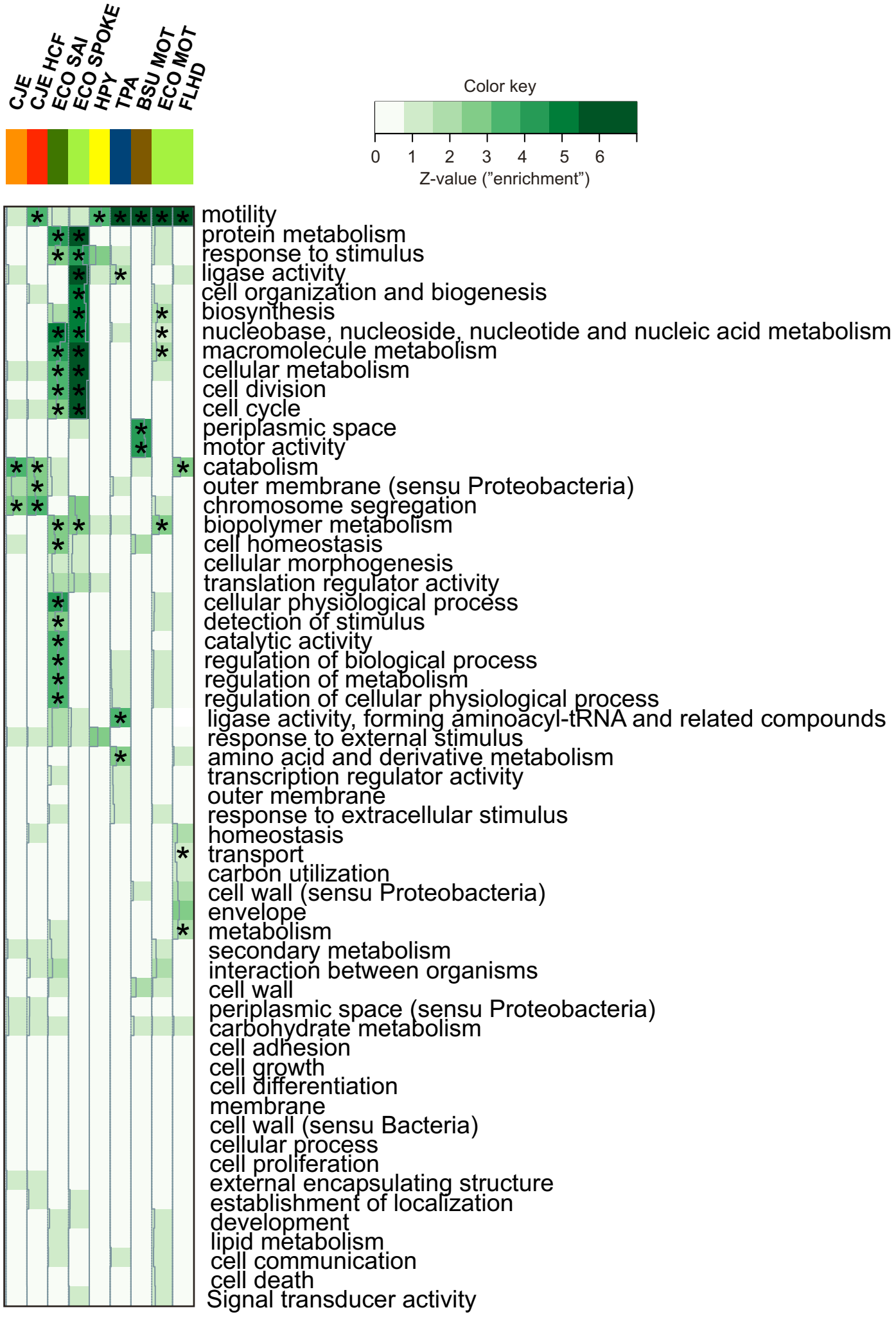
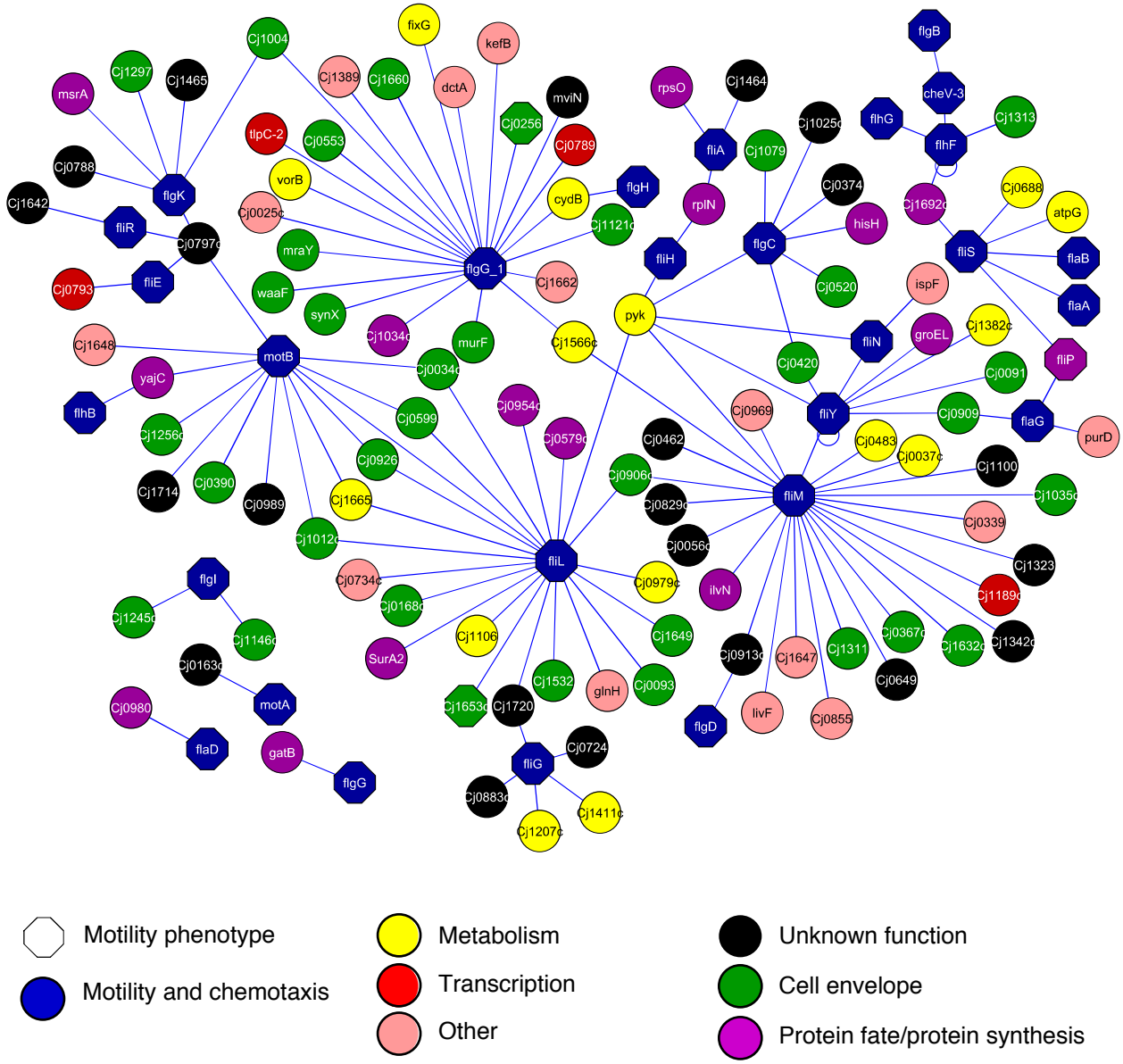
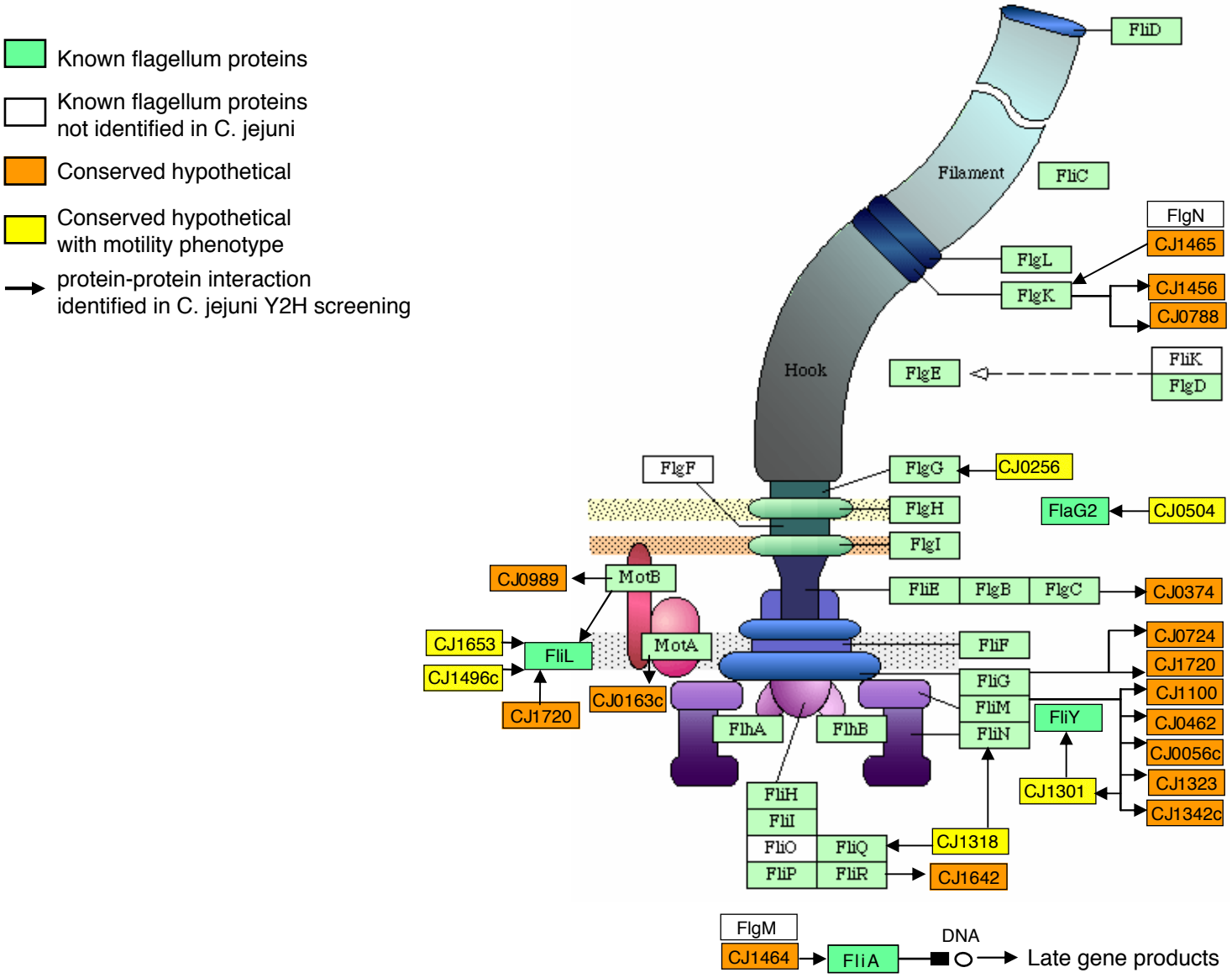
Rajagopala Suppl. Figure S1

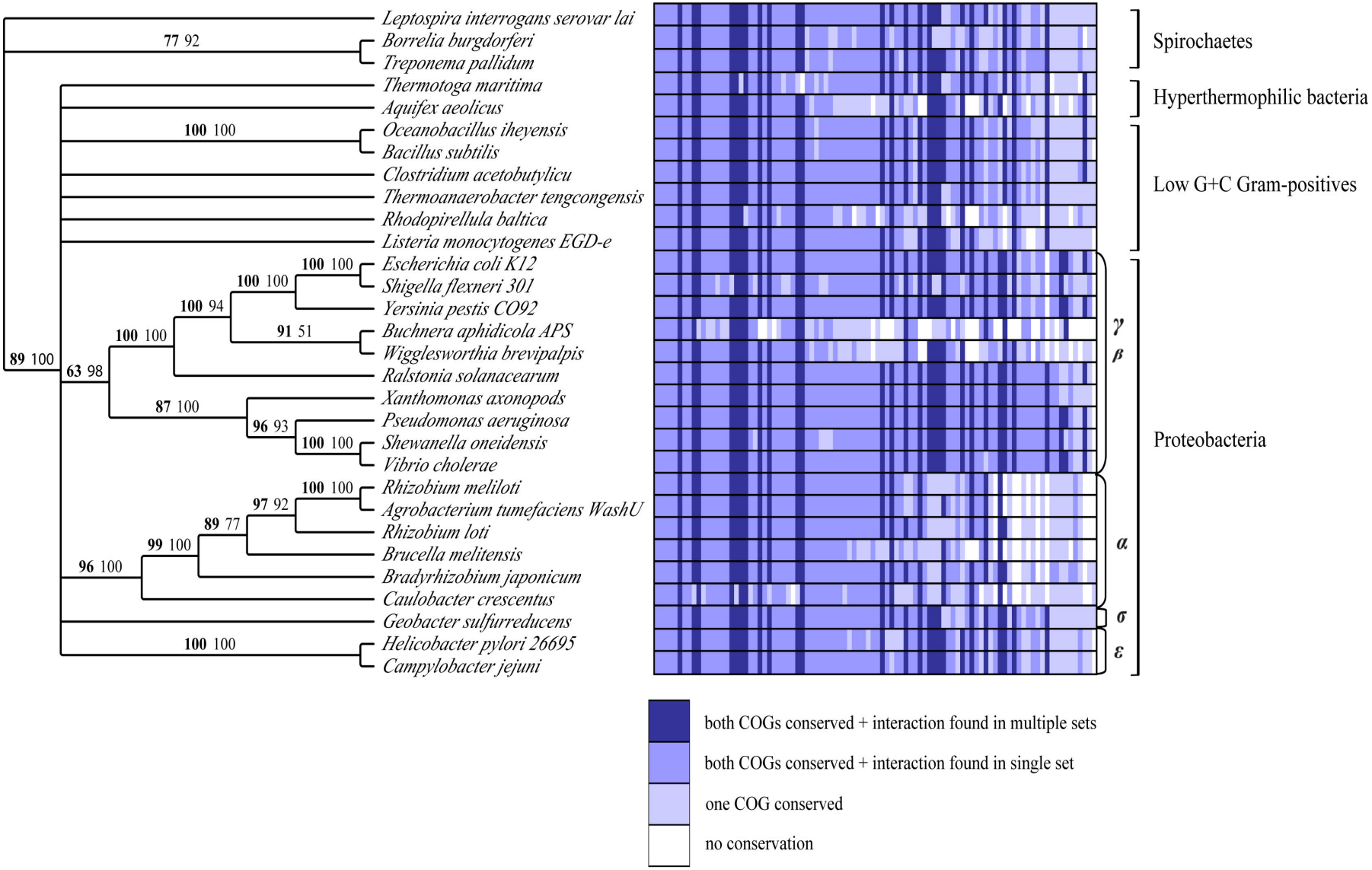



fixG
kefB
Cj1004
Cj1297
Cj1465
Cj1389
Cj1660
dctA
mviN
msrA
Cj0256
Cj0789
tlpC-2
Cj0553
Cj0788
vorB
cydB
flgH
Cj1642
Cj0025c
Cj1121c
fliR
flgK
Cj0797c
mraY
flgG_1
Cj1662
Cj0793
fliE
waaF
synX
Cj1034c
murF
Cj1566c
Cj1648
motB
Cj0034c
yajC
flhB
Cj0599
Cj0954c
Cj1256c
Cj0579c
Cj0926
Cj1714
Cj0390
Cj0989
Cj1665
Cj1012c
flgI
Cj1245c
Cj1146c
Cj0734c
Cj0168c
Cj0163c
SurA2
Cj1106
Cj1649
motA
Cj0980
Cj1532
glnH
Cj0093
Cj1653c
Cj1720
flaD
gatB
flgG
Cj0724
fliG
Cj0883c
Cj1411c
Cj1207c
rpsO
Cj1464
fliA
Cj1025c
Cj1079
rplN
Cj0374
flgC
hisH
fliH
Cj0520
pyk
Cj0462
Cj0969
fliY
Cj0420
fliN
groEL
Cj0483
Cj0037c
Cj0906c
Cj0829c
fliM
Cj0056c
Cj0979c
ilvN
Cj0913c
Cj1647
Cj1311
Cj0649
flgD
livF
Cj0855
flgB
cheV-3
flhG
Cj1313
flhF
Cj1692c
Cj0688
atpG
fliS
flaB
ispF
flaA
Cj1382c
fliP
Cj0091
Cj0909
flaG
purD
Cj1100
Cj1035c
Cj0339
Cj1323
Cj1189c
Cj0367c
Cj1632c
Cj1342c
Motility phenotype
Motility and chemotaxis
Metabolism
Transcription
Other
Unknown function
Cell envelope
Protein fate/protein synthesis

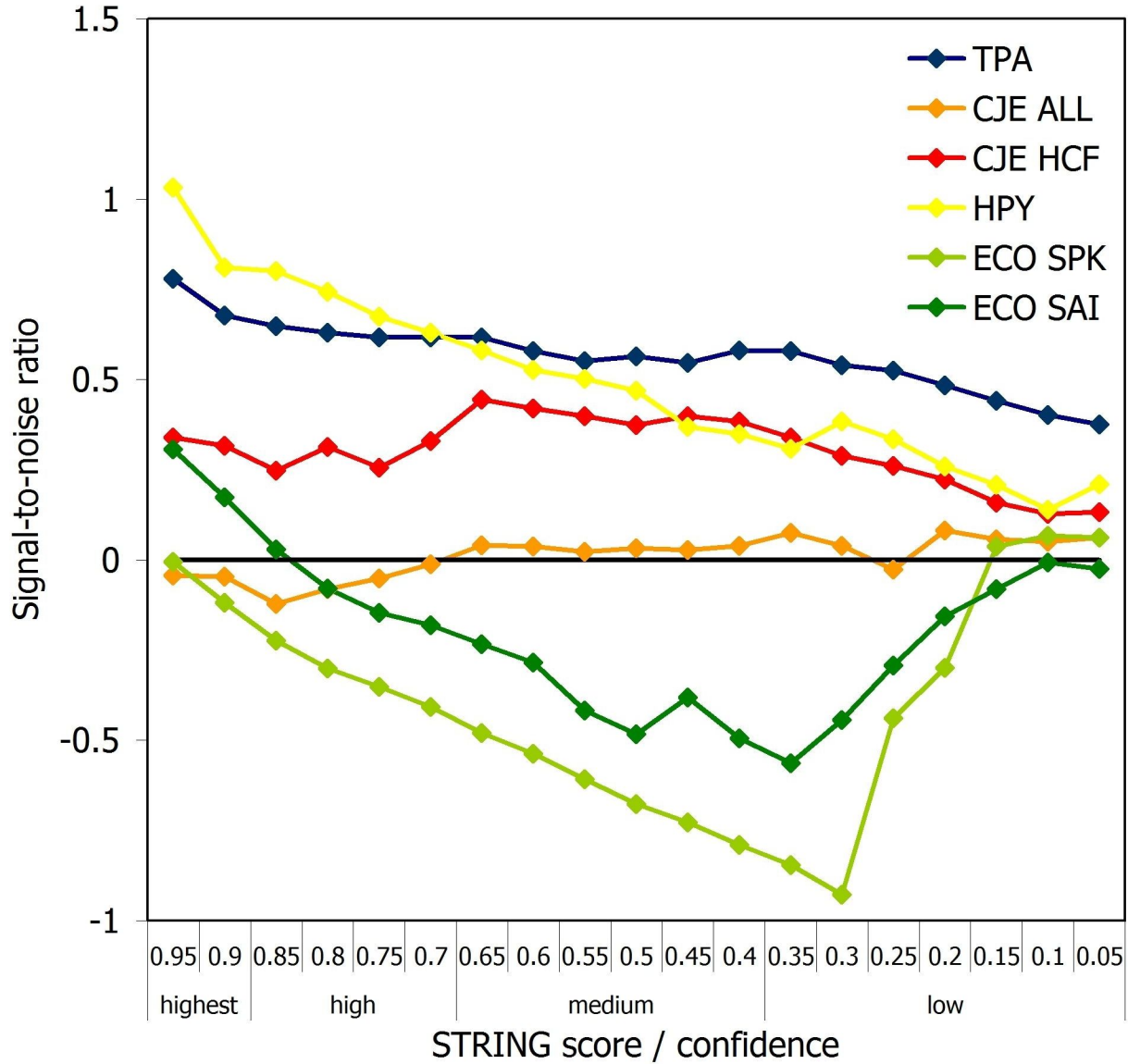
Rajagopala et al., Supplementary Figure S3

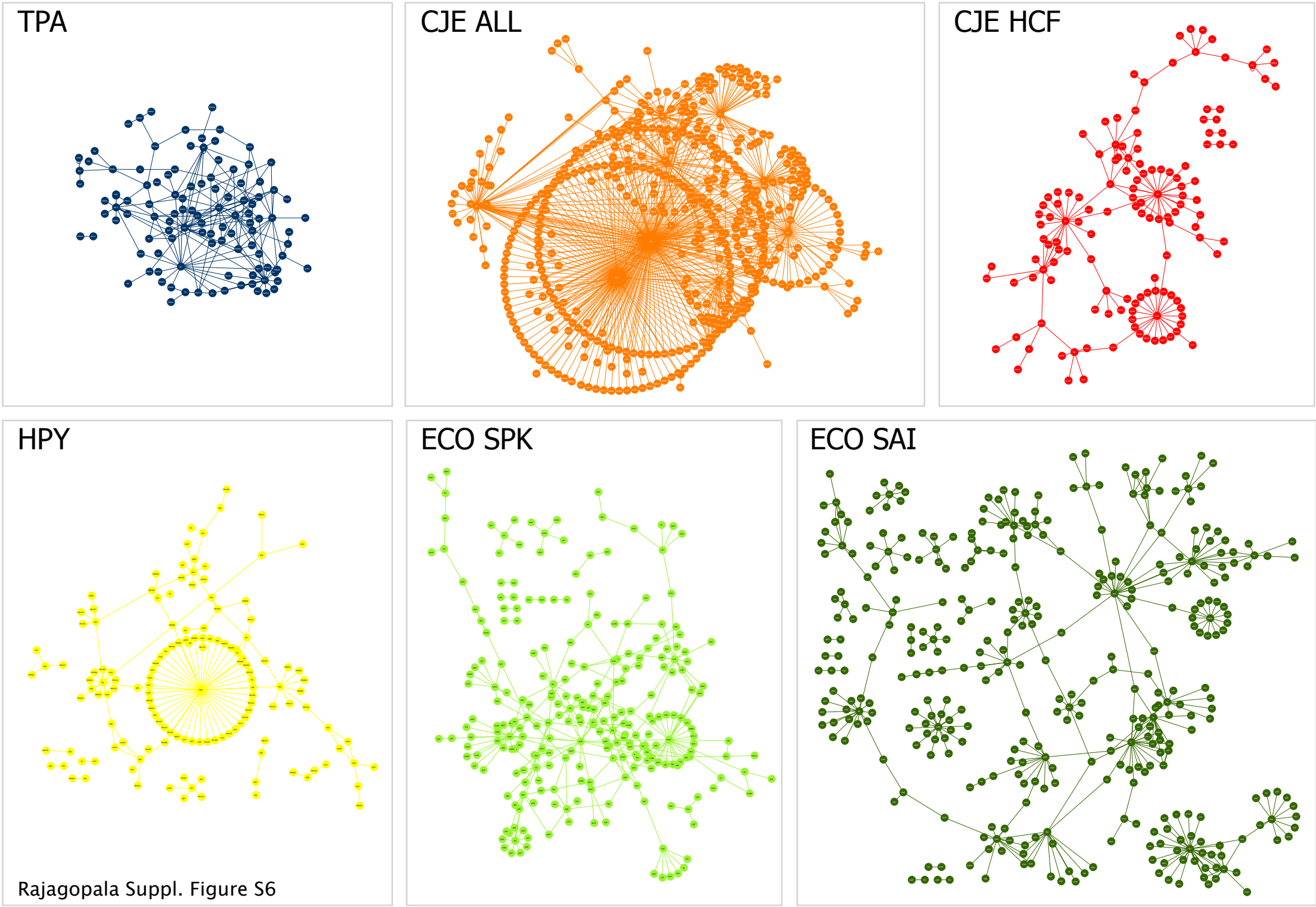


Rajagopala et al., Suppl. Figure S4



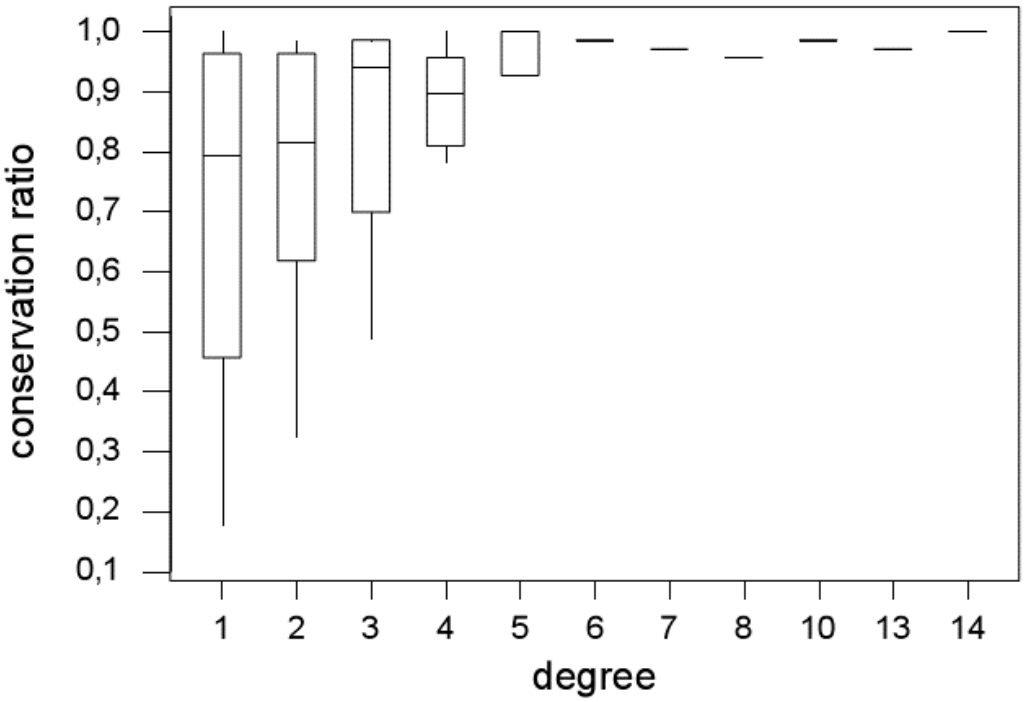
Rajagopala et al., Suppl. Figure S5





Rajagopala Suppl. Figure S6

Rajagopala et al., Supplementary Figure S7



Supplementary Figure Legends

Figure S1. GO Functional Associations with Motility. The associations of known motility proteins with several functional classes (as defined by GO Terms) are shown for the interaction datasets (TPA, CJE, ECO), motility phenotyping sets (BSU MOT, ECO MOT), and for the set regulated by the master regulator of the flagellum, flhD, ([ECO] FlhD). Overrepresentation of a certain GO term in a certain dataset is color-coded according to the shown color key – the color encodes the z-value of the respective association, which corresponds to the times standard deviation distance of the found association to a random set (1000x). Statistical significance of each association is indicated by "*" and is calculated by a ranking statistic for the interaction sets and by a hypergeometric distribution for the others.

Figure S2. The motility protein interaction network of *C. jejuni*. Blue nodes are known motility proteins while black proteins are proteins of yet unknown function. Proteins with a motility phenotype in either *E. coli* or *B. subtilis* or *H. pylori* are indicated in octagons. See legend for other functional assignments. An equivalent Figure for the motility network of *T. pallidum* is shown in as Figure 2 in the main text.

Figure S3. Novel proteins of the *C. jejuni* flagellum projected onto the KEGG "flagellar assembly (02040)" reference pathway (compare to Figure 3 in the main text).

Figure S4. Flagellum supertree with phylogenetic interaction profiles. Bacterial flagellum supertree of 35 flagellar protein families conserved in up to 30 species. Two alternative treeing methods, maximum parsimony (MP) and neighbor-joining (NJ) were used to generate two sets of family trees. For each set, a supertree was constructed. The cladogram shows the consensus tree. Bootstrap values indicate reproducibility (100 replicates) of MP analyses of the supertrees (those obtained for the MP trees are marked in bold, those obtained for the NJ trees are marked in plain).

We used interactions from our integrated network for phylogenetic profiling. Interaction profiles (colored stretches) were mapped onto the supertree. Dark blue stretches reflect conserved interactions found in more than one set (including literature interactions). Strikingly, interaction profiles partially reflect the supertree phylogeny, e. g. the monophyletic group of α proteobacteria.

Figure S5. STRING evaluation of interaction datasets. Percentage of interactions among orthologous groups (signal) which scored greater than a specific STRING-score (x-axis) compared to the percentage expected from the randomized networks (noise). A signal-to-noise ratio (y-axis) above zero indicates that the signal was stronger than the noise, i.e. the observed percentage was higher than the average percentage found in the randomised networks. Confidence as defined by (Stein *et al*, 2005; von Mering *et al*, 2003).

Figure S6. All motility networks used in this study. **TPA** is the Y2H interaction dataset generated in this study. **CJE ALL** and **CJE** are the Y2H datasets generated for *C. jejuni* in this study with all interactions and high confidence interactions, respectively. **HPY** is the Y2H datasets from Rain *et al.* (2001). **ECO SPK** and **ECO SAI** are two interaction datasets for *E. coli* derived from a complex purification study (Arifuzzaman et al. 2006) by extraction of binary interactions. Proteins are indicated as nodes, protein interactions are represented as edges.

Figure S7. Conservation and degree in the integrated protein network. The more interactions a COG has (i.e. the higher its degree), the more conserved it is (measured by the number of species that have this COG, as a fraction of 68 flagellated bacteria; $r = 0.43$, $p < 0.005$).

Methods for Supplementary Figure 4: supertree

Super tree construction of flagellum complex proteins

Proteins involved in the 'Flagellar assembly' pathway (ko02040) were downloaded from KEGG (Kanehisa *et al*, 2006). Protein sequences were aligned using CLUSTAL W 1.83 (Chenna *et al*, 2003) (default parameters). Multiple alignments were submitted to GBLOCKS (Castresana, 2000) (default parameters). If a family contained recent paralogs (paralogs which are more similar to each other than to proteins of other species), one protein wa srandomly chosen and removed. If there were early paralogs (paralogs which are more similar to proteins from other species than to its own), only the most similarcompared to the majority of proteins was retained. Protein families with less than 4 taxa or no conserved GBLOCKS sites were excluded from further analysis.

Construction of maximum parsimony (MP) consensus trees

The PIR formatted GBLOCKS were converted into the NEXUS format by the READSEQ program. The NEXUS files were subjected to phylogenetic analysis using PAUP* win-4b10 (Swofford, 2003). For each protein family, a bootstrap analysis with 100 bootstrap replicates was performed using a heuristic search based on the MP method. In total, 35 bootstrap consensus (50% majority-rule) trees were constructed.

Construction of neighbor-joining (NJ) consensus trees

The PIR formatted GBLOCKS were converted into PHYLIP format by the READSEQ program. The PHYLIP files were bootstrapped with SEQBOOT (Felsenstein, 2005; Schmidt *et al*, 2002) with 100 bootstrap replicates. Maximum likelihood (ML) distance matrices were computed by TREE-PUZZLE 5.2 (Schmidt *et al*, 2002) using the Dayhoff amino acid substitution model incorporating among-site rate variation (gamma law based model, alpha parameter estimated by TREE-PUZZLE, eight gamma rate cat(Felsenstein, 2005)egories) in combination with PUZZLEBOOT 1.035. Trees were generated from these ML distance matrices using NEIGHBOR (Felsenstein, 2005) and summarized into 35 bootsrapped consensus trees (50% majority-rule) using CONSENSE (Felsenstein, 2005).

Construction of supertrees

A matrix representation using parsimony (MRP) approach was used to represent protein family trees as a single binary matrix (only branches with a bootstrap support higher than 50% were considered) (Baum, 1992). The MRP matrices of the 35 MP and NJ bootstrapped consensus trees were constructed with CLANN (Creevey and McInerney, 2005). For each matrix a bootstrapped (100 bootstrap replicates) consensus tree (50% majority-rule) was generated by PAUP* using a heuristic search based on the MP method. The resulting two trees were merged using CLANN (50% majority-rule) and drawn using TreeGraph 6 (Müller and Müller, 2004).

Evaluation based on genomic context

To evaluate the individual interactions based on genomic context we extracted links between orthologous groups (i-COGs) from the individual PPI sets and scored them based on STRING v6.3 (9 February 2006) (von Mering *et al*, 2005) genomic context scores (S-score: gene fusion,

neighborhood, co-occurence) (von Mering et al, 2003). We plotted the percentage of i-COGs (y-axis) found in each set which scored greater than a specific S-score (x-axis) (Supplementary Figure S5). An average percentage distribution was also generated for 1000 randomisations of each set. A signal-to-noise ratio (SNR) was computed from these observed and random percentage distributions (representing signal and noise, respectively) according to the following formula:

$$SNR(S-score) = log_{10} \frac{observed\ percentage\ (S-score)}{avg\ (random\ percentage\ (S-score))}$$

Motility dataset randomization

A randomization procedure for partial interaction networks centered on a selected protein class was devised. The standard rewiring algorithm applied to such partial networks would not yield reliable results, e.g., an overrepresentation of interactions within the selected functional class could not be assessed. The following approach was chosen for the randomization algorithm: retain all proteins of the selected functional class (without addition of additional members of this class), retain the in and out degrees for the members of the selected class, and select interacting proteins randomly until all in and out degrees for the functional class are saturated. The procedure tries to retain the properties of the selected functional class (number of interacting proteins, in and out degrees), however, due to the lack of full information the remaining proteins are randomly sampled.

GO Functional Associations with Motility

In addition to the interaction and phenotyping datasets, a set of genes regulated by the master regulator of the flagellum, FlhD, was obtained from Pruss et al. (Pruss *et al*, 2003).

Functional assignments for proteins were taken from the GOA project (automatically generated GO terms (Camon *et al*, 2004). GO terms were mapped onto GO slim terms (GO slim terms present in prokaryotic GO subset). Known motility proteins (taken from KEGG database, www.genome.jp/kegg/) were defined as an additional functional class ("motility"). The number of interactions between known motility proteins and the other functional categories was counted. Known motility proteins were only counted for the class "motility", not for additional classes to prevent artificial links introduced by intra-motility interactions between proteins annotated with more than one functional class. Overrepresentation of a functional link compared to 1000 randomized networks was assessed by calculating a Z-score

$$Z = \frac{n - < n_{rand} >}{\sigma_{rand}}$$

(n: number of linking interactions, $<n_{rand}>$: average of linking interactions in randomized sets, σ_{rand}: standard deviation in randomized set). In addition, the statistical significance for each association was calculated by a ranking statistic for the interaction sets and by a hypergeometric distribution for the others ($p<0.05$).

References (supertree construction)

Baum BR (1992) Combining trees as a way of combining data sets for phylogenetic inference,and the desirability of combining gene trees. *Taxon* **41:** 3-10.

Camon E, Magrane M, Barrell D, Lee V, Dimmer E, Maslen J, Binns D, Harte N, Lopez R, Apweiler R (2004) The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology. *Nucleic Acids Res* **32:** D262-266.

Castresana J (2000) Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol* **17:** 540-552.

Chenna R, Sugawara H, Koike T, Lopez R, Gibson TJ, Higgins DG, Thompson JD (2003) Multiple sequence alignment with the Clustal series of programs. *Nucleic Acids Res* **31:** 3497-3500.

Creevey CJ, McInerney JO (2005) Clann: investigating phylogenetic information through supertree analyses. *Bioinformatics* **21:** 390-392.

Felsenstein J (2005) Phylip (Phylogeny Inference Package) version 3.6., Seattle.

Kanehisa M, Goto S, Hattori M, Aoki-Kinoshita KF, Itoh M, Kawashima S, Katayama T, Araki M, Hirakawa M (2006) From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res* **34:** D354-357.

Müller J, Müller K (2004) TreeGraph: automated drawing of complex tree figures using an extensible tree description format.
. *Molecular Ecology Notes* **4:** 786-788.

Pruss BM, Campbell JW, Van Dyk TK, Zhu C, Kogan Y, Matsumura P (2003) FlhD/FlhC is a regulator of anaerobic respiration and the Entner-Doudoroff pathway through induction of the methyl-accepting chemotaxis protein Aer. *J Bacteriol* **185:** 534-543.

Schmidt HA, Strimmer K, Vingron M, von Haeseler A (2002) TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics* **18:** 502-504.

Swofford DL (2003) *PAUP*. Phylogenetic Analysis Using Parsimony.* Sunderland, Massachusetts: Sinauer Associates.

von Mering C, Huynen M, Jaeggi D, Schmidt S, Bork P, Snel B (2003) STRING: a database of predicted functional associations between proteins. *Nucleic Acids Res* **31:** 258-261.

von Mering C, Jensen LJ, Snel B, Hooper SD, Krupp M, Foglierini M, Jouffre N, Huynen MA, Bork P (2005) STRING: known and predicted protein-protein associations, integrated and transferred across organisms. *Nucleic Acids Res* **33:** D433-437.