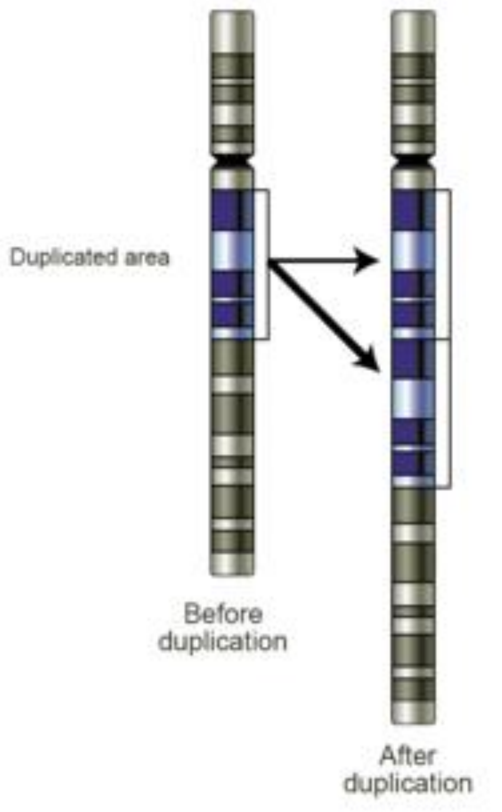
Inparanoid: a comprehensive database of eukaryotic orthologs

Ajay G.H
M.S Computational Biology

Gene Duplication

- Gene duplication or gene amplification
  - Duplication of a region of DNA that contains a gene
  - Key role in evolution

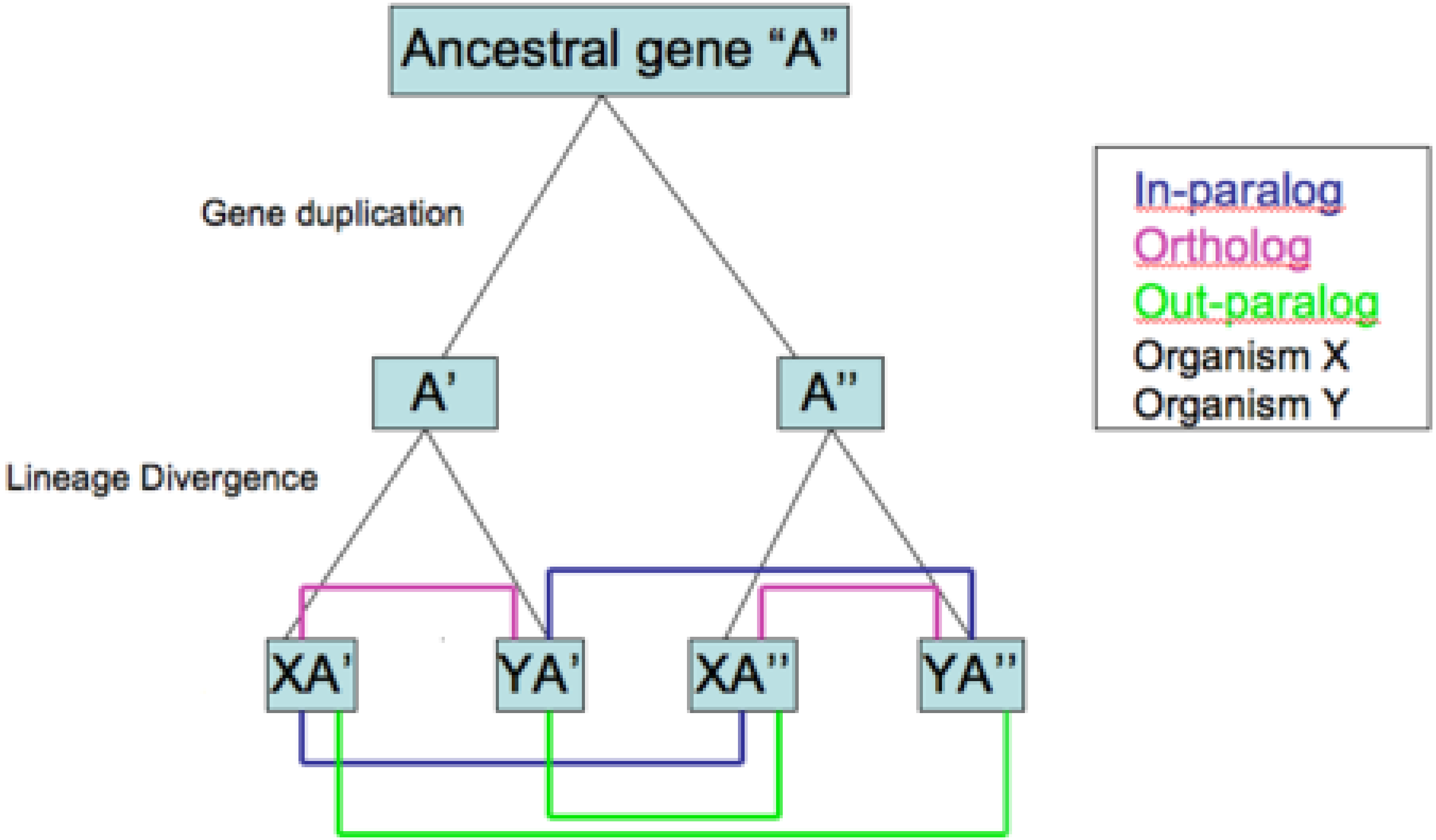
- Reasons:
  - Error in homologous recombinations
  - Retrotransposition events
  - Duplication of an entire chromosome



Significance:

- Second copy of the gene
  - Free from selective pressure
  - Mutations of it have no deleterious effects
  - Gene copy accumulates mutations faster than a functional single-copy gene
  - Plants: very high - maize -hexaploid
- Additional benefits:
  - Increase the fitness of the organism
  - Ice Fish
    - Digestive gene duplication – anti freeze

Orthologs & Paralogs



inParalogs & outParalogs

- Duplication events occur both before and after speciation.
- Inparalogs
    - Paralogs that arose through a gene duplication event after speciation
- Outparalogs
    - Arise following a gene duplication
    - Preceding speciation
    - In different species and derived from a more ancient shared duplication event

Why are we interested?

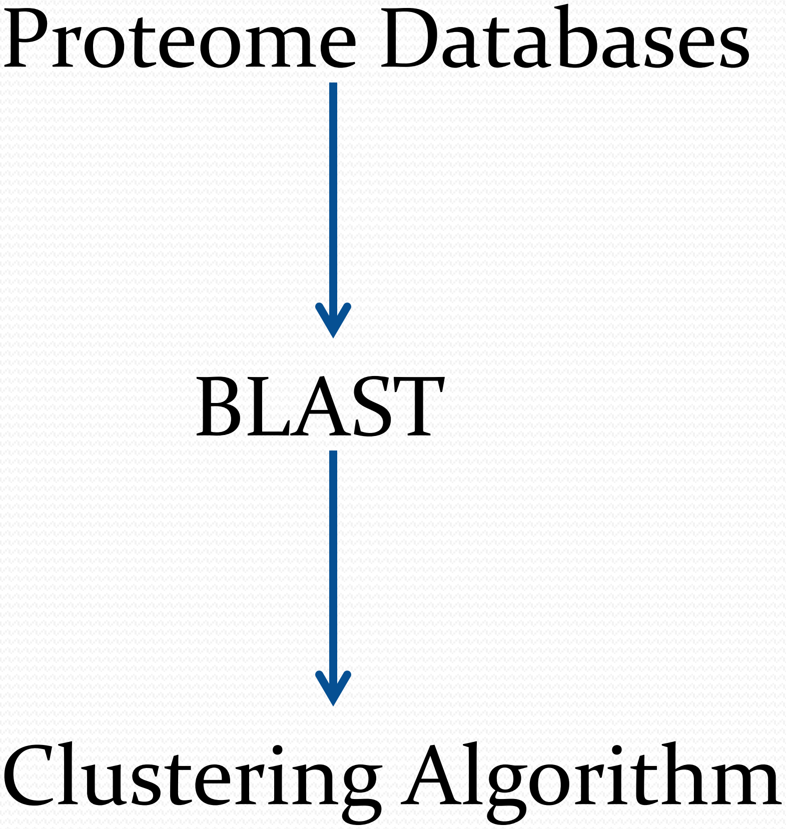
- Inparalogs can form a group of genes that together are orthologous to a gene in another species.
- Cross species modeling advantage?!
- Experiments on a human gene function can often be carried out on other species if an orthologous homolog to the human gene can be found in the genome of that species
- Ex: orthologs in zebrafish, mouse

Inparanoid eukaryotic ortholog database

- Collection of ortholog groups between 17 whole genomes

- *Anopheles gambiae*
- *Caenorhabditis briggsae*
- *Caenorhabditis elegans*
- *Drosophila melanogaster*
- *Danio rerio*
- *Takifugu rubripes*
- *Gallus gallus*
- *Homo sapiens*
- *Mus musculus*
- *Pan troglodytes*
- *Rattus norvegicus*
- *Oryza sativa*
- *Plasmodium falciparum*
- *Arabidopsis thaliana*
- *Escherichia coli*
- *Saccharomyces cerevisiae*
- *Schizosaccharomyces pombe*

Overview

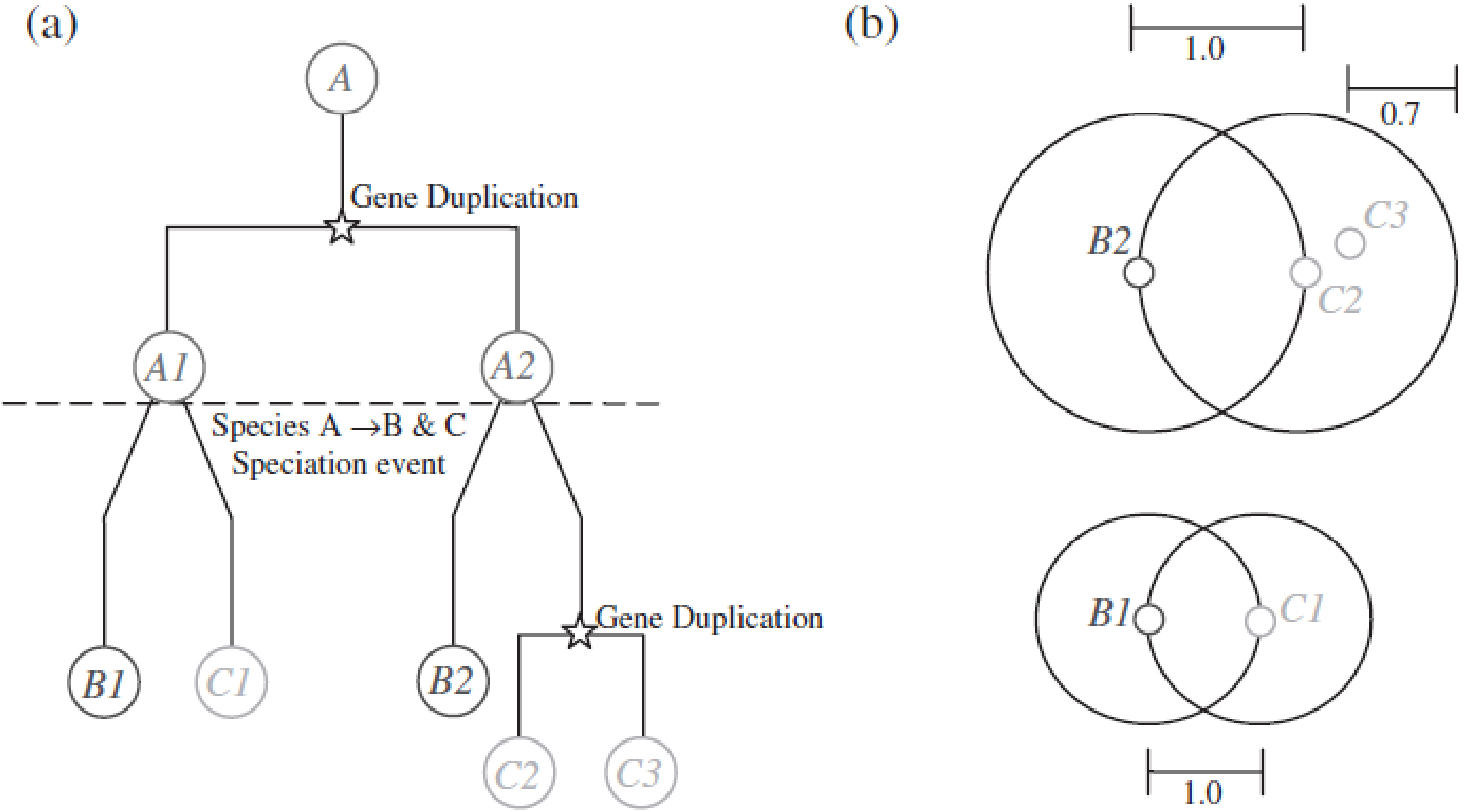


Proteome Databases

BLAST

Clustering Algorithm

Clustering Algorithm



Clustering Algorithm

- Step 1 - Seed identification
  - B2 and C2 are the original seed-ortholog pair
  - Inparalog score of 1.0 assigned
  - All inparalogs are clustered around this pair
  - Identification of seed-inparalog
  - Other inparalogs are scored according to their relative similarity to the seed-inparalog
  - Why is B1 and C1 not considered as seed though they are orthologous????

Clustering Algorithm

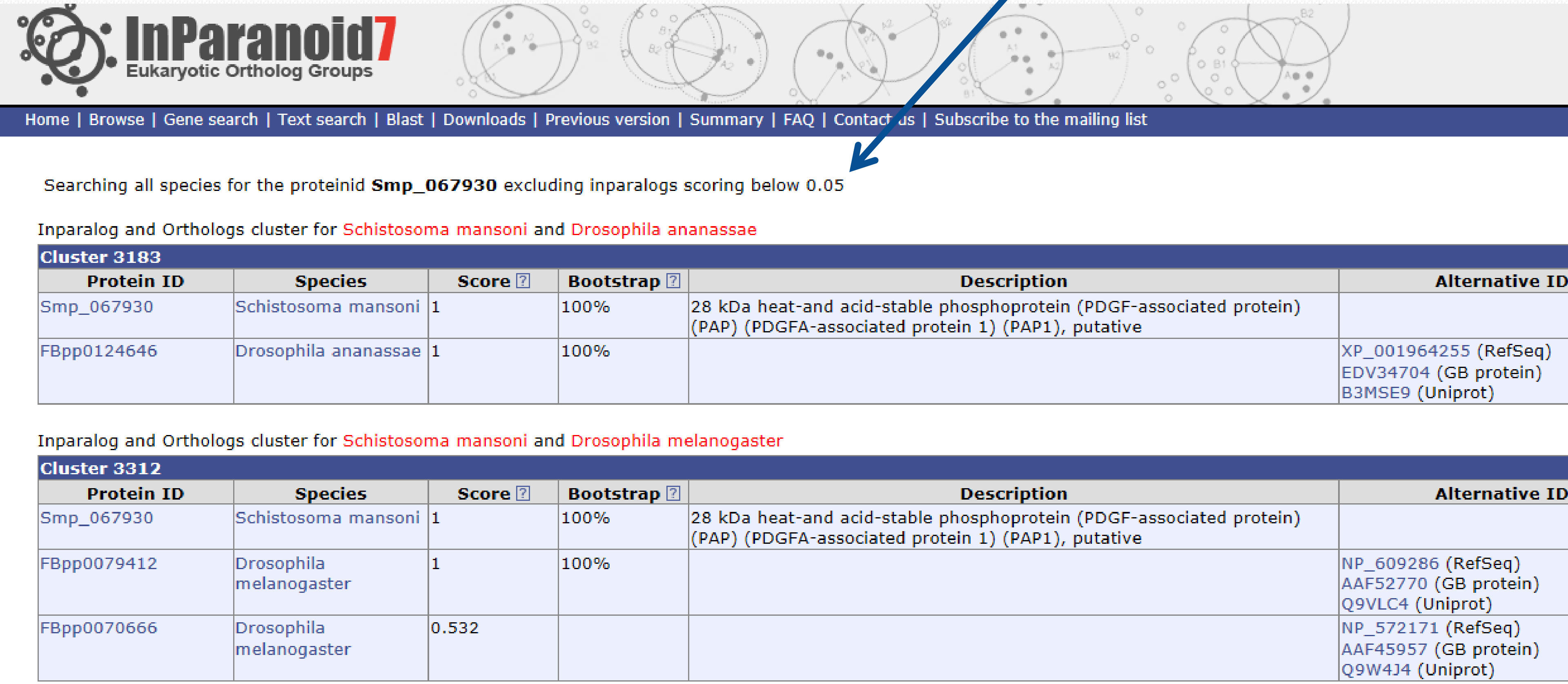
- Step 2: Inparalog score calculation
- Inparalog score of C3 =

(Blast[C2:C3]Blast[C2:B2])/(Blast[C2:C2]Blast[C2:B2])

- where Blast[X:Y] is the averaged blast score between X and Y in bits.
- C1 and B1 are orthologous to each other but are outparalogs of the other cluster and thus form a cluster of their own

Cluster analysis

Cut off score


InParanoid7
Eukaryotic Ortholog Groups
Home | Browse | Gene search | Text search | Blast | Downloads | Previous version | Summary | FAQ | Contact us | Subscribe to the mailing list

Searching all species for the proteinid **Smp_067930** excluding inparalogs scoring below 0.05

Inparalog and Orthologs cluster for Schistosoma mansoni and Drosophila ananassae

Cluster 3183

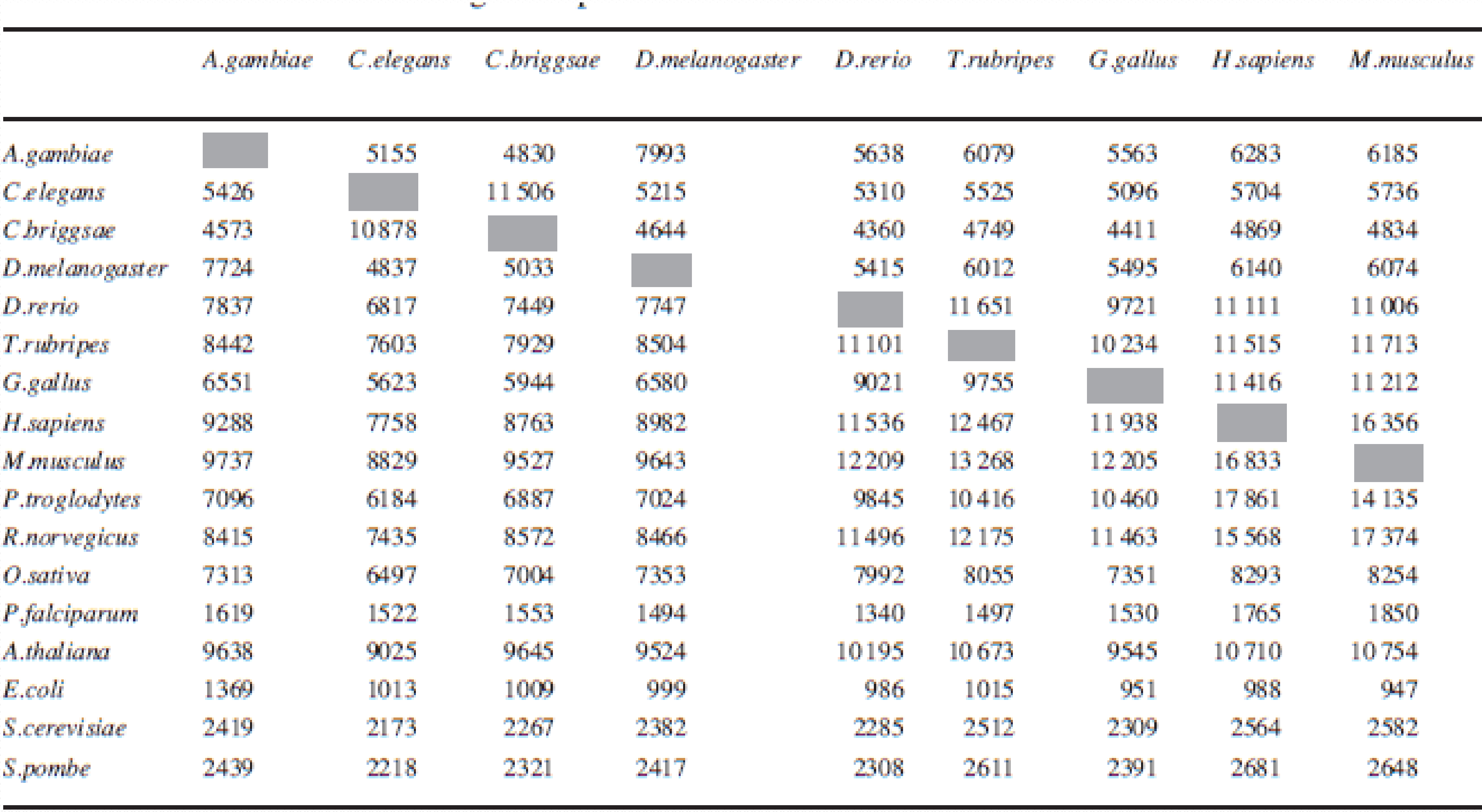
Protein ID	Species	Score ?	Bootstrap ?	Description	Alternative ID
Smp_067930	Schistosoma mansoni	1	100%	28 kDa heat-and acid-stable phosphoprotein (PDGF-associated protein) (PAP) (PDGFA-associated protein 1) (PAP1), putative	
FBpp0124646	Drosophila ananassae	1	100%		XP_001964255 (RefSeq) EDV34704 (GB protein) B3MSE9 (Uniprot)

Inparalog and Orthologs cluster for Schistosoma mansoni and Drosophila melanogaster

Cluster 3312

Protein ID	Species	Score ?	Bootstrap ?	Description	Alternative ID
Smp_067930	Schistosoma mansoni	1	100%	28 kDa heat-and acid-stable phosphoprotein (PDGF-associated protein) (PAP) (PDGFA-associated protein 1) (PAP1), putative	
FBpp0079412	Drosophila melanogaster	1	100%		NP_609286 (RefSeq) AAF52770 (GB protein) Q9VLC4 (Uniprot)
FBpp0070666	Drosophila melanogaster	0.532			NP_572171 (RefSeq) AAF45957 (GB protein) Q9W4J4 (Uniprot)

InParanoid database: Table



	A.gambiae	C.elegans	C.briggsae	D.melanogaster	D.rerio	T.rubripes	G.gallus	H.sapiens	M.musculus
A.gambiae		5155	4830	7993	5638	6079	5563	6283	6185
C.elegans	5426		11 506	5215	5310	5525	5096	5704	5736
C.briggsae	4573	10878		4644	4360	4749	4411	4869	4834
D.melanogaster	7724	4837	5033		5415	6012	5495	6140	6074
D.rerio	7837	6817	7449	7747		11 651	9721	11 111	11 006
T.rubripes	8442	7603	7929	8504	11 101		10 234	11 515	11 713
G.gallus	6551	5623	5944	6580	9021	9755		11 416	11 212
H.sapiens	9288	7758	8763	8982	11 536	12 467	11 938		16 356
M.musculus	9737	8829	9527	9643	12 209	13 268	12 205	16 833	
P.troglodytes	7096	6184	6887	7024	9845	10 416	10 460	17 861	14 135
R.norvegicus	8415	7435	8572	8466	11 496	12 175	11 463	15 568	17 374
O.sativa	7313	6497	7004	7353	7992	8055	7351	8293	8254
P.falciparum	1619	1522	1553	1494	1340	1497	1530	1765	1850
A.thaliana	9638	9025	9645	9524	10 195	10 673	9545	10 710	10 754
E.coli	1369	1013	1009	999	986	1015	951	988	947
S.cerevisiae	2419	2173	2267	2382	2285	2512	2309	2564	2582
S.pombe	2439	2218	2321	2417	2308	2611	2391	2681	2648

Paranoid Database analysis

- How to interpret results
- Protein dataset
  - Total protein set obtained protein databases
- Proteins analyzed
  - Total no of proteins used in inparanoid clustering
- Symmetric?
  - Insight on gene duplication
  - Distance principle not applicable for orthologs

Various tools

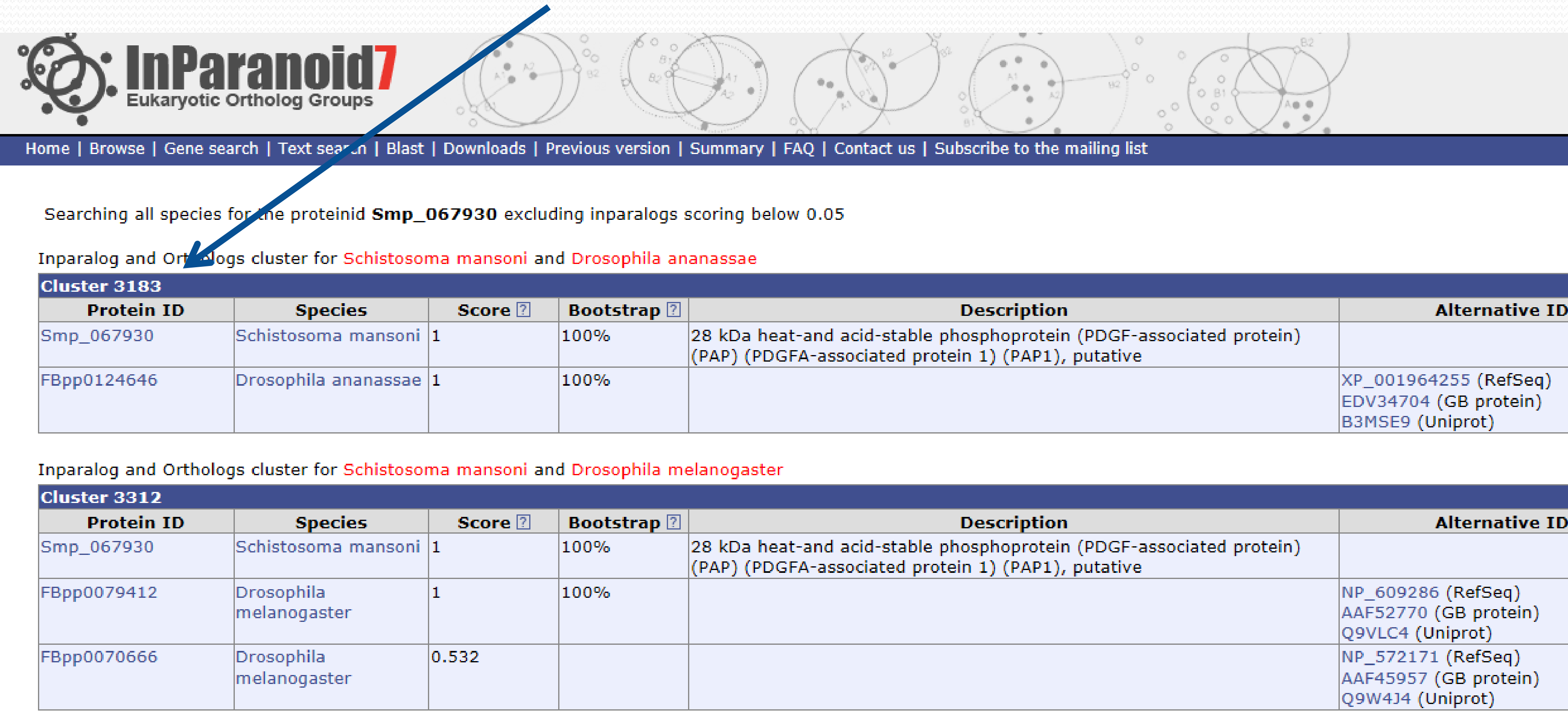
- Human vs All
  - Allows the user to select an organism to display all Inparanoid clusters between it and human
- All species vs All
  - Similar except that one can freely choose which two organisms to pick
- Both approaches displays all possible inparalogs

Tools continued

- Gene Search
  - Requires an identifier
- Text search
  - More flexible search which first outputs a list of genes whose annotation matches the query text string
- Blast Search
  - Allows one to enter a sequence to Blast against the protein datasets

Result Summary

Cluster name


InParanoid7
Eukaryotic Ortholog Groups
Home | Browse | Gene search | Text search | Blast | Downloads | Previous version | Summary | FAQ | Contact us | Subscribe to the mailing list

Searching all species for the proteinid **Smp_067930** excluding inparalogs scoring below 0.05

Inparalog and Orthologs cluster for Schistosoma mansoni and Drosophila ananassae

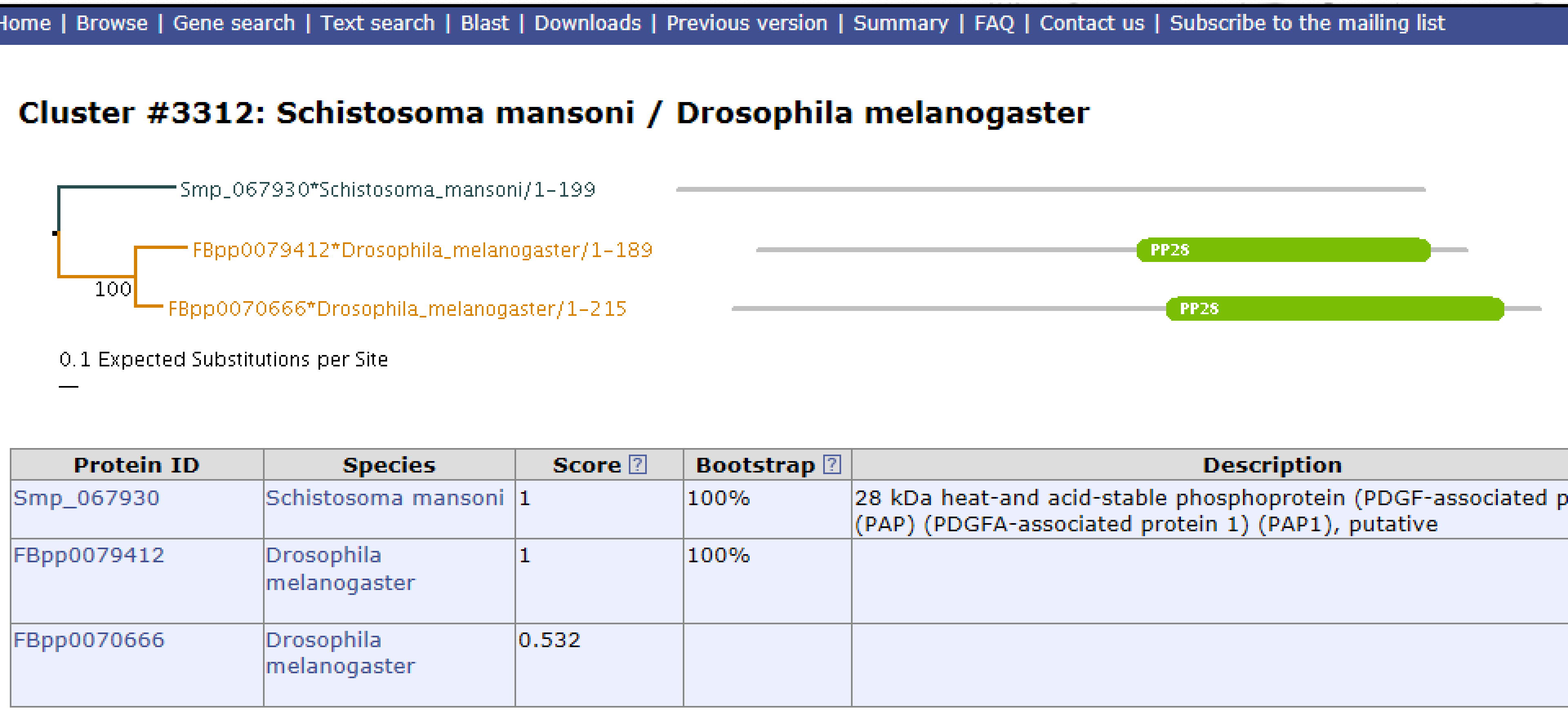
Cluster 3183

Protein ID	Species	Score ?	Bootstrap ?	Description	Alternative ID
Smp_067930	Schistosoma mansoni	1	100%	28 kDa heat-and acid-stable phosphoprotein (PDGF-associated protein) (PAP) (PDGFA-associated protein 1) (PAP1), putative	
FBpp0124646	Drosophila ananassae	1	100%		XP_001964255 (RefSeq) EDV34704 (GB protein) B3MSE9 (Uniprot)

Inparalog and Orthologs cluster for Schistosoma mansoni and Drosophila melanogaster

Cluster 3312

Protein ID	Species	Score ?	Bootstrap ?	Description	Alternative ID
Smp_067930	Schistosoma mansoni	1	100%	28 kDa heat-and acid-stable phosphoprotein (PDGF-associated protein) (PAP) (PDGFA-associated protein 1) (PAP1), putative	
FBpp0079412	Drosophila melanogaster	1	100%		NP_609286 (RefSeq) AAF52770 (GB protein) Q9VLC4 (Uniprot)
FBpp0070666	Drosophila melanogaster	0.532			NP_572171 (RefSeq) AAF45957 (GB protein) Q9W4J4 (Uniprot)

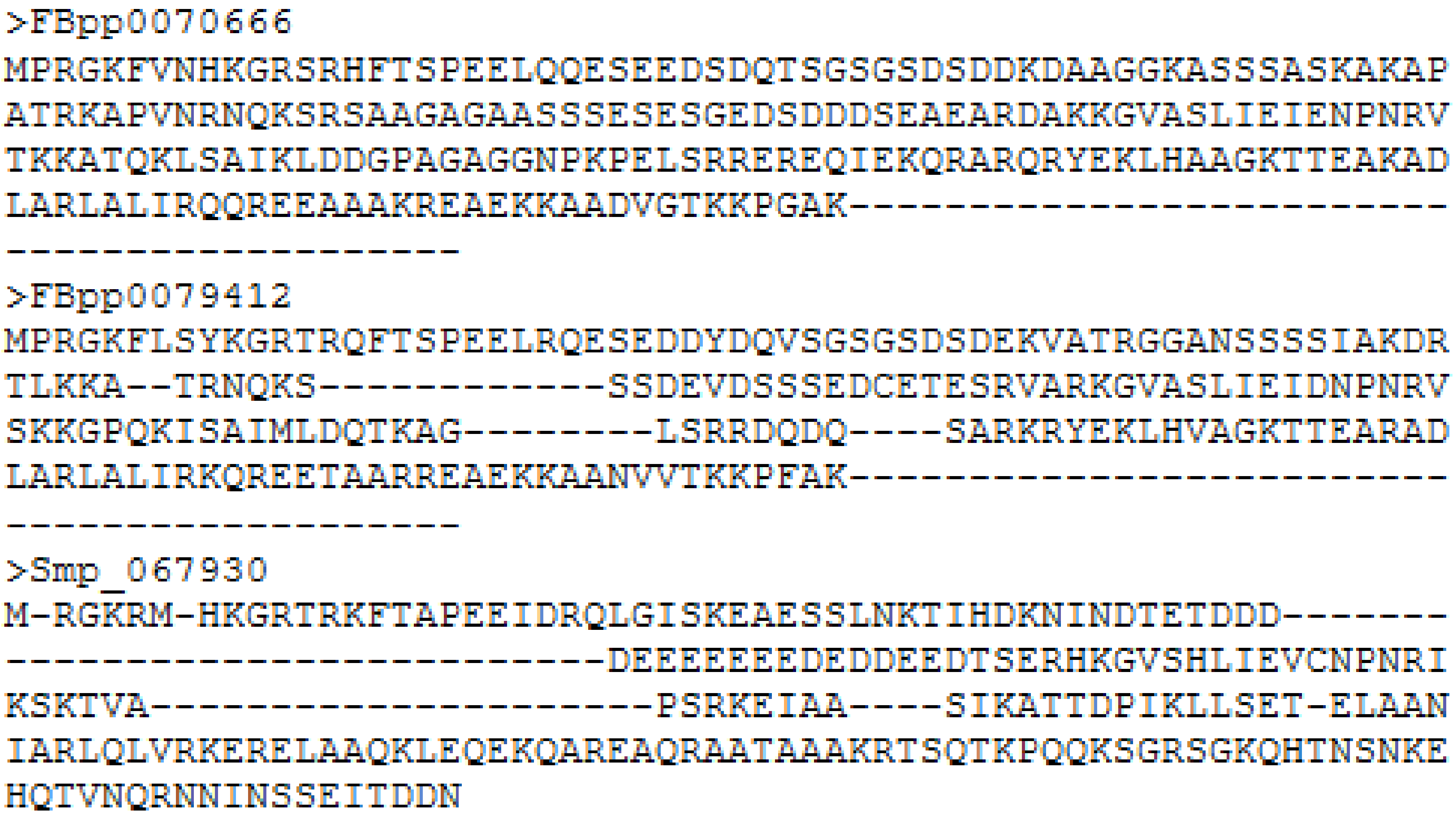


Home | Browse | Gene search | Text search | Blast | Downloads | Previous version | Summary | FAQ | Contact us | Subscribe to the mailing list

Cluster #3312: Schistosoma mansoni / Drosophila melanogaster

Smp_067930*Schistosoma_mansoni/1-199
FBpp0079412*Drosophila_melanogaster/1-189
100
FBpp0070666*Drosophila_melanogaster/1-215
PP28
PP28
0.1 Expected Substitutions per Site

Protein ID	Species	Score ?	Bootstrap ?	Description
Smp_067930	Schistosoma mansoni	1	100%	28 kDa heat-and acid-stable phosphoprotein (PDGF-associated p (PAP) (PDGFA-associated protein 1) (PAP1), putative
FBpp0079412	Drosophila melanogaster	1	100%	
FBpp0070666	Drosophila melanogaster	0.532		



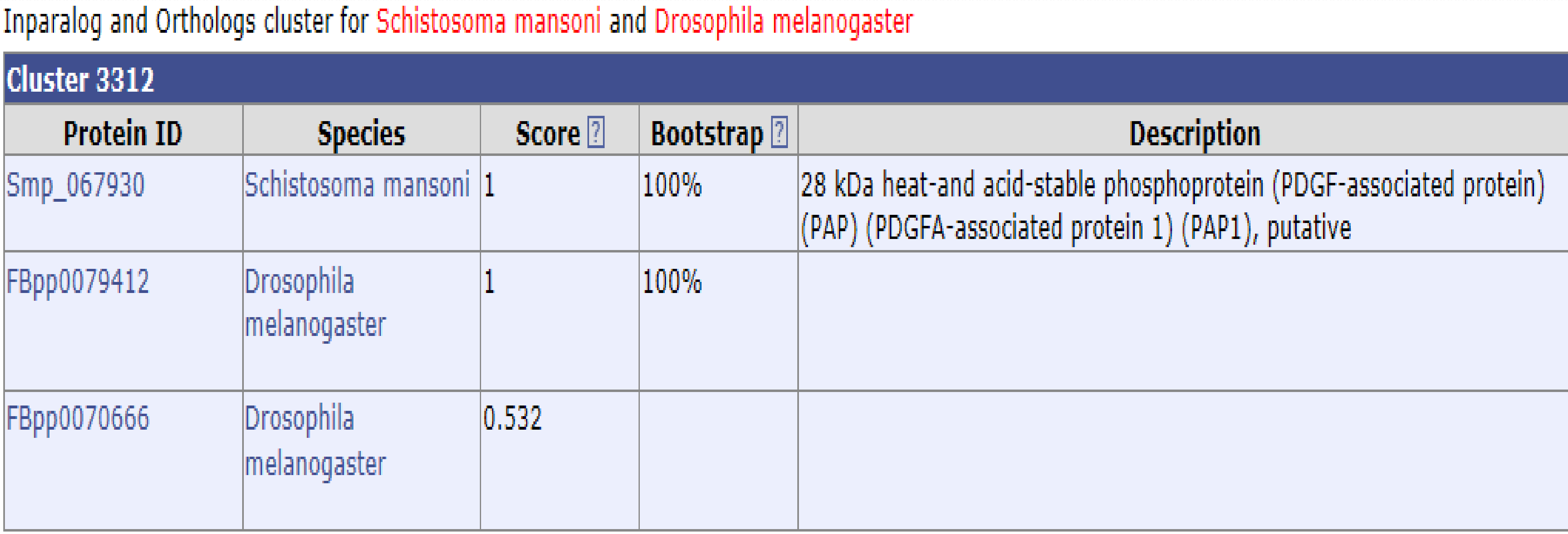
```
>FBpp0070666
MPRGKFVNHKGRSRHFTSPEELQQESEEDSDQTSGSGSDSDDKDAAGGKASSSASKAKAP
ATRKAPVNRNQKSRSAAGAGAASSSESESGEDSDDDSEAEARDAKKGVASLIEIENPNRV
TKKATQKLSAIKLDDGPAGAGGNPKPELSRREREQIEKQRARQRYEKLHAAGKTTEAKAD
LARLALIRQQREEAAAKREAEKKAADVGTKKPGAK-------------------------
-------------------
>FBpp0079412
MPRGKFLSYKGRTRQFTSPEELRQESEDDYDQVSGSGSDSDEKVATRGGANSSSSIAKDR
TLKKA--TRNQKS------------SSDEVDSSSEDCETESRVARKGVASLIEIDNPNRV
SKKGPQKISAIMLDQTKAG--------LSRRDQDQ----SARKRYEKLHVAGKTTEARAD
LARLALIRKQREETAARREAEKKAANVVTKKPFAK-------------------------
-------------------
>Smp_067930
M-RGKRM-HKGRTRKFTAPEEIDRQLGISKEAESSLNKTIHDKNINDTETDDD-------
-------------------------DEEEEEEEDEDDEEDTSERHKGVSHLIEVCNPNRI
KSKTVA---------------------PSRKEIAA----SIKATTDPIKLLSET-ELAAN
IARLQLVRKERELAAQKLEQEKQAREAQRAATAAAKRTSQTKPQQKSGRSGKQHTNSNKE
HQTVNQRNNINSSEITDDN
```

Orthophylogram

- Phylogenetic tree
- Based on the average fraction of InParanoid orthologs between species
- 99 species

Inparanoid -Cross Species Advantage

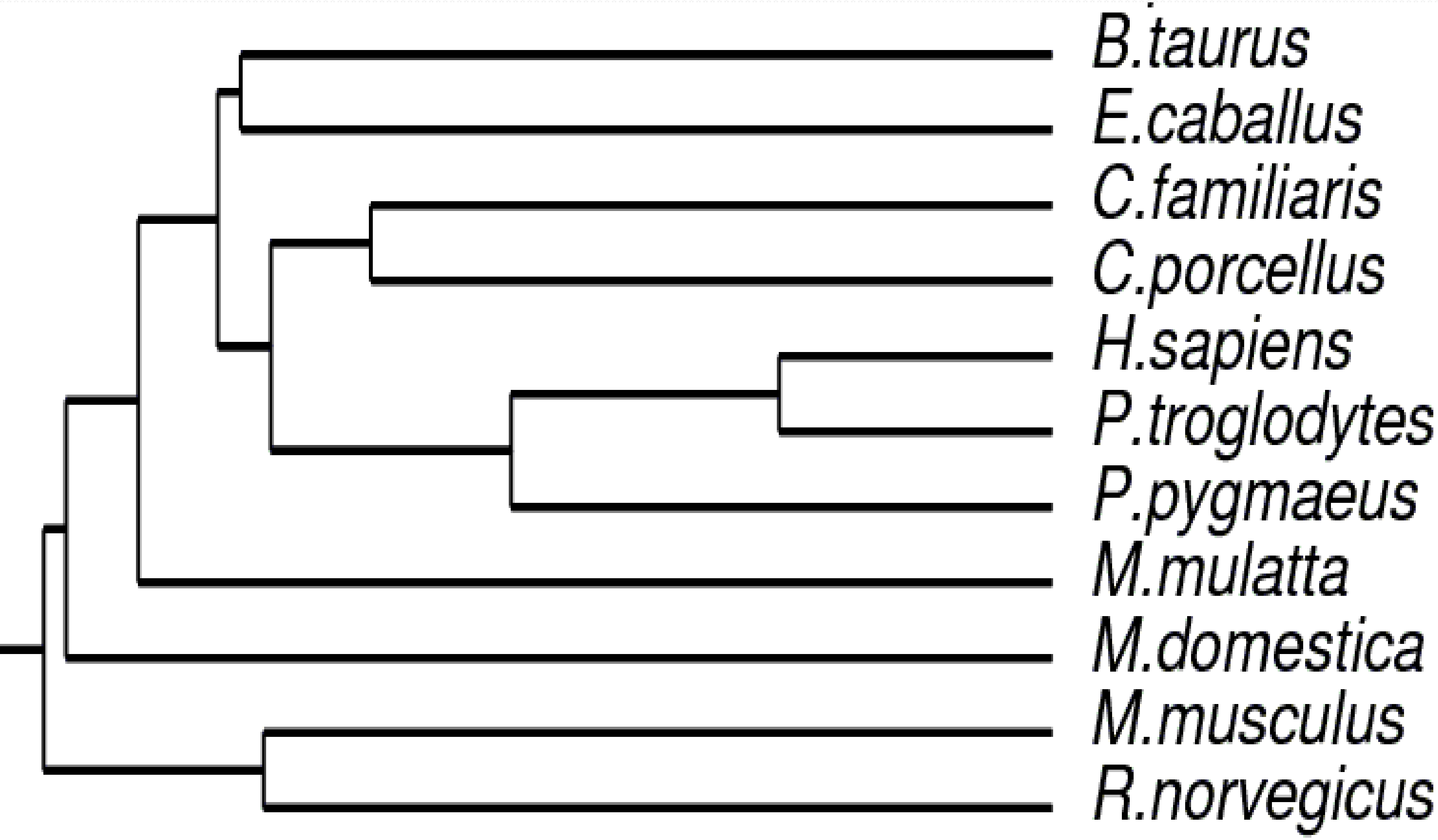
- Orthophylogram – great source for cross species similarity information



Inparalog and Orthologs cluster for Schistosoma mansoni and Drosophila melanogaster

Cluster 3312

Protein ID	Species	Score ?	Bootstrap ?	Description
Smp_067930	Schistosoma mansoni	1	100%	28 kDa heat-and acid-stable phosphoprotein (PDGF-associated protein) (PAP) (PDGFA-associated protein 1) (PAP1), putative
FBpp0079412	Drosophila melanogaster	1	100%	
FBpp0070666	Drosophila melanogaster	0.532		


B.taurus
E.caballus
C.familiaris
C.porcellus
H.sapiens
P.troglodytes
P.pygmaeus
M.mulatta
M.domestica
M.musculus
R.norvegicus

Thank You