

# FUNCTIONAL PREDICTION AT THE (META-)GENOMIC SCALE BY EVOLUTIONARY ANALYSIS

JAIME HUERTA CEPAS  
BORK GROUP, EMBL HEIDELBERG



# **1. PHYLOGENOMIC APPROACHES FOR ORTHOLOGY DETECTION AND FUNCTIONAL PREDICTION**

# ORTHOLOGY AND FUNCTIONAL PREDICTION

## Ortholog conjecture:

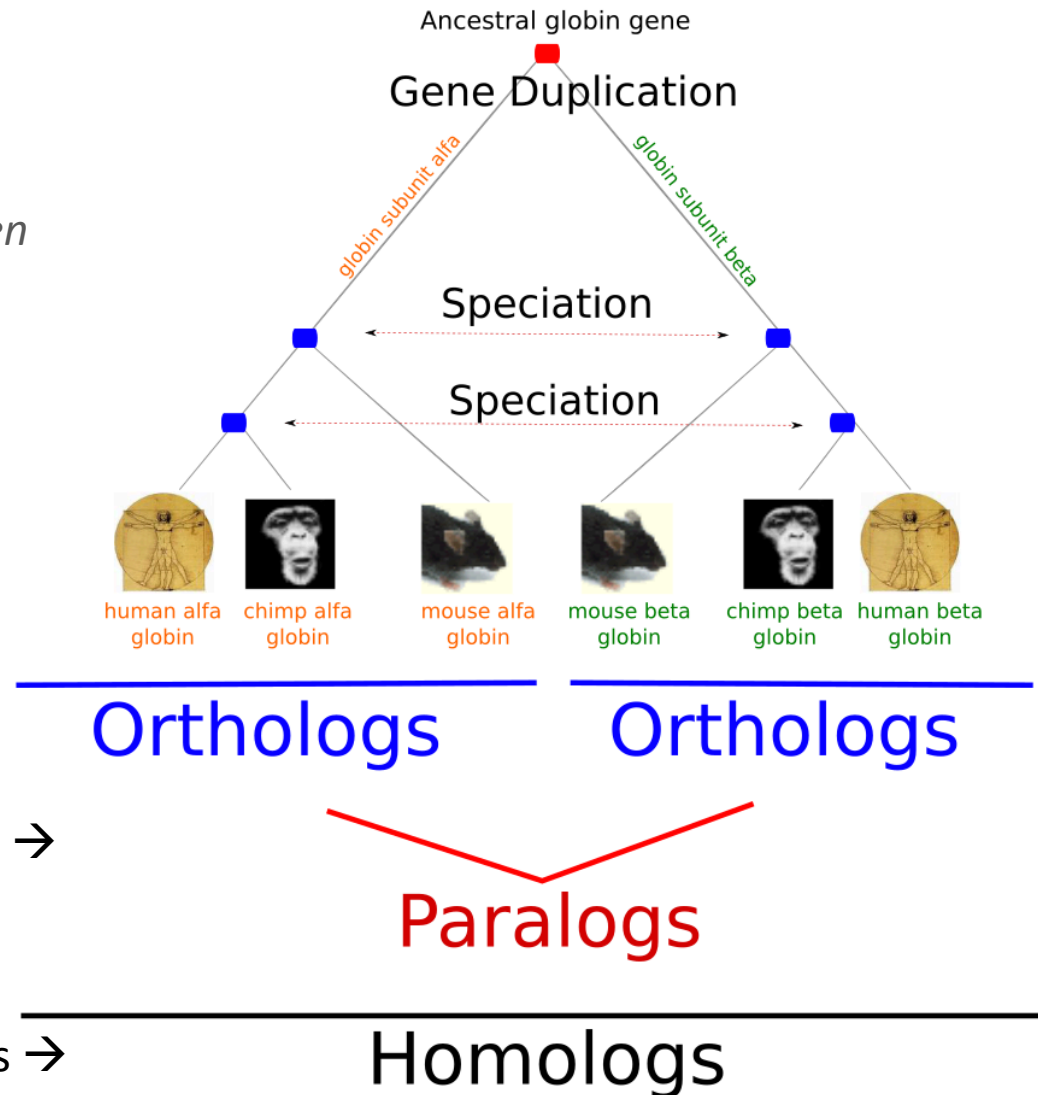
*Function is more conserved between orthologs than between paralogs*



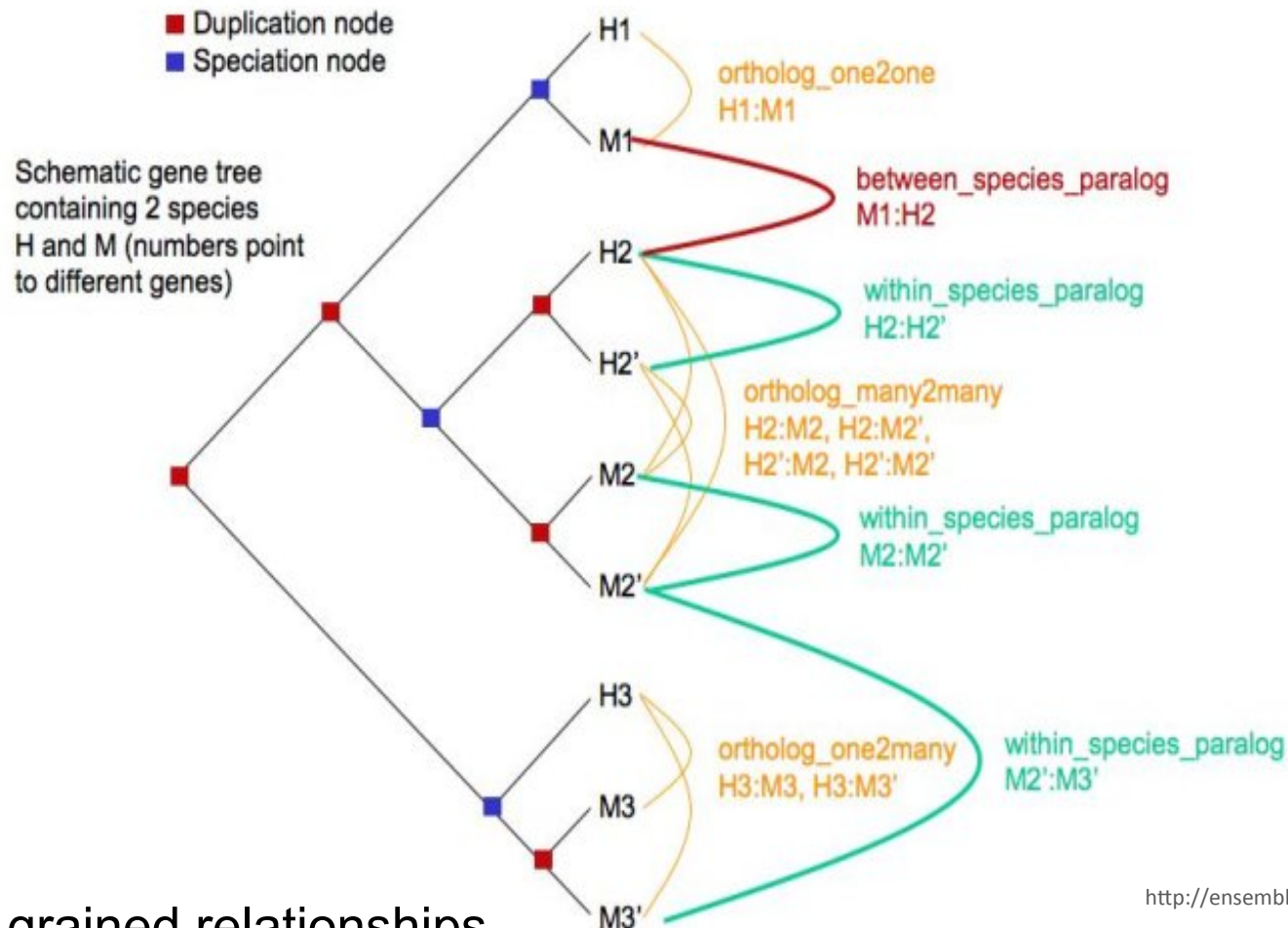
Predicting gene function turns into an orthology prediction problem

Deeper  
Evolutionary →  
analysis

BLAST results →



# ORTHOLOGY DETECTION METHODS (PHYLOGENY)

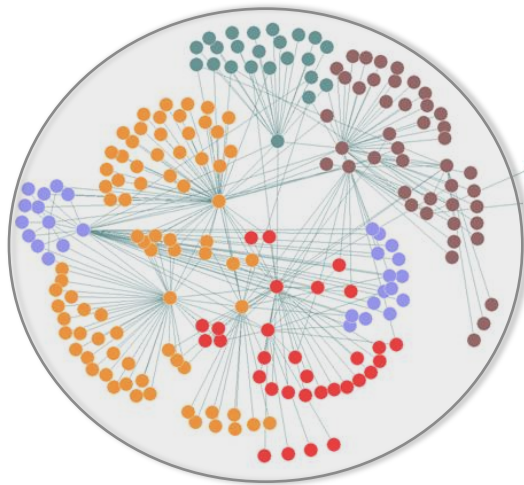


<http://ensembl.org>

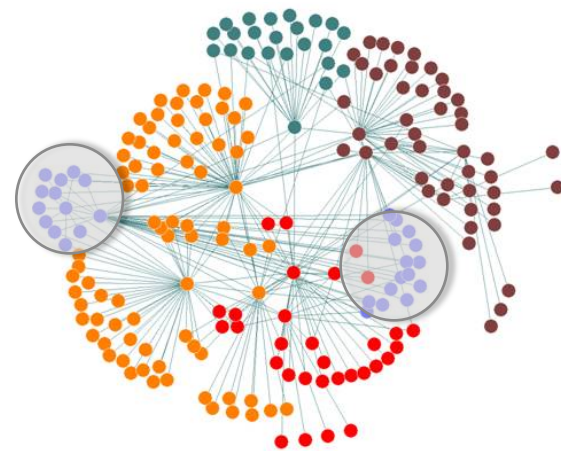
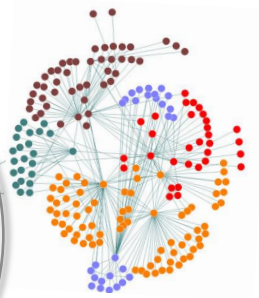
- Infer fine-grained relationships
- Allows tracing the origin of duplication and speciation events

# ORTHOLOGY DETECTION METHODS (CLUSTERING)

- Build Clusters of Orthologous Groups (COGs) based on sequence similarity network
- Tend to be faster and cope for larger amounts of species/sequences
- No co-orthology disambiguation and requires taxonomic constraints



1 COG (all species)



2 COGs (mammal species)

# EggNOG 4.5.1

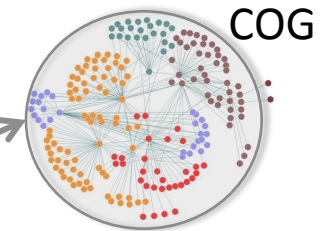
A database of orthologous groups and functional annotation

Organisms  
2,031

Viruses  
352

Orthologous Groups  
190k

Trees & Algs.  
1.9M



- Rapid sequence search
- Find orthologous groups
- Functional annotations
  - Domains
  - Gene Ontology
  - KEGG
  - Taxonomic distribution

<http://eggnog.embl.de>

Huerta-Cepas, et al. 2016 (Nucleic Acid Research)

**EggNOG 4.5.1** Search protein or OG

Found matches in seconds (KB)

Navigation: Home, Sequence search, eggNOG-mapper (New), Downloads, API, Methods, Viral OGs

Query Sequence: 10020.ENSODRP000000 (...)

**ENOG410RTGN** (U) Intracellular trafficking, secretion, and vesicular transport. SEC24 family, member B (S. cerevisiae). 4 proteins, 4 species. evalue:0, score:2,504.5. Orthologs: SEC24B (Pongo abeli, Homo sapiens, Pan troglodytes, Gorilla gorilla).

**ENOG410CFRE** (U) Intracellular trafficking, secretion, and vesicular transport. SEC24 family, member B (S. cerevisiae). 4 proteins, 4 species. evalue:0, score:2,480.1. Orthologs: SEC24B (Felis catus, Ailuropoda melanoleuca, Mustela putorius furo, Canis lupus familiaris).

**ENOG411AU76** (U) Intracellular trafficking, secretion, and vesicular transport family, member B (S. cerevisiae). 18 proteins, 15 species. evalue:0, score:2,478.8. Orthologs: SEC24B (Rattus norvegicus, Pongo abeli, ENSOCUG00000001284, Oryctolagus cuniculus, Cavia porcellus, Rattus norvegicus, Ictidomys, tridecemlineatus, Microcebus murinus, Homo sapiens, 10 more...).

# FINE GRAINED ORTHOLOGY PREDICTIONS

Query gene / protein: **SEC24B in *Mus musculus*** orthologs in Target taxa: **Rodents (5 species)**

Add target taxa...

**ENO4118VD7**
U Intracellular trafficking, secretion, and vesicular transport family, member B (S. cerevisiae)
8 proteins 5 species
★ 1 matches

[Fine-grained Orthologs<sup>beta</sup>](#)
[Orthologous Group](#)
[Taxonomic Profile](#)
[Functional Profile](#)
[Alignment](#)
[Phylogenetic Tree](#)
[Download](#)

Flat tree
PFAM domains
SMART domains
Aligned blocks

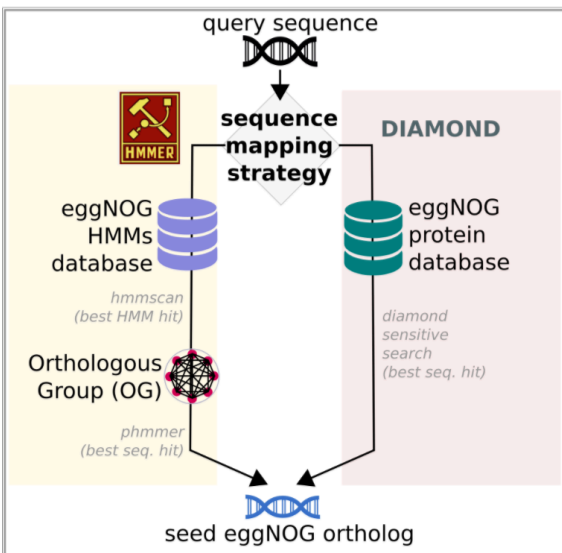
Species	Accession	Protein Name	Domain
<i>Cavia porcellus</i>	(SEC24B)	SEC24B	zf-Sec23
<i>Mus musculus</i>	(SEC24B)	SEC24B	zf-Sec23
<i>Rattus norvegicus</i>	(SEC24B)	SEC24B	zf-Sec23
<i>Ictidomys tridecemlineatus</i>	(SEC24B)	SEC24B	zf-Sec23
<i>Dipodomys ordii</i>	(SEC24B)	SEC24B	zf-Sec23
<i>Cavia porcellus</i>	(SEC24A)	SEC24A	zf-Sec23
<i>Rattus norvegicus</i>	(SEC24A)	SEC24A	zf-Sec23
<i>Mus musculus</i>	(SEC24A)	SEC24A	zf-Sec23

Download orthologous pairs

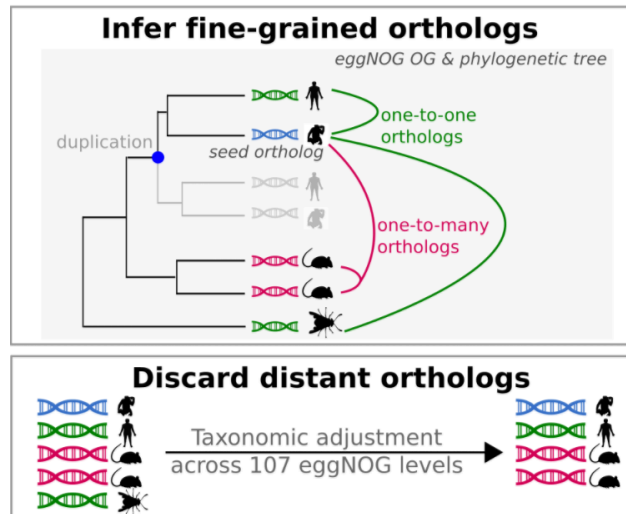
Query protein	Orthologous sequences
<i>Mus musculus</i>	<i>Dipodomys ordii</i> 10020.ENSNDORP00000002855
10090.ENSMUSP00000001079	<i>Rattus norvegicus</i> 10116.ENSRNOP000000037163
	<i>Ictidomys tridecemlineatus</i> 43179.ENSSSTOP00000007202
	<i>Cavia porcellus</i> 10141.ENSCPOP00000008271

<http://egglog.embl.de>

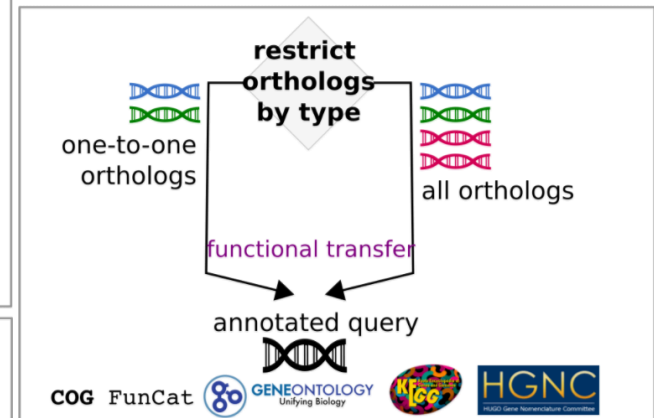
# EFFICIENT ANNOTATION BY ORTHOLOGY MAPPING



*Find query's best match*



*Place it in tree*



*Transfer terms from orthologs*

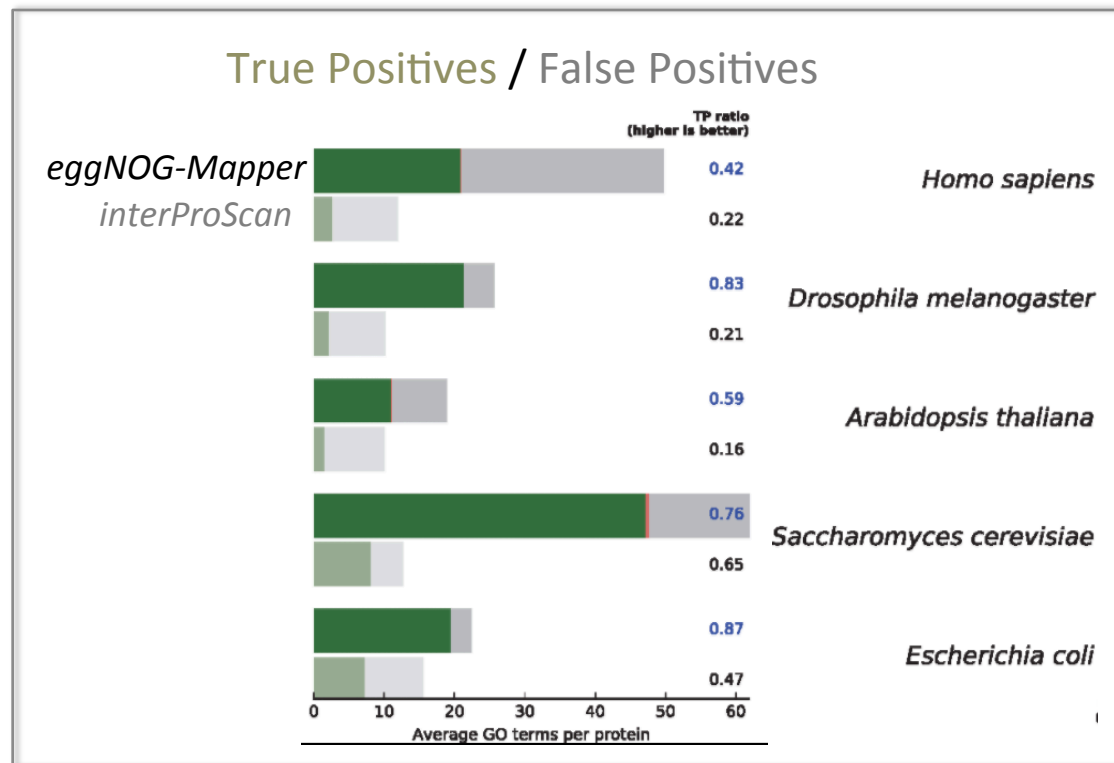
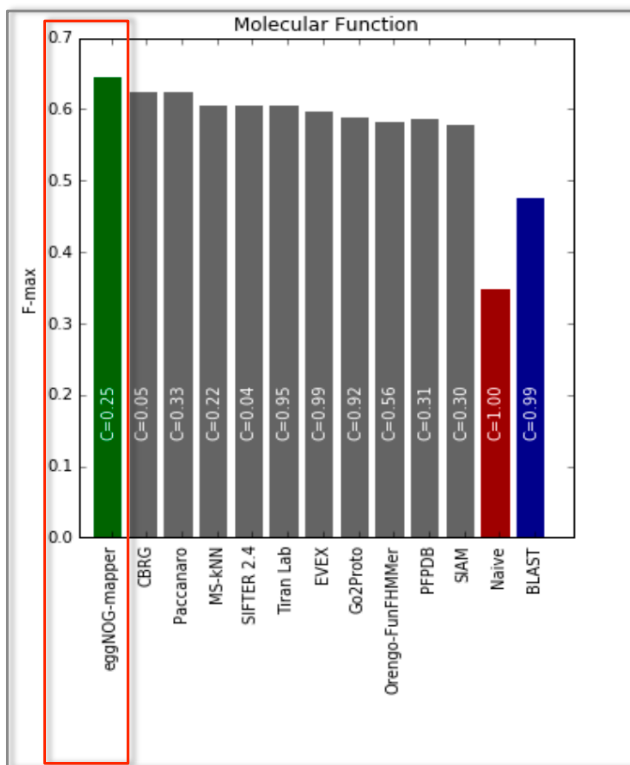
<http://eggno-mapper.embl.de>

*Huerta-Cepas et al. 2016 (submitted)*





# EGGNOG-MAPPER PERFORMANCE



CAFA2: Jiang et al. Genome Biology (2016)

## CAFA2 Challenge:

- eggNOG-mapper ranked **top 5 out of 126 methods** in the 3 GO categories

### **3. FUNCTIONAL EXPLORATION OF METAGENOMIC DATASETS**

# FUNCTIONAL ANNOTATION OF METAGENOMIC DATA

*Objective: “develop cutting edge methods for using large scale data to design cell factories and bacterial communities for biotechnological applications”* <http://dd-decaf.eu>

## @EMBL

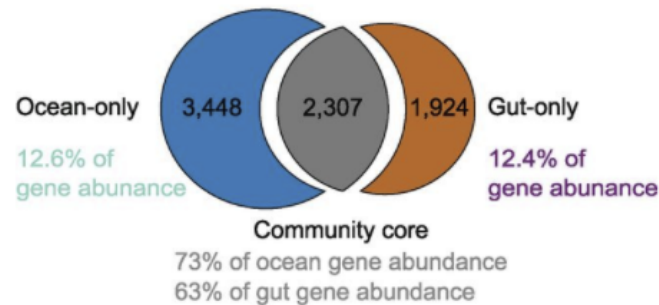
- Mining new enzymes out from genomic and metagenomic data
- Building metabolic models based on functional annotation of bacterial communities



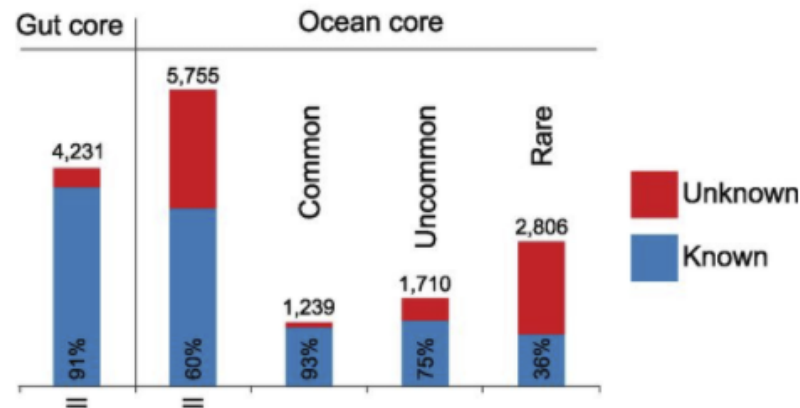
# EXPLORING METAGENOMICS

## UNKNOWN FRACTION

**C** Ocean core vs gut core orthologous groups



**B**

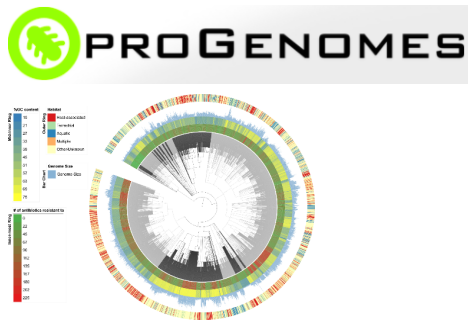


Structure and function of the global ocean microbiome

Sunagawa et al. Science 348 (6237), 1261359



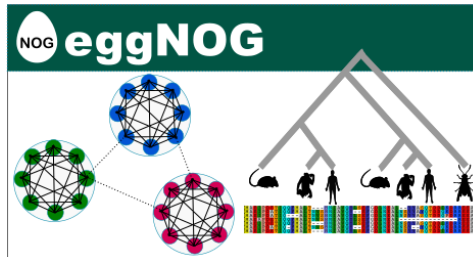
### **3. TOOLS**



Consistently annotated  
prokaryotic genomes

<http://progenomes.embl.de>

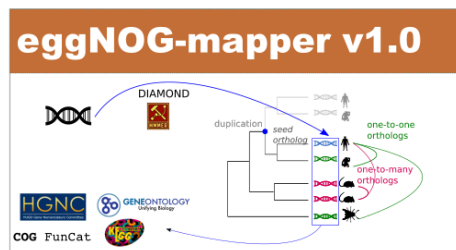
*Mende et al. NAR 45 (2017) D286*



Orthologous groups and  
functional information

<http://egglog.embl.de>

*Huerta-Cepas et al. NAR (2016)*



Fast accurate functional  
annotation

<http://egglog-mapper.embl.de>

*Huerta-Cepas et al. MBE (2017)*

## ETE Toolkit

A Python framework to work with trees

```
from ete3 import Tree
tree = Tree('((A,B), D);')
```

```
print tree
#      /-A
#     /-|
#    --| \-B
#       \-D
```

```
A = tree & "A"
A.up.show()
```

Phylogenetic reconstruction  
and analysis

<http://etetoolkit.org>

*Huerta-Cepas et al. MBE (2016)*