

Additional Material for  
**'SuperPartitions: Detection and Classification of Orthologs'**

**Fredj Tekaiia and Edouard Yeramian**

## Figure S1: Partitioning, MCL clustering and labelling conventions

Rv2787	Rv0530	6e-43
Rv2787	Rv3860	2e-26
Rv2787	Rv3876	6e-22
Rv2787	Rv3888c	2e-21
Rv2787	Rv3884c	3e-18
Rv2787	Rv1798	3e-15
Rv2787	Rv0282	1e-13
Rv0282	Rv3868	2e-77
Rv0282	Rv1798	9e-69
Rv0282	Rv3884c	8e-59
Rv0282	Rv2787	4e-11
Rv1798	Rv3868	3e-69
Rv1798	Rv0282	2e-67
Rv1798	Rv3884c	1e-59
Rv1798	Rv2787	6e-14
Rv3860	Rv3888c	7e-37
Rv3860	Rv0530	2e-28
Rv3860	Rv2787	1e-26
Rv3860	Rv3876	2e-19
Rv3868	Rv0282	6e-77
Rv3868	Rv1798	3e-69
Rv3868	Rv3884c	6e-52
Rv3876	Rv0530	3e-38
Rv3876	Rv2787	1e-27
Rv3876	Rv3888c	1e-27
Rv3876	Rv3860	3e-21
Rv3884c	Rv1798	1e-59
Rv3884c	Rv0282	9e-57
Rv3884c	Rv3868	4e-51
Rv3884c	Rv2787	4e-17
Rv3888c	Rv3860	3e-42
Rv3888c	Rv0530	3e-39
Rv3888c	Rv3876	3e-29
Rv3888c	Rv2787	1e-26
Rv0530	Rv2787	3e-43
Rv0530	Rv3876	2e-38
Rv0530	Rv3888c	8e-35
Rv0530	Rv3860	5e-28

**a**

P9.5	Rv2787
P9.5	Rv3888c
P9.5	Rv3876
P9.5	Rv3860
P9.5	Rv0530
P9.5	Rv3884c
P9.5	Rv1798
P9.5	Rv0282
P9.5	Rv3868

**b** P9.5

C5.4	Rv2787
C5.4	Rv3888c
C5.4	Rv3876
C5.4	Rv3860
C5.4	Rv0530

**c** P9.5.C5.4

C4.32	Rv3884c
C4.32	Rv1798
C4.32	Rv0282
C4.32	Rv3868

**d** P9.5.C4.32

Rv2787	P9.5.C5.4
Rv3888c	P9.5.C5.4
Rv3876	P9.5.C5.4
Rv3860	P9.5.C5.4
Rv0530	P9.5.C5.4
Rv3884c	P9.5.C4.32
Rv1798	P9.5.C4.32
Rv0282	P9.5.C4.32
Rv3868	P9.5.C4.32

**e**

**Panel a:** Randomly chosen example for a subset of similar proteins from *Mycobacterium Tuberculosis* that are linked together with significant blastp e-values indicated on the right column, with the proteins in the subset displaying no significant similarity with proteins not contained in the subset. The e-values in this example vary from 4e-11 to 2e-77. Proteins in this subset are characterized by variable numbers of links with each others. For example the number of links for Rv2787 is 14, and only 6 for Rv3868.

**Panel b:** List of all proteins in subset in Panel a. The 9 members of the subset are merged in the partition denoted P9.5 (the second label, 5, is associated with random indexing).

**Panels c and d:** The MCL algorithm (with inflation index I=3.0) is applied to the whole set of related proteins in *Mycobacterium Tunberculosis*. For the proteins in Panel a, this algorithm

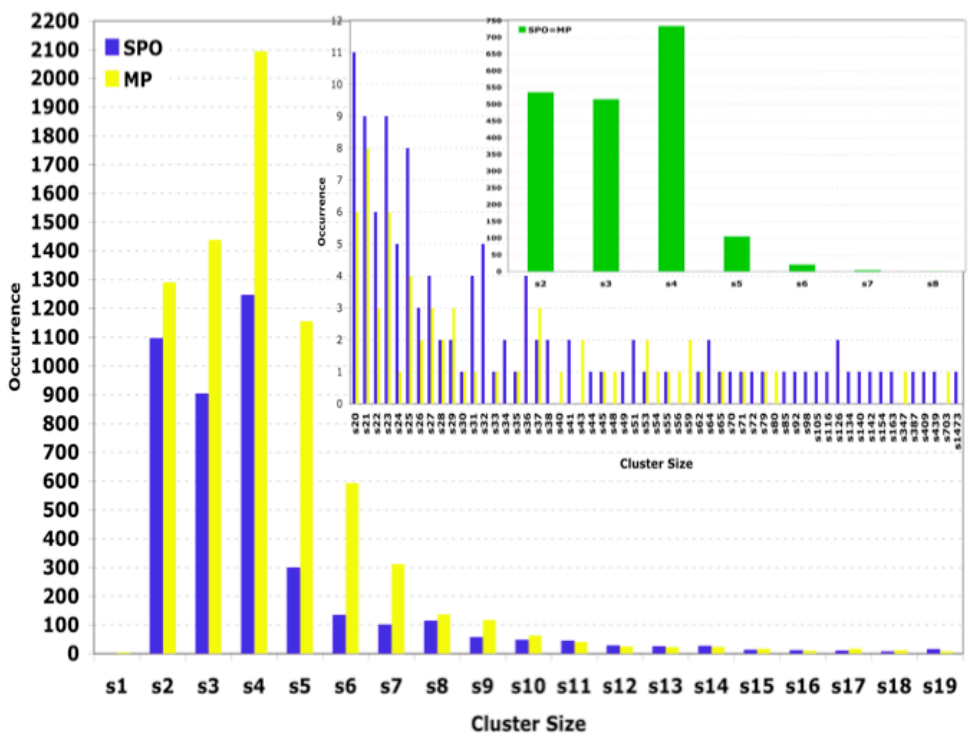
leads to two clusters: C5.4 (Panel c; a cluster containing the 5 proteins represented in this panel; the second label, 4, is associated with random indexing) and C4.32 (Panel d; a cluster containing the 4 proteins represented in the panel; the second label, 32, is associated with random index). The colours in Panels c and d correspond to those in Panel a. Based on the information in Panel a, it appears that in general the e-values corresponding to proteins in C4.32 are more significant than those in C5.4.

Following Panel a we can observe that connected Red and Blue proteins are assigned through MCL respectively C5.4 and C4.32 clusters. Following this observation, based only on MCL clustering, it is possible to miss connections between members of different clusters (such as for those in C5.4 and C4.32). To keep track of the information relative to the partitioning scheme, a Pn.m prefix is appended to each Cp.q cluster, with Pn.m corresponding to the partition to which the members of the Cp.q cluster belong. Accordingly, the MCL clusters C5.4 and C4.32 are respectively labelled as P9.5.C5.4 and P9.5.C4.32.

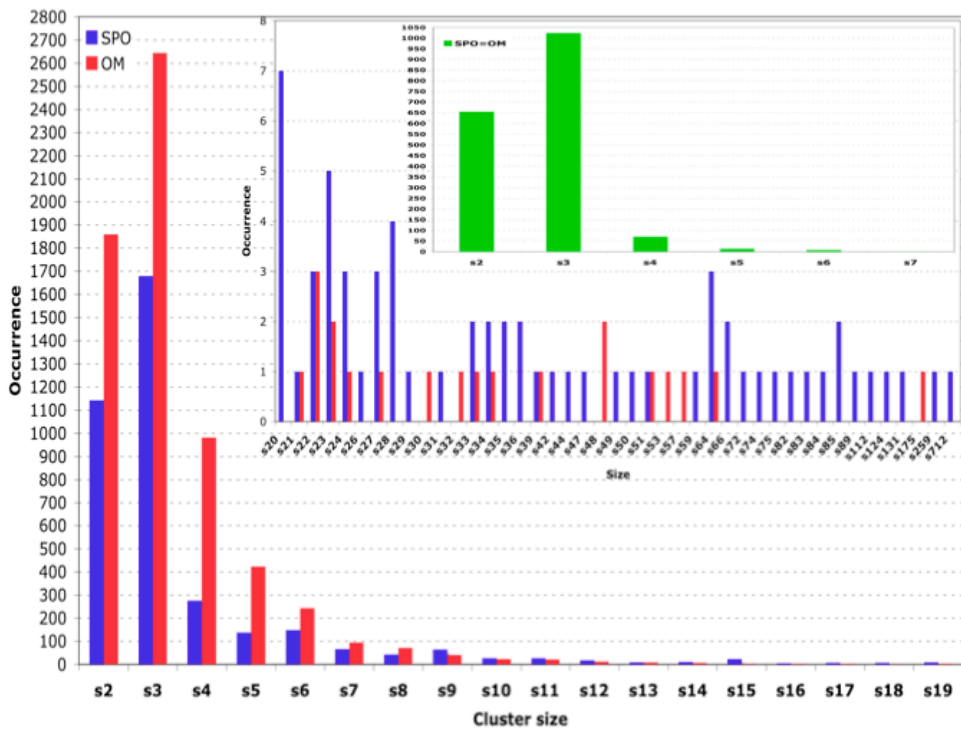
**Panel e:** Final assignment for the classification of proteins in the panels above, illustrating the merging scheme in SPO construction.

---

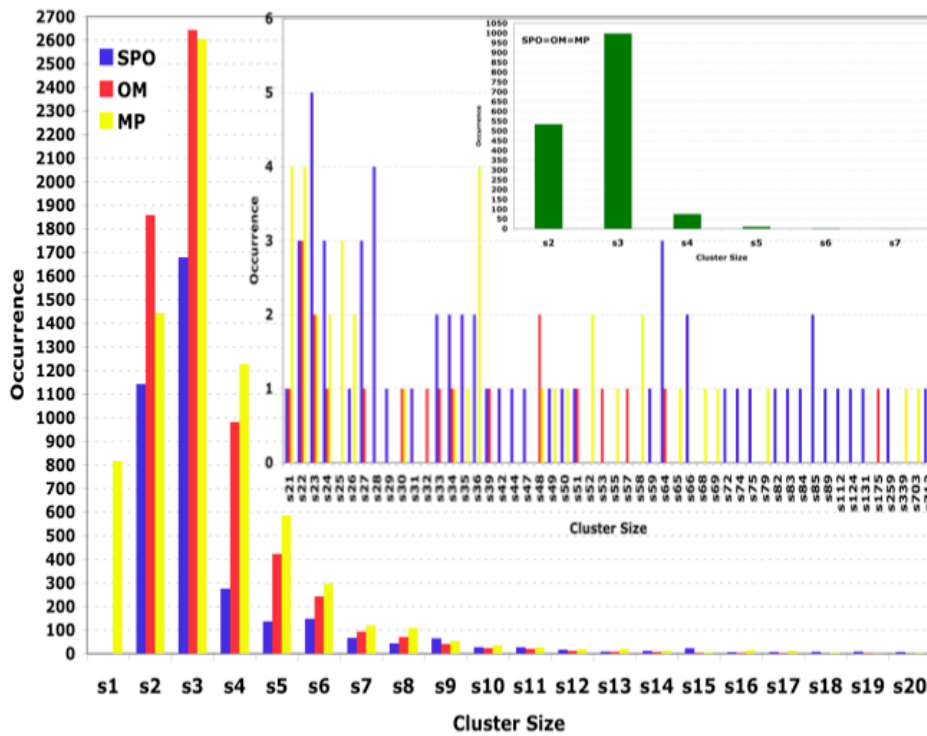
**Figure S2: Distributions of cluster sizes following orthologs methods**



(a)



(b)



(c)

**Panel a:** The distribution of cluster sizes is displayed for the SPO and MultiParanoid (MP) methods. Following such distributions, it appears that small sized clusters (typically up to 10) are more numerous in MP classification than in SPO, whereas larger sizes are more numerous in SPO (typically larger than 20). The inserted histogram displays the distribution of the 1915 identical clusters detected by SPO and MP.

**Panel b:** The distribution of cluster sizes is displayed for the SPO and OrthoMCL (OM) methods. Small sized clusters are much more frequent in OrthoMCL (essentially up to clusters of size 10) than in SPO, whereas larger sizes (essentially clusters larger than 20) are more frequent in SPO. The insert histogram displays the distribution of the 1772 clusters identically detected by SPO and OM.

**Panel c:** The distribution of cluster sizes is displayed for the SPO, MultiParanoid and OrthoMCL methods. The insert histogram displays the distribution of the 1625 clusters (containing 4451 proteins) identically detected by the three methods.

**Figure S3: Detailed illustrative analyses**

**(a) OM cluster COM6.194**

Prot_ID	OM_Cluster	oOM_Cluster	MP_Cluster	MP_ID	SPO	P_RBH	Par_Cluster
ENSP00000037869	COM6.194	OG2_73743	CMI6.12	5836	SPO3.40	P3.2434	P3.225.C3.740
FBpp0085885	COM6.194	OG2_73743	CMI6.12	5836	SPO3.40	P3.2434	P3.45.C3.183
WBGene00014022	COM6.194	OG2_73743	CMI6.12	5836	SPO3.40	P3.2434	singleton
FBpp0085886	COM6.194	OG2_73743	CMI6.12	5836	0	0	P3.45.C3.183
ENSP00000273340	COM6.194	OG2_73743	CMI6.12	5836	0	0	P3.225.C3.740
FBpp0071208	COM6.194	OG2_73743	0	0	0	0	P3.45.C3.183

**(b) MP cluster CMI6.12**

Prot_ID	OM_Cluster	OM_Cluster	MP_Cluster	MP_ID_Cluster	SPO	P_RBH	Par_Cluster
ENSP00000037869	COM6.194	OG2_73743	CMI6.12	5836	SPO3.40	P3.2434	P3.225.C3.740
FBpp0085885	COM6.194	OG2_73743	CMI6.12	5836	SPO3.40	P3.2434	P3.45.C3.183
WBGene00014022	COM6.194	OG2_73743	CMI6.12	5836	SPO3.40	P3.2434	singleton
FBpp0085886	COM6.194	OG2_73743	CMI6.12	5836	0	0	P3.45.C3.183
ENSP00000273340	COM6.194	OG2_73743	CMI6.12	5836	0	0	P3.225.C3.740
ENSP00000330390	0	0	CMI6.12	5836	0	0	P3.225.C3.740

**(c) OM cluster COM4.7**

Prot_ID	OM_Cluster	OM_Cluster	MP_Cluster	MP_ID	SPO	P_RBH	Par_Cluster
ENSP00000367323	COM4.7	OG2_70803	CMI4.344	928	SPO2.6	P2.189	P7.11.C3.263
FBpp0088745	COM4.7	OG2_70803	CMI4.344	928	SPO2.6	P2.189	singleton
ENSP00000369659	COM4.7	OG2_70803	CMI4.344	928	0	0	0
ENSP00000305858	COM4.7	OG2_70803	CMI4.344	928	0	0	0

**Panels a and b:** COM6.194 corresponds to an OrthoMCL cluster of 6 members, 5 of which belong to the MutiParanoid cluster of 6 members CMP6.12. The associated SPO cluster (SPO3.40) contains only 3 members.

The *D. melanogaster* protein FBpp0071208 in COM6.194 (represented in red) belongs to a paralogous cluster (C3.183 labelled P3.45.C3.183). The *H. sapiens* protein ENSP00000330390 in CMP6.12 (represented in red) belongs to a paralogous cluster (C3.740 labelled P3.225.C3.740). The *C. elegans* protein WBGene00014022 in COM6.194 and CMP6.12 corresponds to a singleton.

As the SPO methods handles only Reciprocal Best Hits only 3 distinct proteins are part of the SPO cluster (SPO3.40), corresponding exactly to the RBH partition (P3.2434). In the case of OM and MP methods it would be expected that they cluster 7 members (3 *H. sapiens*, 3 *D. melanogaster* and 1 *C. elegans* proteins). In fact, following this rationale, it is seen that one *D.*

*melanogaster* proteins is missed in the OM clustering whereas one *H. sapiens* protein is missed in the MP clustering.

**Panel c:** COM4.7 corresponds to an OrthoMCL cluster, with identical clustering in OM and MP but not in SPO. The OM and MP clusters contain 4 proteins whereas the SPO cluster contains only 2 proteins. In fact OM and MP retain in the clustering three *H. sapiens* paralogs (corresponding to C3.263) which are significantly similar to the singleton FBpp0088745 in *D. melanogaster*. In such situation the corresponding SPO retains only the two Reciprocal Best Hits.

Such examples demonstrate the importance and utility of various detailed descriptors related to Paralogs (Par\_Cluster) and RBH ( $P_{RBH}$ ) clusters, made available in the SPO procedure. With such descriptors it is possible to understand the observed differences following the different classification methods.

### **Clusters populated predominantly by intra-species members:**

#### **Examples of CMP703.1 and CMP347.1**

For the CMP703.1 cluster (size 703 members; cluster 6195 following the MultiParanoid cluster identification) there were no corresponding detections by SPO and by OM. In fact the 703 proteins in this cluster are composed of 498 *H. sapiens*, 204 *C. elegans* and 1 *D. melanogaster* (FBpp0084363) proteins. BLAST comparisons for these proteins (with the parameters adopted in SPO) displayed no significant similarity with any other protein.

For the CMP347.1 cluster (size 347 members; cluster 967 following the MultiParanoid cluster identification) only 7 proteins were detected by SPO. The 340 other proteins correspond to 332 *H. sapiens*, 6 *C. intestinalis* and 1 *D. melanogaster* (FBpp0082061) proteins. All the human proteins are highly similar to each other, forming an intra-species cluster of paralogs, and highly similar to the unique fly protein.

#### **OM and MP clusters/RBH partitioning:**

Detailed analysis of different classifications shows that, in significant number of cases, the sizes of MP and OM clustering correspond to that of the simple RBH partitions in the SPO scheme. More precisely, from the 17056 proteins that have been detected by SPO (from the set of 30693 proteins considered for the comparisons between the three methods), 8777 proteins display identical OM and  $P_{RBH}$  class sizes and 8488 proteins display identical MP

and P\_RBH class sizes. Accordingly this result indicates that for more than 50% of the proteins detected by SPO, OM and MP classification correspond to the Partitions of RBH ( $P_{RBH}$ ). By comparison, for 3637/8777 and 3765/8488 proteins, the corresponding SPOs merge at least 2 partitions of RBHs.

---

## Additional Data Tables:

### Additional detailed results:

#### a) Detailed results for comparisons between SPO and MultiParanoid (MP):

Detailed results for the detected orthologs and the corresponding clusters are available in the form of an Excel file (MP\_SPO\_Clusters.xls).

Original data relative to *H. sapiens*, *D. melanogaster*, *C. elegans* and *C. intestinalis* were downloaded from:

<http://multiparanoid.sbc.su.se/download/>, Alexeyenko et al., 2006).

#### Headings of columns:

1) **Prot\_ID**: protein Id for *H. sapiens* (HOSA), *D. melanogaster* (DRME), *C. elegans* (CAEL) and *C. intestinalis* (CIIN). Proteins identifications are prefixed by their species code.

2) **MP\_Cluster**: MultiParanoid (MP) cluster identification corresponding to Prot\_ID or 0 if no cluster is assigned by MP.

The cluster is identified by: CMP followed by the number of proteins in the cluster (or size) and by an arbitrary number for indexing.

CMP stands for Cluster following MultiParanoid method.

For example: CMP3.271 corresponds to a MultiParanoid cluster of size 3 (including 3 proteins) and which index is 271.

The size is adapted from the original size by removing the number of proteins from the *C. intestinalis* species that are included in the original classification.

3) **MP\_ID**: original cluster identification (see the above URL), or 0 if no cluster is assigned.

4) **SPO**: SPO cluster, with number of proteins included in the cluster (size) and an arbitrary index, or 0 if no cluster is assigned by SPO.

5) **P\_RBH**: Partition of RBHs (Reciprocal Best Hits). The partition corresponds to the number of proteins it contains and indexed with an arbitrary number (or 0).

6) **Par\_Cluster**: Partition and MCL cluster of the Prot\_ID paralogs (or 0) (see text).

7) **SPO\_C\_Prof**: Conservation profile of the SPO (or 0).

#### b) Detailed results for comparisons between SPO and OrthoMCL (OM):

Detailed results for the detected orthologs and the corresponding clusters are available in the form of an Excel file (OM\_SPO\_Clusters.xls).

Original data relative to *H. sapiens*, *D. melanogaster* and *C. elegans* downloaded from ([http://info.gerstein.org/Ortholog\\_Resources/](http://info.gerstein.org/Ortholog_Resources/), Fang et al. 2010).

Headings of columns:

1) **Prot\_ID:** Protein Id for *H. sapiens* (HOSA), *D. melanogaster* (DRME) and *C. elegans* (CAEL). Original protein identification is prefixed by its species code.

2) **OM\_Cluster:** OrthoMCL (OM) cluster identification corresponding the Prot\_ID, or 0 if no cluster is assigned by OM.

The cluster is identified by: COM followed by the number of proteins in the cluster (or size) and by an arbitrary number for indexing.

COM stands for Clusters following OrthoMCL method.

For example: COM2.1546 corresponds to an OrthoMCL cluster of size 2 (containing 2 proteins) and whose index is 1546 chosen arbitrarily.

3) **OM\_ID:** Original cluster Identification as indicated in the Orthologs Resources URL (see above). 0 if no cluster is assigned.

4) **SPO:** SPO cluster, with number of proteins contained in the cluster (size) and an arbitrary index, or 0 if no cluster is assigned by SPO.

5) **P\_RBH:** Partition of RBHs (Reciprocal Best Hits). The partition corresponds to the number of proteins it includes and indexed by an arbitrary number (or 0).

6) **Par\_Cluster:** Partition and MCL cluster of the ProrId paralogs (or 0 ). (see text)

7) **SPO\_C\_Prof:** Conservation profile of the SPO, ( or 0 ).

**c) Detailed results for comparisons between SPO, MultiParanoid (MP) and OrthoMCL (OM):**

Detailed results for the detected orthologs and the corresponding clusters are available in the form of an Excel file (OM\_MP\_SPO\_Clusters.xls).

Original data relative to *H. sapiens*, *D. melanogaster* and *C. elegans* downloaded from ([http://info.gerstein.org/Ortholog\\_Resources/](http://info.gerstein.org/Ortholog_Resources/), Fang et al. 2010) and from (<http://multiparanoid.sbc.su.se/download/>, Alexeyenko et al., 2006).

Headings of columns:

1) **Prot\_ID:** Protein Id for *H. sapiens* (HOSA), *D. melanogaster* (DRME) and *C. elegans* (CAEL). Original protein identification is prefixed by its species code.

2) **OM\_Cluster:** OrthoMCL (OM) cluster identification corresponding the Prot\_ID, or 0 if no cluster is assigned by OM.

The cluster is identified by: COM followed by the number of proteins in the cluster (or size) and by an arbitrary number for indexing.

COM stands for Clusters following OrthoMCL method.

For example: COM2.1546 corresponds to an OrthoMCL cluster of size 2 (containing 2 proteins) and whose index is 1546, chosen arbitrarily.

3) **OM\_ID:** Original cluster Identification as indicated by Fang et al. 2006 in the Orthologs resources URL (see above), 0 if no cluster is assigned.

4) **MP\_Cluster**: MultiParanoid (MP) cluster identification corresponding to Prot\_ID or 0 if no cluster is assigned by MP.

The cluster is identified by: CMP followed by the number of proteins in the cluster (or size) and by an arbitrary number for indexing.

CMP stands for Cluster following MultiParanoid method.

For example: CMP3.271 corresponds to a MultiParanoid cluster of size 3 (containing 3 proteins) and whose index is 271.

The size is adapted from the original size by removing the number of proteins from the *C. intestinalis* species, included in the original classification.

5) **MP\_ID**: Original cluster identification by Alexeyenko et al., 2006 (see the above URL), or 0 if no cluster is assigned. The original identification corresponds to the one that includes *C. intestinalis*.

6) **SPO**: SPO cluster, with number of proteins contained in the cluster (size) and an arbitrary index, or 0 if no cluster is assigned by SPO.

7) **P\_RBH**: Partition of RBHs (Reciprocal Best Hits). The partition corresponds to the number of proteins it contains and indexed by an arbitrary number (or 0).

8) **Par\_Cluster**: Partition and MCL cluster of the ProrId paralogs (or 0) (see text).

9) **SPO\_C\_Prof**: Conservation profile of the SPO, (or 0).