Explanatory Notes for Supplementary Annotation Table.

Each field is populated with information either generated or collected by the workshop participants. Multiple entries in a field are separated by the symbol "-!-" to assist in parsing. The term "predicted" was included for gene products whose function not was experimentally verified.

Gene Nomenclature and Identifiers

Genes are identified by their feature (CDS or RNA). Pseudogenes are identified in the comment fields for the two strains. Some pseudogenes are fragmented by inserts, others are frameshifts. Gene names are given both in conventional Demerec format ((1); Gene column)) and in a format not restricted to the Demerec format (Locus Name column). Locus names for 2- and 3-part pseudogene fragments are given extensions of "_1", "_2" numbering from the N-terminal to the C-terminal end. Multiple copies of IS proteins are given locus names with "-1", "-2" extensions based on their location on the chromosome relative to the first instance of the specific IS protein. Synonyms of gene names found in the literature and collected from several database sources are provided.

Loci of MG1655 and W3110 are described in terms of their gene boundaries (left end, right end) and direction of transcription (clockwise (+) or counterclockwise (-)). The boundary is defined as the nucleotide number of the start/end codon of a transcript, pseudogene (fragment) or functional RNA. The start and end positions between MG1655 and W3110 differ due to a difference in the start position and inversion, insertion or deletion of regions. Locustags specific to MG1655 (b numbers) and W3110 (JW numbers) are listed. For MG1655 locustags have been assigned to 21 entities representing fused pseudogene fragments (ancestral version of the gene). Locustags were not assigned for the fused pseudogene fragments in W3110.

ECK (*E. coli* K-12) numbers are identifiers assigned to *E. coli* K-12 genes by the workshop participants. ECK numbers are given to unique CDSs, RNAs and pseudogenes. Individual fragments of divided pseudogenes are given the same ECK identifier. The fused pseudogene fragments were assigned the

same ECK identifier as the corresponding fragments. One ECK number is used for multiple copies of an IS protein, resulting in a one to many mapping for these CDSs. This 'one to many' nomenclature is limited to mobile elements and does not include ribosomal RNA genes. The ECK identifiers are numbered sequentially in the order of the MG1655 map beginning with *thrL*.

Gene Product Type.

Assignment of the type of gene product was attempted. Clearly the major types of proteins in *E. coli* are enzymes followed by transporters and regulators. To tally the relative proportions that occupy the genome, gene products were labeled according to type. Assignments are often difficult because of the complexity of biology. A few new categories were added (see Table 2 for complete list), but the most difficult assignments concerned gene products that could be described correctly with more than one function. Examples of such complexities include the phosphotransferases of PTS system (enzymes or transporters), the sigma factors (factors or regulators), DNA polymerase (enzyme or cell process protein), and flagellae (structure of the cell or cell process protein). Complex enzyme subunits such as the four types in the succinate dehydrogenase enzyme could all be labeled as the dehydrogenase enzyme, as is current practice. On the other hand each subunit can be labeled more accurately according to their essential character, such as for instance an inner membrane subunit or an electron transport subunit (carrier). Carrying these properties forward would be more useful to annotation of unknown genes in other organisms. We have made an effort to make assignments that reflect the nature of the individual gene product when it is a part of a larger complex.

Gene Product Descriptions, Comments, Evidence

The assignment of gene product description occupied a major fraction of time and effort by the workshop participants. Starting with existing descriptions from databases and web sites offering full genome

predictions, groups of 1, 2 and 3 participants reexamined data, checked for new information in the literature, new sequence matches, and new kinds of sequence analysis. Web sites new and old were consulted. Whether the product description was derived experimentally or by computation was noted as a measure of the reliability of the assignment. The gene product descriptions were kept succinct. Additional remarks were lodged in the associated comment field.

An attempt was made towards describing the gene products in a uniform format. Enzymes were described by their common name and information on cofactor requirement was included where available. Enzyme complexes were described by the name of the enzyme complex followed by the name of the subunit itself (b0784, MoaD; molybdopterin synthase, small subunit). Some enzymes encode multiple functions either as a result of gene fusion events or as a result of multiple activities encoded at the same site of the protein. The term "fused" was included for the fused proteins and their activities were listed using "-!-" to separate the activities (b0002, ThrA; fused aspartokinase I -!- homoserine dehydrogenase I). The other enzymes with more than one functions were listed as bifunctional (b0025, RibF; bifunctional riboflavin kinase -!- FAD synthetase) or as multifunctional (b0494, TesA; multifunctional acyl-CoA thioesterase I -!- protease I -!- lysophospholipase L1). Transport proteins were listed with the substrate transported (b0336, CodB; cytosine transporter). For the ABC superfamily transport complexes the substrate, complex, and subunit information was listed (b0199, MetN, DL-methionine transporter subunit -!- ATP-binding component of ABC superfamily). Transcriptional regulators were described either as DNA-binding transcriptional, repressor, activator, dual regulator (may act as both activator and regulator), or regulator (not known whether repressor or activator). A uniform format was given to the two component regulatory systems for the response regulators (b0620, CitB; DNA-binding response regulator in two-component regulatory system with CitA), sensory histidine kinases (b0619, CitA; sensory histidine kinase in two-

component regulatory system with CitB) and for the fused two-component regulators (b2218, RcsC; hybrid sensory kinase in two-component system with RcsB and YojN).

For genes encoded in the 10 cryptic prophages or prophage-like elements (9 in W3110 due to lack of CPZ-55), the name of the prophage was listed following the name of the gene product (b0246, YafW; CP4-6 prophage; antitoxin of the YkfI-YafW toxin-antitoxin system). Gene product descriptions for the t-RNAs included information on their anticodon (b0536, ArgU; tRNA-Arg(UCU) (Arginine tRNA4)). For pseudogenes, the term "(pseudogene)" was included in the gene product description as well as "N-ter fragment", "middle fragment" or "C-ter fragment" for fragmented pseudogenes.

Gene product predictions were based on the data collected from several *E. coli* databases and from specialized databases listed in the text (i.e., transmembrane helix predictions, protein family and protein domain predictions, sequence similar homologs, etc.). Gene products whose functions not could be predicted were either annotated as conserved proteins (had homologs beyond *Escherichia* and *Salmonella*) or predicted proteins (did not have homologs outside of *Escherichia* and *Salmonella*).

Literature

Literature given is an incomplete collection derived from GenProtEC and from the Cyber Cell Database.

Cell Location

Locations of the gene products were individually determined through careful evaluation of the literature; transmembrane helix predictions, HMMTOP(2) and TMHMM (3); signal peptide predictions SignalP (4) and LipoP (5) and have been taken from the EchoLocation section of EchoBASE (6). The cell location data were translated into Gene Ontology (GO) terms (7) and are presented in the GO cellular component field.

Context

Names of IS elements and prophages are listed for the loci that belong to these elements.

Enzyme Nomenclature.

EC numbers were collected for the *E. coli* enzymes from EcoCyc (8), GenProtEC (9), BRENDA (10), and from the literature. The IUBMB Enzyme Nomenclature database (http://www.chem.qmw.ac.uk/iubmb/) was consulted for the assigned EC numbers.

Cofactor, Protein Complexes.

Information on cofactors used by enzymes were collected from EcoCyc (8). EcoCyc also provided data on protein complexes (homomultimers and heteromultimers) for over 950 proteins. The name of the complex and its components are provided.

Transporter Classification.

Information on the transport proteins were collected from the Transport Classification Database (http://www.tcdb.org/). Both the Transport Classification (TC) number and Superfamily membership are given.

Regulator Family, Transcriptional Units Regulated.

Data on transcriptional regulators were from RegulonDB (11) and J. Collado-Vides and Heladia Saldago (personal communications). The family membership of regulators and transcriptional units controlled by the regulators are given.

Proteases

Identification of peptide bond hydrolysis characteristics in proteins has allowed prediction of proteases (peptidases). Information on known and predicted proteases of *E. coli* K-12 proteins has been extracted from the MEROPS database (12).

Signal Peptides, Membrane Helices, C-terminus location.

The amino acids predicted to encode the signal peptide according to SignalP(4) are listed. In addition, literature based signal peptide cleavage sites collected from EcoGene (13) are presented. The predicted number of transmembrane helices are provided based on the two algorithms, HMMTOP(2) and TMHMM (3).

Location of the C-terminal end of transmembrane proteins, either outside in the periplasm or inside in the cytoplasm are based on experimental methods (14).

Attenuation Regulation

Information on regulation by transcription-attenuation was included (15). Operons are predicted to be regulated by attenuation based on the presence of possible stem and loop RNA structures in advance of the first gene of the operon. The Attenuation field contains information for the first gene of the operon believed to be regulated by attenuation, and the set of genes in the operon is listed (b0463: AceB; regulated by attenuation (*aceB-aceA*).

Fused Proteins

Fused proteins are encoded by genes which have undergone a gene fusion event. The resulting gene product encodes two or more functions in separate regions of the protein. Such proteins are known to contribute to errors in annotation when alignment regions are not considered for transfer of functions between homologous sequences. The 108 fused *E.* coli proteins(9) are listed with functions and location of functions separated by "-!-".

Structure

Structure data for *E. coli* proteins are presented in the form of PDB IDs from the Protein Data Bank (16).

COG assignments

Membership of proteins in COGs (Clusters of Orthologous Genes; (17)) is presented by the COG IDs and their annotations. Some *E. coli* proteins contain more than one COG. These were provided directly by E.V. Koonin as more than one per gene cannot easily be retrieved from the NCBI Web site.

Superfamily (SCOP domain) assignments

SCOP superfamily domains identify structural elements in protein sequences some of which have known function that can help characterize otherwise unknown proteins. The presence of structural domains based

on similarity to known SCOP superfamily domains are shown in the table. Information on structural domains was obtained from the Superfamily database (18) and is listed with superfamily ID and domain annotation.

Pfam assignments

Information on Pfam assignments for the *E. coli* proteins was obtained from the Pfam database (19). Pfam represents a large collection of multiple sequence alignments and Hidden Markov Models for many common protein domains and families. The Pfam IDs, annotations, e-value, and amino acid range are shown.

TIGRFAM assignments

Membership in TIGRFAM protein families were obtained from the TIGRFAM database (20). TIGRFAMs are curated protein families developed for use in annotation.

GO assignments

Data on cellular component were obtained by translating data from the "Cell Location" field to GO terminology. GO assignments for the cellular process and molecular function levels were obtained by transferring MultiFun cellrole/pathway assignments (9) to Gene Ontology terminology (21). The mapping can be obtained at: http://geneontology.org/external2go/multifun2go. Current MultiFun assignments are present at GenProtEC (http://genprotec.mbl.edu/).

References

1. Demerec,M., Adelberg,E.A., Clark,A.J. and Hartman,P.E. (1966) A proposal for a uniform nomenclature in bacterial genetics. *Genetics*, **54**, 61-76.

2. Tusnady,G.E. and Simon,I. (2001) The HMMTOP transmembrane topology prediction server. *Bioinformatics*, **17**, 849-850.

3. Krogh,A., Larsson,B., Von,H.G. and Sonnhammer,E.L. (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.*, **305**, 567-580.

4. Bendtsen,J.D., Nielsen,H., Von Heijne,G. and Brunak,S. (2004) Improved prediction of signal peptides: SignalP 3.0. *J Mol Biol.*, **340**, 783-795.

5. Juncker,A.S., Willenbrock,H., Von,H.G., Brunak,S., Nielsen,H. and Krogh,A. (2003) Prediction of lipoprotein signal peptides in Gram-negative bacteria. *Protein Sci*, **12**, 1652-1662.

6. Misra,R.V., Horler,R.S., Reindl,W., Goryanin,I.I. and Thomas,G.H. (2005) EchoBASE: an integrated post-genomic database for Escherichia coli. *Nucleic Acids Res.*, **33**, D329-D333.

7. Fujimoto,S. and Clewell,D.B. (1998) Regulation of the pAD1 sex pheromone response of *Enterococcus faecalis* by direct interaction between the cAD1 peptide mating signal and the negatively regulating, DNA-binding TraA protein. *Proc. Natl. Acad. Sci. USA*, **95**, 6430-6435.

8. Keseler,I.M., Collado-Vides,J., Gama-Castro,S., Ingraham,J., Paley,S., Paulsen,I.T., Peralta-Gil,M. and Karp,P.D. (2005) EcoCyc: a comprehensive database resource for *Escherichia coli*. *Nucleic Acids Res.*, **33**, D334-D337.

9. Serres,M.H., Goswami,S. and Riley,M. (2004) GenProtEC: an updated and improved analysis of functions of *Escherichia coli* K-12 proteins. *Nucleic Acids Res.*, **32**, D300-D302.

10. Schomburg,I., Chang,A., Ebeling,C., Gremse,M., Heldt,C., Huhn,G. and Schomburg,D. (2004) BRENDA, the enzyme database: updates and major new developments. *Nucleic Acids Res.*, **32**, D431-D433.

11. Salgado,H., Gama-Castro,S., Martinez-Antonio,A., Diaz-Peredo,E., Sanchez-Solano,F., Peralta-Gil,M., Garcia-Alonso,D., Jimenez-Jacinto,V., Santos-Zavaleta,A., Bonavides-Martinez,C. *et al.* (2004) RegulonDB (version 4.0): transcriptional regulation, operon organization and growth conditions in *Escherichia coli* K-12. *Nucleic Acids Res.*, **32**, D303-D306.

12. Rawlings,N.D., Tolle,D.P. and Barrett,A.J. (2004) MEROPS: the peptidase database. *Nucleic Acids Res.*, **32**, D160-D164.

13. Rudd,K.E. (2000) EcoGene: a genome sequence database for *Escherichia coli* K-12. *Nucleic Acids Res.*, **28**, 60-64.

14. Daley,D.O., Rapp,M., Granseth,E., Melen,K., Drew,D. and Von,H.G. (2005) Global topology analysis of the *Escherichia coli* inner membrane proteome. *Science*, **308**, 1321-1323.

15. Merino,E. and Yanofsky,C. (2005) Transcription attenuation: a highly conserved regulatory strategy used by bacteria. *Trends Genet.*, **21**, 260-264.

16. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235-242.

17. Tatusov,R.L., Fedorova,N.D., Jackson,J.D., Jacobs,A.R., Kiryutin,B., Koonin,E.V., Krylov,D.M., Mazumder,R., Mekhedov,S.L., Nikolskaya,A.N. *et al.* (2003) The COG database: an updated version includes eukaryotes. *BMC. Bioinformatics*, **4**, 41.

18. Madera,M., Vogel,C., Kummerfeld,S.K., Chothia,C. and Gough,J. (2004) The SUPERFAMILY database in 2004: additions and improvements. *Nucleic Acids Res.*, **32**, D235-D239.

19. Bateman,A., Coin,L., Durbin,R., Finn,R.D., Hollich,V., Griffiths-Jones,S., Khanna,A., Marshall,M., Moxon,S., Sonnhammer,E.L. *et al.* (2004) The Pfam protein families database. *Nucleic Acids Res.*, **32**, D138-D141.

20. Haft,D.H., Selengut,J.D. and White,O. (2003) The TIGRFAMs database of protein families. *Nucleic Acids Res.*, **31**, 371-373.

21. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25-29.