Science AAAS

Inferring Genetic Networks and Identifying Compound Mode of Action via Expression Profiling

Timothy S. Gardner, *et al.*

Science **301**, 102 (2003);
DOI: 10.1126/science.1081900

Science (print ISSN 0036-8075; online ISSN 1095-9203) is published weekly, except the last week in December, by the American Association for the Advancement of Science, 1200 New York Avenue NW, Washington, DC 20005. Copyright 2003 by the American Association for the Advancement of Science; all rights reserved. The title *Science* is a registered trademark of AAAS.

been a past history of selection on this trait. Yet, low levels of genetic variation for desiccation resistance appear to be preventing any further increases in resistance in this rainforest species despite ample genetic variation in other traits and at neutral markers as evident from the microsatellite results. Our results show that genetic variation in neutral markers can provide an incomplete picture of the evolutionary potential of populations, consistent with the weak association between genetic diversity as measured by quantitative methods and that measured by molecular methods (*25*). The absence of a selection response for traits linked to climatic stress in this study and in a few other cases (*26*) suggests that levels of variation must be evaluated for ecologically relevant traits in those species that are threatened by climate change and fragmentation, including endangered species (*27*).

References and Notes

1. W. E. Bradshaw, C. M. Holzapfel, *Proc. Natl. Acad. Sci. U.S.A.* **98**, 14509 (2001).
2. L. Hughes, *Trends Ecol. Evol.* **15**, 56 (2000).
3. M. S. Warren *et al.*, *Nature* **400**, 65 (2001).
4. R. C. Lewontin, *The Genetic Basis of Evolutionary Change* (Columbia Univ. Press, New York, 1974).
5. P. A. Parsons, *Evol. Biol.* **14**, 297 (1982).
6. T. E. Lovejoy *et al.*, in *Conservation Biology: The Science of Scarcity and Diversity*, M. E. Soule, Ed. (Sinauer Associates, Sunderland, MA, 1986), pp. 257–285.
7. A. A. Hoffmann, P. A. Parsons, *Biol. J. Linn. Soc.* **37**, 117 (1989).
8. M. W. Blows, A. A. Hoffmann, *Evolution* **47**, 1255 (1993).
9. G. Gibbs, A. K. Chippindale, M. R. Rose, *J. Exp. Biol.* **200**, 1821 (1997).
10. A. A. Hoffmann, P. A. Parsons, *J. Evol. Biol.* **6**, 643 (1993).
11. F. J. Ayala, *Evolution* **19**, 538 (1965).
12. M. J. Hercus, A. A. Hoffmann, *Genetics* **151**, 1493 (1999).
13. N. L. Jenkins, A. A. Hoffmann, *Aust. J. Entomol.* **40**, 41 (2001).
14. Materials and methods are available as supporting material on *Science* Online.
15. D. Karan *et al.*, *Evolution* **52**, 825 (1998).
16. D. Karan, P. Parkash, *Ecol. Entomol.* **23**, 391 (1998).
17. A. Addo-Bediako, S. L. Chown, K. J. Gaston, *J. Insect Physiol.* **47**, 1377 (2001).
18. L. Partridge, N. Prowse, P. Pignatelli, *Proc. R. Soc. London Ser. B* **266**, 255 (1999).
19. L. G. Harshman, A. A. Hoffmann, *Trends Ecol. Evol.* **15**, 32 (2000).
20. R. Frankham, J. D. Ballou, D. A. Briscoe, *Introduction to Conservation Genetics* (Cambridge Univ. Press, Cambridge, 2002).
21. A. A. Hoffmann, in *Adaptive Genetic Variation in the Wild*, T. A. Mousseau, B. Sinervo, J. A. Endler, Eds. (Oxford Univ. Press, New York, 2000), pp. 200–218.
22. D. A. Roff, *Evolutionary Quantitative Genetics* (Chapman & Hall, New York, 1997).
23. M. Sgrò, L. Partridge, *Am. Nat.* **156**, 341 (2000).
24. A. A. Hoffmann, R. Hallas, C. Sinclair, L. Partridge, *Evolution* **55**, 436 (2001).
25. D. H. Reed, R. Frankham, *Evolution* **55**, 1095 (2001).
26. F. Baer, J. Travis, *Evolution* **54**, 238 (2000).
27. K. A. Crandall, C. R. P. Binida-Edmonds, G. M. Mace, R. K. Wayne, *Trends Ecol. Evol.* **15**, 290 (2000).

28. We thank the Australian Research Council for financial support via its Special Research Centre Scheme, as well as the Department of Education, Training and Youth Affairs for molecular infrastructure funding, C. Sgrò and A. Weeks for comments on the manuscript, and M. Higgie for some field lines.

Supporting Online Material
www.sciencemag.org/cgi/content/full/301/5629/100/DC1
Materials and Methods

10 March 2003; accepted 6 May 2003

Inferring Genetic Networks and Identifying Compound Mode of Action via Expression Profiling

Timothy S. Gardner,[1]* Diego di Bernardo,[1,2]* David Lorenz,[1] James J. Collins[1]†

The complexity of cellular gene, protein, and metabolite networks can hinder attempts to elucidate their structure and function. To address this problem, we used systematic transcriptional perturbations to construct a first-order model of regulatory interactions in a nine-gene subnetwork of the SOS pathway in *Escherichia coli*. The model correctly identified the major regulatory genes and the transcriptional targets of mitomycin C activity in the subnetwork. This approach, which is experimentally and computationally scalable, provides a framework for elucidating the functional properties of genetic networks and identifying molecular targets of pharmacological compounds.

Efforts to systematically define the organization and function of gene, protein, and metabolite networks include experimental and computational methods for identifying molecular interactions (*1–3*), global structural properties (*4*, *5*), metabolic limits (*6*), and regulatory modules and characteristics (*7–9*). These methods have provided valuable insights in many applications, but they often provide only structural information or require extensive quantitative information, which is not generally available, particularly for larger regulatory networks. In previous computational studies (*10–12*), alternative methods have been proposed that would enable rapid deduction of network connectivity and functional properties solely from temporal gene-expression data. However, the acquisition of adequate temporal expression data remains difficult, and the practical utility of such approaches has not been determined.

Here, we present a rapid and scalable method that enables construction of a first-order predictive model of a gene and protein regulatory network using only steady-state expression measurements and no previous information on the network structure or function. We use multiple linear regression to determine the model from RNA expression changes resulting from a set of steady-state transcriptional perturbations. The model can be used to identify the regulatory role of individual genes in the network, useful control points in the network, and genes that directly mediate a pharmaceutical compound's bioactivity in the cell. The method, called network identification by multiple regression (NIR), is derived from a branch of engineering called system identification (*13*), in which a model of the connections and functional relations between elements in a network is inferred from measurements of system dynamics (e.g., the response of genes and proteins to external perturbations).

To apply a system-identification method, we assume that the behavior of a gene, protein, and metabolite regulatory network can be modeled by a system of nonlinear differential equations (*14*, *15*). Near a steady-state point (e.g., when gene expression does not change substantially over time), such a nonlinear system may be approximated to the first order by a linear system of equations describing the rate of accumulation of each network species resulting from a transcriptional perturbation:

$$d\mathbf{x}/dt = \mathbf{A}\mathbf{x} + \mathbf{u} \qquad (1)$$

where $\mathbf{x}$ is a vector representing the concentrations of N RNAs, proteins, and metabolites in the network; $d\mathbf{x}/dt$ represents the rate of accumulation of the species in $\mathbf{x}$; $\mathbf{u}$ is a vector representing an external perturbation to the rate of accumulation of the species in $\mathbf{x}$; and $\mathbf{A}$, the network model, is an $N \times N$ matrix of coefficients describing the regulatory interactions between the species in $\mathbf{x}$. Next, we identify the coefficients of $\mathbf{A}$ using only RNA expression changes that result from steady-state transcriptional perturbations. Because we measure RNA but not protein or metabolite species in this study, variables representing proteins and metabolites are not explicitly represented in the network model. Thus, regulatory connections in the model are not, in general, physical connections; rather, they represent effective functional relations between transcripts.

Under the steady-state assumption ($d\mathbf{x}/dt = 0$), Eq. 1 reduces to $\mathbf{A}\mathbf{x} = -\mathbf{u}$. To identify the network model, we could, in principle, make N distinct perturbations, $\mathbf{u}$, to the RNAs in a particular network, recover N sets of RNA concentrations, $\mathbf{x}$, and solve directly for $\mathbf{A}$ (*16*). How-

[1] Center for BioDynamics and Department of Biomedical Engineering, Boston University, 44 Cummington Street, Boston, MA 02215, USA. [2] Telethon Institute for Genetics and Medicine (TIGEM), Via P. Castellino 111, 80131, Naples, Italy.

*These authors contributed equally to this work.
†To whom correspondence should be addressed. E-mail: jcollins@bu.edu

ever, in larger networks it may be impractical to perform a full set of *N* perturbation experiments, and thus our problem would remain underdetermined. Even with a full set of perturbation experiments, RNA expression data are prone to high levels of measurement noise, making the direct solution unreliable. To overcome this problem, we assume that most biochemical networks are not fully connected (*17*, *18*, *19*), that is, some of the coefficients of **A** are zero. Thus, by assuming a maximum of *k* nonzero regulatory inputs to each gene (where $k < N$), we can transform our underdetermined problem into an overdetermined problem, making it robust both to measurement noise and incomplete data sets.

We next apply multiple linear regression (*20*) to calculate the model coefficients for each possible combination of *k* regulatory inputs (*k* coefficients) per gene. The *k* coefficients for each gene that fit the expression data with the smallest error are chosen as the best approximation of **A**. Using the standard errors on the RNA measurement data, the algorithm also computes the statistical significance of each recovered coefficient of **A** and the overall fit of **A**. A complete description of the algorithm is provided in the supporting online text.

We applied the NIR method to a nine-transcript subnetwork of the SOS pathway in *E. coli* (the "test network"). The SOS pathway, which regulates cell survival and repair after DNA damage, involves the *lexA* and *recA* genes, more than 30 genes directly regulated by *lexA* and *recA*, and tens or possibly hundreds of indirectly regulated genes (*21*–*25*). We chose the nine transcripts in our test network (Fig. 1) to include the principal mediators of the SOS response (*lexA* and *recA*), four other regulatory genes with known involvement in the SOS response (*ssb*, *recF*, *dinI*, and *umuDC*), and three sigma factor genes (*rpoD*, *rpoH*, and *rpoS*) whose regulatory role in the SOS response is not fully understood. Because much of the regulatory structure of our test network has been previously mapped, it serves as an excellent subject for the validation of our method. In addition, it serves as an entry point for further study of the SOS pathway, which regulates genes associated with important protective pathways relevant to antibiotic resistance (*23*, *26*).

We applied a set of nine transcriptional perturbations to the test network in *E. coli* cells (*27*). In each perturbation, we overexpressed a different one of the nine genes in the test network with an arabinose-controlled episomal expression plasmid (fig. S1). We grew the cells in batch cultures under constant physiological conditions to their steady state (~5.5 hours after the addition of arabinose). Cells were maintained in the exponential growth phase throughout all experiments. For all nine transcripts, we used quantitative real-time polymerase chain reaction (qPCR) to measure the change in expression relative to that in unperturbed cells. For each transcript, two qPCR reactions from each of eight replicate cultures were obtained, and qPCR data were filtered to eliminate aberrant or inefficient reactions (*27*). The mean expression changes for each transcript in each experiment (**x** in Eq. 1) were calculated (*27*), and only those changes that were greater than their standard error were accepted as significant and used for further analysis (that is, $x_i = 0$ if $|x_i| < S_{x_i}$, where x_i is the mean expression change and S_{x_i} is the standard error for transcript *i*).

Using the nine-perturbation expression data set (the training set, tables S6 to S8) and the NIR algorithm described above, we solved Eq. 1 for **A**, the model of the regulatory interactions in the test network (table S1). The number of input connections per gene (*k*) was chosen such that the solved model provided a statistically significant fit (as determined by an *F* test), was dynamically stable, and provided the best balance between coverage and false-positives (*27*). To evaluate the performance of the algorithm, we determined the number of connections in the test network that were correctly resolved in the model, **A**. A resolved connection was considered correct if there exists a known RNA, protein, or metabolite pathway between the two transcripts and if the sign of the net effect of regulatory interaction (that is, activating or inhibiting) is correct, as determined by the currently known network in Fig. 1.

The algorithm correctly identified the key regulatory connections in the network. For example, the model correctly shows that *recA* positively regulates *lexA* and its own transcription, whereas *lexA* negatively regulates *recA* and its own transcription. In addition, the model correctly identified *recA* and *lexA* as having the greatest regulatory influence on the other genes in the test network (table S5). Overall, the performance (coverage and false-positives) of the NIR algorithm was equivalent to that expected on the basis of simulations of 50 random nine-gene networks (Fig. 2). Moreover, for the subnetwork of six genes typically considered part of the SOS network (*recA*, *lexA*, *ssb*, *recF*, *dinI*, and *umuDC*), the performance of the algorithm improved substantially. This suggests that some of the false-positives identified for the three sigma factors in our model (*rpoD*, *rpoH*, and *rpoS*) may be true connections mediated by genes not included in our test network. Furthermore, our simulation results suggest that even small reductions in the measurement noise observed in our experiments [mean noise level = mean(S_{x_i})/mean(x_i) = 68%] could lead to substantial improvements in coverage and errors in the network model (Fig. 2). Reductions in experimental noise could be achieved with improved RNA measurement technologies such as competitive PCR coupled with matrix-assisted laser desorption/ionization–time-of-flight (MALDI-TOF) mass spectrometry (*28*).

We also tested the performance of the NIR algorithm with an incomplete training set consisting of perturbations to only seven of the nine genes. We solved for network models using all 36 combinations of seven perturbations and found that the algorithm

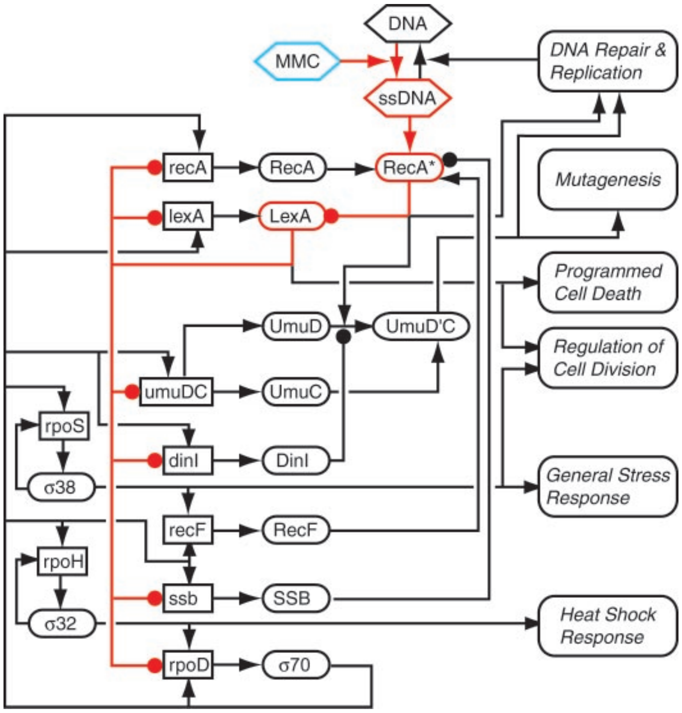

Fig. 1. Diagram of interactions in the SOS network. DNA lesions caused by mitomycin C (MMC) (blue hexagon) are converted to single-stranded DNA during chromosomal replication. Upon binding to ssDNA, the RecA protein is activated (RecA*) and serves as a coprotease for the LexA protein. The LexA protein is cleaved, thereby diminishing the repression of genes that mediate multiple protective responses. Boxes denote genes, ellipses denote proteins, hexagons indicate metabolites, arrows denote positive regulation, filled circles denote negative regulation. Red emphasis denotes the primary pathway by which the network is activated after DNA damage.

also performed comparably to simulations, albeit with slightly reduced performance in comparison with the full nine-perturbation training set (Fig. 2).

Much of the value of the network model lies in its predictive power, that is, its ability to predict expression changes and network behaviors that fall outside the training data set used to solve the model. Here, we demonstrate its predictive power by using it to distinguish the transcripts that are directly targeted by a pharmacological compound (the compound's mode of action) from transcripts that exhibit secondary responses to the expression changes of the direct targets. Thus, the direct targets represent the minimal subset of transcripts in the model that will produce the observed expression pattern if externally perturbed. Because proteins and metabolites are not measured in this study, the compound may not physically interact with transcripts identified as direct targets but instead may interact with protein or metabolite intermediates that are not explicitly represented in the network model.

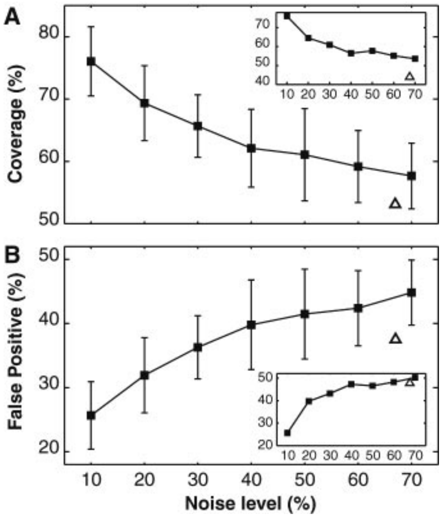

Fig. 2. NIR algorithm performance. (**A**) Coverage (correctly identified connections/total true connections) and (**B**) false-positives (incorrectly identified connections/total identified connections) were calculated for SOS models solved with a nine-perturbation training set (main panels) and a seven-perturbation training set (insets). Error bars are not included in the insets for clarity. Experiment (open triangles): Coverage and false-positives were calculated by comparing the solved model (table S1) to connections described in the literature (table S4 and Fig. 1). Because a nonsignificant fit was obtained for *recF*, the weights for inputs to *recF* were set to zero in the model. The mean noise observed on the mRNA measurements in our experiments was 68% (noise = S_x/μ_x, where S_x is the standard deviation of the mean of x, μ_x). Simulations (filled squares): Simulated perturbations were applied to 50 randomly connected networks of nine genes with an average of five regulatory inputs per gene. For each perturbation to each random network, the mRNA expression changes at steady state were calculated. The noise on the perturbations was set to 20%, equivalent to that observed on perturbations in our experiments. The noise on the mRNA concentrations was varied from 10 to 70%.

To identify direct transcriptional targets of a compound, we first measure RNA expression changes ($\mathbf{x}_p$) resulting from treatment with the compound. The activity of the compound is treated as a set of unknown transcriptional perturbations ($\mathbf{u}_p$) that produce the measured expression changes. From Eq. 1, we calculate the unknown perturbations as $\mathbf{u}_p = -\mathbf{A}\mathbf{x}_p$ (*27*). The direct transcriptional targets of a compound are those that exhibit statistically significant values in $\mathbf{u}_p$. Calculation of the statistical significance of $\mathbf{u}_p$ is described in the supporting online text.

We first applied our scheme to RNA expression changes that result from the simultaneous controlled perturbation of the *lexA* and *recA* genes. This perturbation might represent the effects of a hypothetical compound and serves as a well-defined input for validating the predictive power of our model. Although five of the nine test-network genes responded with statistically significant transcriptional changes (Fig. 3A), application of our network model correctly identified only *lexA* and *recA* as the perturbed genes (2/2 = 100% coverage, 7/7 = 100% specificity) (Fig. 3B).

We next applied a mitomycin C (MMC) perturbation to determine whether our scheme could identify the transcriptional targets of MMC bioactivity in the SOS network. Perturbed cells were grown in 0.75 μg/ml MMC, and transcriptional changes were measured relative to those in control cells grown in the normal baseline condition (0.5 μg/ml MMC). All genes in the test network showed statistically significant transcriptional increases (Fig. 3C). When we applied the network model to the expression data, we correctly identified *recA* as the transcriptional target of MMC bioactivity, with only one false-positive, *umuDC* (1/1 = 100% coverage, 7/8 = 88% specificity) (Fig. 3D). Moreover, *recA* was identified at a higher significance level ($P \leq 0.09$) than was *umuDC* ($P \leq 0.22$), suggesting that it is the more likely, if not the only, true target. It is also possible, however, that *umuDC* interacts with gene, protein, or metabolite targets of the compound that are not represented in our model. Therefore, *umuDC* may have been correctly identified as a target in our model. We also found that a model recovered with a seven-perturbation training set that excludes the *lexA* and *recA* training perturbations performs nearly as well as the model recovered with a full training set (see supporting online text and fig. S3).

The NIR method, a form of system identification based on multiple linear regression analysis of steady-state transcription profiles, provides a framework for rapidly elucidating the structure and function of genetic networks with no prior information. The method is robust to high levels of measurement noise, scalable for larger biochemical networks (*27*), and equally applicable to transcript, protein, and metabolite activity data. With advances in high-throughput measurement methods, it may soon be feasible to include protein and metabolite measurements on a large scale. The model recovered with this method enables the identification of key properties of the network, such as the major regulatory genes, and it provides a mechanism for efficiently identifying the mode of action of uncharacterized pharmacological compounds. These capabilities may facilitate optimization of cellular processes for biotechnology applications and the development of novel classes of therapeutic drugs that account for and utilize the complex regulatory properties of genetic networks.

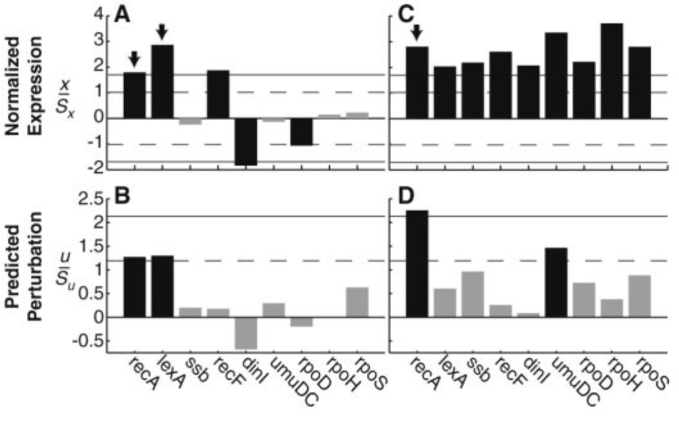

Fig. 3. Cells were perturbed either with a *lexA-recA* double perturbation or with MMC. The mean relative expression changes (x), normalized by their standard deviations (S_x), are illustrated for the *lexA-recA* double perturbation (**A**) and the MMC perturbation (**C**). Arrows indicate the genes known to be targeted by the perturbation. Predicted perturbations in the *lexA-recA* experiment (**B**) and the MMC experiment (**D**) were calculated from the expression data in (A) and (C) using the SOS model solved with the nine-perturbation training set (*27*). The predicted perturbations to each gene (u) were normalized by their standard deviations (S_u) to determine statistical significance. In all panels, black bars indicate statistically significant and gray bars indicate statistically nonsignificant. Horizontal lines denote significance levels: P = 0.3 (dashed), P = 0.1 (solid).

References and Notes

1. T. I. Lee *et al.*, *Science* **298**, 799 (2002).
2. T. Ideker *et al.*, *Science* **292**, 929 (2001).
3. A. Arkin, P. D. Shen, J. Ross, *Science* **277**, 1275 (1997).
4. S. Maslov, K. Sneppen, *Science* **296**, 910 (2002).
5. E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, A.-L. Barabási, *Science* **297**, 1551 (2002).
6. J. S. Edwards, B. O. Palsson, *Proc. Natl. Acad. Sci. U.S.A.* **97**, 5528 (2000).
7. E. H. Davidson *et al.*, *Science* **295**, 1669 (2002).
8. S. S. Shen-Orr, R. Milo, S. Mangan, U. Alon, *Nature Genet.* **31**, 64 (2002).
9. U. S. Bhalla, R. Iyengar, *Science* **283**, 381 (1999).
10. S. Liang, S. Fuhrman, R. Somogyi, *Proc. Pac. Symp. Biocomp.* **3**, 18 (1998).
11. P. D'Haeseleer, X. Wen, S. Fuhrman, R. Somogi, *Proc. Pac. Symp. Biocomp.* **4**, 41 (1999).
12. E. P. van Someren, L. F. A. Wessels, M. J. T. Reinders, E. Backer, *Proceedings of the 2nd International Conference on Systems Biology* (Caltech, Pasadena, CA, 2001), pp. 222–230.
13. L. Ljung, *System Identification: Theory for the User* (Prentice Hall, Upper Saddle River, NJ, 1999).
14. H. H. McAdams, A. Arkin, *Annu. Rev. Biophys. Biomol. Struct.* **27**, 199 (1998).
15. H. de Jong, *J. Comput. Biol.* **9**, 67 (2002).
16. A. de la Fuente, P. Brazhnik, P. Mendes, *Trends Genet.* **18**, 395 (2002).
17. D. Thieffry, A. M. Huerta, E. Pérez-Rueda, J. Collado-Vides, *Bioessays* **20**, 433 (1998).
18. J. Tegner, M. K. Yeung, J. Hasty, J. J. Collins, *Proc. Natl. Acad. Sci. U.S.A.* **100**, 5944 (2003).
19. H. Jeong, S. P. Mason, A.-L. Barabási, Z. N. Oltvai, *Nature* **411**, 41 (2001).
20. D. Montgomery, E. A. Peck, G. G. Vining, *Introduction to Linear Regression Analysis* (Wiley, New York, 2001).
21. J. Courcelle, A. Khodursky, B. Peter, P. O. Brown, P. C. Hanawalt, *Genetics* **158**, 41 (2001).
22. G. C. Walker, in Escherichia coli *and* Salmonella: *Cellular and Molecular Biology* (American Society for Microbiology, Washington, DC, ed. 2, 1996), pp. 1400–1416.
23. W. H. Koch, R. Woodgate, in *DNA Damage and Repair*, vol. 1, *DNA Repair in Prokaryotes and Lower Eukaryotes* (Humana, Totowa, NJ, 1998), pp. 107–134.
24. A. R. Fernández de Henestrosa *et al.*, *Mol. Microbiol.* **35**, 1560 (2000).
25. P. D. Karp *et al.*, *Nucleic Acids Res.* **30**, 56 (2002).
26. K. Lewis, *Microbiol. Mol. Biol. Rev.* **64**, 503 (2000).
27. Materials, methods, and supporting data are available as supporting material on *Science* Online.
28. C. Ding, C. R. Cantor, *Proc. Natl. Acad. Sci. U.S.A.* **100**, 3059 (2003).
29. We thank J. Tegner for his insights on this work, H. Schaeffer for his guidance on qPCR, and the Department of Biology at Boston University for access to their qPCR facility. This work was supported by the Defense Advanced Research Projects Agency, NSF, the Office of Naval Research, the Human Frontiers Science Program, and the Telethon Institute of Genetics and Medicine.

Supporting Online Material
www.sciencemag.org/cgi/content/full/301/5629/102/DC1
Materials and Methods
SOM Text
Figs. S1 to S5
Tables S1 to S8
References

24 December 2002; accepted 3 June 2003

Intracellular Bacterial Biofilm-Like Pods in Urinary Tract Infections

Gregory G. Anderson,[1]* Joseph J. Palermo,[1]* Joel D. Schilling,[1] Robyn Roth,[2] John Heuser,[2] Scott J. Hultgren[1]†

***Escherichia coli* entry into the bladder is met with potent innate defenses, including neutrophil influx and epithelial exfoliation. Bacterial subversion of innate responses involves invasion into bladder superficial cells. We discovered that the intracellular bacteria matured into biofilms, creating pod-like bulges on the bladder surface. Pods contained bacteria encased in a polysaccharide-rich matrix surrounded by a protective shell of uroplakin. Within the biofilm, bacterial structures interacted extensively with the surrounding matrix, and biofilm associated factors had regional variation in expression. The discovery of intracellular biofilm-like pods explains how bladder infections can persist in the face of robust host defenses.**

Urinary tract infections (UTIs) result in $1.6 billion in medical expenditures in the United States each year (*1*), with uropathogenic strains of *Escherichia coli* (UPEC) accounting for 70 to 95% of all UTIs (*2*). With the advance of multi–drug-resistant UPEC (*3*), it is important to determine the pathogenic mechanisms of UPEC. In animal models, UPEC pathogenesis initiates with bacterial binding of superficial bladder epithelial cells via the adhesin FimH at the tips of bacterially expressed type 1 pili (*4*). Initial colonization events activate inflammatory and apoptotic cascades in the epithelium, which is normally inert and only turns over every 6 to 12 months (*5*). Bladder epithelial cells respond to invading bacteria in part by recognizing bacterial lipopolysaccharide (LPS) via the Toll-like receptor 4 (TLR-4)–CD14 pathway, which results in strong neutrophil influx into the bladder (*6*). In addition, FimH-mediated interactions with the bladder epithelium stimulate exfoliation of superficial epithelial cells, causing many of the pathogens to be shed into the urine. Genetic programs are activated that lead to differentiation and proliferation of the underlying transitional cells in an effort to renew the exfoliated superficial epithelium (*7*). Despite the robust inflammatory response and epithelial exfoliation, UPEC are able to maintain high titers in the bladder for several days (*8–13*).

A bacterial mechanism of FimH-mediated invasion into the superficial cells apparently allows evasion of these innate defenses (*9*); subsequent replication as disorganized bacterial clusters inside superficial cells leads to high bacterial titers in the bladder. Bacteria in these intracellular niches [which we termed "bacterial factories" (*9*)] create a chronic quiescent reservoir in the bladder, which can persist undetected for several months without bacteria shedding in the urine. These bacteria are completely resistant to 3- and 10-day courses of antibiotics (*9, 14*). Thus, in addition to the intestine and vagina as reservoirs for UPEC, the bladder itself may serve as the source for recurrent cystitis and asymptomatic bacteriuria seen in a large proportion of women with UTIs (*9, 14, 15*).

To define bacterial-specific effects on UTI progression, we studied acute UTIs initiated by clinically isolated UPEC or laboratory (K-12) strains in TLR-4 mutant C3H/HeJ mice, which lack an intact innate immune response (*16, 17*). C3H/HeJ mice were inoculated with UPEC strain UTI89 (*9*) or type 1–piliated K-12 strain MG1655 (*18*), and numbers of colony-forming units (CFU) were determined in bladders at early time points after inoculation (fig. S1) (*10, 19*). While UTI89 levels increased nearly two orders of magnitude over 24 hours to about 6 × 10^6 CFU per bladder, MG1655 levels decreased over this time period to 10^3 CFU per bladder.

To investigate the increase in UPEC bacterial load at 24 hours, we performed scanning electron microscopy (SEM) (*8, 10*) of infected C3H/HeJ mouse bladders, which revealed numerous, large protrusions, or pods, on the surface of bladders infected with UPEC strain UTI89 (Fig. 1, A to C) (fig. S2). This was a rare event with the K-12 strain of *E. coli*, MG1655, because pods were not detected at this time point (Fig. 1D). In contrast, other clinical isolates such as UPEC strain NU14 (*9, 10*) also elicited abundant pod formation. SEM and hematoxylin and eosin (H&E) staining of the pods revealed that bacterial replication resulted in large bacterial colonies that extended above the lumenal surface (Fig. 1E). Video microscopy revealed that the previously described bacterial factories undergo a maturation process (*20*), whereby the loose collections of UPEC rods converted into a uniform coccoid morphology. This process was coupled with the organization of the bacteria into tightly packed biofilm-like pod structures (Fig. 1E) (*20*), Mutations in *fimH* completely abolish this pathogenic cascade (*10*).

[1]Department of Molecular Microbiology, [2]Department of Cell Biology and Physiology, Washington University School of Medicine, 660 South Euclid Avenue, St. Louis, MO 63110, USA.

*These authors contributed equally to this work.
†To whom correspondence should be addressed. E-mail: hultgren@borcim.wustl.edu