In search of the minimal *Escherichia coli* genome

Darren J. Smalley, Marvin Whiteley and Tyrrell Conway

Advanced Center for Genome Technology, The University of Oklahoma, Norman, OK 73019-0245, USA

Recent plans announced for the systematic cataloging of the minimal *Escherichia coli* gene set, the phenotypes of all mutations, the expression levels of every transcript and gene product, and the interactions of all genetic loci or their gene products point the way towards a new frontier in the biology of model organisms. Powerful tools for this endeavor are emerging, and efforts to organize the *E. coli* community are under way. The anticipated benefit is a functional model of the bacterial cell.

In calling for a 'second human genome project' to compile an inventory of the genomes of the human microflora, Stanley Falkow, Professor of Microbiology and Immunology at Stanford University, stated, 'No one has yet worked out the essence of the biology of why or how *Escherichia coli* colonizes the bowel as a commensal' (BioMedNet News, May 2, 2001; http://www.bmn.com). In the midst of a tidal wave of new functional genomics technologies, a complete understanding of *E. coli* biology appears to be on the horizon. Yet, to date, there has been no organized, worldwide effort to understand *E. coli* at the single-gene level. However, it would appear that this is about to change. Several groups have now begun systematic mutational analysis of the *E. coli* genome, involving the use of new high-throughput tools.

Functional analysis of *E. coli*

In their recent article in *Nature Biotechnology*, Yu *et al.* [1] described a tool for deleting large regions of DNA from the *E. coli* genome. The method involves two pools of mapped Tn*5* insertions that are independently combined on the same genome by transduction, and the excision of large (approximately 60–120 kb) intra-insertional regions by Cre-mediated *loxP* recombination. Several of the deletions were successfully combined in a single strain, including one strain that lacked 287 genes. In other cases, deletions could be made singly, but not combined with others. The suggestion that some mutations are 'mutually exclusive' (can be obtained individually, but not in combination) and the possibility that others could be 'mutually inclusive' (can be obtained in combination, but not individually) is worthy of further discussion. Although the Cre/*loxP* excision system has not been applied systematically, all the indications are that this is another of several powerful tools for analysis of the *E. coli* genome (Table 1).

Where is the scientific community heading with respect to the functional analysis of the *E. coli* genome? Coordination among the many excellent research groups is building rapidly. Recently, the International *E. coli* Alliance (IECA) was announced [2]. This international group of scientists, and a similar, primarily US-based group, the *E. coli* Model Cell Consortium (Emc^2), are being led by Barry Wanner, Professor of Biology at Purdue University. The IECA and Emc^2 seek a complete inventory of the *E. coli* cell. Specifically, this will involve characterizing the function, regulation or interactions of every gene, protein and metabolite. The information generated will form the basis for modeling the cell's dynamic response(s) to real-time changes in its environment.

The systematic approach

In the meantime, several laboratories around the world have contributed to the systematic mutational analysis of the *E. coli* genome. Kolisnychenko *et al.* [3], in collaboration with Frederick Blattner at the University of Wisconsin-Madison, have also developed methods to delete large segments of DNA from the *E. coli* genome. Approximately 40% of the K-islands (strain-specific sequences unique to the non-pathogenic K-12 strain), constituting >8% of the genome, were deleted with no observed effect on *in vitro* growth in either complex or minimal media. Additionally, Blattner's group is attempting to delete each gene from *E. coli* systematically and they have currently examined >25% of the ORFs (http://www.genome.wisc.edu/functional/tnmutagenesis.htm). In Japan, researchers at the Nara Institute are going one step further by deleting each *E. coli* ORF and overproducing and purifying each gene product [4] (http://ecoli.aist-nara.ac.jp/). Of course, systematic mutational analysis must be coupled with high-throughput phenotyping to match unknown genes with their function and to verify annotated gene predictions. The most promising of these is the Phenotype MicroArray [5] offered by Biolog, Inc., which permits testing of 2000 cell properties (e.g. biochemical pathways).

Ongoing efforts with yeast, and other systems, provide the scientific community with a remarkably powerful research tool. In what was a landmark study in genomics, Giaever *et al.*, in a collaboration involving 22 research groups from Europe and North America, combined systematic gene deletion with microarray analysis to profile 96% of all annotated ORFs of *Saccharomyces cerevisiae* [6]. Like *S. cerevisiae*, the ongoing attempts to determine the minimal *E. coli* genome rely on strictly systematic approaches. In other bacteria, shotgun approaches are being used to identify conditionally essential genes. The most widely used approach has been to use transposon (Tn) mutagenesis to reveal genes and DNA sequences that cannot be inactivated under

Corresponding author: Tyrrell Conway (tconway@ou.edu).

http://timi.trends.com 0966-842X/02/$ - see front matter PII: S0966-842X(02)00008-2

Table 1. Minimal genome projects

Year	Organism	Strategy	Ref.
1996	*H. influenzae*; *E. coli*	*In silico* genome comparison	[20]
1998	*H. influenzae*; *S. pneumoniae*	Tn mutagenesis; DNA fingerprinting	[8]
1999	*M. genitalium*	Saturating Tn mutagenesis	[6]
2000	*V. cholerae*	Tn mutagenesis; arabinose promoter	[10]
2001	*S. aureus*	Antisense RNA	[12]
2001	*M. bovis*	Tn mutagenesis; microarray	[7]
2002	*H. influenzae*	Tn mutagenesis; DNA fingerprinting	[9]
2002	*Buchnera* spp.	Sequence comparison	[21]
2002	*S. cerevisiae*	Systematic gene deletion	[5]
2002	*S. aureus*	Antisense RNA	[11]
2002	*E. coli*	Red recombinase excision	[2]
2002	*E. coli*	Cre/*loxP* excision	[1]

particular growth conditions. The limiting factor of saturating Tn mutagenesis is the mapping of insertions, which is being tackled by high-throughput sequencing in *Mycoplasma genitalium* [7], microarrays in *Mycobacterium bovis* [8], and genetic fingerprinting in *Haemophilus influenzae* and *Streptococcus pneumoniae* [9,10]. A modification of saturating Tn mutagenesis was devised by Judson and Mekalanos for use in *Vibrio cholerae* [11]. An outward-facing, arabinose-inducible promoter, carried on a transposon, was inserted randomly throughout the chromosome. In the absence of arabinose many transformants exhibited decreased fitness when compared with the parent, which was negated by the addition of arabinose. Finally, two independent groups identified essential genes in *Staphylococcus aureus* using antisense RNA to effectively prevent the expression of its cognate gene [12,13].

It comes as no surprise that shotgun approaches for the identification of essential genes have focused on pathogenic microorganisms. These strategies allow for high-throughput screening – ideal for searching for essential genes, virulence factors and new targets for antibiotics. By contrast, the approaches used to determine the *E. coli* minimal gene set have and should continue to involve systematic gene knockouts, and large deletions of the chromosome. The goal for *E. coli* studies is the complete cataloging of essential genes and loci, and the interactions between them and their products.

Interesting variables

The methodical approach taken by the *E. coli* community to identify essential genes has taken significantly longer than that in other organisms, but has also led to the elucidation of a relatively unexplored variable in studying essential genes. The identification of 'mutually inclusive' and 'mutually exclusive' genes by combinatorial mutations adds a further dimension to the study of the minimal genome. The term 'mutually exclusive' describes genes that can be deleted independently of one another, but when both genes are deleted the result is a lethal phenotype. For example, Yu *et al.* [1] showed it was not possible to combine some of the large deletions in their study; however, the specific genes involved were not identified. A specific example of mutually exclusive genes would be an *nrdA*/*nrdD* mutant, which would eliminate the ability of *E. coli* to grow both aerobically and anaerobically [14,15]. 'Mutually inclusive' describes cases where a single gene cannot be deleted, however, in combination with a second deletion, the first deletion is tolerated. This is most strikingly observed in cell killing systems such as the chromosomal addiction module genes, in which deletion of the antitoxin gene results in the build-up of toxin and consequent cell death [16]. Deletion of both genes, including that for the toxin, negates the lethality. With this concept of mutually exclusive and inclusive mutations in mind, independent mutations should be combined, perhaps by integration of these mutational strategies, to map gene product interactions systematically.

Another important question raised is the actual number of *E. coli* genes expressed in growing cells. The expression of a gene reflects its importance for the operation of a cell process under a given growth condition. Although not all genes expressed under a given condition are essential, they might in fact contribute substantially to the fitness of the organism for an ecological niche or habitat. Microarray studies of *E. coli* grown on minimal media with glucose as the sole carbon source estimate that between 60–83% of known and putative ORFs are expressed [17,18]. Proteomics studies of *E. coli* indicate that approximately 70% of protein gene products, depending on growth conditions, can be visualized on 2-D gels [19] (http://us.expasy.org/ch2d/). The large discrepancy in these estimates of the number of expressed genes results from technical and analytical differences between experiments. Application of mass spectrometry to proteomics offers the promise of sensitivity approaching one protein copy per cell [20]. Resolution of the issue of how many genes are expressed under a particular condition should be forthcoming.

Concluding remarks

The product of systematic *E. coli* functional genomics will be a database for all expression and essentiality data for every gene in the genome. These data might also be used to understand the mutually exclusive and inclusive genes of *E. coli*. With this resource, our knowledge of the biology of *E. coli* will be advanced significantly. However, it is evident that systematic mutational analysis should go beyond a screen for essential genes and include strategies to evaluate their contribution to fitness and survival in natural environments. For pathogenic bacteria, these natural environments could be within the host or disease vector, whereas for commensal *E. coli*, this would include the mucosal layer of the large intestine and the sewage system. Understanding the functions of all genes that contribute to growth, survival and persistence in these highly competitive ecosystems is the goal of modern microbiology.

References

1 Yu, B.J. *et al.* (2002) Minimization of the *Escherichia coli* genome using a Tn*5*-targeted Cre/*loxP* excision system. *Nat. Biotechnol.* 20, 1018–1023

2 Holden, C. (2002) Alliance launched to model *E. coli*. *Science* 297, 1459–1460

3 Kolisnychenko, V. *et al.* (2002) Engineering a reduced *Escherichia coli* genome. *Genome Res.* 12, 640–647

4 Mori, H. *et al.* (2000) Functional genomics of *Escherichia coli* in Japan. *Res. Microbiol.* 151, 121–128
5 Bochner, B.R. *et al.* (2001) Phenotype microarrays for high-throughput phenotypic testing and assay of gene function. *Genome Res.* 11, 1246–1255
6 Giaever, G. *et al.* (2002) Functional profiling of the *Saccharomyces cerevisiae* genome. *Nature* 418, 387–391
7 Hutchison, C.A. *et al.* (1999) Global transposon mutagenesis and a minimal *Mycoplasma* genome. *Science* 286, 2165–2169
8 Sassetti, C.M. *et al.* (2001) Comprehensive identification of conditionally essential genes in mycobacteria. *Proc. Natl Acad. Sci. USA* 98, 12712–12717
9 Akerley, B.J. *et al.* (1998) Systematic identification of essential genes by *in vitro* mariner mutagenesis. *Proc. Natl Acad. Sci. USA* 95, 8927–8932
10 Akerley, B.J. *et al.* (2002) A genome-scale analysis for identification of genes required for growth or survival of *Haemophilus influenzae*. *Proc. Natl Acad. Sci. USA* 99, 966–971
11 Judson, N. and Mekalanos, J.J. (2000) TnAraOut, a transposon-based approach to identify and characterize essential bacterial genes. *Nat. Biotechnol.* 18, 740–745
12 Forsyth, R.A. *et al.* (2002) A genome-wide strategy for the identification of essential genes in *Staphylococcus aureus*. *Mol. Microbiol.* 43, 1387–1400
13 Ji, Y. *et al.* (2001) Identification of critical staphylococcal genes using conditional phenotypes generated by antisense RNA. *Science* 293, 2266–2269
14 Fuchs, J.A. and Karlstrom, H.O. (1973) A mutant of *Escherichia coli* defective in ribonucleosidediphosphate reductase. 2. Characterization of the enzymatic defect. *Eur. J. Biochem.* 32, 457–462
15 Garriga, X. *et al.* (1996) *nrdD* and *nrdG* genes are essential for strict anaerobic growth of *Escherichia coli*. *Biochem. Biophys. Res. Commun.* 229, 189–192
16 Gotfredsen, M. and Gerdes, K. (1998) The *Escherichia coli relBE* genes belong to a new toxin–antitoxin gene family. *Mol. Microbiol.* 29, 1065–1076
17 Arfin, S.M. *et al.* (2000) Global gene expression profiling in *Escherichia coli* K12. The effects of integration host factor. *J. Biol. Chem.* 275, 29672–29684
18 Tao, H. *et al.* (1999) Functional genomics: expression analysis of *Escherichia coli* growing on minimal and rich media. *J. Bacteriol.* 181, 6425–6440
19 Tonella, L. *et al.* (2001) New perspectives in the *Escherichia coli* proteome investigation. *Proteomics* 1, 409–423
20 Lipton, M.S. *et al.* (2002) Global analysis of the *Deinococcus radiodurans* proteome by using accurate mass tags. *Proc. Natl Acad. Sci. USA* 99, 11049–11054
21 Gil, R. *et al.* (2002) Extreme genome reduction in *Buchnera* spp.: toward the minimal genome needed for symbiotic life. *Proc. Natl Acad. Sci. USA* 99, 4454–4458