ILL Document Delivery

REG-14277567

MDUNMC

NLM -- W1 MO196N (Gen); E-Journal w/ILL access

Mr. Andrew Michaelson
41 Dudley Ct.
Bethesda, MD 20814

ATTN:
PHONE: 301-295-1184
FAX: 206-333-0380
E-MAIL: Andrew.Michaelson@usuhs.mil

SUBMITTED: 2007-11-29 17:15:53
PRINTED: 2007-11-30 09:46:08
REQUEST NO.: REG-14277567
SENT VIA: DOCLINE
DOCLINE NO.: 23818148

REG Copy Journal

TITLE:	MOLECULAR MICROBIOLOGY
PUBLISHER/PLACE:	Blackwell Scientific Publications Oxford, OX ; Boston, MA :
VOLUME/ISSUE/PAGES:	1997 Nov;26(3):417-22 417-22
DATE:	1997
AUTHOR OF ARTICLE:	Hinton JC
TITLE OF ARTICLE:	THE ESCHERICHIA COLI GENOME SEQUENCE: THE END OF A
ISSN:	0950-382X
OTHER NUMBERS/LETTERS:	Unique ID.: 8712028 23818148 9402013
SOURCE:	PubMed
MAX COST:	$12.00
COPYRIGHT COMP.:	Guidelines
CALL NUMBER:	W1 MO196N (Gen); E-Journal w/ILL access
REQUESTER INFO:	Michaelson, Mr. Andrew
DELIVERY:	E-mail: Andrew.Michaelson@usuhs.mil
REPLY:	Mail:

KEEP THIS RECEIPT TO RECONCILE WITH BILLING STATEMENT
For problems or questions, contact NLM at http://wwwcf.nlm.nih.gov/ill/ill_web_form.cfm or phone 301-496-5511.
Include LIBID and request number.

Molecular Microbiology (1997) 25(3), 417–422

MicroCommentary

The *Escherichia coli* genome sequence: the end of an era or the start of the FUN?

Jay C. D. Hinton
Nuffield Department of Clinical Biochemistry, University of Oxford, Institute of Molecular Medicine, Oxford OX3 9DS, UK.

Summary

Our dream of determining the entire *Escherichia coli* K12 genome sequence has been realized. This calls for new approaches for the analysis of gene expression and function in biology's best-understood organism. Comparison of the *E. coli* genome sequence with others will provide important taxonomic insights and have implications for the study of bacterial virulence. Approximately 20% of *E. coli* genes have been designated FUN genes, because they have no known function or homologies to sequence databases. FUN genes promise to have an exciting impact on bacterial research. The post-genome era requires novel strategies that address gene regulation at the level of the entire cell. These strategies need to supersede the reductionist approach to genetic analysis. Only then will the genome sequence lead us to an understanding of how a bacterial cell really works.

Introduction

The recent completion of several bacterial genome sequencing projects signals the end of many traditional strategies for the analysis of gene function and a beginning for exciting new technologies that can now be applied to basic questions concerning cellular function. In this article, I hope to put the importance of the *E. coli* genome sequence into context and to consider its most interesting findings.

Why study *E. coli*?

E. coli is a remarkable organism. It can use an impressive array of molecules as carbon and nitrogen sources and adapt rapidly when moved from benign to toxic environments. For example, infection of mammals by virulent *E. coli* strains often involves aerobic growth on nutrient-rich material at neutral pH, followed by rapid ingestion into the contents of the stomach at a pH of 2. *E. coli* can survive this assault and colonize the intestinal tract; some strains can go on to cause diarrhoea or even death.

Despite its amazing versatility, some regard *E. coli* as little more than a small bag of enzymes designed to propagate cloned DNA. In fact, research on *E. coli* has a long and varied history. It was originally chosen as a model system because of its ability to grow on chemically defined media and its rapid growth rate in the laboratory. In the 1940s and 1950s, the simplicity of genetic analyses and the ease of preparing enzymatically active cell extracts for biochemical analysis made *E. coli* the experimental model of choice. The concerted efforts of several generations of microbiologists, geneticists, biochemists and others have taught us more about *E. coli* and the closely related *Salmonella typhimurium* than about any other organisms on earth. Consequently, the *E. coli* genome sequence offers an unequalled opportunity to relate gene sequence to biology. Up to now, we have focused on what a bacterial cell needs to thrive in a few fairly specialized niches, with too much emphasis on LB agar at 37°C! Now, we need to explore more exotic environments and determine what this organism is really capable of.

The 4 639 221 bp that define *E. coli*

The *E. coli* genome project was the first to be proposed (Blattner, 1983) and has been the proving ground for large-scale sequencing technology. Completion of the sequence has taken 6 years, and the final annotated version was deposited at Genbank in January 1997. Blattner *et al.* (1997) describe the complete *E. coli* genome sequence, which comprises 4 639 221 bp of sequence containing 4288 putative protein-encoding genes. These genes can be classified into several functional groups (Fig. 1). Two new operons for the degradation of aromatic compounds, six new tRNA genes and three cryptic phages were found. The complete repertoire of repeated sequences has been defined, accounting for nearly 2% of the genome. These include new REP, ERIC and Box C sequences scattered throughout the genome. Almost 500 proteins from *E. coli* (11.6% of the total) have known three-dimensional structures or are homologous to proteins with known structures (http://pedant.mips.biochem.mpg.de/frishman/ecoli.

Received 1 September, 1997; accepted 16 September, 1997. E-mail hinton@icrf.icnet.uk; Tel. (01865) 222427; Fax (01865) 222431.

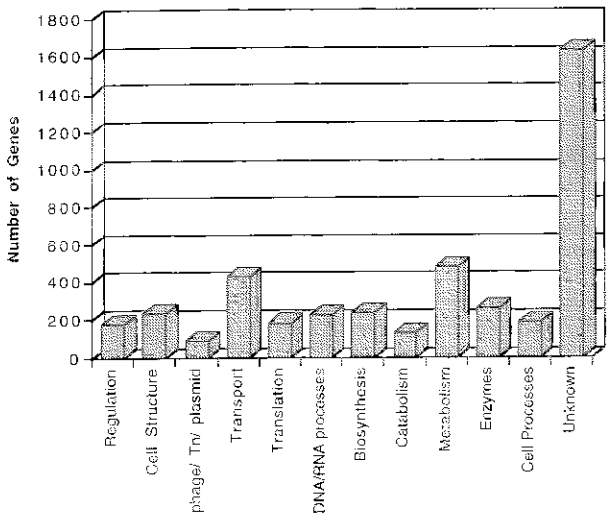

Fig. 1. The function of genes in *E. coli*. Genes have been assigned to functional categories, according to the data of Blattner *et al.* (1997). The number of unknown genes shown here may be overestimated, as discussed in the text.

html). Many genes of 'unknown function' were found for the first time, and these will be discussed below.

At the time of publication, the sequences of more than 10 organisms will be available for analysis through the World-Wide Web (www; Table 1), including the genomes of *Bacillus subtilis*, *Archaeoglobus fulgidus* and *Borrelia burgdorferi*, which have recently been completed. The partial sequences of the closely related *S. typhimurium* and *S. typhi* have been made available at http://genome.wustl.edu/gsc/bacterial/salmonella.html. Other incomplete genomes may be accessed via NCBI (National Center for Biotechnology Information; http://www.ncbi.nlm.nih.gov/) or TIGR (the Institute of Genome Research; http://www.tigr.org/).

Two www sites that have both been updated with the complete genome sequence are particularly useful for analysis of the *E. coli* genome. The excellent *E. coli* databank site (http://genome4.aist-nara.ac.jp/) offers the opportunity of identifying genes by name or location and performing BLAST searches and SWISSPROT analyses with just a few mouse clicks. Alternatively, the ECDC site (http://susi.bio.uni-giessen.de/ecdc.html) allows access to gene and open reading frame (ORF) maps of the entire genome and lists of functional subsets of *E. coli* genes. Another site that is a rich source for the classification of *E. coli* genes is provided by NCBI, but this has yet to be updated with the entire genome sequence (http://www.ncbi.nlm.nih.gov/Complete_Genomes/).

E. coli genomic comparisons

At 4.64 Mb, the *E. coli* genome is the largest bacterium yet sequenced (Table 1). Why is the *E. coli* genome so large? It does not contain many duplicated genes or large non-coding regions; in fact, the proportion of coding regions is broadly similar throughout all sequenced bacterial genomes, suggesting that none contain significant stretches of redundant DNA. A useful concept for considering genome function is that of 'paralogues', which have been defined as 'homologous genes in the same organism whose products perform related but not identical functions' (Koonin *et al.*, 1996b). Around half of the genes in *E. coli* form clusters of paralogues (when analysed from 75% of the genome; Koonin *et al.*, 1996b). The high proportion of paralogous genes in *E. coli* is probably related to its ability to adapt to novel environments. In contrast, only 35% of genes from *Haemophilus influenzae* and 17% of genes from *Helicobacter pylori* (Tomb *et al.*, 1997) are paralogous. This lower level of gene paralogy explains, in part, the smaller genomes of *H. influenzae* and *H. pylori*. One of the smallest bacterial genomes belongs to *Mycoplasma*, as a consequence of its obligate parasitic lifestyle.

We now have the first opportunity to compare entire

Table 1. Comparison of eight microbial genome sequences.

Genome	Genome size (Mb)	Number of proteins per genome	Proportion of genome predicted to encode proteins	Genes with no predicted function (%)
Escherichia coli	4.64	4288	88%	23%[a]
Haemophilus influenzae	1.83	1703	85%	14%[b]
Helicobacter pylori	1.7	1590	91%	31%[a]
Methanococcus jannashii	1.66	1731	NA	28%[b]
Mycoplasma genitalium	0.58	468	88%	13%[b]
Mycoplasma pneumoniae	0.82	677	89%	11%[b]
Saccharomyces cerevisiae	12.1	5885	72%	38%[a]
Synechocystis spp.	3.6	3168	87%	29%[a]

a. The criteria used to assess the proportion of unknown genes come from different published sources and may be different.
b. Data from Koonin *et al.* (1997).
NA, data not available.

eukaryotic and prokaryotic genomes. It was expected that eukaryotes would prove to be more complex than prokaryotes and have a much larger number of proteins owing to the increased complexity of their subcellular organization. However, Table 1 shows that *Saccharomyces cerevisiae* has only 25% more proteins than *E. coli* (5886 vs. 4288 proteins). The proportion of protein functions that are conserved between *E. coli* and yeast remains to be determined.

The end of an era?

To appreciate the significance of the *E. coli* genome sequence, it is useful to consider the information that was previously available. The genetic map of the *E. coli* chromosome already contained 1958 genes (Berlyn *et al.*, 1996), most of which had already been defined biochemically and genetically (Riley and Labedan, 1996). The approximate map location of unknown genes could be obtained by Southern hybridization of an ordered gene library constructed in lambda phage (Kohara *et al.*, 1987). Analysis of global regulation of gene expression in *E. coli* had been started by Chuang *et al.* (1993). The laboratory of Neidhardt had provided us with the Gene–Protein index, which catalogued 1150 spots from two-dimensional gels and identified 400 proteins. This index gives information about the relative expression of these proteins under varied environmental conditions (Van Bogelen *et al.*, 1996). These and other crucial data have been compiled in the second volume of the invaluable bacterial encyclopaedia Escherichia coli *and* Salmonella*; Cellular and Molecular Biology* (Neidhardt, 1996) and represent the end of an era for *E. coli* research.

Research in the post-genome era

Even before the genome sequence was completed, researchers had already drawn important conclusions based on the partial sequence data that had been made available through Genbank. Taxonomists had considered phylogenetic questions (Ochman and Lawrence, 1996), evolutionary biologists had investigated protein evolution (Koonin *et al.*, 1996a), evidence was amassed for horizontal gene transfer within the *E. coli* species (Médigue *et al.*, 1991) and available sequences were analysed in great detail (Hénaut and Danchin, 1996).

Now that the entire genome sequence is available, we can embark on global approaches for the analyses of *E. coli* genes. For example, the yeast two-hybrid system has already proved to be an invaluable tool for identifying *E. coli* proteins that interact with each other (Fields and Song, 1989). In addition, new high-affinity sites for particular DNA-binding proteins can be revealed by the genomic SELEX procedure (Singer *et al.*, 1997). When allied with the genome sequence, these methods will rapidly yield information about novel interacting proteins and simplify the identification of new DNA-binding sites.

Previously, gene functions and regulatory information have been obtained from one operon at a time or, at most, from single regulons. Now, the genome offers us new possibilities: 'chip' technology (De Risi *et al.*, 1996) promises to reveal the patterns of expression of every gene on the *E. coli* genome. The chip contains a grid of up to 64 000 individual oligonucleotides or polymerase chain reaction (PCR) products, which collectively represent the entire genome. Sequential sampling of mRNA after subjecting bacteria to physiological stress, followed by hybridization to the genomic chip, could allow the visualization of regulatory cascades of gene induction. Similar experiments following the inhibition of transcription should allow the simultaneous determination of the rate of mRNA degradation for every gene. Analysis of mRNA isolated from strains mutated in individual global gene regulators should lead quickly to the complete characterization of the members of each regulon. An alternative approach involves the new IVET and STM techniques, which could be used to find genes expressed in novel environmental situations (Hensel and Holden, 1996, Heithoff *et al.*, 1997).

It is clear that the response of bacteria to external stimuli involves more than just mRNA expression. Increased mRNA levels do not always lead to greater protein expression, because the translation of mRNA is influenced by factors such as codon usage, mRNA stability and protein turnover. Consequently, approaches that consider the proteome (defined as the 'protein complement expressed by a genome') will be crucial. All approaches to the study of protein expression by *E. coli* will build on the work of Van Bogelen *et al.* (1996) who used two-dimensional gel analysis to examine environmental effects on protein expression. New electrospray mass spectrometry techniques should allow the identification of the remaining 750 spots from two-dimensional gels (Mann and Wilm, 1995), allowing genes to be associated with the protein regulation profiles that already exist. Similar techniques will be used to identify post-translational modifications of proteins, such as phosphorylation or acetylation.

Genetic analyses will be revolutionized by the availability of the genome sequence. In future, the location of an insertion mutation in the *E. coli* genome will only require the sequencing of less than 100 bp of flanking DNA. There will be no need to use traditional genetic means of gene mapping or to use conventional hybridization approaches for the identification of cloned genes. It will be rare to clone genes by digesting DNA with restriction enzymes, because all genes can now be cloned in one step by PCR-related technology. We will no longer need to sequence novel genes from *E. coli*, which should give us more time to design interesting experiments.

Implications for the study of bacterial virulence

Emerging problems with microbial infections and increased bacterial antibiotic resistance lend impetus to our goal of understanding how a bacterial cell works. Unlike virulent *E. coli*, *Erwinia*, *Shigella*, *Salmonella*, *Yersinia* and other members of the Enterobacteriaceae, K-12 was originally a commensal member of the gut microflora and does not cause disease in animals or plants. *E. coli* can cause six distinct syndromes of diarrhoea, as well as urinary tract and other infections. It is known that the *E. coli* K-12 genome contains many genes that are important pathogenicity factors in other species, but it also lacks many significant virulence genes (Groisman and Ochman, 1994). The *E. coli* genome sequence will allow such virulence genes to be identified and studied in the genetically amenable context of K-12. The different *E. coli* strains that cause varied diseases have acquired sets of specific virulence genes on plasmids, phages or within pathogenicity islands. A simple approach to identifying pathogenicity islands involves the construction of a lambda gene library of a pathogenic *E. coli* strain and sequencing the ends of each clone. Subsequent alignment with the *E. coli* K-12 sequence would reveal gene insertions of interest.

The availability of the *E. coli* and other bacterial genome sequences is already influencing microbial vaccine development. Genome sequences of virulent bacteria reveal new cell surface proteins or virulence factors that are currently being tested for vaccine production.

The start of the FUN

Blattner *et al.* (1997) found a large proportion (38%) of *E. coli* genes that could not be assigned a function. However, Blattner *et al.* made no attempt to predict functions of putative proteins that did not have similarity to well-characterized protein families. The authors note that 'when the functions of the hit sequences were varied and there was no solid agreement even for type of function, or when only one sequence was hit, no function was assigned to the query ORF, and it was classified "unknown"'. Consequently, this estimate of 'unknown' genes is likely to be an overestimate. Koonin *et al.* (1996b) have developed a more sophisticated procedure, which relies on improved database searching programs that can identify weaker, but still statistically significant, similarities. This approach has been used on an incomplete version (75%) of the genome sequence to show that genes of unknown function occupy 23% of the genome sequence available in 1996. Unknown genes were defined as having no established function and having negligible or weak database similarity, and so represent completely new classes of proteins. (Koonin *et al.*, 1996b; see http://www.ncbi.nlm.nih.gov/cgi-bin/Complete_Genomes/ectable?entrez).

I propose that we follow the example of the yeast genome and designate this class of genes 'FUN' (function unknown; Dujon, 1996). Koonin *et al.* will publish a complete survey of FUN genes in *E. coli*, which are expected to occupy around 20% of the genome, accounting for approximately 860 genes (E. V. Koonin, personal communication). The proportions of FUN genes in other bacteria range from 13% to 29% (Table 1).

FUN genes already offer new strategies for antibiotic development. A number of FUN genes are conserved throughout the currently available bacterial genomes but are not found in eukaryotes (represented by *S. cerevisiae*). Such gene products make ideal candidates as targets for antibiotic therapy. Alternatively, the identification of essential FUN genes by insertional mutagenesis may reveal crucial proteins, which could be the basis of novel antibacterial therapies.

What will FUN genes tell us?

The high proportion of FUN genes offers an unexpected challenge to the *E. coli* community. It had been expected that most of the genes involved in basic biochemical and metabolic processes had already been identified and that the genome sequence would simplify the identification of the few remaining genes. Now, we have the first opportunity to identify the 860 FUN genes, as until now they have been effectively invisible.

Visual inspection of the genetic map of *E. coli* shows that FUN genes are dispersed throughout the chromosome, many being part of operons, others appearing to be transcribed as individual genes. Are FUN genes conserved between pathogenic and non-pathogenic bacteria? between Gram-negative and Gram-positive species? Such *in silico* (i.e. computer-based) analysis could reveal important new gene families.

What might these FUN genes be? Some will be previously uncharacterized players in metabolic and regulatory processes that have already been identified. Some of the FUN genes may be evolutionary relics, which are no longer expressed under any conditions. Others will be so-called cryptic genes, like the *rcs* and *bgl* genes (Bender, 1996). Genes are termed cryptic when we have not found conditions under which they are expressed; often particular environmental stimuli may be identified by doing extensive studies with a barrage of unusual chemicals and substrates. An example of a cryptic gene that was subsequently shown to be inducible is *celF*. Originally, expression of the *celF* gene could only be shown at high temperatures, which caused cell death (Droffner and Yamamoto, 1992). Subsequently, a screen of random luciferase fusions showed that *celF* could be induced by nickel, a completely unexpected result (Guzzo and DuBow, 1994). Similar strategies could reveal patterns of expression of FUN genes.

The 860 FUN genes could be the first indication that *E coli* has completely novel abilities and previously unsuspected properties. These may include survival in novel environments, cell–cell communication, the existence of discrete subcellular microenvironments, novel sensing mechanisms involving unknown signalling molecules or even hitherto unknown sources of energy. Phenomena that could be explained by FUN genes include the intriguing and complex morphogenesis observed during the growth of *E. coli* colonies (Shapiro, 1995).

Does the high number of FUN genes represent a failure for the process of genetical and biochemical research in *E. coli*? No, it is not the research process itself that is flawed; indeed, it has been extremely successful in explaining many aspects of gene expression and regulation. For instance, detailed genetic analyses can identify all genes involved in simple metabolic systems, such as the *lac* operon. However, it is more difficult to identify genes involved in novel processes. The problem is not related to the tools we have at our disposal but to our limited imagination. Given a particular question, a geneticist can design a screen to identify most types of mutations in structural genes or to dissect more complex systems by the isolation of genes that suppress particular mutant phenotypes. Obviously, we have not been asking the right questions. The genome sequence gives us the opportunity to study all FUN genes and to characterize their function using the power of 'reverse genetics' or to analyse their patterns of expression with chip technology (De Risi *et al.*, 1996).

We have already identified many global regulatory systems in *E. coli* involving the CRP, FIS, H-NS, IHF, LRP and RpoS proteins. Nevertheless, it is conceivable that gene expression could be co-ordinated in ways that Jacob and Monod never dreamed. FUN genes could encode a hypothetical family of 'integrative' regulatory proteins, which co-ordinate other cellular regulatory systems under certain unknown environmental conditions. Such regulators could have been disguised from genetic identification by in-built redundancy or plasticity of function. Integrative regulatory proteins could function as a higher level of 'government' within *E. coli*, controlling and organizing behind the scenes.

Towards an integrative approach?

Until now, *E coli* research has been essentially reductionist, involving the understanding of particular genes and proteins. Much of this work is now complete, with metabolic pathways being understood down to the level of three-dimensional protein structures. However, there are clear limits to reductionism for molecular biologists (Bray, 1997). The challenge now is to integrate our current knowledge by considering how regulatory systems communicate, how proteins interact and how gene expression is co-ordinated. This approach offers our first hope of understanding how an entire organism functions at the most intimate level.

Conclusion

We now have everything we need to complete our dissection of *E. coli* and to synthesize this knowledge into understanding the workings of an entire cell, rather than of isolated regulatory systems. Now is the time for creative ideas rather than new technology. *E. coli* research presents the opportunity to move from a reductionist position towards an integrative approach for the first time. Let us hope that a concerted effort, with open collaboration throughout the bacterial community, will help us unlock the secrets suggested by the genome sequence.

Acknowledgements

I would like to thank Chris Burns, Dan Forbes-Ford, Mike DuBow, Stephen Gunn, Chris Higgins, Derek Hood, Bart Jordi, Nigel Saunders, Daniel Shoemaker and Jean-Francois Tomb for creative discussions. I am grateful to Nicky Langston for secretarial assistance.

References

Bender, R.A. (1996) Variations on a theme by *Escherichia*. In Escherichia coli *and* Salmonella: *Cellular and Molecular Biology*, 2nd edn. Neidhardt, F. (ed.). Washington DC: American Society for Microbiology Press, pp. 4–12.

Berlyn, M.K.B., Brookes-Low, K., Rudd, K.E., and Singer, M. (1996) Linkage map of *Escherichia coli* K12, edition 9. In Escherichia coli *and* Salmonella: *Cellular and Molecular Biology*, 2nd edn. Neidhardt, F. (ed.). Washington DC: American Society for Microbiology Press, pp. 1715–1902.

Blattner, F.R. (1983) Biological frontiers. *Science* **222:** 719–720.

Blattner, F.R., Plunkett, III, G., Bloch, C.A., Perna, N.T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J.D., Rode, C.K., Mayhew, G.F., Gregor, J., Davis, N.W., Kirkpatrick, H.A., Goeden, M.A., Rose, D.J., Mau, B., and Shau, Y. (1997) The complete genome sequence of *Escherichia coli* K12. *Science* **277:** 1453–1462.

Bray, D. (1997) Reductionism for biochemists: how to survive the protein jungle. *Trends Biochem Sci* **22:** 325–326.

Chuang, S.-E., Daniels, D.L., and Blattner, F.R. (1993) Global regulation of gene expression in *Escherichia coli*. *J Bacteriol* **175:** 2026–2036.

De Risi, J., Penland, L., Brown, P.O., Bittner, M.L., Meltzer, P.S., Ray, M., Chen, Y., Su, Y.A., and Trent, J.M. (1996) Use of a cDNA microarray to analyse gene expression patterns in human cancer. *Nature Genet* **14:** 457–60.

Droffner, M.L., and Yamamoto, N. (1992) Demonstration of cell operon expression of *Escherichia. coli, Salmonella typhimurium* and *Pseudomonas aeruginosa* at elevated temperatures refractory to their growth. *Appl Environ Microbiol* **58:** 1784–1785.

Dujon, B. (1996) The Yeast genome project: what did we learn? *Trends Genet* **12:** 263–270.

Fields, S., and Song, O. (1989) A novel genetic system to detect protein–protein interactions. *Nature* **340:** 245–246.

Groisman, E.A., and Ochman, H. (1994) How to become a pathogen. *Trends Microbiol* **2:** 289–294.

Guzzo, A., and DuBow, M.S. (1994) A *luxAB* transcriptional fusion to the cryptic *celF* gene of *Escherichia coli* displays increased luminescence in the presence of nickel. *Mol Gen Genet* **242:** 455–460.

Hénaut, A., and Danchin, A. (1996) Analysis and predictions from *Escherichia coli* sequences, or *E. coli in silico*. In Escherichia coli *and* Salmonella: *Cellular and Molecular Biology*, 2nd edn. Neidhardt, F. (ed.). Washington DC: American Society for Microbiology Press, pp. 2047–2066.

Helthoff, D.M., Conner, C.P., and Mahan, M.J. (1997) Dissecting the biology of a pathogen during infection *Trends Microbiol* **5:** (in press).

Hensel, M., and Holden, D.W. (1996) Molecular genetic approaches for the study of virulence in both pathogenic bacteria and fungi. *Microbiology* **142:** 1049–1058.

Kohara, Y., Akiyama, K., and Isono, K. (1987) The physical map of the whole *E. coli* chromosome: application of a new strategy for rapid analysis and sorting of a large genomic library. *Cell* **50:** 495–508.

Koonin, E.V., Tatusov, R.L., and Rudd, K.E. (1996a) *Escherichia coli* protein sequences: functional and evolutionary implications. In Escherichia coli *and* Salmonella: *Cellular and Molecular Biology*, 2nd edn. Neidhardt, F. (ed.). Washington DC: American Society for Microbiology Press, pp. 2203–2217.

Koonin, E.V., Tatusov, R.L., and Rudd, K.E. (1996b) Protein sequence comparison at genome scale. *Methods Enzymol* **266:** 295–322.

Koonin, E.V., Mushegian, A.R., Galperin, M.Y., and Walker, D.R. (1997) Comparison of archaeal and bacterial genomes: computer analysis of protein sequences predicts novel functions and suggests a chimeric origin for the archaea. *Mol Microbiol* **25:** 619–637.

Mann, M., and Wilm, M. (1995) Electrospray mass spectrometry for protein characterization. *Trends Biochem Sci* **20:** 219–224.

Médigue, C., Rouxel, T., Vigier, P., Hénaut, A., and Danchin, A. (1991) Evidence for horizontal gene transfer in *Escherichia coli* speciation. *J Mol Biol* **222:** 851–856.

Neidhardt, F. (ed.) (1996) Escherichia coli *and* Salmonella: *Cellular and Molecular Biology*, 2nd edn. Washington DC: American Society for Microbiology Press.

Ochman, H., and Lawrence, J.G. (1996) Phylogenetics and the amelioration of bacterial genomes. In Escherichia coli *and* Salmonella: *Cellular and Molecular Biology*, 2nd edn. Neidhardt, F. (ed.). Washington DC: American Society for Microbiology Press, pp. 2627–2637.

Riley, M., and Labedan, B. (1996) *Escherichia coli* gene products: physiological functions and common ancestries. In Escherichia coli *and* Salmonella: *Cellular and Molecular Biology*, 2nd edn. Neidhardt, F. (ed.). Washington DC: American Society for Microbiology Press, pp. 2118–2202.

Shapiro, J.A. (1995) The significances of bacterial colony patterns. *Bioessays* **17:** 597–607.

Singer, B.S., Shtatland, T., Brown, D., and Gold, L. (1997) Libraries for genomic SELEX. *Nucleic Acids Res* **25:** 781–786.

Tomb, J.-F., White, O., Kerlavage, A.R., Clayton, R.A., Sutton, G.G., Fleischmann, R.D., Ketchum, K.A., Klenk, H.P., Gill, S., Dougherty, B.A., Nelson, K., Quackenbush, J., Zhou, L., Kirkness, E.F., Peterson, S., Loftus, B., Richardson, D., Dodson, R., Khalak, H.G., Glodek, A., McKenney, K., Fitzgerald, L.M., Lee, N., Adams, M.D., Hickey, E.K., Berg, D.E., Gocayne, J.D., Utterback, T.R., Peterson, J.D., Kelley, J.M., Cotton, M.D., Weidman, J.M., Fujii, C., Bowman, C., Watthey, L., Wallin, E., Hayes, W.S., Borodovsky, M., Karp, P.D., Smith, H.O., Fraser, C.M., and Venter, J.C. (1997) The complete genome sequence of the gastric pathogen *Helicobacter pylori*. *Nature* **388:** 539–547.

Van Bogelen, R.A., Abshire, K.Z., Pertsemlidis, A., Clark, R.L., and Neidhardt, F.C. (1996) Gene–protein database of *Escherichia coli* K-12, edition 6. In Escherichia coli *and* Salmonella: *Cellular and Molecular Biology*, 2nd edn. Neidhardt, F. (ed.). Washington DC: American Society for Microbiology Press, pp. 2067–2117.