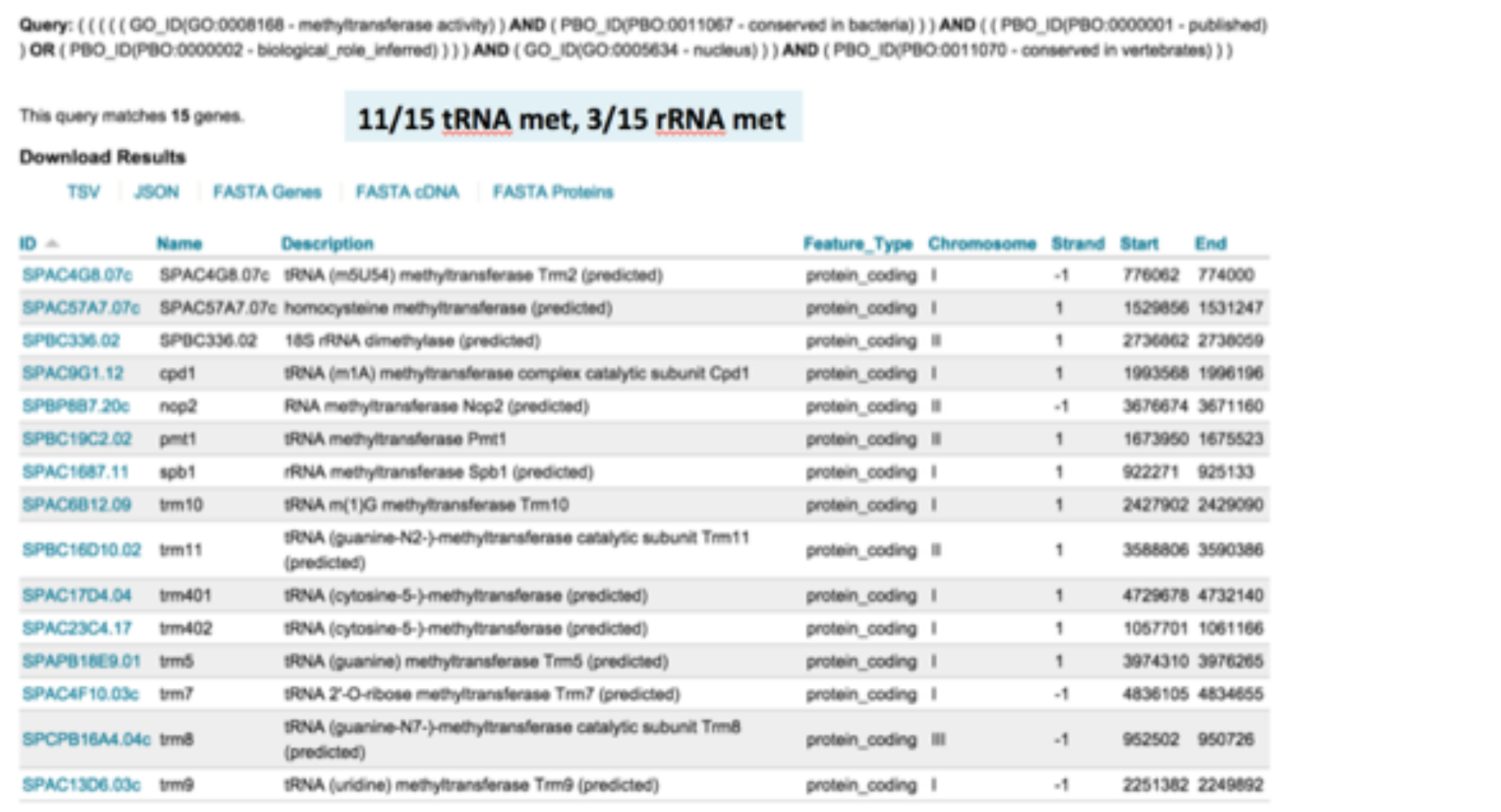
Predicting processes E.g. 2



SPAC27D7.08c	y	y	y	c/n	🟧	🟥
SPAC806.04c	y	y		c/n	🟧	

1. nuclear localization
2. methyltransferase domain
3. conserved in bacteria
4. Conserved in vertebrates

“which known genes best match these profiles?”



Query: (((((GO_ID(GO:0008168 - methyltransferase activity)) AND (PBO_ID(PBO:0011067 - conserved in bacteria))) AND ((PBO_ID(PBO:0000001 - published)) OR (PBO_ID(PBO:0000002 - biological_role_inferred)))) AND (GO_ID(GO:0005634 - nucleus))) AND (PBO_ID(PBO:0011070 - conserved in vertebrates)))

This query matches 15 genes.

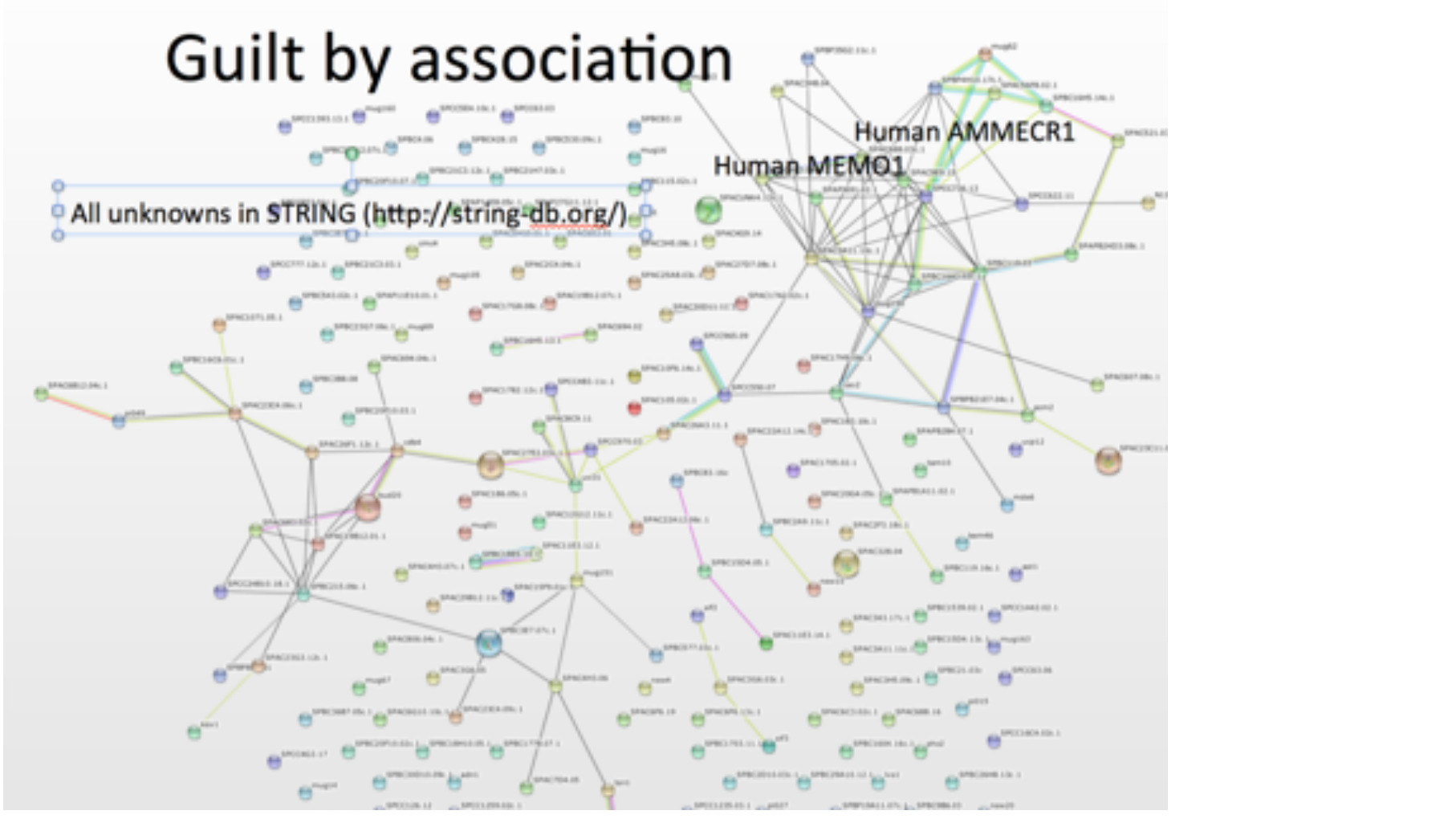
11/15 tRNA met, 3/15 rRNA met

Download Results

TSV | JSON | FASTA Genes | FASTA cDNA | FASTA Proteins

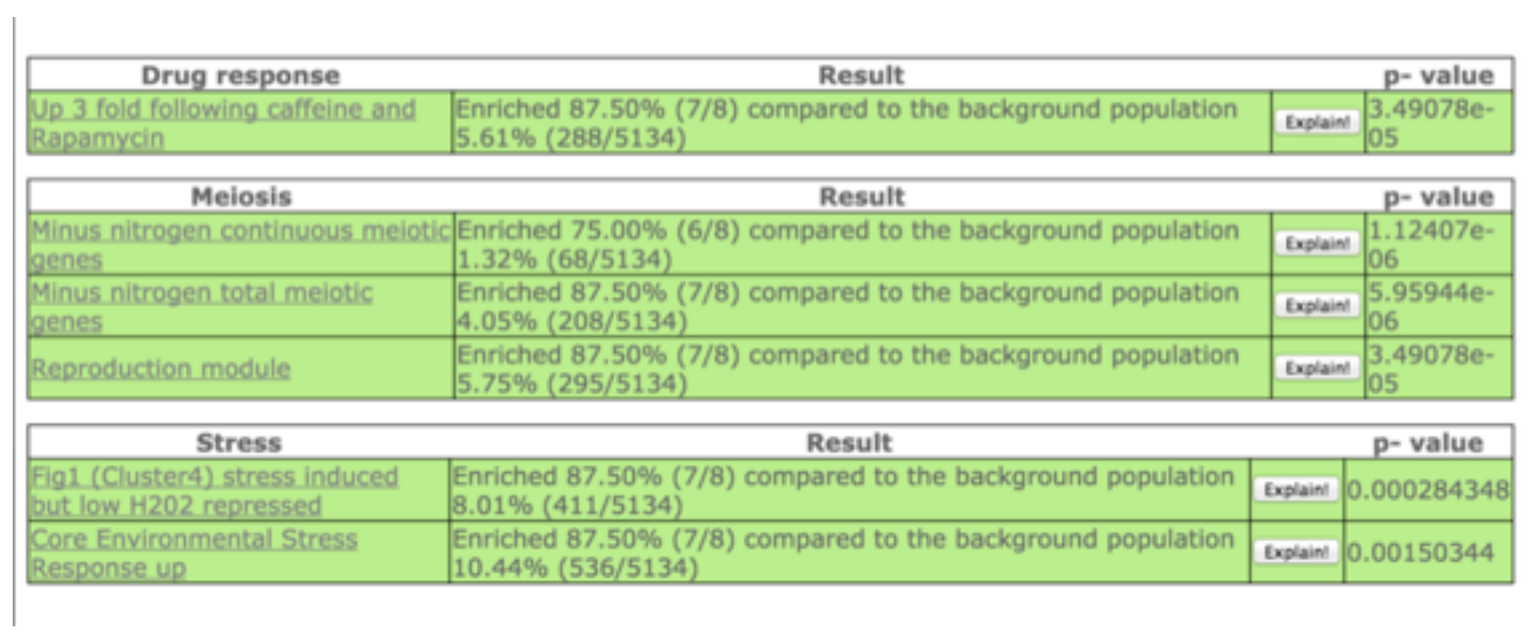
ID	Name	Description	Feature_Type	Chromosome	Strand	Start	End
SPAC4G8.07c	SPAC4G8.07c	tRNA (m5U54) methyltransferase Trm2 (predicted)	protein_coding	I	-1	776062	774000
SPAC57A7.07c	SPAC57A7.07c	homocysteine methyltransferase (predicted)	protein_coding	I	1	1529856	1531247
SPBC336.02	SPBC336.02	18S rRNA dimethylase (predicted)	protein_coding	II	1	2736862	2738059
SPAC9G1.12	cpd1	tRNA (m1A) methyltransferase complex catalytic subunit Cpd1	protein_coding	I	1	1993568	1996196
SPBP8B7.20c	nop2	RNA methyltransferase Nop2 (predicted)	protein_coding	II	-1	3676674	3671160
SPBC19C2.02	pmt1	tRNA methyltransferase Pmt1	protein_coding	II	1	1673950	1675523
SPAC1687.11	spb1	rRNA methyltransferase Spb1 (predicted)	protein_coding	I	1	922271	925133
SPAC6B12.09	trm10	tRNA m(1)G methyltransferase Trm10	protein_coding	I	1	2427902	2429090
SPBC16D10.02	trm11	tRNA (guanine-N2-)-methyltransferase catalytic subunit Trm11 (predicted)	protein_coding	II	1	3588806	3590386
SPAC17D4.04	trm401	tRNA (cytosine-5-)-methyltransferase (predicted)	protein_coding	I	1	4729678	4732140
SPAC23C4.17	trm402	tRNA (cytosine-5-)-methyltransferase (predicted)	protein_coding	I	1	1057701	1061166
SPAPB18E9.01	trm5	tRNA (guanine) methyltransferase Trm5 (predicted)	protein_coding	I	1	3974310	3976265
SPAC4F10.03c	trm7	tRNA 2'-O-ribose methyltransferase Trm7 (predicted)	protein_coding	I	-1	4836105	4834655
SPCPB16A4.04c	trm8	tRNA (guanine-N7-)-methyltransferase catalytic subunit Trm8 (predicted)	protein_coding	III	-1	952502	950726
SPAC13D6.03c	trm9	tRNA (uridine) methyltransferase Trm9 (predicted)	protein_coding	I	-1	2251382	2249892

Predicting processes E.g. 3



"

Using Angeli http://bahlerweb.cs.ucl.ac.uk/cgi-bin/GLA/GLA_input
AMMECR subnetwork has connections to meiosis, possibly signalling



Drug response	Result		p- value
Up 3 fold following caffeine and Rapamycin	Enriched 87.50% (7/8) compared to the background population 5.61% (288/5134)	Explain!	3.49078e-05

Meiosis	Result		p- value
Minus nitrogen continuous meiotic genes	Enriched 75.00% (6/8) compared to the background population 1.32% (68/5134)	Explain!	1.12407e-06
Minus nitrogen total meiotic genes	Enriched 87.50% (7/8) compared to the background population 4.05% (208/5134)	Explain!	5.95944e-06
Reproduction module	Enriched 87.50% (7/8) compared to the background population 5.75% (295/5134)	Explain!	3.49078e-05

Stress	Result		p- value
Fig1 (Cluster4) stress induced but low H202 repressed	Enriched 87.50% (7/8) compared to the background population 8.01% (411/5134)	Explain!	0.000284348
Core Environmental Stress Response up	Enriched 87.50% (7/8) compared to the background population 10.44% (536/5134)	Explain!	0.00150344

AUTHOR AFFILIATIONS
1. PomBase, Cambridge University, Tennis Court Rd, Cambridge, CB2 1QW, UK
2. PomBase, University College London, Darwin Building, Gower Street, London WC1E 6BT, UK