Supplementary Information

- Supplementary results- Examples of annotation conflicts

- List of Supplementary Figures

Supplementary Figure 1 – Gene Ontology Tree of corresponding GO terms of STEP nomenclature.

Supplementary Figure 2 – Prediction tools and criteria scheme

Supplementary Figure 3 – Decision tree of the annotation process

Supplementary Figure 4 – Screenshot of the “more info” slide panel

Supplementary Figure 5 – Screenshot of Flagellum protein complex as it is drawn in the STEPdb web interface

- List of Supplementary Tables

Supplementary Table 1 - Mobile elements *E.coli* K-12 and BL21

a. Summary comparison
b. Detailed list of K-12 and BL21 mobile elements

Supplementary Table 2 - Analysis of *E.coli* K-12 pseudogenes

c. Summary comparison
d. Detailed list of K-12 pseudogenes

Supplementary Table 3 – Correlation between Gene Ontology terms and STEP sub-cellular topology nomenclature.

Supplementary Table 4 - Synopsis of previous protein annotation in *E.coli* K-12 and the newly proposed STEP annotation.

Supplementary Table 5 – Annotation Process **A.** Criteria **B.** Annotation path

Supplementary Table 6 – Manually curated *E.coli* K-12 proteins and protein complexes

a. List of literature references emerged from the manual annotation of K-12 proteins

b. Manually curated *E.coli* K-12 complexes

c. High-throughput studies used to identify proteins with verified sub-cellular topology

Supplementary Table 7 – Proteins of unknown sub-cellular localization

Supplementary Table 8 - Conflicts in current proteome databases concerning the annotation of sub-cellular topology of *E.coli* K-12 proteins

Supplementary Table 9 - Re-evaluation of integral IM proteins

Supplementary Table 10 - Common core proteome of *E.coli* Strains

Supplementary Results

Annotation Conflicts

Here we give some detailed examples of protein localizations that have been resolved:

a. Uniprot annotates proteins that are part of the flagellar filament (e.g. FliC and FliD) and hook (FlgK and FlgL) with the specified GO term ”flagellum” with no clear sub-cellular location (Supplementary Table 4B). STEPdb classifies them as OM proteins facing the extra-cellular space (Figure 1).

b. YdeS, a protein of unknown function, is homologous to FimF that is a structural element of type 1 fimbriae (1). FimF along with the major fimbrial subunit protein, FimA, and FimH are required for mannose-specific adhesion (2). Therefore, YdeS was also classified as a peripheral outer membrane protein facing the extra-cellular space in STEPdb (F4; Figure 1).

c. The proposed integral IM proteome (3) was re-evaluated making use of the last generation bioinformatics tools and/or existing experimental evidence. Specifically, 222 of 1108 proposed to be IM proteins have been re-assigned in STEPdb to other sub-cellular locations (Supplementary Table 9). Most of them had been predicted as bitopic IM proteins (e.g. DapB and IlvA) (4) but are no longer predicted as IM proteins by the latest bioinformatics tools and do not have obvious TMs. These proteins were re-classified as cytoplasmic. Four proteins (Pta, HofP, YbfL and YncH) had their N-terminal regions predicted as a TM in past version of TMHMM whereas in TMHMM v2.0 it is predicted as a signal peptide (3). This is an inherent problem of signal peptide and TM helices prediction due to some similarity of a TM region with the hydrophobic core (H-domain) of the signal peptide (5,6). These four proteins were re-classified as periplasmic.

d. 220 cytoplasmic proteins were re-classified as nucleoid-associated, based on experimental evidence (7-9).

e. Five proteins (RpoE, Psd, ArtP, GlnK and GlnQ) were registered in EchoLOCATION as cytoplasmic “experimental”. These are now proposed as PIM proteins in STEPdb due to strong evidence of their association with the inner membrane. RpoE apart from its role as a sigma factor (σE) is a PIM protein that senses miss-folding of periplasmic proteins by interacting with membrane protein RseA (10).

Psd is an interesting enzyme that is processed in a post-translational event into the heterodimer. This enzyme is unique among the known pyruvoyl-dependent decarboxylases because it is membrane-associated; this has been shown by both proteomic (11) and biochemical processes (12-14).

ArtP is the ATP carrier protein of arginine-uptake system (artPIQM); it has been detected in association with the inner membrane (11). GlnK binds to the membrane in an AmtB-dependent manner and acts as a negative regulator of the

transport activity of AmtB (15). Finally GlnQ has been isolated in association with the inner membrane in to proteomic studies (11,16)

f. Another functional class is the DNA-binding proteins that are localized on the nucleoid. STEPdb lists 264 nucleoid-localized proteins, amongst them seven sigma factors (RpoD, N, S, H, F, E and FecI) and other transcription factors that have been identified via genomic searches for nucleic-acid protein interactions (7).

De novo **discovered experimentally verified proteins**

Here we give some examples of newly identified proteins with validated topologies:

a. YbgT and YneM are small membrane proteins that have been found to co-fractionate with the inner membrane and have their orientation on the IM identified experimentally (17).

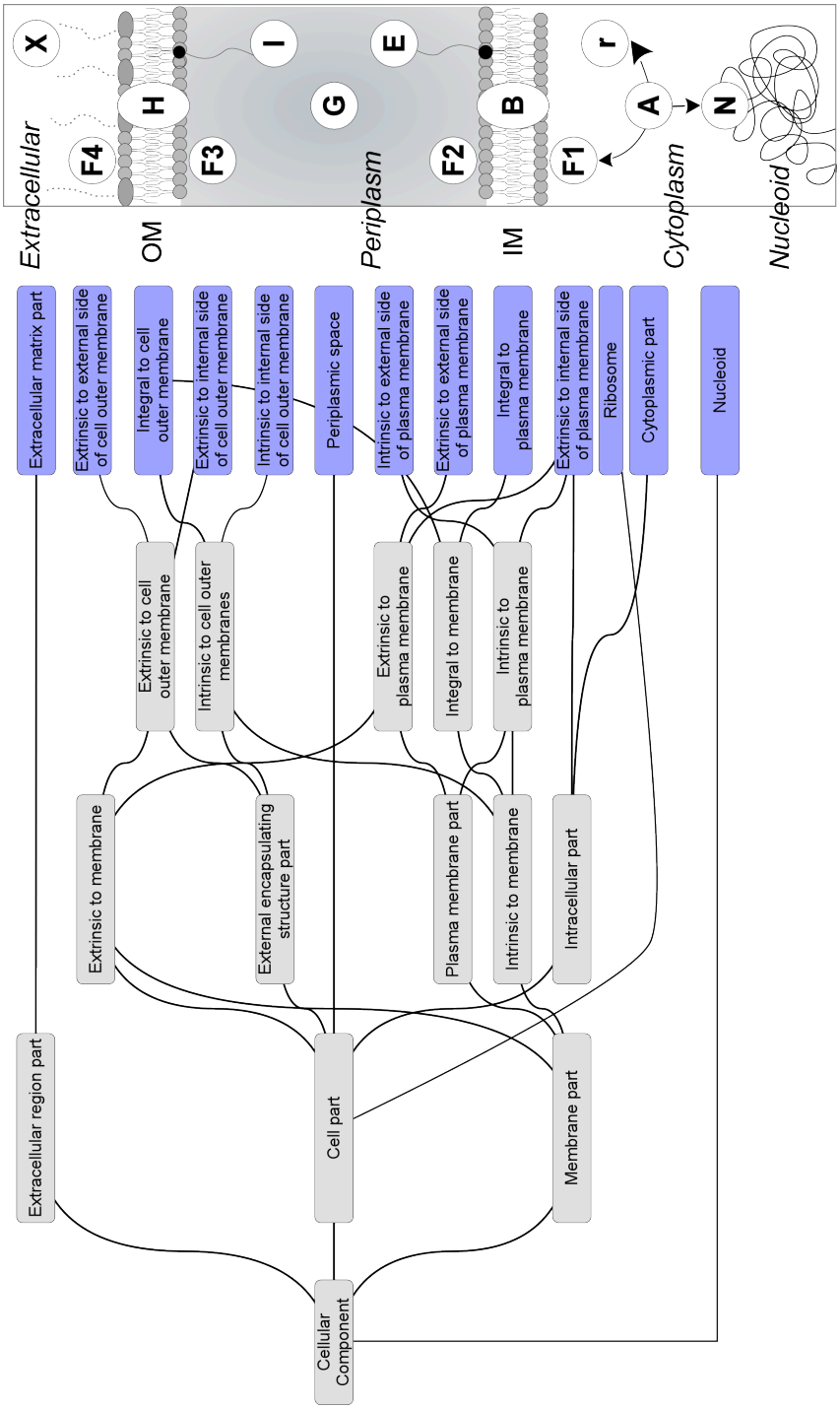
b. SodM and SodF are E.coli proteins that are close homologues of the RISodA (Rhizobium leguminosarum) that lack any known signal peptide motif. RISodA in is targeted to the periplasm in a SecA-dependent manner by a novel mechanism (18). Therefore by similarity these proteins are annotated as periplasmic proteins

c. YagT protein contains a 49 amino acid Tat leader peptide that allows the export of the active heterotrimer to the periplasm. Tat substrates are matured in the cytoplasm prior to their translocation (19,20).

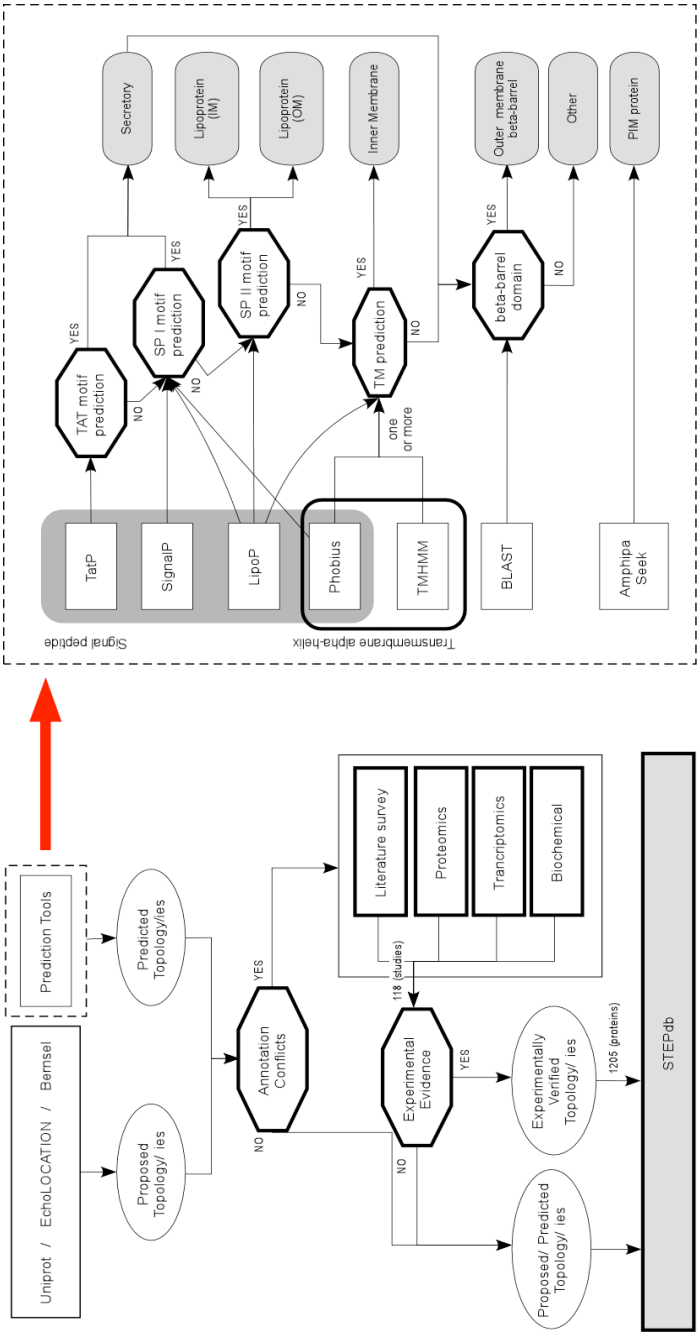
d. PyrH is a PIM protein that has been shown via fluorescence microscopy to predominantly localized near the bacterial membranes of different cells undergoing septation and division (21,22).

e. **Supplementary Figures**

Supplementary Figure 1 – Gene Ontology Tree of corresponding GO terms of STEP nomenclature. This is a schematic representation of the GO tree of only the subset of GO terms that correspond to STEP topological categories. A cartoon of a bacterial cell is drawn alongside the GO tree. Here the sub-cellular compartments are in a bottom up order from the cytoplasm to the extra-cellular space.



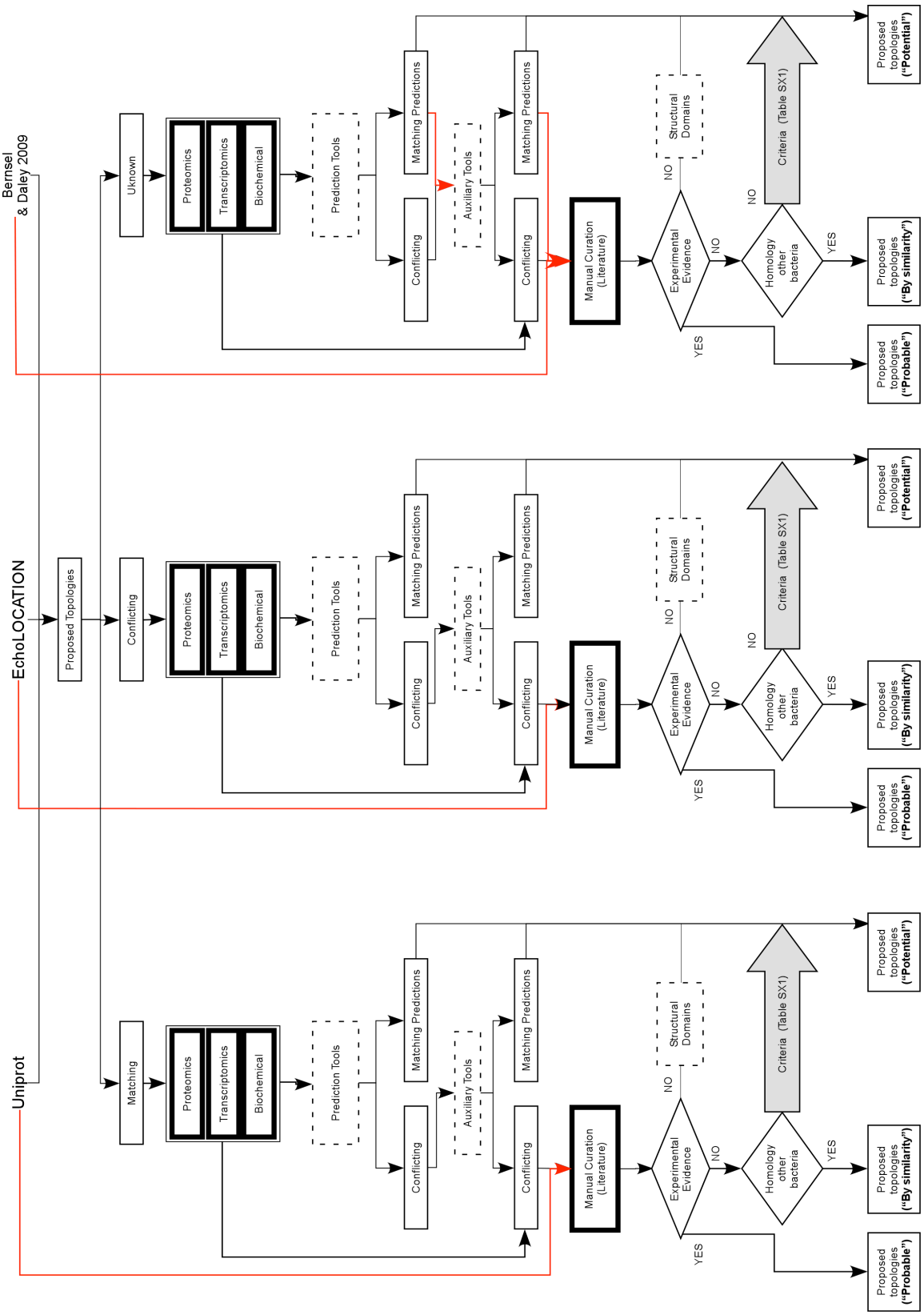
Supplementary Figure 2 – Prediction tools and criteria scheme
Here we describe in more detail the criteria we used in order to combine the predictions of the bioinformatics tools. Some of the bioinformatics tools that we utilized have overlapping prediction abilities. For example three tools, TMHMM, Phobius and LipoP can predict transmembrane regions. Tools that predict transmembrane alpha-helices lead to direct assignment to the IM protein class. However the prediction of a Sec signal can only lead to the assumption that the protein is localized somewhere across the IM (periplasm, in the OM or even secreted fully from the cell). To further clarify the unique sub-cellular localization we relied on experimental data (trancriptomic, proteomic and biochemical).



Supplementary Figure 3 – Decision tree of the annotation process

In this flowchart we summarize the steps that we followed for the topological annotation of *E.coli* K-12 proteome. Three primary sources (Uniprot, EchoLOCATION, theoretical IM proteome (3)) were combined and the proposed sub-cellular topologies were compared. Proteins were separated into three cases "Matching", "Conflicting", "Uknown". For proteins that were annotated as experimentally verified by Uniprot (23) or EcholOCATION (24) we accepted the proposed topologies. To *de novo* annotate the "Uknown" and resolve the "Conflicting" we combined bioinformatic prediction tools, high throughput proteomic, genomic and biochemical data and manual literature survey.

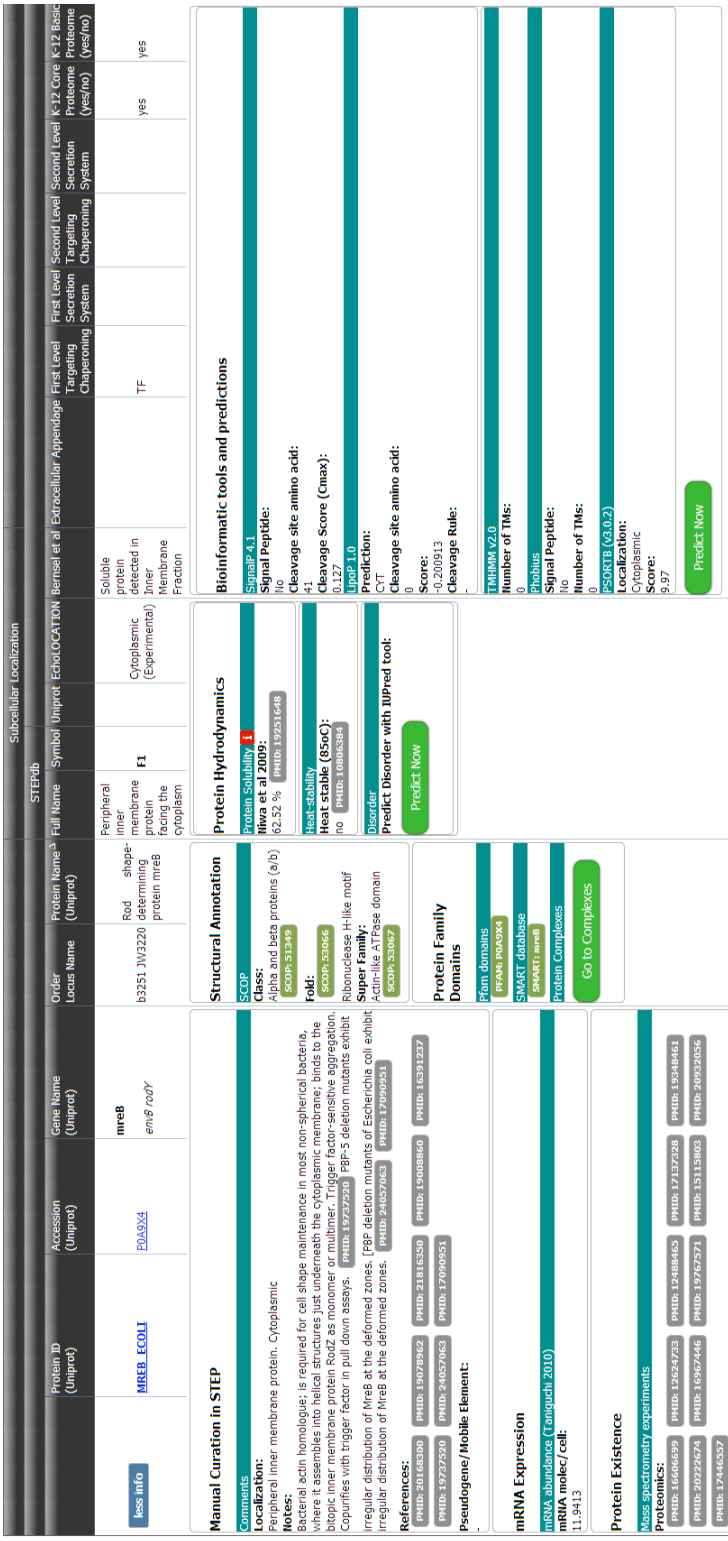
131


Bernsel & Daley 2009
EchoLOCATION
Uniprot
Uknown
Matching
Proposed Topologies
Conflicting
Proteomics
Transcriptomics
Biochemical
Prediction Tools
Auxiliary Tools
Matching Predictions
Conflicting
Manual Curation (Literature)
Experimental Evidence
Structural Domains
Homology other bacteria
Criteria (Table SX1)
NO
YES
Proposed topologies ("Potential")
Proposed topologies ("By similarity")
Proposed topologies ("Probable")

132
133

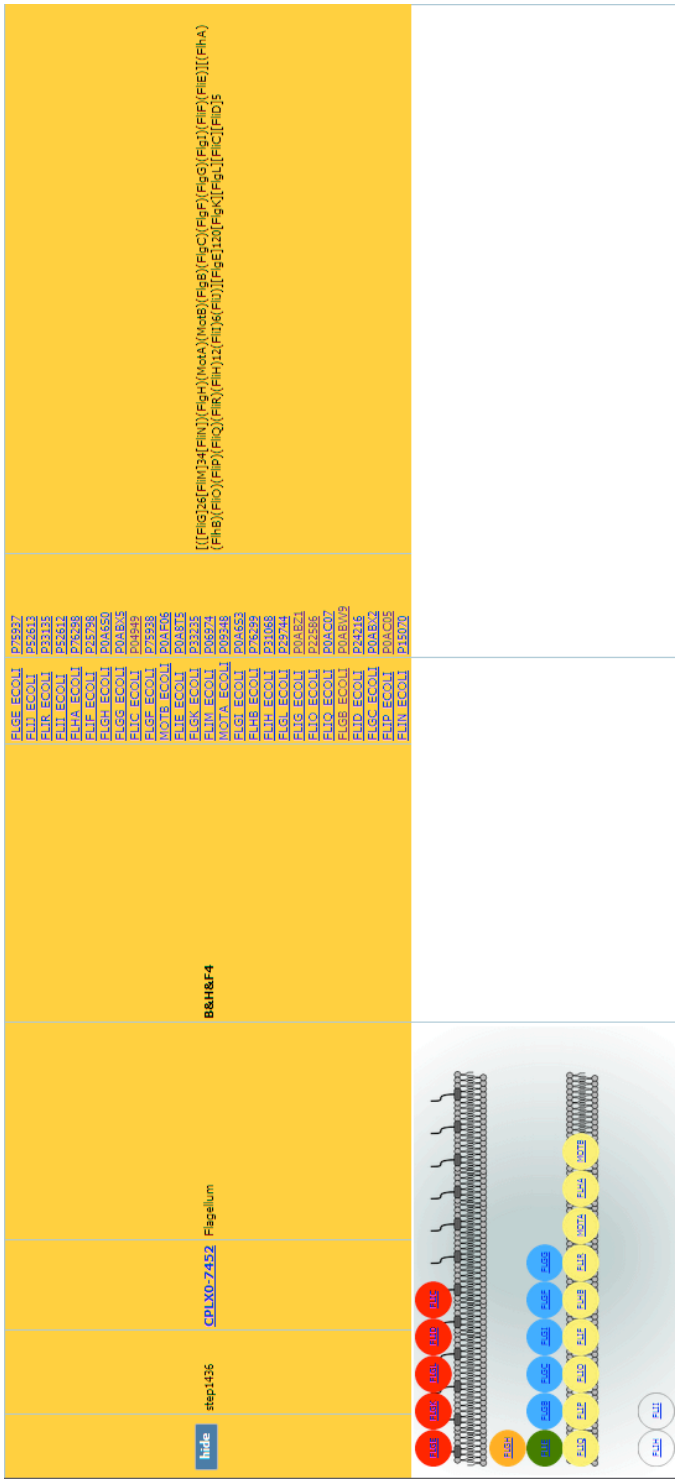
Supplementary Figure 4 – Screenshot of the “more info” slide panel

This is a screenshot of the information incorporated in the “more info” sliding panel accompanying each protein in the web-interface.



Supplementary Figure 5 – Screenshot of Flagellum protein complex as it is drawn in the STEPdb web interface

Protein complexes can be drawn dynamically in STEPdb web interface. This screenshot shows an example of a macromolecular complex, the flagellum, which consists of various types of proteins from cytoplasmic to surface appendages.



144

Supplementary Tables

Supplementary Table 1 - Mobile elements *E.coli* K-12 and BL21
a. Summary comparison: A comparison between the two *E. coli* strains regarding their proteomes, mobile elements and pseudogenes. Mobile elements include prophage integrases, transposases, insertion and Rhs elements (for further explanation, refer to the actual table). A set of 381 genes (out of a total of 624) located within gene islands that upon deletion have no effect in the growth of the bacterium in LB medium (25) are also considered as mobile elements in this study. The *E.coli* "core proteome", as defined by proteome comparison between 43 *E.coli* strains (Supplementary Table 10) is also shown here. A similar comparison was performed for the K-12 and BL21(DE3) proteomes were 4037 and 4483 proteins correspondingly were found to belong to the "common proteome" between the two strains. In both strains there were proteins with multiple homologs, 464 and 52 *E.coli* K-12 and BL21 proteins correspondingly were matched to more than one homologue (data not shown in table).
b. Detailed list of *E. coli* K-12 and BL21(DE3) mobile elements.

Supplementary Table 2 - Analysis of *E.coli* K-12 pseudogenes

Summary comparison:

a. Breakdown table of possible pseudogenes as reported in Uniprot , EcoGene (26) and by Ochman (27). The potential pseudogenes have been differentiated according to their protein existence annotation in Uniprot (evidence at protein and transcript level, predicted, inferred from homology and uncertain).

b. Detailed list of K-12 pseudogenes. Pseudogenes marked with purple have been reported to synthesize protein (28) indicating that these are functional genes.

Supplementary Table 3 – Correlation between GO terms and STEP nomenclature.

STEP classifies proteins into thirteen categories which correspond to different sub-cellular localizations. This table maps STEP categories to the most closely related gene ontology terms. GO name and ID are provided (29).

Supplementary Table 4 - Synopsis of previously existing protein annotation and newly proposed STEP annotation.

a. STEPdb and summary of the sub-cellular annotations of the two *E.coli* strains **b,c.** Existing topological annotations of Uniprot (23) and EchoLOCATION (24) are summarized in terms of STEPdb nomenclature. Some of the existing Uniprot or EchoLOCATION annotations could not be discriminated in order to be assigned to a distinct STEPdb class. **d.** theoretical IM proteome as defined in Bernsel and Daley (3). In total, 1108 out of 1133 proteins where mapped to the

K-12 reference proteome maintained in Uniprot. **e.** List of predicted proteins in EchoLOCATION. 4243 of the 4345 listed proteins were mapped to the reference *E.coli* proteome. The remaining proteins are unknown coding sequences, pseudogenes not existing or deleted from Uniprot and in some case duplicates of the same protein.

Supplementary Table 5 - Annotation Criteria

List of formalized rules that describe the way we handled the contradictions among the bioinformatic tools. These criteria were applied only at the last step of the annotation process and if during manual curation step no indication of experimental evidence was discovered.

Supplementary Table 6 – Manually curated *E.coli* K-12 proteins and protein complexes

a. Inventory of 1205 experimentally verified proteins that have been discovered during the manual curation process. Each verified protein is accompanied by the corresponding list of pmids that justified the proposed location in STEPdb. The proposed topology is associated with a characterization of the level of evidence. There are two levels of evidence: “Experimental” and “Probable”. The term “Experimental” is used in cases of strong experimental data whereas the term “Probable” indicates that there exists some experimental evidence.

b. Manually curated *E.coli* K-12 complexes

List of 61 protein complexes discovered during manual curation process.

c. High-throughput studies used to identify proteins with verified sub-cellular topology

List of high throughput studies and the corresponding experimentally identified proteins. Among these studies: 11 proteomic studies coupled with sub-fractionation methods, two genomic studies, two biochemical analysis and one fluorescent microscopy study.

Supplementary Table 7 – Proteins of unknown sub-cellular localization

List of 36 proteins that were previously of unknown sub-cellular topology, based on the three resources (Uniprot, EchoLOCATION, Bernsel (10)). Proteins that their sub-cellular location has been experimentally confirmed are highlighted with light yellow. The remaining proteins were assigned to sub-cellular class based on the prediction of the tools.

Supplementary Table 8 - Conflicts in current proteome databases concerning the annotation of sub-cellular topology of *E.coli* proteins

Topological annotation conflicts initially existed even between the two databases, Uniprot and EchoLOCATION and/or the experimental set of IM proteins (3,23,24). This table lists 601 proteins with conflicting topologies among the three resources (Uniprot, EchoLOCATION, Bernsel and Daley, 2009) exhibited contradicting sub-cellular topologies.

Supplementary Table 9 - Re-evaluation of IM proteins

This table lists the predicted IM proteins that have been proposed by Bernsel and Daley (2009) that STEP re-classifies them as non IM proteins. This reconsideration of topology was based on next generation prediction tools: TMHMM v2.0, Phobius, LipoP, SignalP4.0 (5,30-32), current annotation in Uniprot and EchoLOCATION (23,24) and existing experimental data found in literature.

Supplementary Table 10 - Common core *E.coli* Strains

Detailed names of the 43 *E.coli* strains used to define the core proteome (see also “experimental procedures”).

References

1. Waksman, G., Hultgren, S.J. (2009) Structural biology of the chaperone-usher pathway of pilus biogenesis. *Nat Rev Microbiol*, **7**, 765-774.
2. Klemm, P., Christiansen, G. (1987) Three fim genes required for the regulation of length and mediation of adhesion of Escherichia coli type 1 fimbriae. *Mol Gen Genet*, **208**, 439-445.
3. Bernsel, A., Daley, D.O. (2009) Exploring the inner membrane proteome of Escherichia coli: which proteins are eluding detection and why? *Trends Microbiol*, **17**, 444-449.
4. Bernsel, A., Viklund, H., Falk, J.*, et al.* (2008) Prediction of membrane-protein topology from first principles. *Proc Natl Acad Sci U S A*, **105**, 7177-7181.
5. Kall, L., Krogh, A., Sonnhammer, E.L. (2007) Advantages of combined transmembrane topology and signal peptide prediction--the Phobius web server. *Nucleic Acids Res*, **35**, W429-432.
6. Kall, L., Krogh, A., Sonnhammer, E.L. (2004) A combined transmembrane topology and signal peptide prediction method. *J Mol Biol*, **338**, 1027-1036.
7. Ishihama, A. (2012) Prokaryotic genome regulation: a revolutionary paradigm. *Proc Jpn Acad Ser B Phys Biol Sci*, **88**, 485-508.
8. Metzner, M., Germer, J., Hengge, R. (2004) Multiple stress signal integration in the regulation of the complex sigma S-dependent csiD-ygaF-gabDTP operon in Escherichia coli. *Mol Microbiol*, **51**, 799-811.
9. Liu, X., Matsumura, P. (1995) An alternative sigma factor controls transcription of flagellar class-III operons in Escherichia coli: gene sequence, overproduction, purification and characterization. *Gene*, **164**, 81-84.
10. Collinet, B., Yuzawa, H., Chen, T.*, et al.* (2000) RseB binding to the periplasmic domain of RseA modulates the RseA:sigmaE interaction in the cytoplasm and the availability of sigmaE.RNA polymerase. *J Biol Chem*, **275**, 33898-33904.
11. Papanastasiou, M., Orfanoudaki, G., Koukaki, M.*, et al.* (2013) The Escherichia coli peripheral inner membrane proteome. *Mol Cell Proteomics*, **12**, 599-610.
12. Li, Q.X., Dowhan, W. (1988) Structural characterization of Escherichia coli phosphatidylserine decarboxylase. *J Biol Chem*, **263**, 11516-11522.
13. Dowhan, W., Li, Q.X. (1992) Phosphatidylserine decarboxylase from Escherichia coli. *Methods Enzymol*, **209**, 348-359.
14. Tyhach, R.J., Hawrot, E., Satre, M.*, et al.* (1979) Increased synthesis of phosphatidylserine decarboxylase in a strain of Escherichia coli bearing a hybrid plasmid. Altered association of enzyme with the membrane. *J Biol Chem*, **254**, 627-633.

15. Coutts, G., Thomas, G., Blakey, D.*, et al.* (2002) Membrane sequestration of the signal transduction protein GlnK by the ammonium transporter AmtB. *EMBO J*, **21**, 536-545.
16. Stenberg, F., Chovanec, P., Maslen, S.L.*, et al.* (2005) Protein complexes of the Escherichia coli cell envelope. *J Biol Chem*, **280**, 34409-34419.
17. Fontaine, F., Fuchs, R.T., Storz, G. (2011) Membrane localization of small proteins in Escherichia coli. *J Biol Chem*, **286**, 32464-32474.
18. Krehenbrink, M., Edwards, A., Downie, J.A. (2011) The superoxide dismutase SodA is targeted to the periplasm in a SecA-dependent manner by a novel mechanism. *Mol Microbiol*, **82**, 164-179.
19. Tullman-Ercek, D., DeLisa, M.P., Kawarasaki, Y.*, et al.* (2007) Export pathway selectivity of Escherichia coli twin arginine translocation signal peptides. *J Biol Chem*, **282**, 8309-8316.
20. Neumann, M., Mittelstadt, G., Iobbi-Nivol, C.*, et al.* (2009) A periplasmic aldehyde oxidoreductase represents the first molybdopterin cytosine dinucleotide cofactor containing molybdo-flavoenzyme from Escherichia coli. *FEBS J*, **276**, 2762-2774.
21. Landais, S., Gounon, P., Laurent-Winter, C.*, et al.* (1999) Immunochemical analysis of UMP kinase from Escherichia coli. *J Bacteriol*, **181**, 833-840.
22. Watt, R.M., Wang, J., Leong, M.*, et al.* (2007) Visualizing the proteome of Escherichia coli: an efficient and versatile method for labeling chromosomal coding DNA sequences (CDSs) with fluorescent protein genes. *Nucleic Acids Res*, **35**, e37.
23. Dimmer, E.C., Huntley, R.P., Alam-Faruque, Y.*, et al.* (2012) The UniProt-GO Annotation database in 2011. *Nucleic Acids Res*, **40**, D565-570.
24. Horler, R.S., Butcher, A., Papangelopoulos, N.*, et al.* (2009) EchoLOCATION: an in silico analysis of the subcellular locations of Escherichia coli proteins and comparison with experimentally derived locations. *Bioinformatics*, **25**, 163-166.
25. Pósfai, G., Plunkett, G., 3rd, Feher, T.*, et al.* (2006) Emergent properties of reduced-genome Escherichia coli. *Science*, **312**, 1044-1046.
26. Rudd , K.E. (2000) EcoGene: a genome sequence database for *Escherichia coli* K-12. *Nucleic Acids Res*, **28**, 60-64.
27. Ochman, H., Davalos, L.M. (2006) The nature and dynamics of bacterial genomes. *Science*, **311**, 1730-1733.
28. Iwasaki, M., Miwa, S., Ikegami, T.*, et al.* (2010) One-Dimensional Capillary Liquid Chromatographic Separation Coupled with Tandem Mass Spectrometry Unveils the *Escherichia coli* Proteome on a Microarray Scale. *Analytical Chemistry*, **82**, 2616-2620.
29. Ashburner, M., Ball, C.A., Blake, J.A.*, et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*, **25**, 25-29.

30. Krogh, A., Larsson, B., von Heijne, G.*, et al.* (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J Mol Biol*, **305**, 567-580.
31. Juncker, A.S., Willenbrock, H., Von Heijne, G.*, et al.* (2003) Prediction of lipoprotein signal peptides in Gram-negative bacteria. *Protein Sci*, **12**, 1652-1662.
32. Petersen, T.N., Brunak, S., von Heijne, G.*, et al.* (2011) SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nat Methods*, **8**, 785-786.