JOURNAL OF BACTERIOLOGY, Jan. 2010, p. 26–28 Vol. 192, No. 1
0021-9193/10/$12.00 doi:10.1128/JB.01316-09

Global Approaches for Finding Small RNA and Small Open Reading Frame Functions▽

Karen M. Wassarman[1] and Patricia J. Kiley[2]*

Department of Bacteriology[1] and Department of Biomolecular Chemistry,[2] University of Wisconsin—Madison, Madison, Wisconsin 53706

Sequencing of the first *Escherichia coli* (K-12) genome revealed that there were ~4,300 open reading frames (ORFs) expected to encode proteins and many stable RNA genes that encode rRNA and tRNA (3). Surprisingly, at that time only 43% of the ORFs had been previously described, and 38% could not be assigned even predicted functions. These observations suggested that there was a wealth of information still awaiting discovery even in this extremely well-studied model organism, which spurred great interest and focused studies on the ORFs of unknown function. Over the past decade it also has become increasingly apparent that there is an abundance of additional genes that play important roles in cellular physiology; these genes encode small RNAs (sRNAs) and small proteins (small ORFs [sORFs], defined as ≤50 amino acids) that were overlooked in the initial annotations due to their small size and/or lack of open reading frames (8, 16). Most of the sRNA and sORF genes are currently of unknown function; however, sRNAs and sORFs of known function often have regulatory roles, frequently in signal transduction pathways and in coordinating regulatory networks (6, 8, 16). Therefore, there is great interest in these gene classes and the identification of their functions, and two papers in this issue of the *Journal of Bacteriology*, by Hobbs et al. (9) and Hemm et al. (7), present new genome-based approaches to search for functions of sRNAs and sORFs.

Although directed genetic and genomic approaches (1, 2) have been quite successful in identifying functions for protein-coding genes and in reducing the number of ORFs of unknown function (10), these approaches have been less profitable for identifying sRNA and sORF gene functions. Loss of function of regulatory sRNA-encoding genes typically leads to more subtle phenotypes than commonly observed when regulatory proteins are mutated, perhaps due to the observation that sRNA regulation is usually more modulatory in nature than that of their protein counterparts that commonly direct large-scale changes (16). As a consequence, phenotypes associated with mutations in sRNA genes can be difficult to recognize without additional information pointing toward a specific time, event, or condition to explore. In addition, sRNA and sORF genes are significantly smaller targets for traditional genetic mutation, and sRNA genes are not subject to frameshift or nonsense mutations (since they do not code for proteins), making it more difficult to generate loss-of-function alleles by general mutation. Therefore, it is perhaps not surprising that sRNA and sORF genes have not been identified readily in classical genetic screens or selections. More global approaches, such as mRNA expression profiling and proteomic studies, also have been extremely powerful for characterization of many ORFs and have led to identification of many cellular functions. However, once again such approaches have not been as successful for studies of sRNAs and sORFs, due in part to the fact that most commercially available microarrays do not include representation of these genes, as they have only been recently identified, and that general protocols used for proteomic studies, including mass spectrometry or two-dimensional gel electrophoresis, are not well suited for studying small proteins.

The two highlighted papers from Gisela Storz's group, the reports of Hobbs et al. (9) and Hemm et al. (7), tackle the problem of characterization and identification of function for sRNAs and sORFs, respectively, and present their progress in large-scale directed studies especially suited to the study of these genes. Importantly, the approaches described in these papers are also generally applicable to functional studies of any unknown gene in *E. coli* or other microorganisms.

In the first paper by Hobbs et al. (9), the authors demonstrate the power of analyzing DNA bar-coded mutants in mixed population studies by uncovering mutant phenotypes associated with several small RNAs and small proteins. For these experiments, strains were generated in which genes of interest were individually replaced by bar codes that were originally designed for construction of mutant libraries in yeast (5, 11). The bar code contains two 20-mer sequences (UP and DN) that are unique for each mutation generated. Flanking the specific UP and DN sequences are three additional sequence elements (Fig. 1, common sequences com1, -2, and -3) that are identical for all bar codes. These "common" sequences serve as priming sites for PCR amplification of the UP or DN bar codes from each strain (Fig. 1, arrows 1a and 1b for UP and 2a and 2b for DN). Thus, isolation of genomic DNA from a mixed population of mutant cells allows the amplification of the UP and DN barcodes using just two sets of primers (Fig. 1). The relative abundance of each bar code within the PCR products should be representative of the relative abundance of each strain within the population. In this way, each mutant can be independently tracked, even in a culture containing many different genotypes, by hybridization of amplified bar code DNA to a commercially available DNA array designed to detect the bar codes. Note that each bar code has two sequence elements (UP and DN) that can be analyzed independently.

The bar-coded mutant approach clearly facilitates analysis of a large number of mutants at once, providing quantitative representation of each strain present. However, in addition, it

* Corresponding author. Mailing address: Department of Biomolecular Chemistry, 1300 University Avenue, University of Wisconsin—Madison, Madison, WI 53706. Phone: (608) 262-6632. Fax: (608) 262-5253. E-mail: pjkiley@wisc.edu.

▽ Published ahead of print on 23 October 2009.

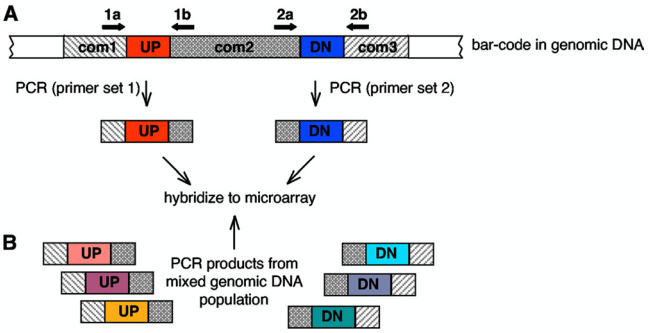

FIG. 1. Schematic representation of a bar code and the regions that will be amplified and detected for analysis (A) and the regions amplified from bar codes from different mutants (B). UP and DN are the unique sequences that will be different for every mutation generated. com1, com2, and com3 are common sequences that are the same sequence for every mutation made and serve as priming sites. Primers are indicated by black arrows above the bar codes (1a, 1b, 2a, and 2b). Note that the com2 sequence from the Hobbs et al. report contains a gene to confer kanamycin resistance to facilitate mutant strain generation, but drug resistance is not used for selection when screening mutant libraries for phenotypes. In some cases, the kanamycin resistance gene was removed to prevent any polar effects, but primer sites 1b and 2a were retained. In panel B, each UP and DN bar code represented comes from a different strain, as indicated by the change in color of the unique UP or DN sequence.

is the competition for growth within a mixed population that is likely to be critical for enhancing detection of subtle mutant phenotypes commonly associated with loss of sRNAs. For example, competitive growth experiments were required to reveal altered phenotypes for cells lacking 6S RNA in previous work (13). Thus, the use of bar-coded mutants dramatically expands the potential to search for phenotypes, even when individual mutant cells are a minority population in a complex mixture of cells, demonstrating an important advance in genetic approaches for identifying sRNA and small protein function.

Tracking individual DNA bar codes in a mixed cell population allowed the authors to specifically screen their library for strains that showed either increased or decreased resistance to their test conditions: acid or cell envelope stress. Previous studies showed that several sRNAs regulate synthesis of outer membrane proteins (15) and 70% of the small proteins are membrane localized (8), providing a rationale that the selected stress conditions (acid or envelope stress) might target small proteins and sRNAs. Indeed, this study revealed 15 genes, encoding 6 sRNAs and 9 small proteins, with previously unknown roles in these pathways. Surprisingly, one sRNA of known function, tmRNA (SsrA), was shown to play a previously unanticipated role in cell envelope stress, demonstrating that even knowing something about how an sRNA works does not fully elucidate its full physiological potential.

The library of 125 DNA bar-coded mutants also provides an important new resource for the *E. coli* K-12 toolbox. The 125 engineered mutations specify deletions of genes encoding 49 sRNAs, 50 small proteins of 50 amino acids or less, 13 small proteins of 50 to 75 amino acids, DppA (a target of GcvB sRNA), 8 known stress survival proteins, SmpA, GadE, TrpA, UspA, UspB, UspD, UspE, and OxyR, and two repetitive loci, *ldr* and *sib*. Recently, a collection of an additional 99 DNA bar-coded mutations (largely in genes associated with DNA repair) was also reported (12), creating an even larger resource for the community. One can imagine that expanding this library to include all ORFs would be a desirable genetic tool. However, creation of a comprehensive ORF library would be labor-intensive, particularly if the internal kanamycin cassette initially used to insert the bar code were removed from each strain to eliminate any possible polar effects on downstream gene expression, as was done for several of the strains in the Storz collection. Nevertheless, the ease of screening DNA bar-coded mutant libraries for fitness under any kind of growth or stress conditions should make construction of a mutant library of at least genes of unknown function a priority. In addition, this method could be easily adapted to high-throughput approaches using robots to simplify strain handling and dispensing, once appropriately sized strain libraries are generated. Finally, the approach described here of sampling a large number of mutants at once for competition for growth provides a nice complement to recent genetic approaches developed by Butland et al. (4) and Typas et al. (14), in which identification of gene function was guided by identification of interacting genes that either enhanced or inhibited growth.

In the second paper, by Hemm et al. (7), the authors describe a systematic approach for assaying and defining conditions that induce the synthesis of low-molecular-weight proteins, which have been difficult to study. The rationale is that understanding when a gene is expressed may give insight into gene function. For this goal, the authors generated strains in which sORF genes of interest have been modified to encode a small, in-frame tag (sequential peptide affinity [SPA] tag) to facilitate protein analysis with a commercially available antibody specific to the tag. Differences in expression of SPA-tagged proteins were analyzed under a variety of carbon source or stress conditions to assess when these small proteins might be functional. In addition, for several of these proteins, the mRNA transcripts were also studied and transcription factors were identified that mediate the observed regulation. Overall, the authors found that 21 of 51 proteins tested were induced under at least one condition tested (heat shock, oxidative stress, zinc limitation, oxygen limitation, acid or envelope stress, or changes in carbon source). Remarkably, the levels of over half of these proteins were increased during heat shock, suggesting that these small proteins may play a specific role in the response to increased temperatures.

In summary, the two highlighted papers from this issue demonstrate the successful use of two approaches that can be applied globally to learn more about the functions of sRNAs and small proteins. Extension of these approaches to additional growth or stress conditions should provide an even richer data set to fully comprehend the physiological function of these exciting gene products.

REFERENCES

1. **Baba, T., T. Ara, M. Hasegawa, Y. Takai, Y. Okumura, M. Baba, K. A. Datsenko, M. Tomita, B. L. Wanner, and H. Mori.** 2006. Construction of *Escherichia coli* K-12 in-frame, single-gene knockout mutants: the Keio collection. Mol. Sys. Biol. **2:**2006.0008.
2. **Baba, T., H. C. Huan, K. Datsenko, B. L. Wanner, and H. Mori.** 2008. The applications of systematic in-frame, single-gene knockout mutant collection of *Escherichia coli* K-12. Methods Mol. Biol. **416:**183–194.
3. **Blattner, F. R., G. Plunkett, 3rd, C. A. Bloch, N. T. Perna, V. Burland, M. Riley, J. Collado-Vides, J. D. Glasner, C. K. Rode, G. F. Mayhew, J. Gregor, N. W. Davis, H. A. Kirkpatrick, M. A. Goeden, D. J. Rose, B. Mau, and Y.**

Shao. 1997. The complete genome sequence of *Escherichia coli* K-12. Science **277:**1453–1462.

4. **Butland, G., M. Babu, J. J. Diaz-Mejia, F. Bohdana, S. Phanse, B. Gold, W. Yang, J. Li, A. G. Gagarinova, O. Pogoutse, H. Mori, B. L. Wanner, H. Lo, J. Wasniewski, C. Christopolous, M. Ali, P. Venn, A. Safavi-Naini, N. Sourour, S. Caron, J. Y. Choi, L. Laigle, A. Nazarians-Armavil, A. Deshpande, S. Joe, K. A. Datsenko, N. Yamamoto, B. J. Andrews, C. Boone, H. Ding, B. Sheikh, G. Moreno-Hagelseib, J. F. Greenblatt, and A. Emili.** 2008. eSGA: *E. coli* synthetic genetic array analysis. Nat. Methods **5:**789–795.
5. **Giaever, G., A. M. Chu, L. Ni, C. Connelly, L. Riles, S. Veronneau, S. Dow, A. Lucau-Danila, K. Anderson, B. Andre, A. P. Arkin, A. Astromoff, M. El-Bakkoury, R. Bangham, R. Benito, S. Brachat, S. Campanaro, M. Curtiss, K. Davis, A. Deutschbauer, K. D. Entian, P. Flaherty, F. Foury, D. J. Garfinkel, M. Gerstein, D. Gotte, U. Guldener, J. H. Hegemann, S. Hempel, Z. Herman, D. F. Jaramillo, D. E. Kelly, S. L. Kelly, P. Kotter, D. LaBonte, D. C. Lamb, N. Lan, H. Liang, H. Liao, L. Liu, C. Luo, M. Lussier, R. Mao, P. Menard, S. L. Ooi, J. L. Revuelta, C. J. Roberts, M. Rose, P. Ross-Macdonald, B. Scherens, G. Schimmack, B. Shafer, D. D. Shoemaker, S. Sookhai-Mahadeo, R. K. Storms, J. N. Strathern, G. Valle, M. Voet, G. Volckaert, C. Y. Wang, T. R. Ward, J. Wilhelmy, E. A. Winzeler, Y. Yang, G. Yen, E. Youngman, K. Yu, H. Bussey, J. D. Boeke, M. Snyder, P. Philippsen, R. W. Davis, and M. Johnston.** 2002. Functional profiling of the *Saccharomyces cerevisiae* genome. Nature **418:**387–391.
6. **Gottesman, S., C. A. McCullen, M. Guillier, C. K. Vanderpool, N. Majdalani, J. Benhammou, K. M. Thompson, P. C. FitzGerald, N. A. Sowa, and D. J. FitzGerald.** 2006. Small RNA regulators and the bacterial response to stress. Cold Spring Harbor Symp. Quant. Biol. **71:**1–11.
7. **Hemm, M. R., B. J. Paul, J. Miranda-Rios, A. Zhang, N. Soltanzad, and G. Storz.** 2010. Small stress response proteins in *Escherichia coli*: proteins missed by classical proteomic studies. J. Bacteriol. **192:**46–58.
8. **Hemm, M. R., B. J. Paul, T. D. Schneider, G. Storz, and K. E. Rudd.** 2008. Small membrane proteins found by comparative genomics and ribosome binding site models. Mol. Microbiol. **70:**1487–1501.
9. **Hobbs, E. C., J. L. Astarita, and G. Storz.** 2010. Small RNAs and small proteins involved in resistance to cell envelope stress and acid shock in *Escherichia coli*: analysis of a bar-coded mutant collection. J. Bacteriol. **192:**59–67.
10. **Hu, P., S. C. Janga, M. Babu, J. J. Diaz-Mejia, G. Butland, W. Yang, O. Pogoutse, X. Guo, S. Phanse, P. Wong, S. Chandran, C. Christopoulos, A. Nazarians-Armavil, N. K. Nasseri, G. Musso, M. Ali, N. Nazemof, V. Eroukova, A. Golshani, A. Paccanaro, J. F. Greenblatt, G. Moreno-Hagelsieb, and A. Emili.** 2009. Global functional atlas of *Escherichia coli* encompassing previously uncharacterized proteins. PLoS Biol. **7:**e96.
11. **Pierce, S. E., R. W. Davis, C. Nislow, and G. Giaever.** 2007. Genome-wide analysis of barcoded *Saccharomyces cerevisiae* gene-deletion mutants in pooled cultures. Nat. Prot. **2:**2958–2974.
12. **Rooney, J. P., A. Patil, M. R. Zappala, D. S. Conklin, R. P. Cunningham, and T. J. Begley.** 2008. A molecular bar-coded DNA repair resource for pooled toxicogenomic screens. DNA Repair (Amst.) **7:**1855–1868.
13. **Trotochaud, A. E., and K. M. Wassarman.** 2004. 6S RNA function enhances long-term cell survival. J. Bacteriol. **186:**4978–4985.
14. **Typas, A., R. J. Nichols, D. A. Siegele, M. Shales, S. R. Collins, B. Lim, H. Braberg, N. Yamamoto, R. Takeuchi, B. L. Wanner, H. Mori, J. S. Weissman, N. J. Krogan, and C. A. Gross.** 2008. High-throughput, quantitative analyses of genetic interactions in *E. coli*. Nat. Methods **5:**781–787.
15. **Valentin-Hansen, P., J. Johansen, and A. A. Rasmussen.** 2007. Small RNAs controlling outer membrane porins. Curr. Opin. Microbiol. **10:**152–155.
16. **Waters, L. S., and G. Storz.** 2009. Regulatory RNAs in bacteria. Cell **136:** 615–628.

The views expressed in this Commentary do not necessarily reflect the views of the journal or of ASM.