COGs

Phylogenetic classification of proteins encoded in complete genomes

NCBI

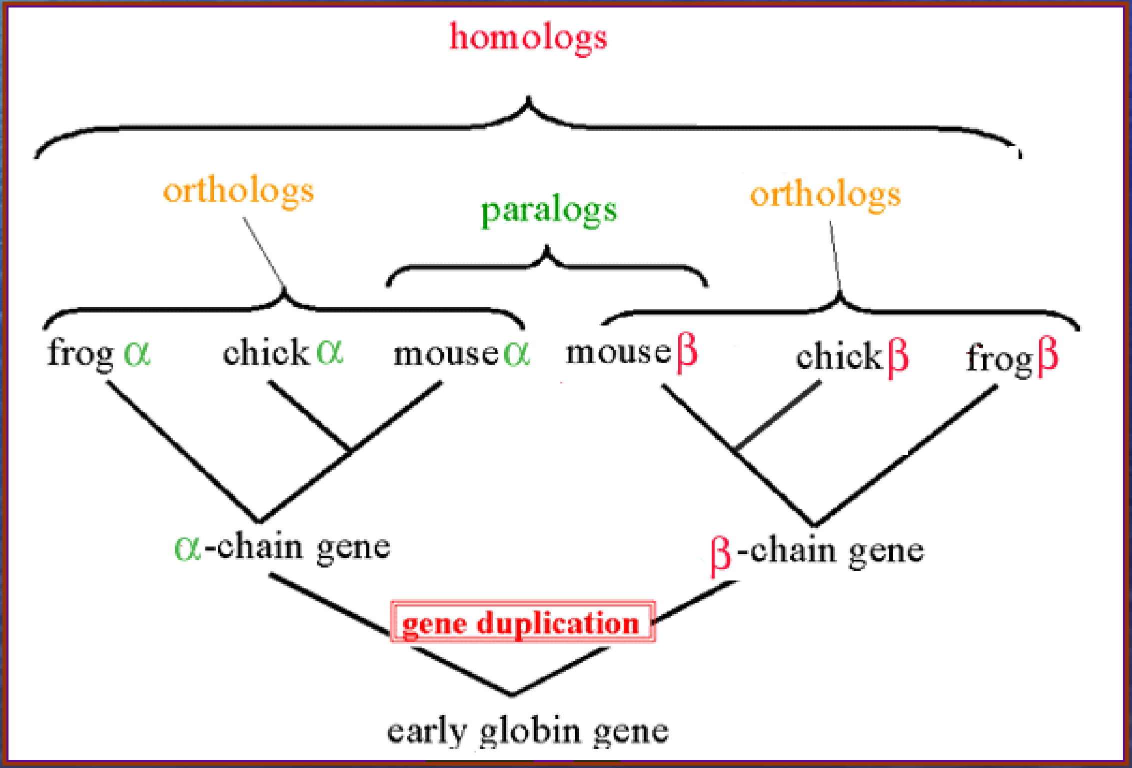
Cluster of Orthologous Groups (COGs)

Harold Ardila

Carlos A. Niño

COGs

- Ortólogos son genes que se encuentran en diferentes especies y provienen de un gen ancestral común. Generalmente presentan la misma función a través de la evolución.



COGs

- Cada COG se define como el grupo de tres o más proteínas que son **Ortólogas**
- Han sido desarrollados mediante la comparación, todos contra todos, de secuencias de proteínas codificadas en genomas completos secuenciados, utilizando el programa BLAST.

- 5666 COGs (unicelulares)
- 4852 KOGs (eucariotas)

What does NCBI do?

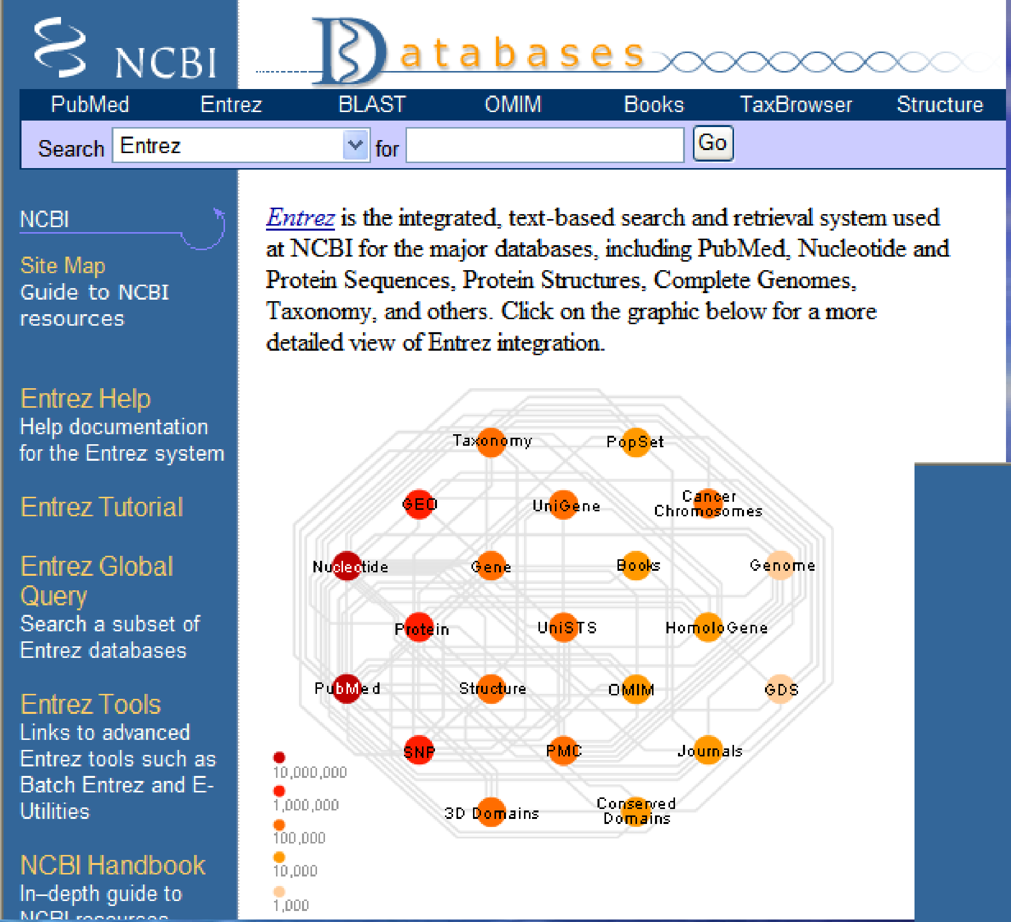
Established in 1988 as a national resource for molecular biology information, NCBI creates public databases, conducts research in computational biology, develops software tools for analyzing genome data, and disseminates biomedical information - all for the better understanding of molecular processes affecting human health and disease. More...

New dbGaP

NCBI's dbGaP Genome Wide Association Database

NCBI's dbGaP (database of Genotype and Phenotype) provides data from Genome Wide Association (GWA) studies. The resource is intended to help elucidate the link between genes and disease. For each study, users have access to detailed information about the phenotypic variables measured and pre-computed associations between subjects' phenotypes and genotypes. Click here to read the press release. To read more about GWA projects, see NCBI's GWA resource page

ACCESO



Site Map
Guide to NCBI resources

Entrez Help
Help documentation for the Entrez system

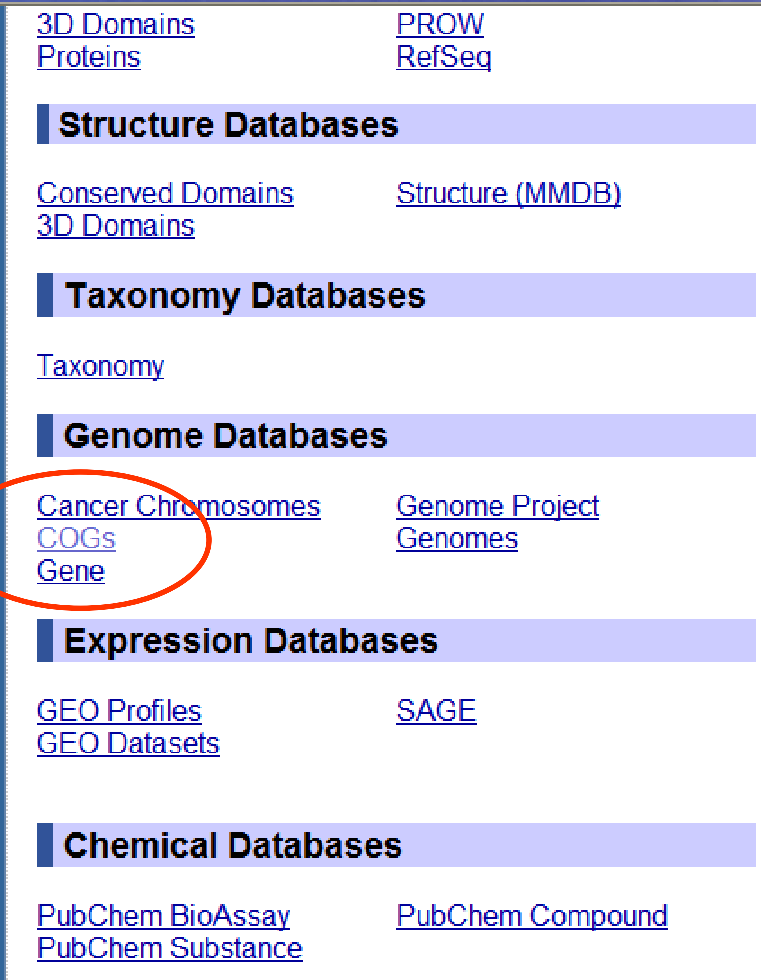
Entrez Tutorial

Entrez Global Query
Search a subset of Entrez databases

Entrez Tools
Links to advanced Entrez tools such as Batch Entrez and E-Utilities

NCBI Handbook
In–depth guide to

Entrez is the integrated, text-based search and retrieval system used at NCBI for the major databases, including PubMed, Nucleotide and Protein Sequences, Protein Structures, Complete Genomes, Taxonomy, and others. Click on the graphic below for a more detailed view of Entrez integration.



3D Domains Proteins	PROW RefSeq

Structure Databases

Conserved Domains 3D Domains	Structure (MMDB)

Taxonomy Databases

Taxonomy

Genome Databases

Cancer Chromosomes COGs Gene	Genome Project Genomes

Expression Databases

GEO Profiles GEO Datasets	SAGE

Chemical Databases

PubChem BioAssay PubChem Substance	PubChem Compound

COGs

Phylogenetic classification of proteins encoded in complete genomes

Clusters of Orthologous Groups of proteins (COGs) were delineated by comparing protein sequences encoded in complete genomes, representing major phylogenetic lineages. Each COG consists of individual proteins or groups of paralogs from at least 3 lineages and thus corresponds to an ancient conserved domain.

1

66 genomes
38 *orders*
28 *classes*
14 *phyla*

Unicellular clusters **FTP**

Science 1997 Oct 24;278(5338):631-7.
BMC Bioinformatics 2003 Sep 11;4(1):41.

Initial version

Euryarchaeota	
Methanobacteriales	Mth
Methanococcales	Mja
Halobacteriales	Hbs
Thermoplasmatales	Tac Tvo
Thermococcales	Pho Pab
Archaeoglobales	Afu
Methanopyrales	Mka
Methanosarcinales	Mac

Crenarchaeota	
Thermoproteales	Pya
Sulfolobales	Sso
Desulfurococcales	Ape

Aquificae	
Aquificales	Aae

Thermotogae	
Thermotogales	Tma

Cyanobacteria	
Nostocales	Nos
Chroococcales	Syn

Deinococcus-Thermus	
Deinococcales	Dra

Fusobacteria	
Fusobacterales	Fnu

Actinobacteria	
Actinomycetales	Cgl Mtu MtC Mle

Firmicutes	
Clostridiales	Cac
Bacillales	Sau Lin Bsu Bha
Lactobacillales	Lla Spy Spn
Mycoplasmatales	Uur Mpu Mpn Mge

Proteobacteria	
Pseudomonadales	Pae
Enterobacteriales	Eco EcZ Ecs Ype Sty Buc
Xanthomonadales	Xfa
Vibrionales	Vch
Pasteurellales	Hin Pmu

2

Eukaryotic Clusters **FTP**

Code	*Name*	Abbreviation
A	*Arabidopsis thaliana* (thale cress)	***ath***
C	*Caenorhabditis elegans* (worm)	***cel***
D	*Drosophila melanogaster* (fruit fly)	***dme***
H	*Homo sapiens* (human)	***hsa***
Y	*Saccharomyces cerevisiae* (baker yeast)	***sce***
P	*Schizosaccharomyces pombe* (fission yeast)	***spo***
E	*Encephalitozoon cuniculi* (Microsporidia)	***ecu***

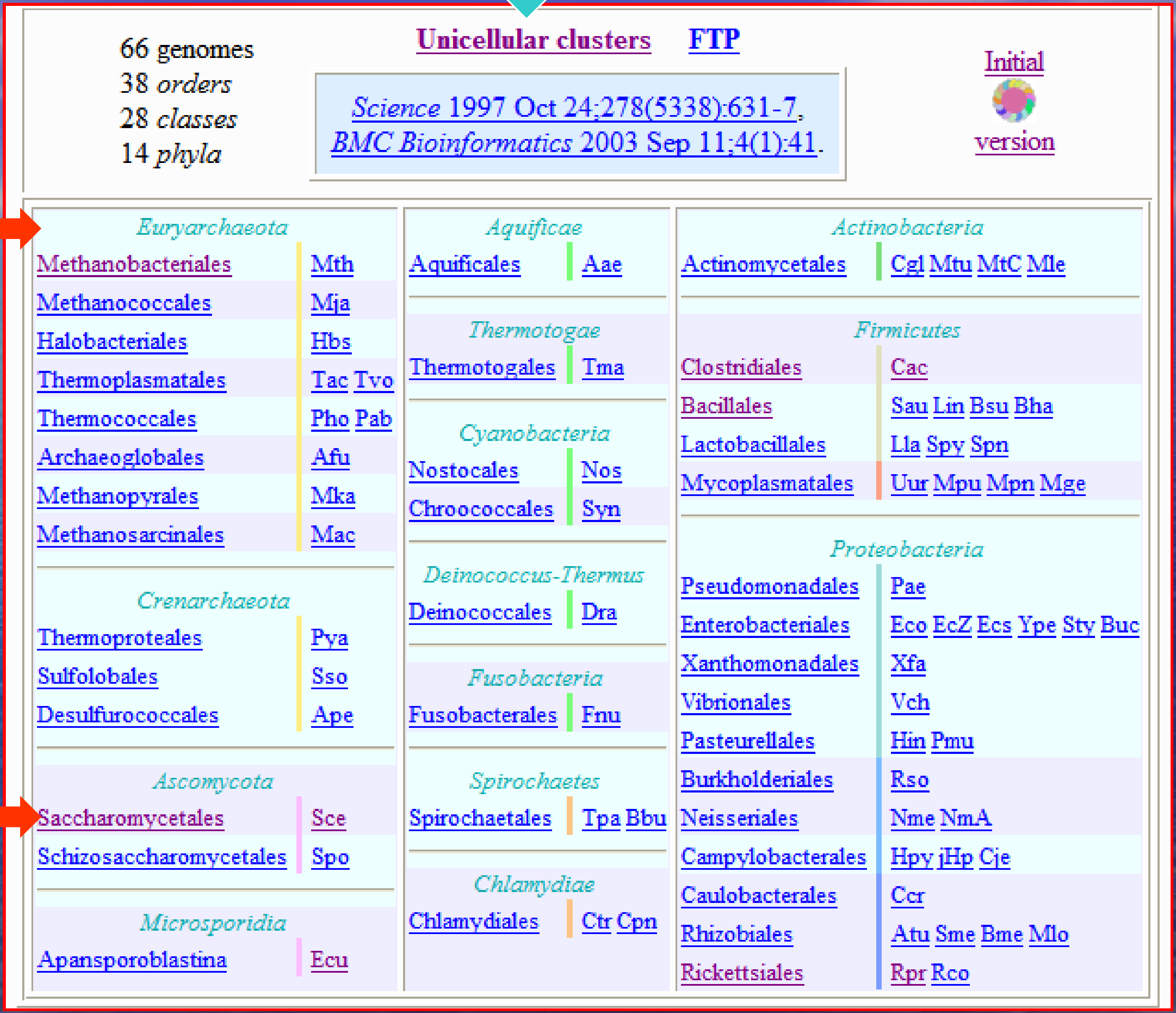
Upcoming eukaryotic genomes

1. **COGs – UNICELULARES**
2. **KOGs - EUCARIOTES**

Phyla

ORDEN

Taxonomy DB



66 genomes
38 *orders*
28 *classes*
14 *phyla*

Unicellular clusters **FTP**

Science 1997 Oct 24;278(5338):631-7,
BMC Bioinformatics 2003 Sep 11;4(1):41.

Initial version

Euryarchaeota	
Methanobacteriales	Mth
Methanococcales	Mja
Halobacteriales	Hbs
Thermoplasmatales	Tac Tvo
Thermococcales	Pho Pab
Archaeoglobales	Afu
Methanopyrales	Mka
Methanosarcinales	Mac

Crenarchaeota	
Thermoproteales	Pya
Sulfolobales	Sso
Desulfurococcales	Ape

Ascomycota	
Saccharomycetales	Sce
Schizosaccharomycetales	Spo

Microsporidia	
Apansporoblastina	Ecu

Aquificae	
Aquificales	Aae

Thermotogae	
Thermotogales	Tma

Cyanobacteria	
Nostocales	Nos
Chroococcales	Syn

Deinococcus-Thermus	
Deinococcales	Dra

Fusobacteria	
Fusobacterales	Fnu

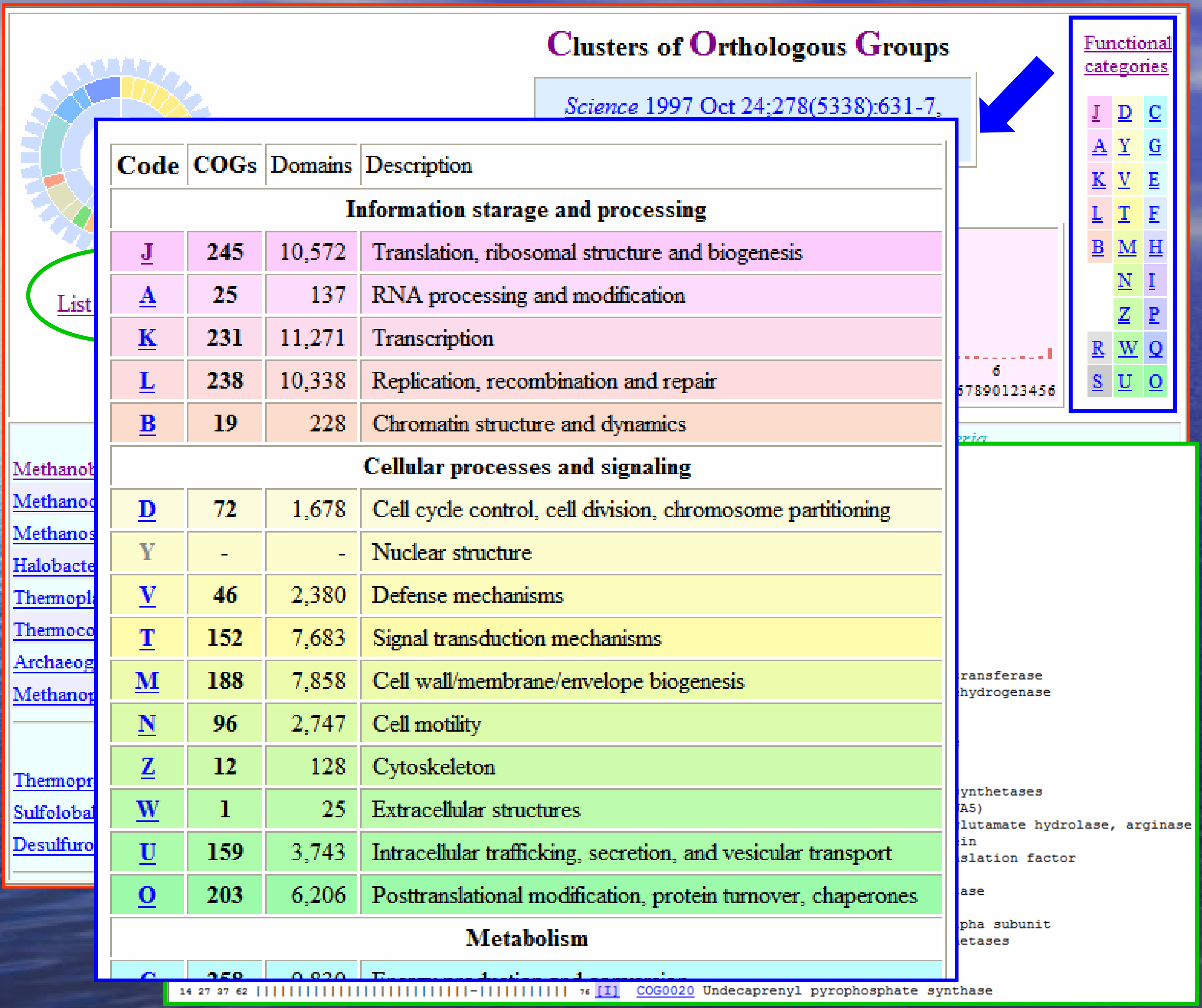
Spirochaetes	
Spirochaetales	Tpa Bbu

Chlamydiae	
Chlamydiales	Ctr Cpn

Actinobacteria	
Actinomycetales	Cgl Mtu MtC Mle

Firmicutes	
Clostridiales	Cac
Bacillales	Sau Lin Bsu Bha
Lactobacillales	Lla Spy Spn
Mycoplasmatales	Uur Mpu Mpn Mge

Proteobacteria	
Pseudomonadales	Pae
Enterobacteriales	Eco EcZ Ecs Ype Sty Buc
Xanthomonadales	Xfa
Vibrionales	Vch
Pasteurellales	Hin Pmu
Burkholderiales	Rso
Neisseriales	Nme NmA
Campylobacterales	Hpy jHp Cje
Caulobacterales	Ccr
Rhizobiales	Atu Sme Bme Mlo
Rickettsiales	Rpr Rco



Clusters of Orthologous Groups

Science 1997 Oct 24;278(5338):631-7,

Functional categories

J	D	C
A	Y	G
K	V	E
L	T	F
B	M	H
	N	I
	Z	P
R	W	Q
S	U	O

Code	COGs	Domains	Description
Information starage and processing			
J	245	10,572	Translation, ribosomal structure and biogenesis
A	25	137	RNA processing and modification
K	231	11,271	Transcription
L	238	10,338	Replication, recombination and repair
B	19	228	Chromatin structure and dynamics
Cellular processes and signaling			
D	72	1,678	Cell cycle control, cell division, chromosome partitioning
Y	-	-	Nuclear structure
V	46	2,380	Defense mechanisms
T	152	7,683	Signal transduction mechanisms
M	188	7,858	Cell wall/membrane/envelope biogenesis
N	96	2,747	Cell motility
Z	12	128	Cytoskeleton
W	1	25	Extracellular structures
U	159	3,743	Intracellular trafficking, secretion, and vesicular transport
O	203	6,206	Posttranslational modification, protein turnover, chaperones
Metabolism			

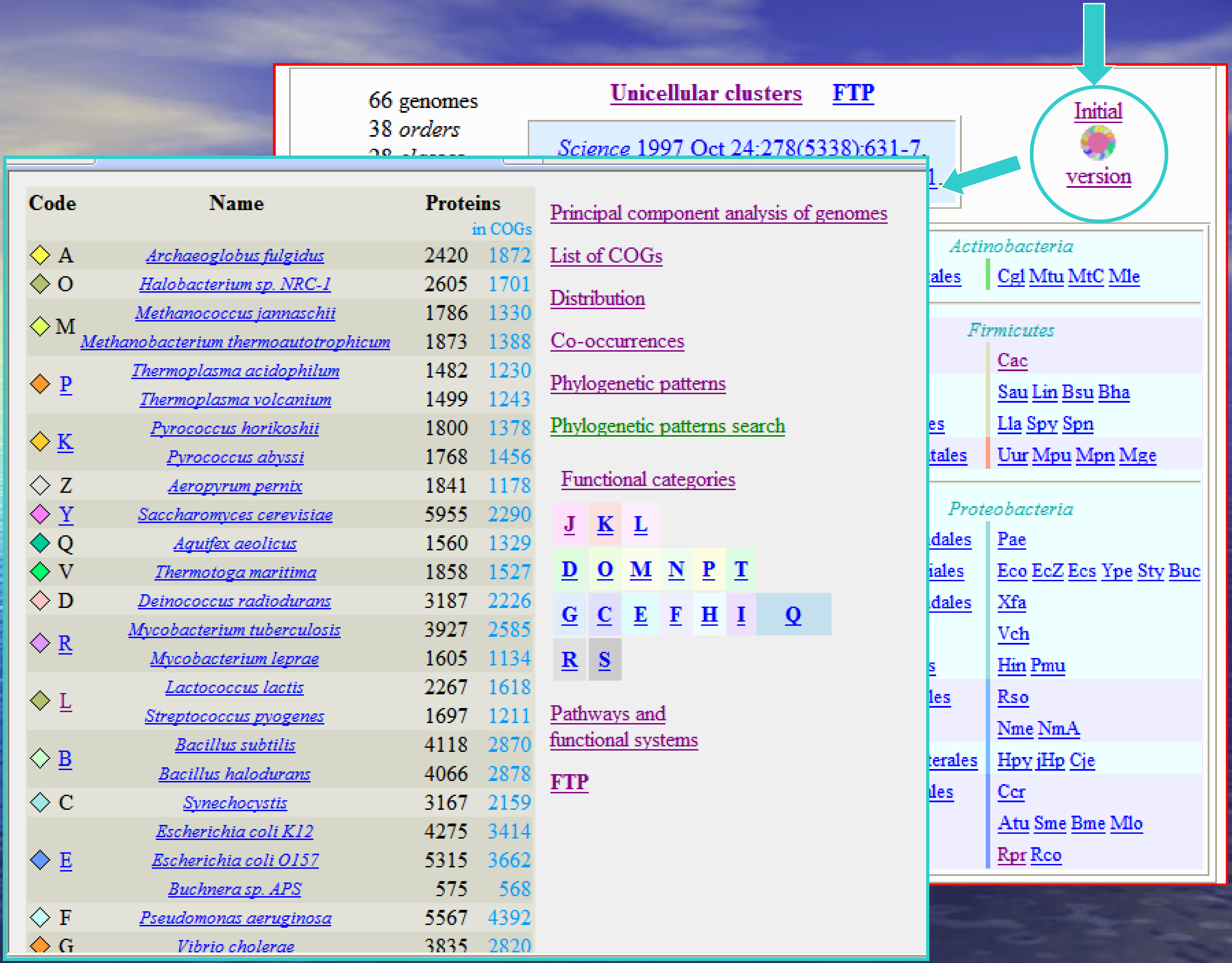
```
ransferase
hydrogenase

ynthetases
A5)
lutamate hydrolase, arginase
in
slation factor

ase

pha subunit
etases

14 27 37 62 ||||||||||||||||||||||||||||-||||||||||| 76 [I]  COG0020 Undecaprenyl pyrophosphate synthase
```



66 genomes
38 orders
28 classes

Unicellular clusters FTP

Science 1997 Oct 24;278(5338):631-7.

Initial version

Code	Name	Proteins	in COGs
◆ A	*Archaeoglobus fulgidus*	2420	1872
◆ O	*Halobacterium sp. NRC-1*	2605	1701
◆ M	*Methanococcus jannaschii*	1786	1330
	Methanobacterium thermoautotrophicum	1873	1388
◆ P	*Thermoplasma acidophilum*	1482	1230
	Thermoplasma volcanium	1499	1243
◆ K	*Pyrococcus horikoshii*	1800	1378
	Pyrococcus abyssi	1768	1456
◇ Z	*Aeropyrum pernix*	1841	1178
◆ Y	*Saccharomyces cerevisiae*	5955	2290
◆ Q	*Aquifex aeolicus*	1560	1329
◆ V	*Thermotoga maritima*	1858	1527
◇ D	*Deinococcus radiodurans*	3187	2226
◆ R	*Mycobacterium tuberculosis*	3927	2585
	Mycobacterium leprae	1605	1134
◆ L	*Lactococcus lactis*	2267	1618
	Streptococcus pyogenes	1697	1211
◇ B	*Bacillus subtilis*	4118	2870
	Bacillus halodurans	4066	2878
◇ C	*Synechocystis*	3167	2159
◆ E	*Escherichia coli K12*	4275	3414
	Escherichia coli O157	5315	3662
	Buchnera sp. APS	575	568
◇ F	*Pseudomonas aeruginosa*	5567	4392
◆ G	*Vibrio cholerae*	3835	2820

Principal component analysis of genomes

List of COGs

Distribution

Co-occurrences

Phylogenetic patterns

Phylogenetic patterns search

Functional categories

J	**K**	**L**				
D	**O**	**M**	**N**	**P**	**T**	
G	**C**	**E**	**F**	**H**	**I**	**Q**
R	**S**					

Pathways and functional systems

FTP

Actinobacteria

ales Cgl Mtu MtC Mle

Firmicutes

Cac

Sau Lin Bsu Bha

es Lla Spy Spn

tales Uur Mpu Mpn Mge

Proteobacteria

dales Pae

iales Eco EcZ Ecs Ype Sty Buc

dales Xfa

Vch

s Hin Pmu

les Rso

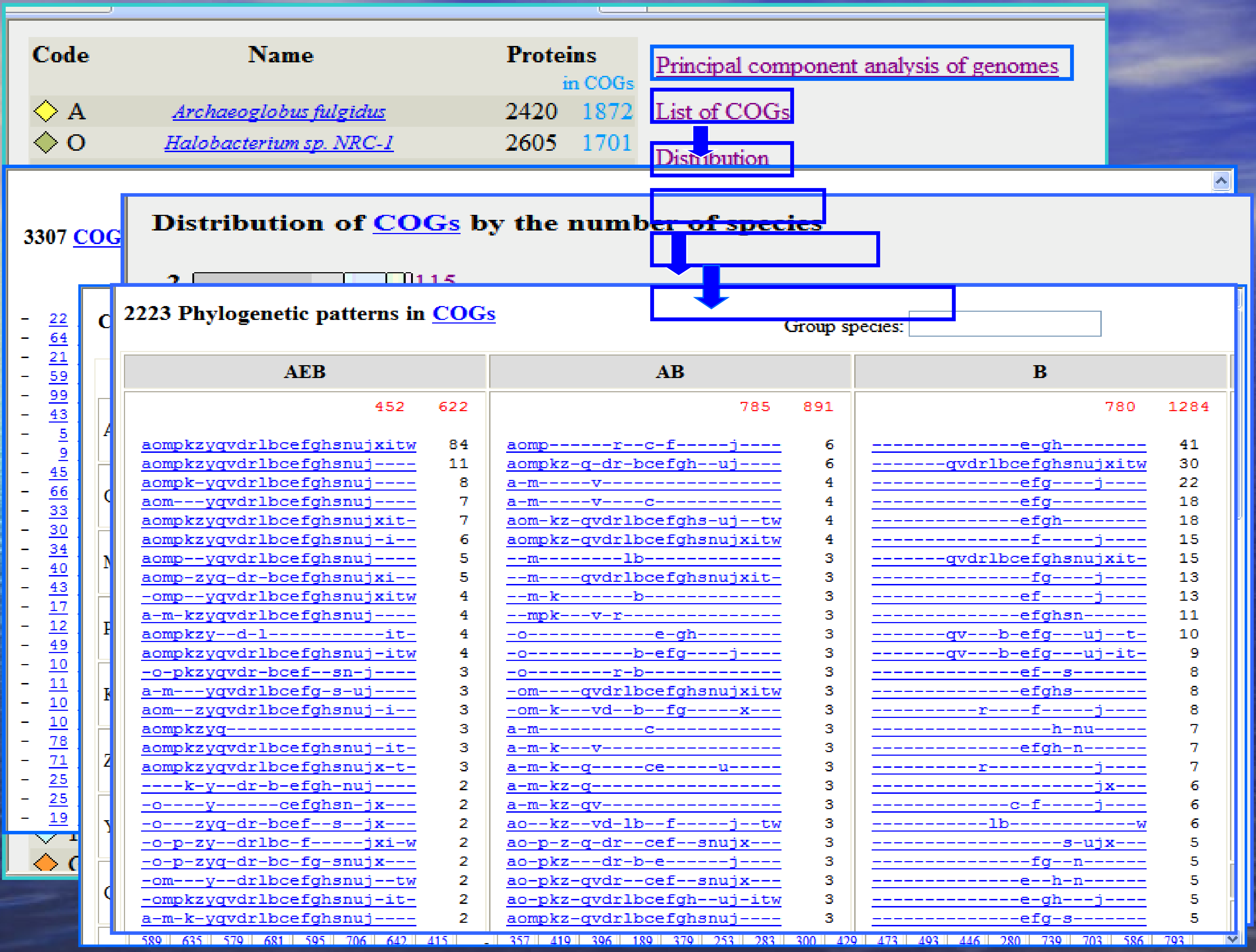
Nme NmA

terales Hpy jHp Cje

les Ccr

Atu Sme Bme Mlo

Rpr Rco



Code	Name	Proteins	in COGs
◇ A	*Archaeoglobus fulgidus*	2420	1872
◇ O	*Halobacterium sp. NRC-1*	2605	1701

Principal component analysis of genomes

List of COGs

Distribution

3307 COG

Distribution of COGs by the number of species

2 115

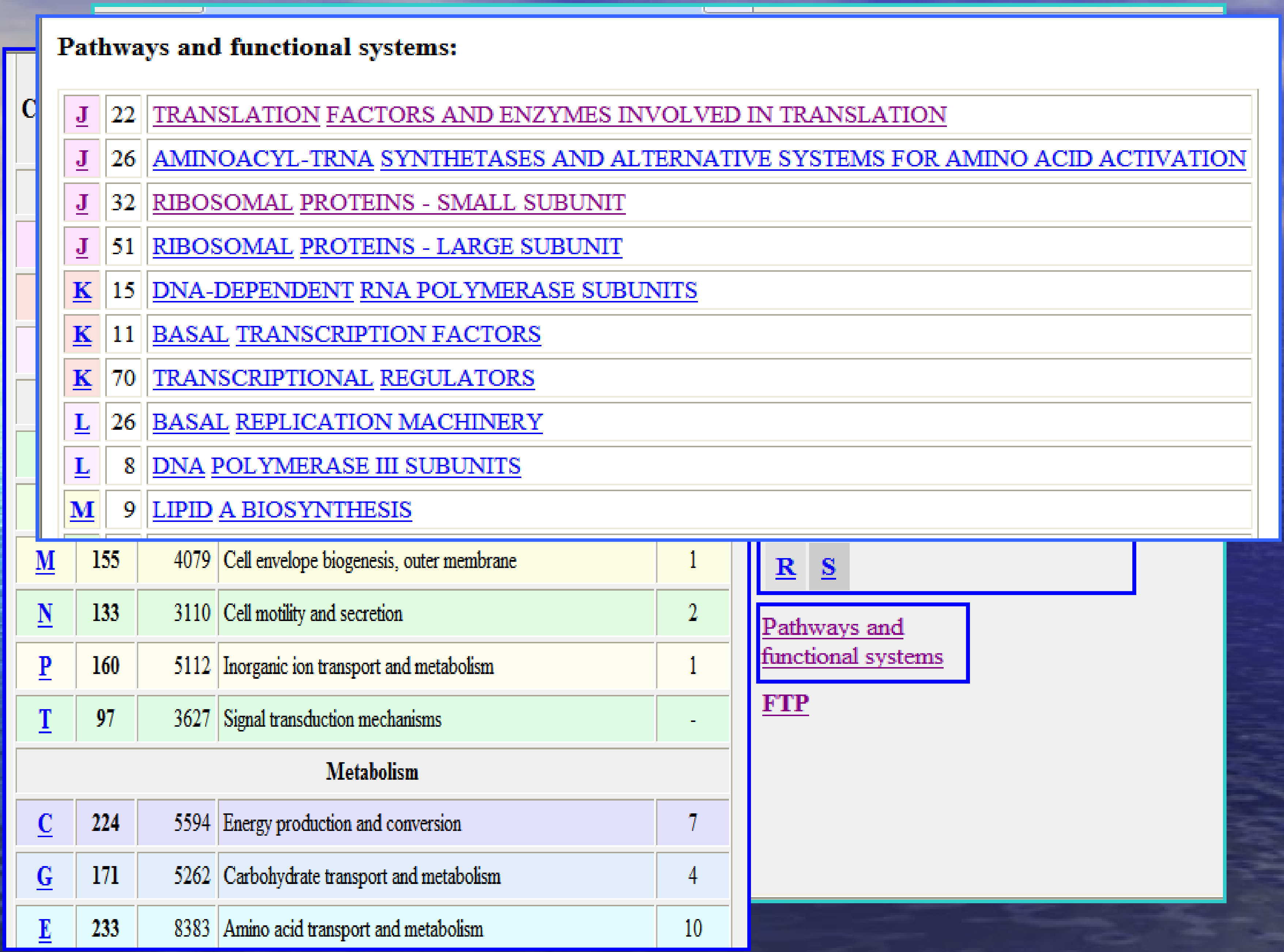
– 22
– 64
– 21
– 59
– 99
– 43
– 5
– 9
– 45
– 66
– 33
– 30
– 34
– 40
– 43
– 17
– 12
– 49
– 10
– 11
– 10
– 10
– 78
– 71
– 25
– 25
– 19

2223 Phylogenetic patterns in COGs

Group species:

AEB		AB		B	
452	622	785	891	780	1284
aompkzyqvdrlbcefghsnujxitw	84	aomp------r--c-f-----j----	6	--------------e-gh--------	41
aompkzyqvdrlbcefghsnuj----	11	aompkz-q-dr-bcefgh--uj----	6	-------qvdrlbcefghsnujxitw	30
aompk-yqvdrlbcefghsnuj----	8	a-m-----v-----------------	4	--------------efg----j----	22
aom---yqvdrlbcefghsnuj----	7	a-m-----v----c------------	4	--------------efg---------	18
aompkzyqvdrlbcefghsnujxit-	7	aom-kz-qvdrlbcefghs-uj--tw	4	--------------efgh--------	18
aompkzyqvdrlbcefghsnuj-i--	6	aompkz-qvdrlbcefghsnujxitw	4	---------------f-----j----	15
aomp--yqvdrlbcefghsnuj----	5	--m--------lb-------------	3	-------qvdrlbcefghsnujxit-	15
aomp-zyq-dr-bcefghsnujxi--	5	--m----qvdrlbcefghsnujxit-	3	---------------fg----j----	13
-omp--yqvdrlbcefghsnujxitw	4	--m-k-------b-------------	3	--------------ef-----j----	13
a-m-kzyqvdrlbcefghsnuj----	4	--mpk---v-r---------------	3	--------------efghsn------	11
aompkzy--d-l-----------it-	4	-o------------e-gh--------	3	-------qv---b-efg---uj--t-	10
aompkzyqvdrlbcefghsnuj-itw	4	-o----------b-efg----j----	3	-------qv---b-efg---uj-it-	9
-o-pkzyqvdr-bcef--sn-j----	3	-o--------r-b-------------	3	--------------ef--s-------	8
a-m---yqvdrlbcefg-s-uj----	3	-om----qvdrlbcefghsnujxitw	3	--------------efghs-------	8
aom--zyqvdrlbcefghsnuj-i--	3	-om-k---vd--b--fg-----x---	3	----------r----f-----j----	8
aompkzyq------------------	3	a-m----------c------------	3	-----------------h-nu-----	7
aompkzyqvdrlbcefghsnuj-it-	3	a-m-k---v-----------------	3	--------------efgh-n------	7
aompkzyqvdrlbcefghsnujx-t-	3	a-m-k--q-----ce-----u-----	3	----------r----------j----	7
----k-y--dr-b-efgh-nuj----	2	a-m-kz-q------------------	3	---------------------jx---	6
-o----y------cefghsn-jx---	2	a-m-kz-qv-----------------	3	-------------c-f-----j----	6
-o---zyq-dr-bcef--s--jx---	2	ao--kz--vd-lb--f-----j--tw	3	-----------lb------------w	6
-o-p-zy--drlbc-f-----jxi-w	2	ao-p-z-q-dr--cef--snujx---	3	------------------s-ujx---	5
-o-p-zyq-dr-bc-fg-snujx---	2	ao-pkz---dr-b-e------j----	3	---------------fg--n------	5
-om---y--drlbcefghsnuj--tw	2	ao-pkz-qvdr--cef--snujx---	3	--------------e--h-n------	5
-ompkzyqvdrlbcefghsnuj-it-	2	ao-pkz-qvdrlbcefgh--uj-itw	3	--------------e-gh---j----	5
a-m-k-yqvdrlbcefghsnuj----	2	aompkz-qvdrlbcefghsnuj----	3	--------------efg-s-------	5

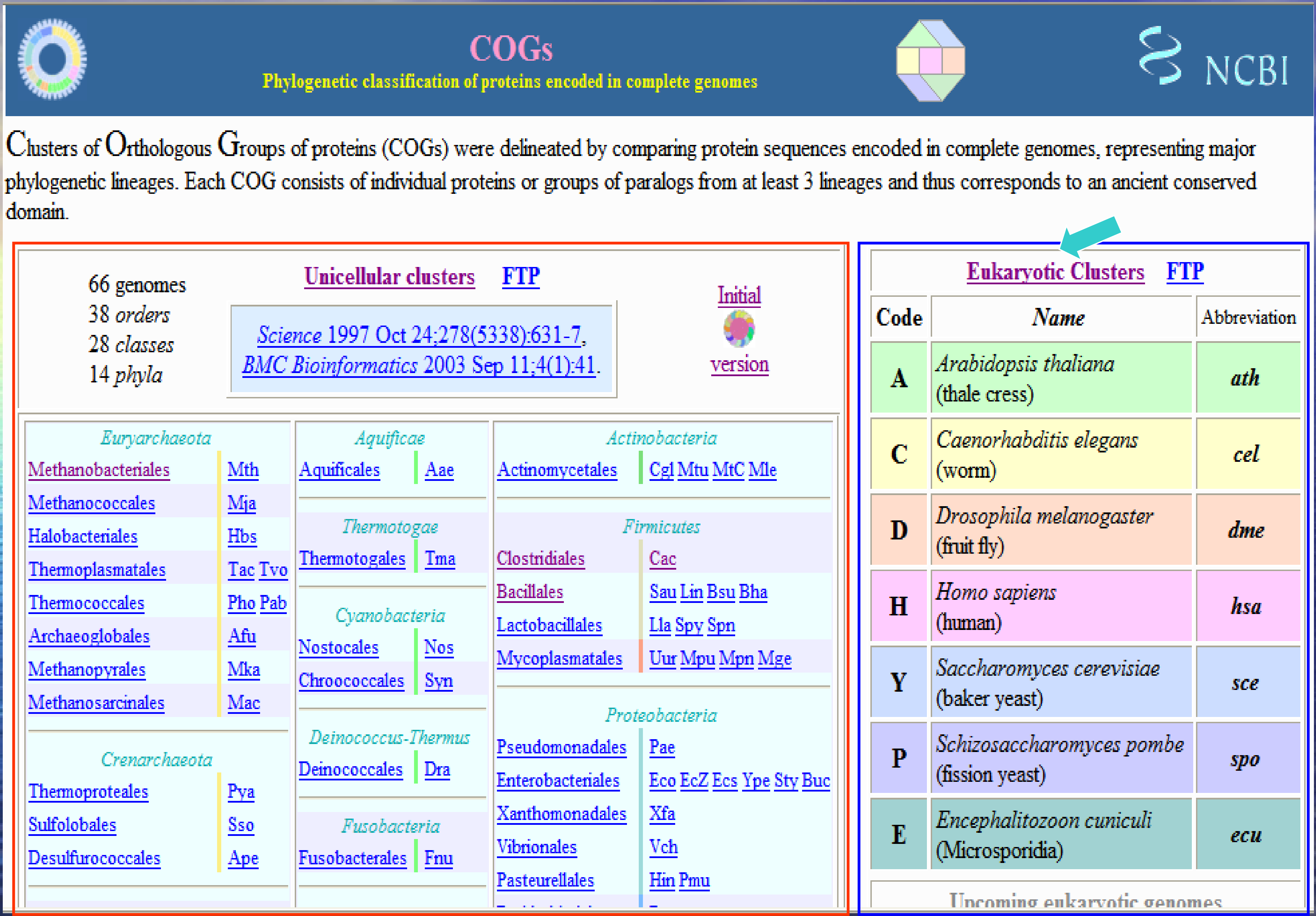
589 635 579 681 595 706 642 415 - 357 419 396 189 379 253 283 300 429 473 493 446 280 739 703 586 793



Pathways and functional systems:

J	22	TRANSLATION FACTORS AND ENZYMES INVOLVED IN TRANSLATION
J	26	AMINOACYL-TRNA SYNTHETASES AND ALTERNATIVE SYSTEMS FOR AMINO ACID ACTIVATION
J	32	RIBOSOMAL PROTEINS - SMALL SUBUNIT
J	51	RIBOSOMAL PROTEINS - LARGE SUBUNIT
K	15	DNA-DEPENDENT RNA POLYMERASE SUBUNITS
K	11	BASAL TRANSCRIPTION FACTORS
K	70	TRANSCRIPTIONAL REGULATORS
L	26	BASAL REPLICATION MACHINERY
L	8	DNA POLYMERASE III SUBUNITS
M	9	LIPID A BIOSYNTHESIS

M	155	4079	Cell envelope biogenesis, outer membrane	1
N	133	3110	Cell motility and secretion	2
P	160	5112	Inorganic ion transport and metabolism	1
T	97	3627	Signal transduction mechanisms	-
			Metabolism	
C	224	5594	Energy production and conversion	7
G	171	5262	Carbohydrate transport and metabolism	4
E	233	8383	Amino acid transport and metabolism	10



COGs

Phylogenetic classification of proteins encoded in complete genomes

Clusters of Orthologous Groups of proteins (COGs) were delineated by comparing protein sequences encoded in complete genomes, representing major phylogenetic lineages. Each COG consists of individual proteins or groups of paralogs from at least 3 lineages and thus corresponds to an ancient conserved domain.

66 genomes
38 *orders*
28 *classes*
14 *phyla*

Unicellular clusters **FTP**

Science 1997 Oct 24;278(5338):631-7,
BMC Bioinformatics 2003 Sep 11;4(1):41.

Initial version

Euryarchaeota	
Methanobacteriales	Mth
Methanococcales	Mja
Halobacteriales	Hbs
Thermoplasmatales	Tac Tvo
Thermococcales	Pho Pab
Archaeoglobales	Afu
Methanopyrales	Mka
Methanosarcinales	Mac

Crenarchaeota	
Thermoproteales	Pya
Sulfolobales	Sso
Desulfurococcales	Ape

Aquificae	
Aquificales	Aae

Thermotogae	
Thermotogales	Tma

Cyanobacteria	
Nostocales	Nos
Chroococcales	Syn

Deinococcus-Thermus	
Deinococcales	Dra

Fusobacteria	
Fusobacterales	Fnu

Actinobacteria	
Actinomycetales	Cgl Mtu MtC Mle

Firmicutes	
Clostridiales	Cac
Bacillales	Sau Lin Bsu Bha
Lactobacillales	Lla Spy Spn
Mycoplasmatales	Uur Mpu Mpn Mge

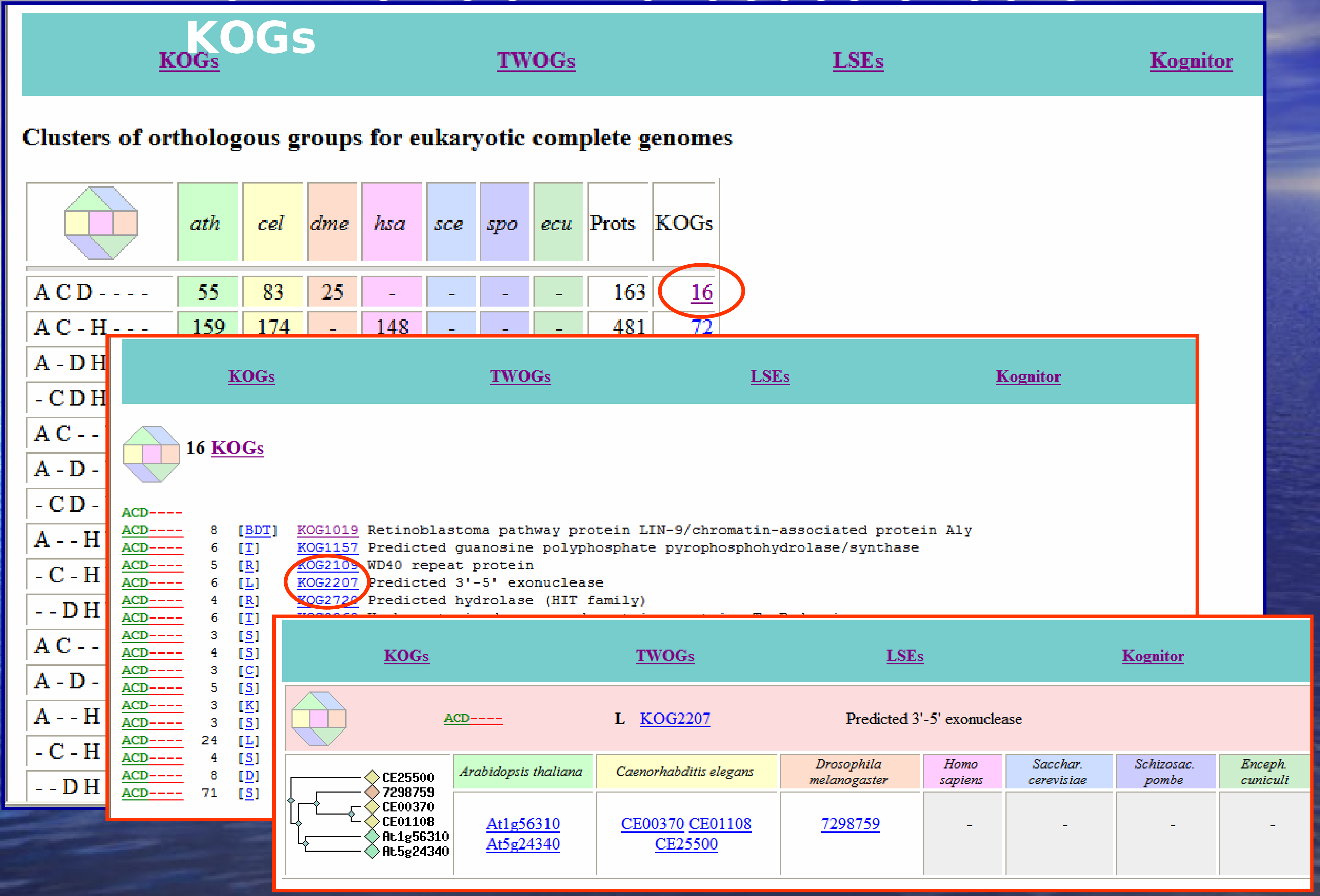
Proteobacteria	
Pseudomonadales	Pae
Enterobacteriales	Eco EcZ Ecs Ype Sty Buc
Xanthomonadales	Xfa
Vibrionales	Vch
Pasteurellales	Hin Pmu

Eukaryotic Clusters **FTP**

Code	*Name*	Abbreviation
A	*Arabidopsis thaliana* (thale cress)	***ath***
C	*Caenorhabditis elegans* (worm)	***cel***
D	*Drosophila melanogaster* (fruit fly)	***dme***
H	*Homo sapiens* (human)	***hsa***
Y	*Saccharomyces cerevisiae* (baker yeast)	***sce***
P	*Schizosaccharomyces pombe* (fission yeast)	***spo***
E	*Encephalitozoon cuniculi* (Microsporidia)	***ecu***

Upcoming eukaryotic genomes

EUKARIOTIC ORTHOLOGOUS GROUPS KOGs



KOGs **TWOGs** **LSEs** **Kognitor**

Clusters of orthologous groups for eukaryotic complete genomes

	ath	*cel*	*dme*	*hsa*	*sce*	*spo*	*ecu*	Prots	KOGs
A C D - - - -	55	83	25	-	-	-	-	163	16
A C - H - - -	159	174	-	148	-	-	-	481	72

A - D H
- C D H
A C - -
A - D -
- C D -
A - - H
- C - H
- - D H
A C - -
A - D -
A - - H
- C - H
- - D H

KOGs **TWOGs** **LSEs** **Kognitor**

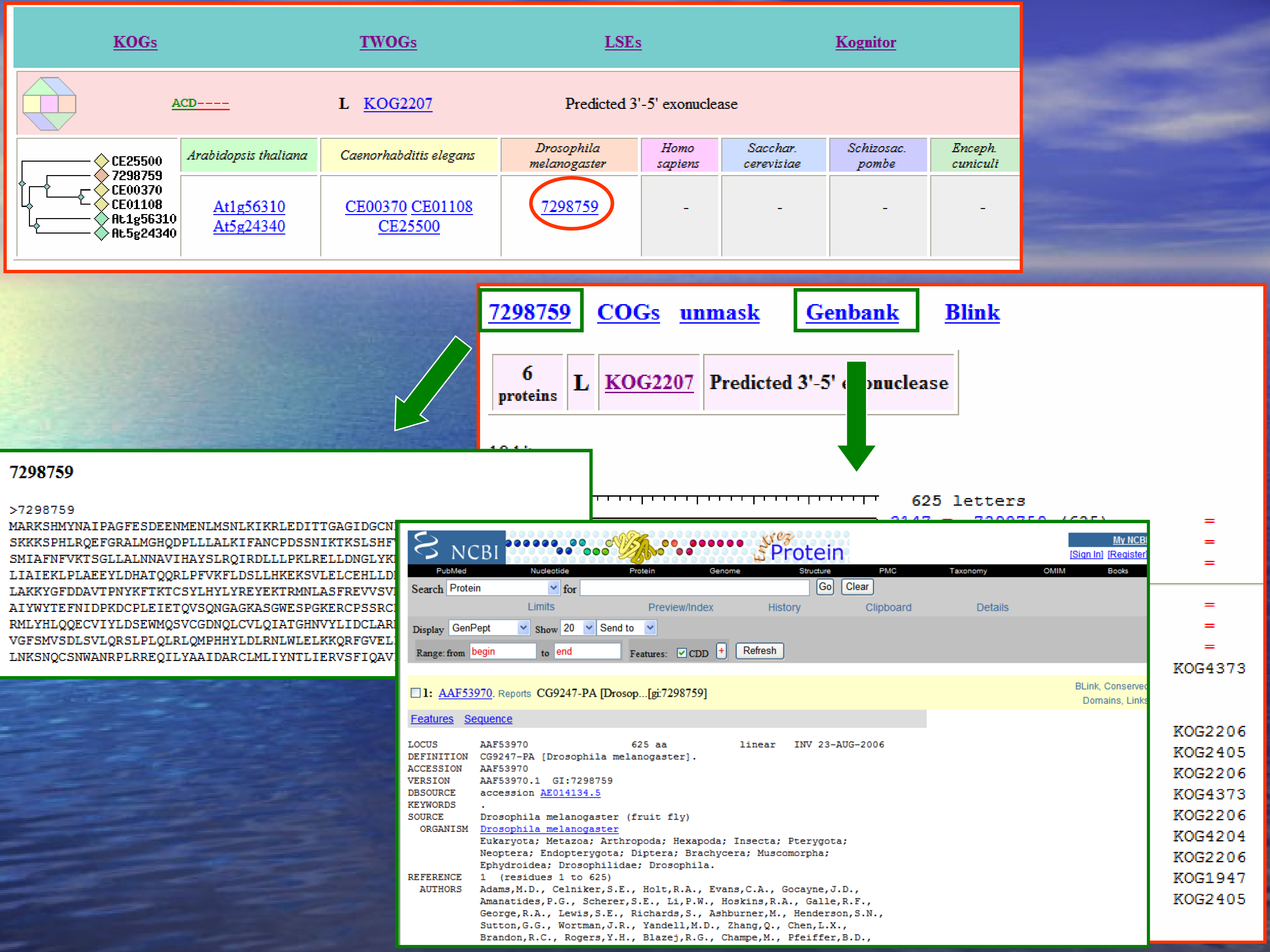
16 **KOGs**

```
ACD----
ACD----     8   [BDT]   KOG1019 Retinoblastoma pathway protein LIN-9/chromatin-associated protein Aly
ACD----     6   [T]     KOG1157 Predicted guanosine polyphosphate pyrophosphohydrolase/synthase
ACD----     5   [R]     KOG2109 WD40 repeat protein
ACD----     6   [L]     KOG2207 Predicted 3'-5' exonuclease
ACD----     4   [R]     KOG2720 Predicted hydrolase (HIT family)
ACD----     6   [T]
ACD----     3   [S]
ACD----     4   [S]
ACD----     3   [C]
ACD----     5   [S]
ACD----     3   [K]
ACD----     3   [S]
ACD----    24   [L]
ACD----     4   [S]
ACD----     8   [D]
ACD----    71   [S]
```

KOGs **TWOGs** **LSEs** **Kognitor**

ACD---- L KOG2207 Predicted 3'-5' exonuclease

	Arabidopsis thaliana	*Caenorhabditis elegans*	*Drosophila melanogaster*	*Homo sapiens*	*Sacchar. cerevisiae*	*Schizosac. pombe*	*Enceph. cuniculi*
CE25500 7298759 CE00370 CE01108 At1g56310 At5g24340	At1g56310 At5g24340	CE00370 CE01108 CE25500	7298759	-	-	-	-



L	KOG2207	Predicted 3'-5' exonuclease

Arabidopsis thaliana	*Caenorhabditis elegans*	*Drosophila melanogaster*	*Homo sapiens*	*Sacchar. cerevisiae*	*Schizosac. pombe*	*Enceph. cuniculi*
At1g56310 At5g24340	CE00370 CE01108 CE25500	7298759	-	-	-	-

7298759 COGs unmask **Genbank** Blink

6 proteins	L	KOG2207	Predicted 3'-5' exonuclease

625 letters

KOG4373
KOG2206
KOG2405
KOG2206
KOG4373
KOG2206
KOG4204
KOG2206
KOG1947
KOG2405

7298759

```
>7298759
MARKSHMYNAIPAGFESDEENMENLMSNLKIKRLEDITTGAGIDGCN
SKKKSPHLRQEFGRALMGHQDPLLLALKIFANCPDSSNIKTKSLSHF
SMIAFNFVKTSGLLALNNAVIHAYSLRQIRDLLLPKLRELLDNGLYK
LIAIEKLPLAEEYLDHATQQRLPFVKFLDSLLHKEKSVLELCEHLLD
LAKKYGFDDAVTPNYKFTKTCSYLHYLYREYEKTRMNLASFREVVSV
AIYWYTEFNIDPKDCPLEIETQVSQNGAGKASGWESPGKERCPSSRC
RMLYHLQQECVIYLDSEWMQSVCGDNQLCVLQIATGHNVYLIDCLAR
VGFSMVSDLSVLQRSLPLQLRLQMPHHYLDLRNLWLELKKQRFGVEL
LNKSNQCSNWANRPLRREQILYAAIDARCLMLIYNTLIERVSFIQAV
```

Search Protein for

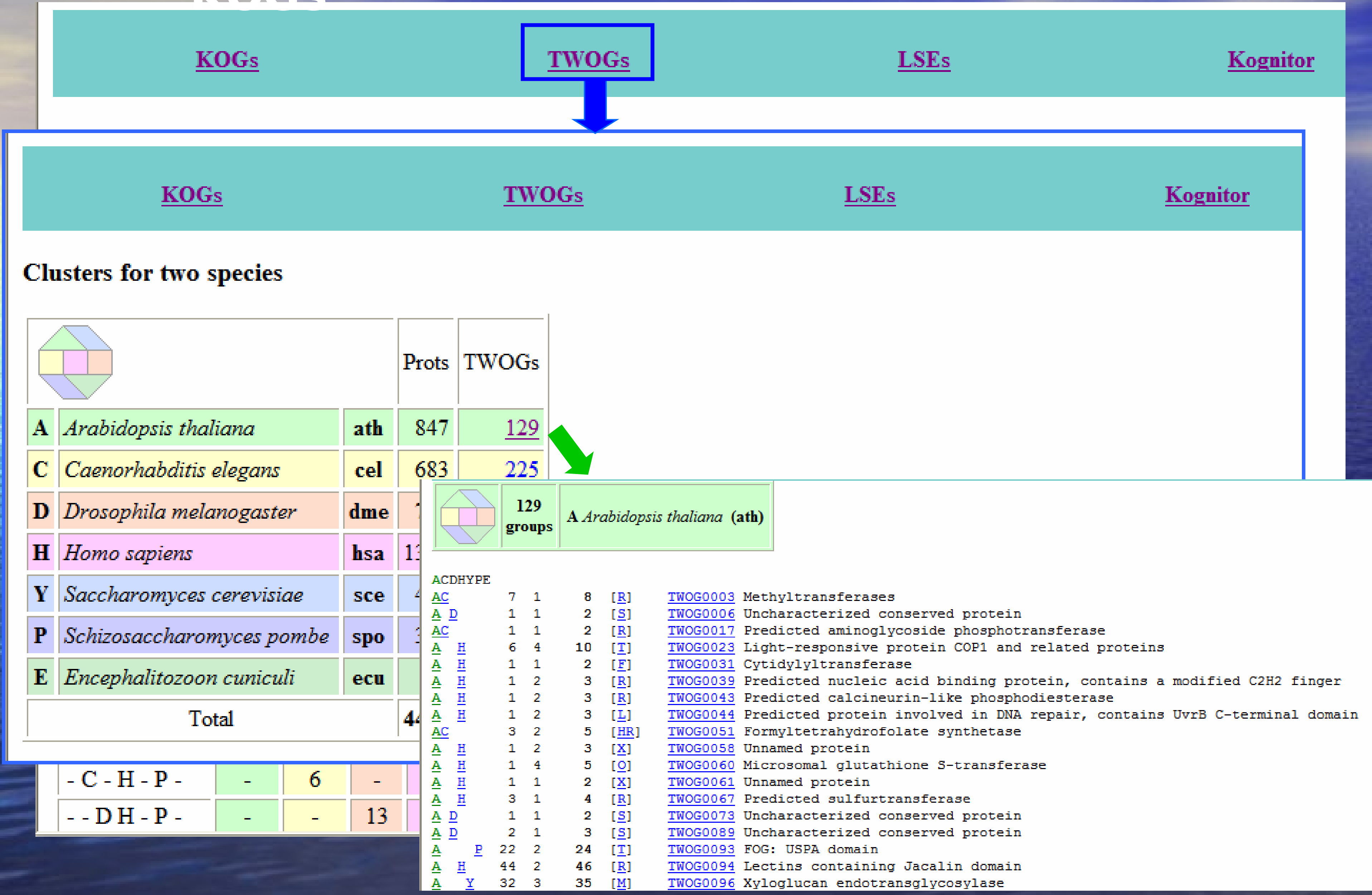
Display GenPept Show 20 Send to

Range: from begin to end Features: ☑ CDD

☐ 1: AAF53970. Reports CG9247-PA [Drosop...[gi:7298759]

```
LOCUS       AAF53970                 625 aa            linear   INV 23-AUG-2006
DEFINITION  CG9247-PA [Drosophila melanogaster].
ACCESSION   AAF53970
VERSION     AAF53970.1  GI:7298759
DBSOURCE    accession AE014134.5
KEYWORDS    .
SOURCE      Drosophila melanogaster (fruit fly)
  ORGANISM  Drosophila melanogaster
            Eukaryota; Metazoa; Arthropoda; Hexapoda; Insecta; Pterygota;
            Neoptera; Endopterygota; Diptera; Brachycera; Muscomorpha;
            Ephydroidea; Drosophilidae; Drosophila.
REFERENCE   1  (residues 1 to 625)
  AUTHORS   Adams,M.D., Celniker,S.E., Holt,R.A., Evans,C.A., Gocayne,J.D.,
            Amanatides,P.G., Scherer,S.E., Li,P.W., Hoskins,R.A., Galle,R.F.,
            George,R.A., Lewis,S.E., Richards,S., Ashburner,M., Henderson,S.N.,
            Sutton,G.G., Wortman,J.R., Yandell,M.D., Zhang,Q., Chen,L.X.,
            Brandon,R.C., Rogers,Y.H., Blazej,R.G., Champe,M., Pfeiffer,B.D.,
```

EUKARIOTIC ORTHOLOGOUS GROUPS KOGs



Clusters for two species

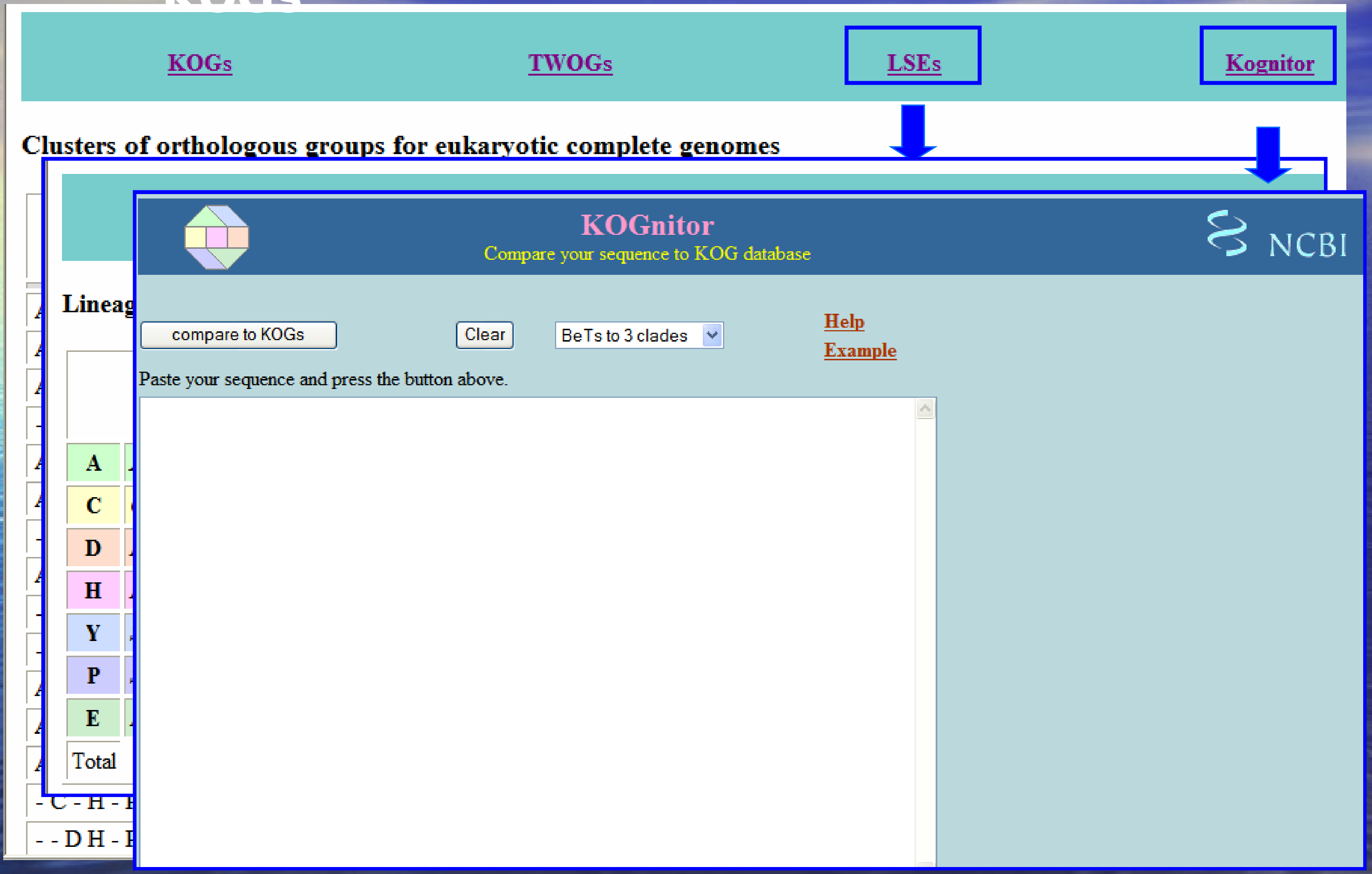
			Prots	TWOGs
A	*Arabidopsis thaliana*	**ath**	847	129
C	*Caenorhabditis elegans*	**cel**	683	225
D	*Drosophila melanogaster*	**dme**		
H	*Homo sapiens*	**hsa**		
Y	*Saccharomyces cerevisiae*	**sce**		
P	*Schizosaccharomyces pombe*	**spo**		
E	*Encephalitozoon cuniculi*	**ecu**		
Total				

- C - H - P -	-	6	-
- - D H - P -	-	-	13

	129 groups	**A** *Arabidopsis thaliana* **(ath)**

```
ACDHYPE
AC         7   1      8   [R]     TWOG0003 Methyltransferases
A D        1   1      2   [S]     TWOG0006 Uncharacterized conserved protein
AC         1   1      2   [R]     TWOG0017 Predicted aminoglycoside phosphotransferase
A  H       6   4     10   [T]     TWOG0023 Light-responsive protein COP1 and related proteins
A  H       1   1      2   [F]     TWOG0031 Cytidylyltransferase
A  H       1   2      3   [R]     TWOG0039 Predicted nucleic acid binding protein, contains a modified C2H2 finger
A  H       1   2      3   [R]     TWOG0043 Predicted calcineurin-like phosphodiesterase
A  H       1   2      3   [L]     TWOG0044 Predicted protein involved in DNA repair, contains UvrB C-terminal domain
AC         3   2      5   [HR]    TWOG0051 Formyltetrahydrofolate synthetase
A  H       1   2      3   [X]     TWOG0058 Unnamed protein
A  H       1   4      5   [O]     TWOG0060 Microsomal glutathione S-transferase
A  H       1   1      2   [X]     TWOG0061 Unnamed protein
A  H       3   1      4   [R]     TWOG0067 Predicted sulfurtransferase
A D        1   1      2   [S]     TWOG0073 Uncharacterized conserved protein
A D        2   1      3   [S]     TWOG0089 Uncharacterized conserved protein
A    P    22   2     24   [T]     TWOG0093 FOG: USPA domain
A  H      44   2     46   [R]     TWOG0094 Lectins containing Jacalin domain
A   Y     32   3     35   [M]     TWOG0096 Xyloglucan endotransglycosylase
```

EUKARIOTIC ORTHOLOGOUS GROUPS KOGs



Clusters of orthologous groups for eukaryotic complete genomes

Compare your sequence to KOG database

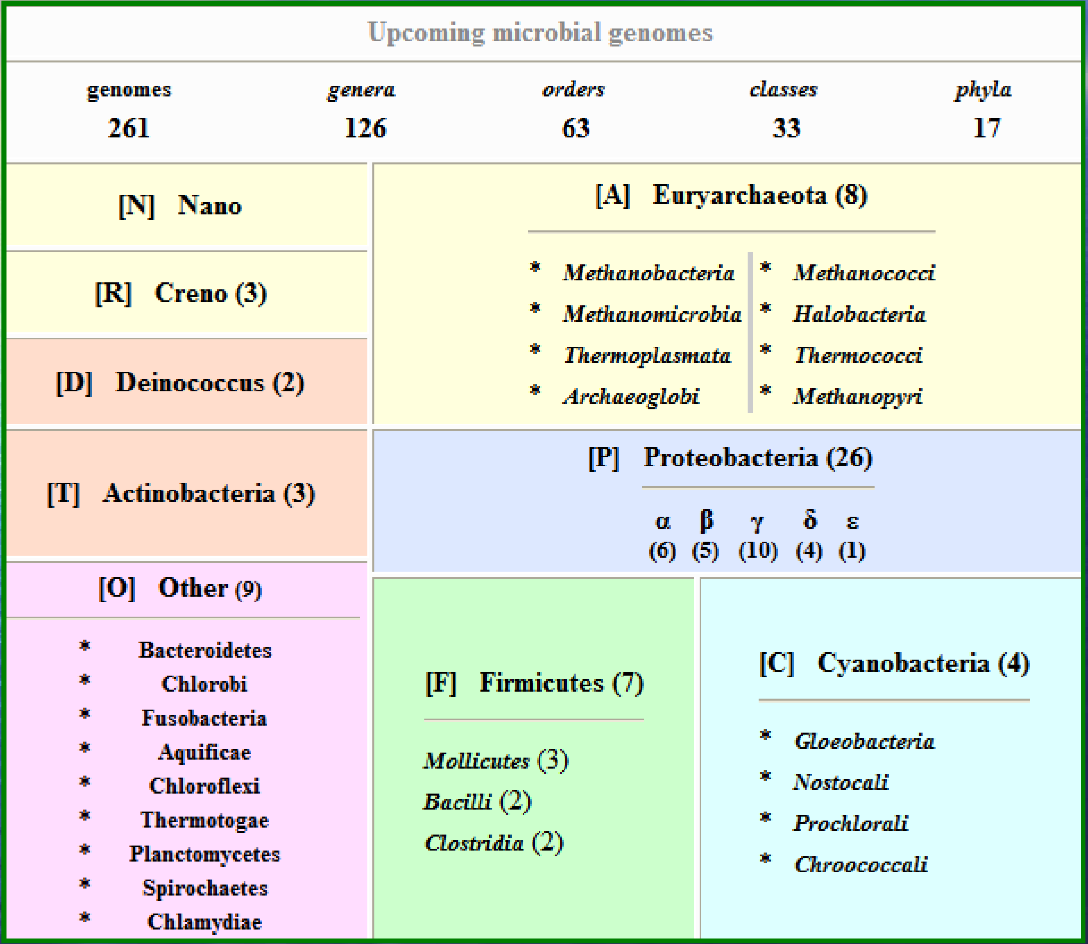
Paste your sequence and press the button above.

COGs

Phylogenetic classification of proteins encoded in complete genomes

NCBI

próximamente ...



Upcoming microbial genomes

genomes	*genera*	*orders*	*classes*	*phyla*
261	**126**	**63**	**33**	**17**

[N] Nano

[R] Creno (3)

[D] Deinococcus (2)

[T] Actinobacteria (3)

[O] Other (9)

- Bacteroidetes
- Chlorobi
- Fusobacteria
- Aquificae
- Chloroflexi
- Thermotogae
- Planctomycetes
- Spirochaetes
- Chlamydiae

[A] Euryarchaeota (8)

- *Methanobacteria*
- *Methanomicrobia*
- *Thermoplasmata*
- *Archaeoglobi*
- *Methanococci*
- *Halobacteria*
- *Thermococci*
- *Methanopyri*

[P] Proteobacteria (26)

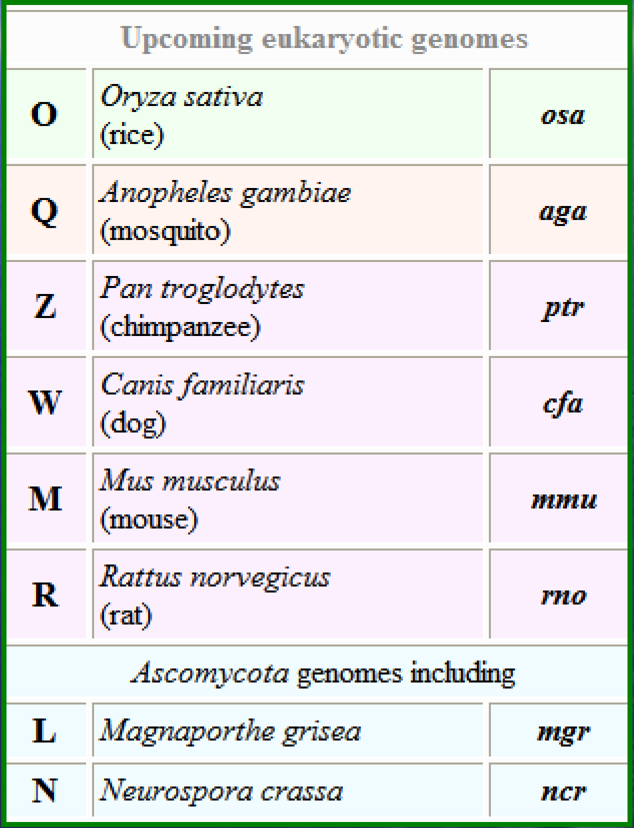
α	β	γ	δ	ε
(6)	(5)	(10)	(4)	(1)

[F] Firmicutes (7)

Mollicutes (3)
Bacilli (2)
Clostridia (2)

[C] Cyanobacteria (4)

- *Gloeobacteria*
- *Nostocali*
- *Prochlorali*
- *Chroococcali*



Upcoming eukaryotic genomes

O	*Oryza sativa* (rice)	*osa*
Q	*Anopheles gambiae* (mosquito)	*aga*
Z	*Pan troglodytes* (chimpanzee)	*ptr*
W	*Canis familiaris* (dog)	*cfa*
M	*Mus musculus* (mouse)	*mmu*
R	*Rattus norvegicus* (rat)	*rno*
	Ascomycota genomes including	
L	*Magnaporthe grisea*	*mgr*
N	*Neurospora crassa*	*ncr*

Equipo COG

Eugene V. Koonin
Senior Investigator
National Center for Biotechnology Information (NCBI)
National Library of Medicine (NLM)
National Institutes of Health (NIH)