

100848
2
2003

This is to certify that the

dissertation entitled

THE EVOLUTION OF A BALANCED POLYMORPHISM IN A LONG-TERM
LABORATORY POPULATION OF ESCHERICHIA COLI

presented by

DANIEL E. ROZEN

has been accepted towards fulfillment
of the requirements for

Ph.D. degree in ZOOLOGY

Richard E. Unzicker

Major professor

Date December 13, 2000

LIBRARY
Michigan State
University

PLACE IN RETURN BOX to remove this checkout from your record.
TO AVOID FINES return on or before date due.
MAY BE RECALLED with earlier due date if requested.

DATE DUE	DATE DUE	DATE DUE

THE EVOLUTION OF A BALANCED POLYMORPHISM IN A LONG-TERM
LABORATORY POPULATION OF *ESCHERICHIA COLI*

BY

DANIEL E. ROZEN

A DISSERTATION

SUBMITTED TO
MICHIGAN STATE UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

DEPARTMENT OF ZOOLOGY

2000

ABSTRACT

THE EVOLUTION OF A BALANCED POLYMORPHISM IN A LONG-TERM LABORATORY POPULATION OF *ESCHERICHIA COLI*

By

Daniel E. Rozen

Attempts to understand the origin and maintenance of ecologically important genetic variation in natural populations is hampered by the fact that the generation times of most organisms prevent the direct observation and experimental manipulation of the long-term processes that influence such variation. In recent years, experiments with bacteria have been undertaken so that the ecological and evolutionary factors that influence this variation can be directly observed and manipulated. In this dissertation, I describe work on one such microbial system, wherein I examine the dynamical history, ecological mechanisms, and genetic bases of a balanced polymorphism that evolved in an experimental population of *Escherichia coli*. In Chapter 1, I describe the two variants, designated S and L, as well as elucidate the ecological mechanisms that enable them to coexist. S and L were isolated after 18,000 generations of laboratory evolution, although they are derived from divergent clades that are substantially older, having arisen between 3,000 and 6,000 generations. The S and L clones differ in average cell size and their maximum growth rate on glucose, the sole substrate provided. L's maximum growth rate exceeds S's by nearly 20%, yet S and L coexist in a frequency-dependent fashion because: S metabolizes one of more products that L secretes into the medium; and L exhibits higher mortality during periods of starvation, an effect that is increased by the presence of S.

When grown together, S and L achieve a stable equilibrium. However, over the long-term, their relative frequencies oscillated between about 10% and 90%. In Chapter 2, I examine the phylogenetic history and dynamics of adaptation of S and L. Phylogenetic reconstruction shows that S and L belong to different monophyletic clades. Following their respective origins, competition experiments and the dynamics of genetic variation within each clade show evidence for continued adaptation. Such adaptation appears to be responsible for fluctuations in their relative frequencies through evolutionary time. In Chapter 3, I identify five IS mutations that went to fixation in either S or L. Genotypes that differed in the allelic state of two mutations were constructed, and their direct effects on competitive fitness, and on the balanced polymorphism, were determined. I found that one of the two mutations is beneficial in S whereas the other is neutral, and neither mutation markedly influences the frequency-dependent coexistence of S and L. However, the fitness effects of both mutations are dependent on genetic background (epistatic), leaving open the possibility that the observed fitness effects do not fully reflect the significance of these mutations when they arose in the S clade. In Chapter 4, I use DNA microarrays to identify genes and pathways that are involved in the adaptive evolution of S and L. I conducted three paired comparisons: 1) S and L during exponential growth; 2) S grown alone and S grown in the presence of L secretions; and 3) S and L during stationary phase. Relative gene expression differs dramatically in the first two comparisons. This evidence suggests that genes with global regulatory effects, and thus extensive pleiotropy at the level of gene expression, may be important for the genetic divergence and ecological coexistence of S and L.

ACKNOWLEDGEMENTS

This dissertation could not have been completed without the continuous encouragement and advice provided by my thesis advisor, Dr. Richard Lenski. Over the years Rich has been hugely generous with his time, extensive (though always constructive!) in his critiques, and unwavering in his support. He has given me enormous and unprecedented independence to think and work on a variety of projects, often on topics quite distinct from those considered in this dissertation. Such independence can come at a cost, of course, and I am deeply grateful that Rich saw my "side-projects" as essentially fruitful rather than essentially distracting. This generous view allowed me the freedom to both succeed and fail, and offered me a most realistic view of my scientific future. I hope that I am able to provide similarly valuable mentorship to my own students, when (and if) such a time comes.

In addition to the direct advice that Rich offered, he also provided exceptional indirect support by surrounding me with a uniformly superb group of labmates. Both in and out of the lab, this group of extremely fun and talented people has made my experience at Michigan State University remarkable. Santi Elena and Arjan de Visser played an early, and important, role on my ability to ask and answer questions. Judy Mongold taught me the necessity for patience in my judgement of colleagues and ideas. Vaughn Cooper was critical at all stages--intellectually, socially, and athletically. Susi Remold continuously reminded me to consider the broad relevance of our work; I may learn something about the "sheep on the mountain" yet. Paco Moore, Dominique Schneider, and Tim Cooper

have been magnificent collaborators. Phil Gerrish, my lifelong math-guy, showed me the importance of mathematical rigor (despite my complaints) and repeatedly steered me from the folly of my intuition. More recently, Charles Ofria and Elizabeth Ostrowski forced me to reexamine old issues, and to consider new routes towards their solutions. Lynette Ekunwe and Neerja Hajela provided invaluable technical and logistical support. And many thanks to all other labmates who made my daily trip to work something that I anticipated with pleasure rather than dread.

Life outside lab was enriched considerably by a number of great friends. Of particular note are Sam Hazen, Heather Rowe and Erin O'Bryant. They have each seen me at my worst, and yet stuck around for more of my nonsense! A "Thank you" is hardly sufficient. Thanks also to the running group: Jim Hancock, Hal Prince, Vaughn Cooper, and Judy Kolkman; and to squash group: Tim Cooper, Rich Lenski, Paco Moore, and Matthew Collett. I could not have made it without these daily opportunities to forget about work entirely.

Finally, I offer my most heartfelt thanks to my parents, Jack and Rosalie Rozen for their years of encouragement. It has not always been clear that I would ever obtain my Ph.D. Their continuous support and patience made my efforts seem worthwhile and reasonable.

TABLE OF CONTENTS

List of Tables	viii
List of Figures	ix
Introduction	1
Chapter 1: The evolution and maintenance of a balanced polymorphism in a long-term evolving population of <i>Escherichia coli</i>	11
Methods	15
Results	19
Discussion	34
Literature cited	45
Chapter 2: The phylogenetic history of a balanced polymorphism in a long-term evolving population of <i>Escherichia coli</i>	48
Methods	52
Results	57
Discussion	68
Literature cited	71
Chapter 3: The role of IS mutations in the evolution of a balanced Polymorphism in a laboratory population of <i>Escherichia coli</i>	76
Methods	79
Results	86
Discussion	102
Literature cited	109

Chapter 4: Exploring the utility of microarrays for identifying causes of adaptive differences between S and L	112
Methods	114
Results	119
Discussion	148
Literature cited	154

LIST OF TABLES

Table 1. Bootstrap support for monophyly of S and L using only clones collected from specified time points.	61
Table 2. Relevant properties of strains used to examine the role of <i>menC</i> and b2875	80
Table 3. Genomic location of IS mutations that became fixed in S and L	87
Table 4. Analysis of covariance for fitness of S competed against L, with and without supplemented menaquinone.	92
Table 5. Analysis of epistatic effects between two IS mutations, <i>menC</i> and b2875	98
Table 6. Three-way analysis of variance of fitness of S and S-derived mutants when competed against L.	101
Table 7. Comparison of expression differences between S and L during exponential growth in DM25.	129
Table 8. Comparison of expression differences between S growing in DM25 and S growing in DM25 conditioned by L cells.	131
Table 9. Comparison of expression differences between S and L during stationary phase.	133
Table 10. Expression differences between S and L during exponential growth in DM25 for genes involved in "Transport and Binding".	135
Table 11. Genes regulated by cAMP that differ between S and L during growth in DM25.	139
Table 12. Genes regulated by cAMP that differ between S growing in DM25 and S growing in DM25 conditioned by L cells.	143
Table 13. Genes regulated by <i>rpoS</i> that differ between S growing in DM25 and S growing in DM25 conditioned by L cells.	146

LIST OF FIGURES

Figure 1. Relative fitness of S and L during short term competition.	22
Figure 2. S and L convergence on a stable equilibrium during 20 growth cycles (~ 130 generations).	23
Figure 3. Frequency-dependent advantage of S versus L during both growth and stationary phase.	25
Figure 4. Rates of change of S and L densities during stationary phase.	27
Figure 5. Maximum growth rate of S and L when growing on metabolites secreted into the culture medium.	30
Figure 6. Long-term dynamics of the S and L polymorphism.	33
Figure 7. Hypothetical models for the evolutionary history of S and L	50
Figure 8. Neighbor-joining phylogeny of S and L based upon 5,000 bootstrap replicates.	59
Figure 9. Trajectories of invasion of S and L.	63
Figure 10. Time course of genetic variation of S and L, as calculated from pairwise genetic distances within samples.	65
Figure 11. Change in mean fitness between 13,000 and 17,000 generations within S and L.	67
Figure 12. Frequency of new IS mutations in the S clade.	89
Figure 13. Frequency-dependent fitness of S versus L, conducted in DM25 and in medium supplemented with menaquinone.	91
Figure 14. Fitness effects of <i>menC</i> allelic replacement in Anc, S, and L.	94
Figure 15. Fitness effects of b2875 allelic replacement in Anc, S, and L.	96
Figure 16. Fitness effects of double mutants containing both <i>menC</i> and b2875 allelic replacements in Anc, S, and L.	97
Figure 17. Frequency-dependent relative fitness of S, S/ <i>menC</i> ⁺ , S/b2875 ⁺ , and S/ <i>menC</i> ⁺ /b2875 ⁺ , each competed against L.	100

Figure 18. Scatter plots of expression values, and histograms of relative expression of S and L growing exponentially in DM25.	121
Figure 19. Scatter plots of expression values, and histograms of relative expression of S grown in DM25 and S grown in DM 25 that has been conditioned by L cells.	123
Figure 20. Scatter plots of expression values, and histograms of relative expression of S and L during stationary phase in DM25.	125

INTRODUCTION

A major research emphasis in evolutionary ecology is to understand the origin and persistence of the abundant ecological and genetic variation found in natural populations (Futuyma 1998; Hartl 1988). In few cases, however, has it been possible to study both the evolutionary factors that influence the emergence of polymorphism and their ecological consequences. This results from the fact that the long generation times of most organisms limit study of the fate and consequences of new mutations to the short term. However, evolutionary and ecological factors are inextricably tied and interact over long time periods to generate extant patterns of diversity. It is thus necessary to develop systems that allow ecological and evolutionary factors to be considered simultaneously and over the long term. Recent experimental work with microbes has been developed with these aims in mind (Rainey et al. 2000).

Microbes are ideal for addressing questions of fundamental ecological and evolutionary importance (Dykhuizen 1990; Lenski 1995). Among the reasons for this are that microorganisms are simple to culture, have large populations with short generation times, and have relatively simple genetic systems that are easy to manipulate. These attributes allow one to create defined genotypes in order to measure performance and fitness (Chao and Levin 1981; Dykhuizen and Dean 1990; Elena and Lenski 1997) and to study natural selection acting on newly arising mutations (Dykhuizen, 1990; Helling et al. 1988; Korona et al. 1994; Lenski et al. 1991; Velicer et al. 1998). One can also construct and

examine the long or short term dynamics of controlled ecological interactions (Bohannan and Lenski 1997; Chao et al. 1977). A final advantage of such systems is that they reduce the complexity of the "real world" while not eliminating it altogether. It is consequently easier to identify the mechanisms that have resulted in specific ecological and evolutionary outcomes (Rainey et al. 2000).

Populations of asexual organisms that are evolving in unstructured environments provisioned with a single limiting resource are predicted to remain effectively monomorphic, though not evolutionarily static (Atwood et al. 1951; Levin 1981). Increases in population fitness occur via the process of periodic selection whereby beneficial mutations arise and increase to fixation in a sequential manner. This is also known as a selective sweep. Although this process is not specific to asexual populations, its consequences differ between sexual and asexual systems. Whereas selective sweeps in sexual species will only cause local reductions in genetic diversity (i.e. in genomic regions closely linked to the beneficial mutation) (Begun and Aquadro 1992; Hudson et al. 1997), selective sweeps in asexual species cause the population-wide elimination of genetic variation (Dykhuizen 1990) because the entire asexual genome is a single linkage unit. Because of periodic selection, genetic variation in strictly asexual populations is presumed to be transient. The ecological principle of "competitive exclusion" states that complete competitors cannot coexist (Hardin 1960), which also leads to the assumption that simple microbial populations will remain monomorphic.

Certain situations, however, can promote the evolution of stable polymorphism in asexual populations. For example, Chao et al.(1977) observed the evolution of *E. coli* mutants that were resistant to viral infections, which were then able to coexist with their susceptible progenitors in a predator mediated fashion. Rainey and Travisano (1998) demonstrated the evolution of stable polymorphism in structured populations of *Pseudomonas fluorescens*, which resulted from differential competitive ability of evolved morphs in distinct spatial niches. And Helling et al. (1988) and Turner et al. (1996) observed the evolution of *E. coli* genotypes that coexisted in a frequency dependent manner resulting from cross-feeding, where metabolites secreted by a competitively dominant genotype were selectively utilized by a second genotype.

My dissertation work examined a balanced polymorphism that arose during a long-term evolution experiment with *E. coli*. In this experiment twelve replicate populations, initiated from a single genotype, have been serially propagated in a glucose limited minimal medium for more than 20,000 generations (Cooper and Lenski 2000; Lenski et al. 1991; Lenski and Travisano 1994). A survey conducted after 10,000 generations of evolution found that populations harbored substantially more genetic variation for fitness than would be expected based upon the mutation rates of *E. coli* and on the rate of population adaptation observed at that time (Elena and Lenski 1997). Instead, most variation could be attributed to frequency-dependent selection of the sort known to enable stable polymorphism (Levin 1988). In this dissertation, I describe my work on the single population that exhibited the most extreme frequency-dependence.

I isolated two clones, called S and L, from this population that had been evolving for 18,000 generations. S and L differ in a number of heritable traits, such as cell size and an approximately 20% higher maximum growth rate of L on the sole substrate provided during this long term experiment. The finding that L had a significantly higher maximum growth rate than S argued against the possibility that S and L could coexist, because maximum growth rate is an especially important component of fitness in the experimental environment (Vasi et al. 1994). However, when S and L were competed versus one another, we found that the fitness of both morphs was frequency-dependent; that is, both S and L could invade one another from initial rarity. Indeed, over the course of a few weeks, not only could S coexist with L, but it attained a slightly higher frequency when S and L achieved equilibrium.

I identified two important factors that enabled S and L to coexist: cross-feeding and differential death during stationary phase. During growth, L cells (and to a lesser extent S cells) secrete one or more products upon which S cells can grow and increase their growth rates. This phenomenon is termed cross-feeding. L cells do not use the products that they secrete. Also, during stationary phase, L cells die at a higher rate than S cells, an effect that is increased by the presence of S. I did not determine whether S was producing an allelopathic substance that was toxic to L, or if S was removing a substance that contributed to the viability of L during stationary phase. However, allelopathic production of a toxin, by itself, would not provide a selective advantage to an invading genotype in a mass-action environment (Chao and Levin 1981), which may indirectly

favor the hypothesis that S depletes some nutrient necessary for the survival of L.

My initial work studied S and L isolated after 18,000 generations of evolution. However, I found that S and L are derived from divergent clades that are substantially older, having arisen between 3,000 and 6,000 generations. By using RFLP genetic fingerprinting with Insertion Sequences (IS) (Lawrence et al. 1989; Papadopoulos et al. 1999), I examined the phylogenetic history of S and L and found that both morphs were monophyletic. In addition, despite the stability observed during short term competition, S and L frequencies have been dynamic through time with their relative frequencies shifting repeatedly between 10% and 90%.

Following their respective origins, the dynamics of genetic variation within each clade show evidence for continued, independent adaptation. Competition experiments between S and L clones sampled from different time points confirmed the continued adaptation of each clade following divergence. This continued adaptation may be responsible for fluctuations in their relative frequencies through evolutionary time.

In the phylogenetic study of S and L, IS were used as markers to provide information about the history of both groups. However, three features suggested that the IS mutations might themselves be causally linked to the adaptive changes that have occurred in both clades. First, a series of IS mutations were derived and became fixed within each clade. Second, the time of appearance of some IS mutations coincided closely with the first

observation of the S clade. Finally, recent work in this (Cooper 2000) and other systems has found evidence for IS mediated beneficial mutations (Treves et al. 1998).

Consequently, I sought evidence that IS mutations were causally involved in the evolutionary dynamics of S and L.

Genotypes that differed in the allelic state of two of five characterized mutations were constructed using allelic replacement, and their direct effects on competitive fitness and on the balanced polymorphism, were determined. I found that one of the two mutations is beneficial in S whereas the other is neutral and neither mutation markedly influences the frequency-dependent coexistence of S and L. However, the fitness effects of both mutations are highly dependent on genetic background (epistatic), leaving open the possibility that the observed fitness effects do not fully reflect the evolutionary significance of these mutations when they first arose in the S clade.

In addition to the approaches above, I examined the utility of DNA microarrays for identifying genes and pathways that may be involved in the divergence and adaptive evolution of S and L. Though microarrays have been primarily used to discover the function and regulation of newly identified genes (Arfin et al. 2000; Chu et al. 1988; deRisi et al. 1997; Duggan et al. 1999; Richmond et al. 1999; Tao et al. 1999), they have also been used by biologists to gain insight into the mechanistic basis of evolution (Ferea et al. 1999). Using this approach, gene expression can be monitored and compared across genotypes with distinct evolutionary histories and fitness levels. Genes whose

expression is increased or decreased across genotypes are genes whose products may be causally associated with fitness differences and are candidates for further manipulation.

I conducted three paired comparisons with gene arrays in order to begin to understand the genetic and phenotypic bases of S and L coexistence. Each comparison corresponded to factors that contribute to the coexistence of the S and L clades. In Experiment 1, I compared the expression profiles of S and L during exponential growth. In Experiment 2, I compared the expression profiles of S cells grown alone and in the presence of L secretions (L conditioned media). Finally, in Experiment 3, I compared gene expression of S and L during stationary phase.

Relative gene expression differs dramatically between S and L during growth, and between S and S grown in L secretions. The vastness of these differences impaired my ability to identify single loci that were critical for the evolution of S and L. Evidence suggests that genes with global regulatory effects may be important for some S and L differences.

In this dissertation, I have described the processes that have occurred over the nearly 20,000 generation history of a balanced polymorphism that has evolved in a simple laboratory environment. A variety of approaches--ecological, traditional genetic and molecular genetic-- have been used to understand both the origin of S and L as well as their dynamic persistence through time.

Literature Cited

- Arfin, S. M., A. D. Long, E. T. Ito, L. Toller, M. M. Riehle, E. S. Paegle, and G. W. Hatfield. 2000. Global gene expression profiling in *Escherichia coli* K12: The effects of integration host factor. *Journal of Biological Chemistry* 275:29672-29684.
- Atwood, K. C., L. K. Schneider, and F. J. Ryan. 1951. Periodic selection in *Escherichia coli*. *Proceedings of the National Academy of Sciences of the USA* 37:146-155.
- Begun, D. J., and C. F. Aquadro. 1992. Levels of naturally occurring DNA polymorphism correlate with recombination rates in *D. melanogaster*. *Nature* 356:519-520.
- Bohannon, B. J. M., and R. E. Lenski. 1997. Effect of resource enrichment on a chemostat community of bacteria and bacteriophage. *Ecology* 78:2303-2315.
- Chao, L., and B. R. Levin. 1981. Structured habitats and the evolution of anticompetitor toxins in bacteria. *Proceedings of the National Academy of Sciences, USA* 78:6324-6328.
- Chao, L., B. R. Levin, and F. M. Stewart. 1977. Complex community in a simple habitat: an experimental study with bacteria and phage. *Ecology* 58:369-378.
- Chu, S., J. DeRisi, M. Eisen, J. M. Holland, D. Botstein, P. O. Brown, and I. Herskowitz. 1988. The transcriptional program of sporulation in budding yeast. *Science* 282:699-705.
- Cooper, V. S. 2000. Consequences of ecological specialization in long-term evolving populations of *Escherichia coli*. Ph.D. dissertation. Michigan State University, East Lansing, MI.
- Cooper, V. S., and R. E. Lenski. 2000. The population genetics of ecological specialization in evolving *E. coli* populations. *Nature* 407:736-739.
- deRisi, J. L., V. R. Iyer, and P. O. Brown. 1997. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* 278:680-686.
- Duggan, D. J., M. Bittner, Y. Chen, P. Meltzer, and J. M. Trent. 1999. Expression profiling using cDNA microarrays. *Nature Genetics* (suppl.) 21:10-14.

- Dykhuizen, D. E. 1990. Experimental studies of natural selection in bacteria. *Annual Review of Ecology and Systematics* 21:373-398.
- Dykhuizen, D. E., and A. M. Dean. 1990. Enzyme activity and fitness--evolution in solution. *Trends in Ecology & Evolution* 5:257-262.
- Elena, S. F., and R. E. Lenski. 1997. Long-term experimental evolution in *Escherichia coli* .VII. Mechanisms maintaining genetic variability within populations. *Evolution* 51:1058-1067.
- Ferea, T. L., D. Botstein, P. O. Brown, and R. F. Rosenzweig. 1999. Systematic changes in gene yeast expression patterns following adaptive evolution in yeast. *Proceedings of the National Academy of Sciences, USA* 96:9721-9726.
- Futuyma, D. J. 1998. *Evolutionary Biology*. Sinauer Associates. Sunderland, Mass.
- Hardin, G. 1960. The competitive exclusion principle. *Science (Washington, D. C.)* 131:1292-1297.
- Hartl, D. L. 1988. *A primer of population genetics*. Sinauer Associates, Inc., Sunderland, Mass.
- Helling, R. B., C. N. Vargas, and J. Adams. 1988. Evolution of *Escherichia coli* during growth in a constant environment. *Genetics* 116:349-358.
- Hudson, R. R., A. G. Saez, and F. J. Ayala. 1997. DNA variation at the *Sod* locus of *Drosophila melanogaster*: An unfolding story of natural selection. *Proceedings of the National Academy of Sciences of the USA* 94:7725-7729.
- Lawrence, J. G., D. E. Dykhuizen, R. F. Dubose, and D. L. Hartl. 1989. Phylogenetic Analysis Using Insertion-Sequence Fingerprinting in *Escherichia-Coli*. *Molecular Biology and Evolution* 6:1-14.
- Lenski, R. E. 1995. Molecules are more than markers: new directions in molecular microbial ecology. *Molecular Ecology* 4:643-651.
- Lenski, R. E., M. R. Rose, S. C. Simpson, and S. C. Tadler. 1991. Long-term experimental evolution in *Escherichia coli* .I. Adaptation and divergence during 2,000 generations. *American Naturalist* 138:1315-1341.
- Lenski, R. E., and M. Travisano. 1994. Dynamics of adaptation and diversification-a 10,000-generation experiment With bacterial populations. *Proceedings of the National Academy of Sciences, USA* 91:6808-6814.

- Levin, B. R. 1981. Periodic selection, infectious gene exchange and the genetic structure of *E. coli* populations. *Genetics* 99:1-23.
- Levin, B. R. 1988. Frequency-dependent selection in bacterial populations. *Philosophical Transactions of the Royal Society of London B, Biological Sciences* 319:459-472.
- Papadopoulos, D., D. Schneider, J. Meier-Eiss, W. Arber, R. E. Lenski, and M. Blot. 1999. Genomic evolution during a 10,000-generation experiment with bacteria. *Proceedings of the National Academy of Sciences, USA* 96:3807-3812.
- Rainey, P. B., A. Buckling, R. Kassen, and M. Travisano. 2000. The emergence and maintenance of diversity: insights from experimental bacterial populations. *Trends in Ecology & Evolution* 15:243-247.
- Rainey, P. B., and M. Travisano. 1998. Adaptive radiation in a heterogeneous environment. *Nature* 394:69-72.
- Richmond, C. S., J. D. Glasner, R. Mau, H. Jin, and F. R. Blattner. 1999. Genome-wide expression profiling in *Escherichia coli* K-12. *Nucleic Acids Research* 27:3821-3835.
- Tao, H., C. Bausch, C. Richmond, F. R. Blattner, and C. Conway. 1999. Functional genomics: Expression analysis of *Escherichia coli* growing on minimal and rich media. *Journal of Bacteriology* 181:6425-6440.
- Treves, D. S., S. Manning, and J. Adams. 1998. Repeated evolution of an acetate-crossfeeding polymorphism in long-term populations of *Escherichia coli*. *Molecular Biology and Evolution* 15:789-797.
- Turner, P. E., V. Souza, and R. E. Lenski. 1996. Tests of ecological mechanisms promoting the stable coexistence of two bacterial genotypes. *Ecology* 77:2119-2129.

Chapter I

THE EVOLUTION AND MAINTENANCE OF A BALANCED POLYMORPHISM IN A LONG-TERM EVOLVING POPULATION OF *ESCHERICHIA COLI*

The use of bacteria and other microorganisms to address questions of fundamental ecological and evolutionary importance has substantially increased in recent years (Dykhuizen 1990; Lenski 1995). Among the reasons for this are that certain microbes are easy to culture, they exist in large populations with short generation times, and they have simple genetic systems that are relatively easy to manipulate. These attributes allow one to create defined genotypes in order to measure their performance and fitness (Dykhuizen and Hartl 1980; Chao and Levin 1981; Dykhuizen and Dean 1990; Elena and Lenski 1997), as well as to study natural selection acting on spontaneous mutants (Helling et al. 1987; Lenski et al. 1991; Bennett et al. 1992; Lenski and Travisano 1994; Velicer et al. 1998). One can also construct communities to examine the dynamics and stability of ecological interactions (Chao et al. 1977; Rosenzweig et al. 1994; Bohannan and Lenski 1997). A potential concern is that these systems are so artificial that they may prevent the emergence of complexity and thereby limit the insights that can be drawn from them. In this paper, and following earlier studies (Helling et al. 1987; Rosenzweig et al. 1994; Turner et al. 1996), we demonstrate the emergence of a stable polymorphism even in a simple environment. Moreover, we show that the dynamics of this polymorphism, while

fairly simple over short intervals, become very complex over the long term.

Bacteria reproduce asexually, and it is often assumed that their evolution on a single limiting resource will consist of a temporal series of replacements by ever more fit genotypes, via the process of “periodic selection” (Atwood et al. 1951; Koch 1974; Levin 1981). Each selective replacement creates a bottleneck of one contributor in an asexual population, eliminating all genetic variation. Accordingly, any polymorphisms that are ecologically significant (in contrast to neutral or deleterious alleles maintained by recurrent mutation) are presumed to be transient and indicative of selective sweeps in progress. The competitive exclusion principle, according to which two competitors cannot coexist indefinitely on one limiting resource (Hardin 1960), also implies that microbial populations evolving to become better competitors for a single limiting resource will be monomorphic. Thus, two processes – one genetic (periodic selection) and the other ecological (competitive exclusion) – should maintain monomorphism in asexual microbial populations as they evolve under simple and uniform laboratory regimes.

Certain circumstances, however, can promote the evolution of ecologically stable polymorphisms in asexual populations. For example, Chao et al. (1977) observed the evolution of *E. coli* mutants that were resistant to viral infections, which then stably coexisted with their sensitive progenitors in a predator-mediated manner. Helling et al. (1987) reported the emergence of a stable polymorphism in *E. coli* populations that were

propagated in a chemostat on a single resource; they showed that the polymorphisms were maintained by cross-feeding interactions, in which secondary resources are secreted as metabolic by-products of a primary resource (Rosenzweig et al. 1994). Turner et al. (1996) observed the coexistence of two *E. coli* strains in a serial transfer regime; a cross-feeding interaction and a tradeoff in relative growth rate at high and low resource concentrations were both implicated (see also Levin 1972).

The present study examines the emergence, ecological mechanisms, and evolutionary dynamics of a stable polymorphism that arose during a long-term evolution experiment with *E. coli* (Lenski et al. 1991; Lenski and Travisano 1994; Vasi et al. 1994; Travisano et al. 1994; Elena et al. 1996; Travisano and Lenski 1996; Elena and Lenski 1997). In that experiment, twelve replicate populations were serially propagated in a glucose-limited minimal medium in a constant batch-culture environment. Previous papers in this series reported on the dynamics of genetic adaptation, and on the extent of variation *within* and *among* the evolving populations. The replicate populations exhibit substantial differences from one another in certain phenotypic traits, such as average cell size and performance in novel environments. By contrast, they are very similar (but not identical) to one another in the extent of their fitness improvement measured in the selective environment itself.

Throughout the 20,000 generations of this experiment, there emerged an interesting temporal pattern of fitness variation within the evolving populations. During the initial

2000 generations – the period of most rapid adaptation – the extent of within-population variation in fitness corresponded closely to the level predicted by Fisher’s fundamental theorem from the observed rate of adaptation (Lenski et al. 1991). In other words, it was unnecessary to invoke any ecologically significant polymorphism during this early phase, beyond the transient variation that must occur whenever beneficial mutations sweep through a population. After 10,000 generations, however, the situation had become much more complex and interesting; the rate of genetic adaptation declined substantially relative to the earlier phase, whereas the variation in performance among clones within a population remained high (Elena and Lenski 1997). Only ~1% of the within-population variation could then be explained by on-going selective sweeps, whereas previously all the variation could be thus explained. Another modest fraction, about 10%, of the within-population variation for fitness at generation 10,000 could be attributed to deleterious mutations, which had become more common after some of the replicate populations had evolved much higher mutation rates during this period (Sniegowski et al. 1997). Most of the variation in performance was attributed instead to frequency-dependent selection of the form that promotes balanced polymorphism. Negative frequency-dependent selection occurs when the fitness of a genotype is highest when that genotype is rare. This form of frequency-dependence has often been invoked to explain the maintenance of stable polymorphisms in nature (Ayala and Campbell 1974; Levin 1988). By performing experiments in which marked clones were re-introduced at variable frequencies into the populations from which they had been sampled, Elena and Lenski (1997) showed that the marked clones had, on average, higher fitness when they

were rare than when they were common in all six of the populations they studied. In five of the populations, the average advantage when rare was small ($\sim 1\%$), but in one population the fitness advantage when rare was much greater ($\sim 5\%$).

The present paper focuses on the population that showed the most extreme frequency-dependent selection. We demonstrate that there are two predominant morphs in this population, and we confirm that each morph does indeed have a strong selective advantage when it is rare, such that there exists a stable polymorphism. We show that the two morphs can be distinguished on the basis of several phenotypic differences, and we examine which differences can explain the stable polymorphism. Finally, we document when the balanced polymorphism arose during the population's history, and we show that this seemingly stable polymorphism has in fact exhibited unexpectedly complex dynamics over the duration of its existence.

Materials and Methods

Bacterial Strains

The genotypes used in this study were derived from a single clone of *Escherichia coli* B that has been serially propagated for almost 20,000 generations in glucose-limited batch culture (see Lenski et al. 1991 for description of the ancestral strain). Throughout the course of this long-term experiment, samples taken from the evolving populations have

been periodically spread as individual cells onto petri plates to estimate population size, and to examine the populations for possible contamination. During a routine examination of the populations at generation 18,000, two morphotypes could be distinguished in one of the evolving populations (designated Ara-2). This same population had previously been shown to harbor significant genetic variation in performance that was maintained by frequency-dependent selection (Elena and Lenski 1997). The two morphotypes differed in their colony size and time of appearance on tetrazolium-arabinose (TA) indicator agar plates at 37°C. Per Liter, TA plates contain 10 g Tryptone, 1 g yeast extract, 5 g NaCl, 16 g agar, 10 g arabinose, and 1 mL of a 5% stock of tetrazolium (2,3,4-triphenyltetrazolium chloride). The large type (L) produced visible colonies ~24 hours after plating, whereas colonies of the small type (S) were visible only after ~48 hours. Representative clones of each morphotype from generation 18,000 were sub-cultured on two TA plates prior to storage in 15% glycerol at -80°C. This procedure ensured that we had obtained a single clone of each type, and it also indicated that the distinctive colony morphologies of L and S were heritable and stable. We determined that neither S nor L were contaminants by examining their phenotypes with respect to markers specific to the experimental populations (Lenski et al. 1991). We will present additional data in the Results to show that phenotypic differences between L and S are genetically based.

The L and S clones were isolated from a population that was founded by an ancestral strain unable to metabolize arabinose; consequently, both S and L were phenotypically

Ara⁻. To facilitate counting the L and S clones during competition experiments, we isolated Ara⁺ mutants of both types by plating about 10⁸ cells on minimal-arabinose agar (Lenski 1988). The Ara⁺ mutants retained their characteristic colony morphologies. They are designated S⁺ and L⁺, and they too were stored in 15% glycerol at -80°C. When samples from competition experiments are spread on TA agar, Ara⁻ strains produce red colonies, whereas Ara⁺ colonies are white (Miller 1992).

Growth Conditions

Unless otherwise noted, bacteria were cultured in Davis minimal medium supplemented with thiamine hydrachloride (at 2×10^{-3} µg/mL) and glucose at 25 µg/mL (hereafter, DM25). This medium supports a stationary-phase population density of $\sim 5 \times 10^7$ cells/mL. In all experiments, 10 mL cultures were maintained in 50mL Erlenmeyer flasks placed in a rotary shaker at 37°C and 120 rpm. Each day, 0.1mL was transferred from the stationary-phase culture into 9.9 mL of fresh medium. This 100-fold dilution and re-growth allowed ~ 6.64 generations of binary fission per day ($\log_2 100 \approx 6.64$). In competition experiments, specified ratios of each genotype were mixed and then diluted, so that each culture received the same initial density of cells as in the long-term evolution experiment.

Competition Experiments and Fitness Estimation

Competition experiments were performed to determine the relative fitness of the L and S clones. Before doing so, however, it was necessary to test the neutrality of the Ara⁺ mutants of both L and S relative to their isogenic Ara⁻ progenitors. Prior to each fitness assay, each competitor was separately grown for one full day in DM25; this acclimation step was used to ensure that both competitors were in similar physiological states and at similar cell densities. Following the acclimation step, either S⁺ and S⁻ or L⁺ and L⁻ were mixed at a 1:1 ratio and diluted into fresh DM25. Initial and final (after 1 d) densities of each competitor were determined from counts on TA agar. The fitness of one genotype relative to the other was calculated as the ratio of their Malthusian parameters, which for each genotype was estimated by $m_i = \ln[N_i(1) / N_i(0)] / (1d)$, where $N_i(0)$ and $N_i(1)$ are initial and final densities, respectively (Lenski et al. 1991). We performed 13 replicate competition experiments for each morphotype. The fitness of S⁻ relative to S⁺ was 1.004 ± 0.012 (mean \pm SE), and the fitness of L⁻ relative to L⁺ was 1.022 ± 0.022 . Neither value is significantly different from 1.0 (S: $t = 0.32$, $df = 12$, $P = 0.755$; L: $t = 1.04$, $df = 12$, $P = 0.320$), indicating that the Ara marker is effectively neutral on each background. In all subsequent competition experiments, we used only S⁺ and L⁻, which henceforth are denoted simply as S and L. To examine frequency-dependent interactions between S and L, we used the same basic protocol as described above, except that the competition experiments were inoculated at three different initial ratios of S and L, which were 9:1, 1:1, and 1:9. Each treatment was replicated ten-fold.

Measurements of Maximum Growth Rate and Average Cell Size

The maximum growth rate, V_m , of each genotype was measured under standard culture conditions using a Coulter electronic particle counter (model ZM and channelyzer model 256). The glucose concentration in DM25 (25 $\mu\text{g/mL}$) has been shown to be well above that which limits growth rate (Vasi et al. 1994). Four replicate cultures of each clone were grown to stationary phase in DM25, and each culture was then diluted 100-fold into fresh DM25. Beginning 2 h after transfer, cell counts were obtained every half-hour until the rate of population growth began to slow appreciably due to depletion of the limiting glucose. The maximum growth rate of each culture was estimated by regressing the natural logarithm of population density against time, using only those time points over which population density increased log-linearly.

The Coulter counter was also used to measure average cell size during stationary phase in DM25. For each genotype, ten replicate cultures were allowed to complete the standard 24-h propagation cycle; in this cycle glucose is typically exhausted from the medium between 8-12 hours of growth. The individual cell volumes of 10^4 - 10^5 bacteria were obtained from each culture; however, for the purpose of statistical analysis, the mean cell size from each independent culture was the unit of replication.

Results

L and S Differ in Average Cell Size

The two genotypes, L and S, were originally distinguished by the size of their colonies and the time of their appearance on agar plates. They also differ in the average volume (1

fL = 10^{-15} L) of individual cells measured at stationary phase. The average cell size for L is 1.251 ± 0.051 fL (mean \pm SE, based on ten replicate cultures), whereas for S it is 0.700 ± 0.008 . This difference is highly significant using Welch's approximate t -test, which takes into account their unequal variances ($t = 10.619$, 9 df, $P < 0.0001$). Thus, the S genotype has smaller individual cells, as well as smaller colonies, than does the L type. This significant morphological differentiation at the level of cells further indicates the heritable nature of the polymorphism.

Maximum Growth Rate of L is Greater than that of S

We measured the maximum exponential growth rates of both genotypes in DM25, with four-fold replication of paired cultures. The maximum growth rate for L is 1.079 ± 0.003 h⁻¹ (mean \pm SE), while that for S is only 0.896 ± 0.009 h⁻¹. This difference is highly significant (paired $t = 17.80$, 3 df, $P = 0.0004$).

Stability of the Polymorphism

The population sample from which the clones L and S were isolated contained substantial frequencies of both morphotypes. The finding that L has a significantly higher maximum growth rate than does S tends to argue against the stable maintenance of the polymorphism, because maximum growth rate is an especially important component of fitness in the serial transfer regime used during the experimental evolution

experiment(Vasi et al. 1994). One reasonable interpretation, therefore, is that the polymorphism is transient, with the superior genotype L having been caught in the middle of its selective replacement of the inferior genotype S. Alternatively, S may have some countervailing advantage at some other stage of the growth cycle, which may allow it to persist despite its lower maximum growth rate. If that is the case, then the relative fitness of S and L in competition over the entire growth cycle may depend on their initial frequencies and the polymorphism may be maintained at some stable equilibrium.

Relative Fitness is Frequency-Dependent. To test whether the fitness of S relative to L depends on their relative abundance, we performed competition experiments at three different starting ratios of S and L (1:9, 1:1, and 9:1), each with ten-fold replication. Figure 1 shows that the fitness of S relative to L is greater than 1.0 when S is rare, whereas the same relative fitness is less than 1.0 when S is common. An ANOVA indicates that the effect of initial frequency on relative fitness is significant ($F = 19.35$, 2 and 27 df, $P < 0.0001$). Thus, each genotype has a selective advantage when it is rare, which implies a stable equilibrium.

Existence of a Stable Equilibrium. We then propagated mixtures of S and L by daily serial dilution in DM25 for 20 days (~ 130 generations). Samples from each mixed population were spread onto TA agar every day to determine the frequency of both types. Figure 2 shows that each genotype can invade the other when it is initially rare, and that

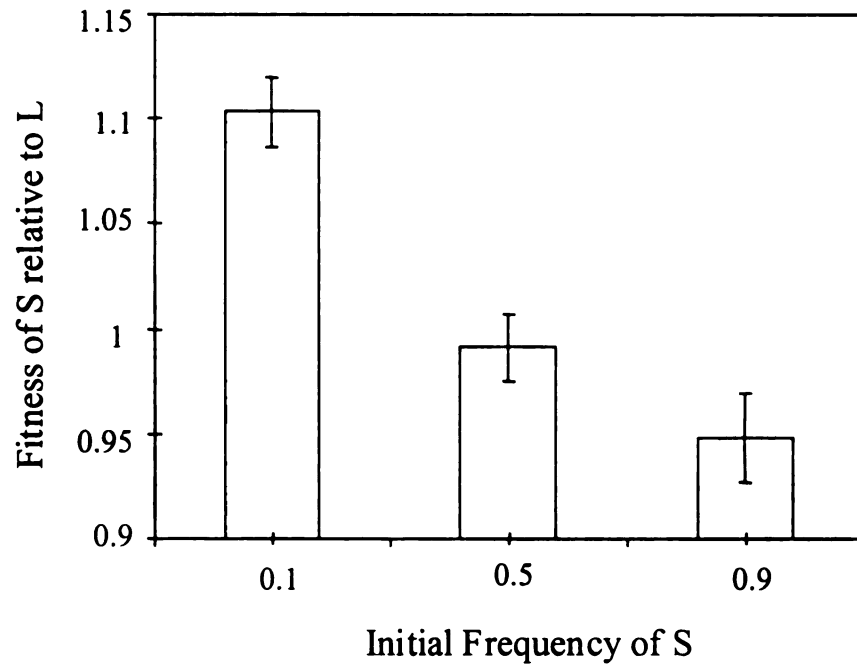


Figure 1: Relative fitness is frequency-dependent in short-term competition experiments. The fitness of genotype S relative to genotype L is shown as a function of the initial frequency of S. Each value is the mean of ten observations; error bars are standard errors. See text for ANOVA.

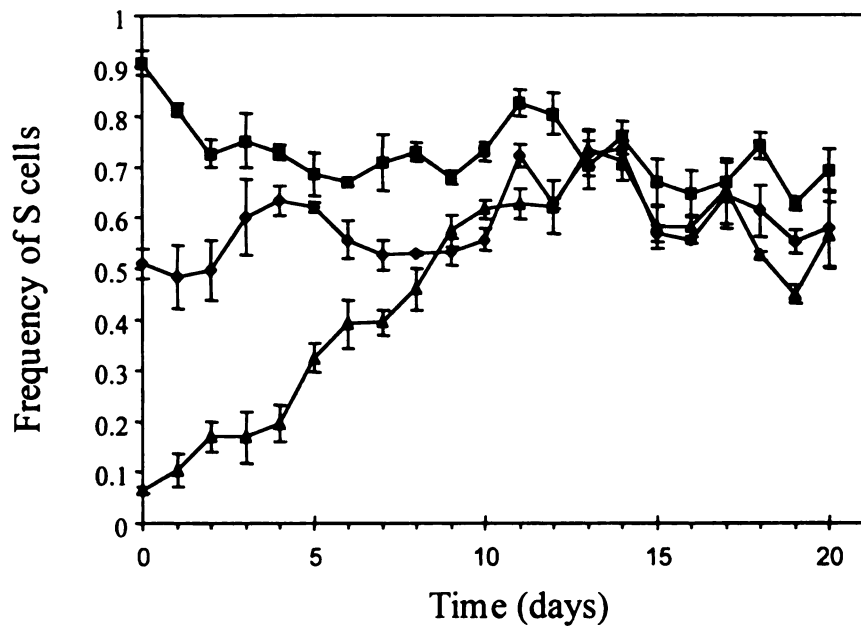


Figure 2: Convergence on a stable equilibrium during 20 dilution and growth cycles (≈ 130 generations). Genotype S is able to invade when rare, but it declines in frequency when it is initially very common, leading to a balanced polymorphism. Three replicate trajectories were run starting from each of three initial conditions; error bars are standard errors.

the two genotypes converge on a stable equilibrium. The final frequency of the S genotype was 0.612 ± 0.043 (mean \pm SE, based on all nine mixtures at day 20). Thus, it is clear that frequency-dependent selection maintains a stable polymorphism, despite the large advantage that accrues to the L genotype during exponential growth. Evidently, the S genotype must have an off-setting advantage elsewhere in the population growth cycle.

S has Advantages in both Growth and Stationary Phases. We sought to determine when, during the population cycle, S compensates for its lower maximum growth rate. To that end, we sampled from the same competition experiments (used to infer frequency-dependence) at an intermediate time point, 8 h, as well as at the start and finish of the daily growth cycle. We chose 8 h because that is the approximate duration of the growth phase, given an initial lag period of 1-2 h and 6-7 cell divisions with a doubling time of about 1 h. Thus, at about 8 h, the cultures would exhaust the glucose in the medium and enter stationary phase, where they would remain for the rest of the daily cycle (Vasi et al. 1994). We then computed the fitness of S relative to L over just the first 8 h of growth as well over the entire 24-h cycle.

Figure 3 shows the fitness of S relative to L over these two time intervals for the three initial ratios. There are three conclusions from this experiment. First, the fitness of S relative to L is frequency-dependent, with S having an advantage only when it is rare. This pattern is true over just the first 8 h ($F = 10.67$, 2 and 27 df, $P = 0.0003$; non-

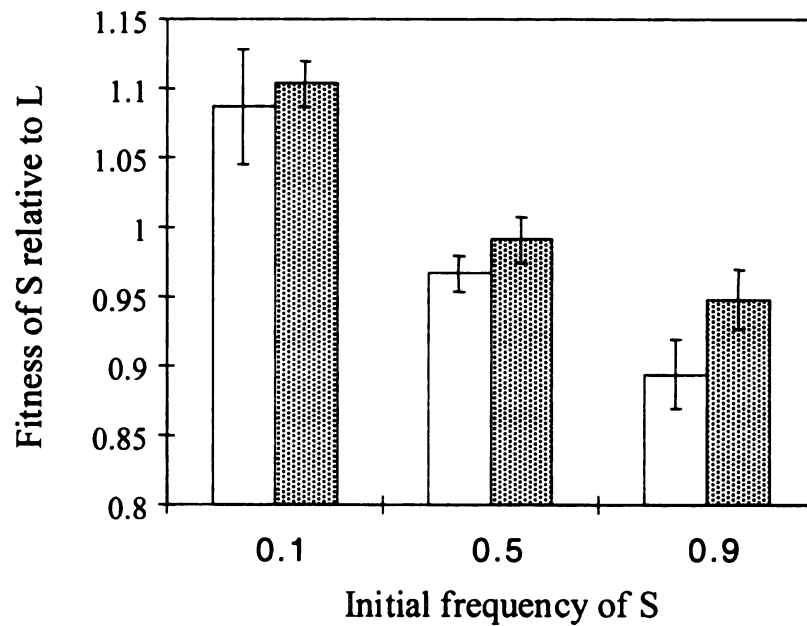


Figure 3: Genotype S has frequency-dependent advantages in both growth and stationary phases. The fitness of S relative to L is shown over two different portions of the population growth cycle and as a function of its initial frequency. Open bars: Growth phase (0-8 h). Filled bars: Growth and stationary phases combined (0-24 h, shown previously in Fig. 1). Each value is the mean of ten observations; error bars are standard errors. See text for statistical analyses.

parametric Kruskal-Wallis test, $P = 0.008$) as well as over the full 24-h cycle ($P < 0.0001$ as reported above). Second, S appears to have an advantage when rare even in those first 8 h ($t = 2.084$, 9 df, $P = 0.0668$), despite its much lower growth rate when grown by itself in DM25. Third, S gains a further advantage between 8-24 h, when growth has diminished due to glucose depletion. Paired comparisons of the fitness values obtained over the two different intervals are significant with all three initial frequencies combined (mean difference = 0.0316, $t = 2.5313$, 29 df, $P = 0.0170$). This late-arising advantage is especially strong when S was initially common (for $S = 0.1$: mean difference = 0.0167, $t = 0.5424$, 9 df, $P = 0.6007$; for $S = 0.5$: mean difference = 0.0244; $t = 1.9899$, 9 df, $P = 0.0778$; for $S = 0.9$: mean difference = 0.0537, $t = 2.9856$, 9 df, $P = 0.0153$). Evidently, S has advantages in both growth and stationary phases that offset its lower rate of exponential growth in pure culture.

S Affects the Death of L in Stationary Phase. The preceding analysis does not show whether the stationary-phase advantage of S relative to L is a consequence of differential growth or death. To address that issue, we analyzed the same data in terms of absolute (rather than relative) rates of change in population density between 8 and 24 h. Figure 4 shows that the changes over this interval are mostly due to the death of L, rather than continued growth by S, at least when S is initially abundant. An ANOVA indicates a significant effect of initial frequency on the rate of numerical change for L ($F_{2,27} = 4.742$, $P = 0.017$), but not for S ($F_{2,27} = 2.034$, $P = 0.150$). The fact that L declined in

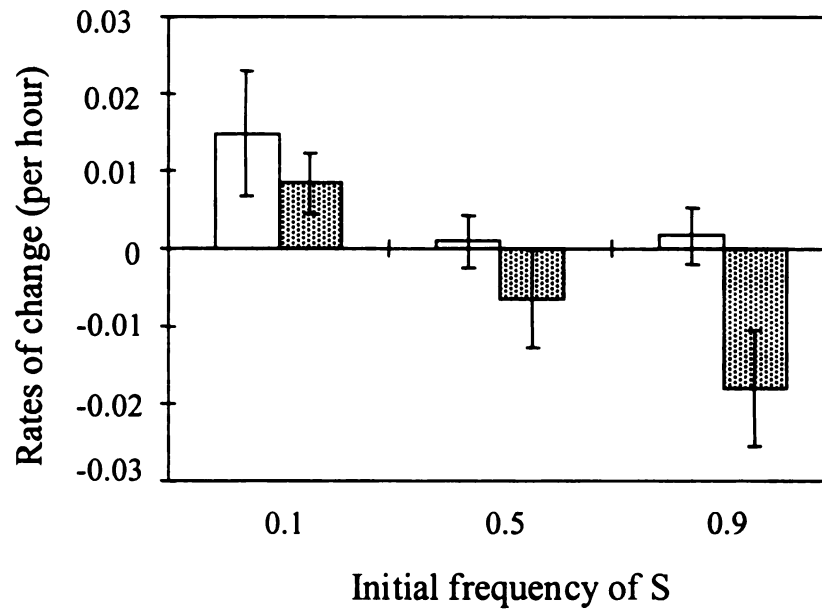


Figure 4: Genotype S promotes the death of L in stationary phase. The absolute rates of change in both population densities during stationary phase (8-24 h) are shown as a function of the initial frequency of S. Open bars: Genotype S. Filled bars: Genotype L. Each value is the mean of ten observations; error bars are standard errors. See text for statistical analyses.

density only when S was abundant suggests that S produces some metabolite that is toxic to L or that S removes a substance that promotes the survival of L, by removing a substance that is critical to L survival, may be indirectly resulting in increased L death. Our results cannot distinguish between these hypotheses.

Cross-Feeding of S on Metabolites during Growth Phase. In addition to its survival advantage in stationary phase, S also has an advantage when rare during the growth phase, which offsets its slower growth in pure culture. A plausible explanation is cross-feeding, whereby S may be able to use (more effectively than L) one or more byproducts of glucose metabolism that L secretes into the medium. To examine this possibility, we prepared conditioned media that contained the secretions of each genotype, and we then measured the maximum growth rate of each type in these media. Cultures of each genotype were grown separately to stationary phase (either 8 or 24 h) and then filtered through 0.45 μm filters to remove all cells. Because the time course of the accumulation and degradation of metabolites during stationary phase is unknown, we prepared conditioned media using filtrates made near the start (8 h) and at the end (24 h) of stationary phase. In all cases, the conditioned media comprised a filtrate reconstituted with glucose to 25 $\mu\text{g/mL}$ (the same concentration as in fresh DM25). We prepared five different media in all: fresh DM25 (which serves as a control, the results for which were given earlier), L8, L24, S8, and S24 (the letter indicates the genotype that produced the filtrate, and the numeral the number of hours the genotype spent to produce the filtrate).

Each medium was prepared in four independent batches to preclude any spurious effects of variation among batches.

Following the usual acclimation step, all five media were separately inoculated with each genotype, with four-fold replication (corresponding to the independently prepared batches and treated as blocks in the statistical analyses). The maximum growth rate of each genotype in every medium was obtained as before. Figure 5 summarizes the data, which support three important findings. First, L has a significantly higher maximum growth rate than does S in the unconditioned DM25 medium, as reported above. Second, the growth rate of L is unaffected by any conditioning of the media by either genotype ($F = 0.965$, 4 and 12 df, $P = 0.4615$). Third, by contrast, the growth rate of S is significantly influenced by conditioning of the media ($F = 61.19$, 4 and 12 df, $P < 0.0001$). A Tukey-Kramer test indicates that eight of ten pairwise contrasts are significant ($P < 0.05$). The growth rates of S in the different media can therefore be ranked as follows: $L24 = L8 > S24 = S8 > DM25$. Evidently, both S and especially L secrete metabolites that promote the growth of S; but L does not effectively use these metabolites, and so the cross-feeding occurs specifically from L to S.

Long-Term Dynamics of the Polymorphism

The preceding experiments demonstrate that two clones, S and L, isolated at generation 18,000 of an evolution experiment, can stably coexist with one another. These experiments also reveal two different ecological mechanisms, involving cross-feeding

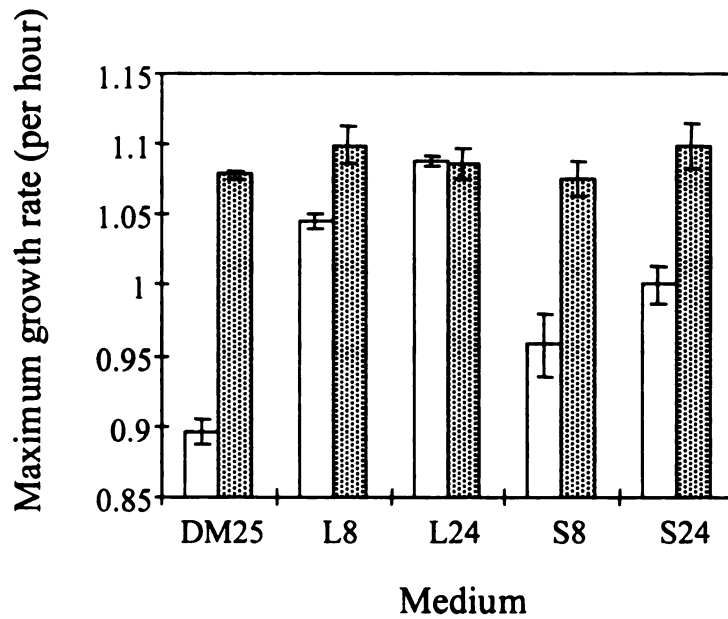


Figure 5: Cross-feeding of genotype S, but not genotype L, on metabolites secreted into the culture medium. The maximum growth rates of S (open bars) and L (filled bars) are shown in five different culture media. DM25 is the control medium, whereas the other four have been supplemented with filtrates obtained by growing either L or S for either 8 or 24 h. Each value is the mean of four observations; error bars are standard errors. See text for statistical analyses.

and possibly and differential death during stationary phase, that allow S to persist despite the much faster exponential growth by L in pure culture. Given the rapidity with which the two clones approach their joint equilibrium (Fig. 2), one might imagine that these two types have been at this equilibrium for a long time. However, one cannot exclude alternative scenarios; for example, their relative abundance may fluctuate over time due to further evolution of one or both types. To examine this issue, and to ascertain when the polymorphism arose, we examined the “fossil record” of this population, from which large samples were obtained every 500 generations, then stored frozen at -80°C (Lenski et al. 1991; Lenski and Travisano 1994).

Aliquots from the frozen stocks were revived, acclimated to growth conditions, and then spread on the same TA agar plates on which the polymorphism was noted at 18,000 generations. For each 500-generation interval, five separate plates (several hundred colonies) were scored as either S or L based on the timing of their appearance: as noted previously, L colonies generally appear after 24 h, whereas S colonies become visible only after 48 h. (For some 200 colonies, we confirmed that assignments based on colony appearance were corroborated by differences in average cell size measured with a Coulter counter. For the same 200 clones, we also confirmed the assignments by running restriction digests and using insertion sequences as genetic probes; we observed characteristic differences in the genetic “fingerprints” of the S and L morphotypes using this approach (D. E. Rozen, D. Schneider, M. Blot, and R. E. Lenski, unpublished data).

Figure 6 shows the frequency of the S type between 0 and 19,500 generations at 500-generation intervals. These data indicate that the S morphotype initially invaded L, rather than the other way around. They also show that the polymorphism is quite ancient, with the S type being common by generation 6,500 and remaining so throughout the duration. Of course, the S type must have arisen earlier in order to have become common by that time; when a mutant lineage first appears, its frequency is $1/N_e$, where the effective population size (adjusted for the bottlenecks during serial dilution) in the long-term evolution experiment is $\sim 3 \times 10^7$ (Lenski et al. 1991). If we assume that S had a relative fitness of 1.1 during its initial invasion, as it does when rare at generation 18,000 (see Fig. 1), then it would have taken ~ 200 generations for S to have increased from a single individual to $\sim 3\%$ at generation 6000. Further extrapolating from the convergence on the stable equilibrium (see Fig. 2), it should then have taken another 100 generations (15 d) or so for S to have increased to its equilibrium frequency of $\sim 60\%$. But that approach to the equilibrium did not occur; the frequency of S did not even reach 50% in the next 1000 generations, yet it then continued to increase to more than 80%. We can also estimate the time-averaged fitness of S relative to L during the period between 6000 and 7500 generations (Dykhuizen 1990), when its frequency increased from $\sim 3\%$ to $\sim 86\%$. That estimate is 1.005, and extrapolating back assuming this much lower fitness yields an estimated origin several thousand generations earlier, around generation 2000 or so. Thus, while we know that both types were present by generation 6000, we remain ignorant of the time of origin of the S morphotype due to the uncertainty about its

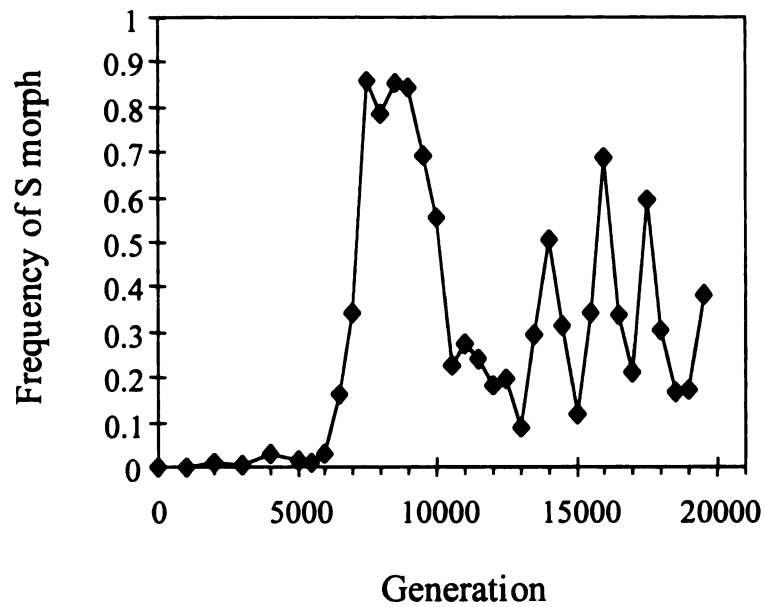


Figure 6: Long-term dynamics of the S-L polymorphism. Each point reflects scoring several hundred individuals. Based on the binomial distribution, 95% confidence limits extend at most a few percent in either direction. Despite the short-term stability of the polymorphism (Fig. 2), it is clearly unstable over much longer intervals. See Discussion for four alternative explanations for these fluctuations.

selective advantage when it first invaded.

It is also clear from these data that the polymorphism is very dynamic through time and has not simply remained at the equilibrium frequency that obtains from ecological interactions over the relatively short-term (~130 generations: Fig. 2). These fluctuations are not merely statistical noise. Each datum in Figure 6 is based on a sample size of several hundred colonies; from the binomial distribution, the 95% confidence intervals should in every case encompass only a few percent in each direction. Yet, the observed frequencies of the S morphotype vary between ~10% and ~85% after generation 6500. Despite these dramatic oscillations in relative frequency, calculations indicate that differences in relative fitness of less than 1% would be sufficient to explain even the most rapid of these fluctuations, given that they are manifest over thousands of generations. In the Discussion, we propose four different scenarios that might explain the apparent changes in the relative fitness of the S and L types.

Discussion

We observed the emergence of two distinct morphotypes, L and S, in an evolving population of the bacterium *E. coli*. This population was founded from a single haploid cell, and it lacks any mechanism for genetic exchange; hence it is strictly asexual (Lenski et al. 1991). The two types show a number of heritable differences, including the appearance of their colonies on agar plates, the average size of their individual cells, and

several important demographic properties. We showed that the S type invaded the ancestral L type, with the S type achieving polymorphic frequency ($>1\%$) at generation 6000 (Fig. 6). We calculate that the S type may have arisen, by mutation, anywhere from hundreds to thousands of generations earlier, depending on different assumptions about its initial rate of invasion. Between generations 6000 and 19,500 (the latest data available), the two types have coexisted. Such coexistence is not without precedent (see Helling et al. 1987; Rosenzweig et al. 1994; Treves et al. 1998) but is nonetheless unexpected on simple ecological and population genetic grounds. In ecological terms, the culture medium used for the experimental evolution contained glucose as the sole carbon and energy input, and it was density-limiting (Hansen and Hubbell 1980; Tilman 1982). On population genetic grounds, the asexual condition of the bacteria implies that each successive sweep of a beneficial mutation should purge all genetic variation from the evolving population (Muller 1932; Atwood et al. 1951).

Using L and S clones isolated after 18,000 generations, we examined first the dynamical stability of their coexistence and then the ecological mechanisms responsible for the interaction. We showed that the interaction between these clones was dynamically stable by two different approaches. First, one-day competition experiments showed that each type, when rare, had a fitness greater than one relative to the other (Fig. 1). Second, over the course of a few weeks (~ 100 generations), the two types converged on the same final relative abundance (~ 3 S to 2 L) regardless of their initial frequencies (Fig. 2).

The L genotype has a much higher maximum growth rate in the culture medium, DM25, than does S (Fig. 5, left-most pair). Maximum growth rate is an extremely important component of fitness in the serial batch regime employed during our long-term evolution experiments (Vasi et al. 1994). If all else had been equal, this difference would have led to the competitive exclusion of S by L. However, the S clone had two opposing advantages that allowed it to invade and coexist with the L clone (indeed, S was numerically dominant at the resulting equilibrium). One of these advantages is that L dies during stationary phase (i.e., after the glucose has been exhausted), whereas S does not (Fig. 4). In fact, the death rate of L increases when S is more abundant, which suggests that S may produce some metabolite that is toxic to L or that S removes some factor from the medium that sustains the viability of L. We cannot distinguish between these two possibilities based on the evidence at hand. However, allelopathic production of a toxin, by itself, would not provide a selective advantage to an invading genotype in a mass-action environment (Chao and Levin 1981). This consideration may indirectly favor the hypothesis that S depletes some nutrient necessary for the survival of L. Second, both L and (to a lesser degree) S secrete one or more metabolites into the medium that increase the growth rate of S, but which do not promote the growth of L (Figure 5).

This latter mechanism echoes the earlier findings of Helling et al. (1987), Rosenzweig et al. (1994), and Treves et al (1998) who demonstrated the evolutionary emergence of cross-feeding interactions among *E. coli* genotypes growing in chemostat culture. The physiological mechanism of the cross-feeding in their populations involved an increased

rate of glucose uptake coupled with the secretion of acetate; a mutation causing semiconstitutive overexpression of acetyl CoA synthetase then allows a second genotype to persist as a specialist on the secreted acetate (Rosenzweig et al. 1994). Treves et al. (1998) have found that the mutation resulting in overexpression of acetyl CoA synthetase has occurred repeatedly across replicate chemostat cultures.

This reproducibility contrasts with our own work, wherein strong frequency-dependence apparently evolved in only one of six replicate populations examined (Elena and Lenski 1997). There may be a simple ecological explanation for this difference in the two studies in the propensity of evolving populations to give rise to balanced polymorphisms based on cross-feeding interactions. Such interactions are sensitive to the concentration of metabolites in the medium, which in turn depend on bacterial density and ultimately on the amount of resource put into the system. In the chemostat experiments, bacteria were propagated on a medium that contained fivefold more glucose than the one used in our experiments; moreover, cells were diluted 100-fold each day in our serial transfer regime, whereas the bacteria were continuously maintained at their maximum density in the chemostat populations (Helling et al. 1987; Lenski et al. 1991). This hypothesis could be tested by varying the glucose concentration and examining its effect on the emergence of stable polymorphisms mediated by cross-feeding interactions.

Consideration of physiological mechanisms may also help explain the difference between chemostat and serial transfer regimes in their propensity to promote cross-feeding

interactions. Catabolite repression is a physiological process in bacteria that causes sequential rather than simultaneous use of multiple substrates for growth (Harder and Dijkhuizen 1982). In *E. coli*, catabolite repression ensures that available glucose is exploited before other less profitable resources are used. The strength of repression increases with the concentration of preferred resource as well as the growth rate of the population. In chemostats, bacteria hold the glucose concentration to a much lower level than the concentration experienced during the growth phase in the serial transfer regime; and the chemostat populations grow much more slowly than their counterparts during the exponential growth phase of the serial transfer regime. This difference in the strength of catabolite repression between chemostat and serial transfer populations may influence the phenotypic expression of mutants that can exploit metabolic by-products, perhaps amplifying the selective effect of metabolite concentration noted earlier.

In our experiment, as in the chemostat studies, the stable coexisting types evolved from a common ancestor and diverged while they were sympatric (indeed, in a thoroughly mixed environment). The ecological opportunity for this evolutionary divergence evidently depended on the generation, by the organisms themselves, of a diverse resource base from one that was otherwise homogeneous (Rosenzweig et al. 1994; Rainey and Travisano 1998). The S type emerged from the L type, and the cross-feeding interaction clearly benefits S at the expense of L. However, such strong frequency-dependent interactions and polymorphism do *not* appear to have evolved in five other replicate lines that were founded with the same ancestral strain and evolved under identical conditions

(Elena and Lenski 1997). This difference in outcome was evident even though the effective population size and number of generations were so large that all simple mutations should have occurred multiple times – but in different chronological order – in each evolving population (Lenski and Travisano 1994). Taken together, these observations suggest that two or more genetic events may have been necessary for the emergence of this balanced polymorphism. First, the lineage that gave rise to L may have had a mutation that increased its rate of glucose utilization, but at the expense of efficient metabolism, which led to the coincident loss of metabolites to the medium (due either to enhanced secretion or diminished reacquisition). Then the lineage that produced S may have benefited from a mutation that enabled it to scavenge and use these metabolites. Perhaps the properties of S that increase the death of L evolved still later. This scenario is similar to one model of ecological succession, where earlier successional species alter the environment so as to facilitate invasion by later species. An important difference is that, in our experimental system, the invader evolves *in situ*. In both cases, however, the frequency of the earlier species or genotype is depressed by the invader, with the outcome, either extinction or coexistence, depending on the specifics of their interaction.

Long-Term Dynamics of a “Stable” Polymorphism

The sequence of events that we presented in the previous paragraph is merely a scenario, at present, but it serves to illustrate two points. First, it points out the interest in

determining the number and timing of the genetic events that led to the emergence of the balanced polymorphism of the L and S types. Second, it emphasizes the fact that a polymorphism that is stable over the short term (Figure 2) may exhibit more complex dynamics on a longer time scale (Figure 6); indeed, that is what we observed. The ratio of the S and L types fluctuated ~60-fold over several thousand generations, whereas clones of these types that were isolated at one point in time converged on a stable equilibrium (Fig. 2). Our study is the first one with sufficient temporal duration to show such pronounced fluctuations in a “stable” polymorphism. A major focus of our future research on this polymorphism will be to determine the cause of these fluctuations. We can formulate four distinct hypotheses to account for the fluctuations, which we will seek to distinguish by appropriate experiments.

H1: Environmental Fluctuations. The fluctuations in relative abundance could reflect fluctuations in environmental variables – in the absence of any further genetic change in either L or S – despite our best effort to maintain a constant environment. For example, the equilibrium frequency of S might vary from 10% to 90% even over a very slight temperature range (say, 1°C). The samples that were characterized in Fig. 6 were analyzed at the same point in time, so that variation in conditions at the time of analysis is not a factor. But the samples were taken at different points in time, and so the fluctuations in relative abundance could reflect subtle fluctuations in the environment. If this strictly ecological hypothesis were true, then (a) L and S clones isolated from various

time points should give the same equilibrium when they are *run at the same time*, (b) but blocks of such experiments *run at different times* may give different equilibria. In some sense, this is the null hypothesis. The three alternative hypotheses below all have an evolutionary component, such that L and S clones isolated at one time point must have heritable differences (in ecologically relevant properties) from their counterparts isolated at other time points.

H2: Multiple Origins of S. The derived morphotype, S, may not be monophyletic, but instead it may have been repeatedly derived from the L lineage. Thus, one can imagine that L1 gave rise to S1, and that the two types achieved a balanced polymorphism based on cross-feeding for some period. Then, a beneficial mutation arose in L1 that created L2, and the advantage of L2 in terms of competing for glucose was so strong that it not only replaced L1 but also caused the extinction of S1. Nonetheless, L2 may have continued to secrete useful metabolites, so that a cross-feeding mutant S2 – derived from L2 – could readily invade. And so on and so forth. This hypothesis can be tested by finding enough molecular genetic markers to construct a phylogeny that resolves whether (a) S clones isolated later in the experiment are more closely related to S clones from early in the experiment, supporting monophyly, or (b) S clones from different time points are more closely related to various L clones than to one another, which implies multiple origins of S.

H3: Adaptation to General Conditions. The derived type, S, may be monophyletic, but

both L and S continually adapt to general aspects of their environment, such as temperature or pH. These adaptations allow L2 to replace L1, and they shift the equilibrium away from S toward L, but they must not cause the extinction of S1. Later, S also adapts genetically to the environment, giving rise to S2 and shifting the equilibrium relative abundance back toward S, but without driving L to extinction. Repeated rounds of adaptation thus produce fluctuations in relative abundance. This hypothesis can be tested by competing genetically marked variants of strains isolated at earlier and later time points. For example, L2 should outcompete L1, and S2 should outcompete S1, under this hypothesis. However, the fitness advantage should presumably be small relative to the advantage that each type (S or L) has when rare, so that neither type drives the other extinct.

H4: Coevolutionary Red Queen. This hypothesis is essentially the same as H3, except that instead of independent genetic adaptation of each lineage to the general culture environment, the adaptations are coevolutionary in nature. For example, L2 might replace L1, not because L2 is any better in competition with L1 in isolation, but instead because L2 is better at resisting an allelopathic effect of S. Distinguishing between the evolutionary and coevolutionary hypotheses (H3 vs. H4) will require comparing, for example, the fitness of L2 relative to L1 in the absence of any S, in the presence of S1, and in the presence of S2.

A related line of inquiry concerns the fact that the polymorphism emerged after 2000

generations, by which time most of the overall adaptation relative to the ancestral strain had already taken place. Several beneficial mutations of large effect swept through each evolving population during the first 2000 generations of the long-term experiment, whereas later sweeps were more infrequent and had less dramatic effect on fitness (Lenski and Travisano 1994). This change presumably occurred because the evolving populations, as they became better adapted, had fewer avenues available for further improvements of a similar magnitude. It is possible that S-type mutants started to invade the population in a frequency-dependent manner, well before the successful invasion around generation 6000, but these early invaders might have been purged by mutations of strong beneficial effect that continued to sweep through the L background. Only after the strongest beneficial mutations were already incorporated into L--such that further generally beneficial mutations would be insufficient to disrupt an emerging polymorphism--could the S type become common enough to be detected and, moreover, to persist by its own further evolution (or coevolution). In effect, the actual history might be some composite of hypotheses H2 and H3 (or H4). More generally, we intend to perform experiments across all of the replicate evolving populations to determine whether frequency-dependent interactions became more important over time, as this composite scenario would suggest.

Coda

Many ecological and genetic simplifications are made in an experimental investigation

such as this one. These include environmental constancy, the lack of any other species, a single founding genotype, the absence of sexual recombination, and the focus on an organism that is much simpler than many others. Yet, despite these simplifications, rather complex dynamics can and do emerge, even over relatively short periods, and further complexities become evident over somewhat longer time scales. As in previous studies of bacterial evolution (Helling et al. 1988; Rosenzweig et al. 1994; Treves et al. 1998; Turner et al. 1996; Rainey and Travisano 1998), we observed the evolution of ecologically stable interactions among genotypes that had evolved from a common ancestor. But unlike these earlier studies, we showed that these stable interactions could be destabilized over longer periods by subtle environmental or genetic changes. That such long-term complexities are seen even in simple model systems suggests that they might help to illuminate the evolution of polymorphism, and even speciation, in macro- and micro-organisms alike (Schluter 1996; Reznick et al. 1997; Rainey and Travisano 1998; Wilson 1998).

Literature Cited

- Atwood, K. C, L. K. Schneider, and F. J. Ryan. 1951. Periodic selection in *Escherichia coli*. Proceedings of the National Academy of Sciences of the USA 37:146-155.
- Ayala, F. J., and C. A. Campbell. 1974. Frequency dependent selection. Annual Review of Ecology and Systematics 5:115-138.
- Bennett, A. F., R. E. Lenski, and J. E. Mittler. 1992. Evolutionary adaptation to temperature. I. Fitness responses of *Escherichia coli* to changes in its thermal environment. Evolution 46:16 30.
- Bohannon, B. J. M., and R. E. Lenski. 1997. Effect of resource enrichment on a chemostat community of bacteria and bacteriophage. Ecology 78:2303-2315.
- Chao, L., and B. R. Levin. 1981. Structured habitats and the evolution of anticompetitor toxins in bacteria. Proceedings of the National Academy of Sciences of the USA 78:6324-6328.
- Chao, L., B. R. Levin, and F. M. Stewart. 1977. A complex community in a simple habitat: an experimental study with bacteria and phage. Ecology 58: 369-378.
- Dykhuizen, D. E., and A. M. Dean. 1990. Enzyme activity and fitness: evolution in solution. Trends in Ecology & Evolution 5:257-262.
- Dykhuizen, D. E., and D. L. Hartl. 1980. Selective neutrality of 6PGD allozymes in *Escherichia coli* and the effects of genetic background. Genetics 96:801-817.
- Dykhuizen, D. E. 1990. Experimental studies of natural selection in bacteria. Annual Review of Ecology and Systematics 21: 373-398.
- Elena, S. F., and R. E. Lenski. 1997. Long-term experimental evolution in *Escherichia coli*. VII. Mechanisms maintaining genetic variability within populations. Evolution 51:1058-1067.
- Elena, S. F., V. S. Cooper, and R. E. Lenski. 1996. Punctuated evolution caused by selection of rare beneficial mutations. Science (Washington, D.C.) 272:1802-1804.
- Hansen, S. R., and S. P. Hubbell. 1980. Single-nutrient microbial competition: qualitative agreement between experimental and theoretically forecast outcomes. Science (Washington D.C.) 207:1491-1493.
- Hardin, G. 1960. The competitive exclusion principle. Science (Washington, D.C.)

- 131:1292-1297.
- Helling, R. B., C. N. Vargas, and J. Adams. 1987. Evolution of *Escherichia coli* during growth in a constant environment. *Genetics* 116:349-358.
- Koch, A. L. 1974. The pertinence of the periodic selection phenomenon to prokaryotic evolution. *Genetics* 77:127-142.
- Lenski, R. E. 1988. Experimental studies of pleiotropy and epistasis in *Escherichia coli*. I. Variation in competitive fitness among mutants resistant to virus T4. *Evolution* 42:425-433
- Lenski, R. E. 1995. Molecules are more than markers: new directions in molecular microbial ecology. *Molecular Ecology* 4:643-651.
- Lenski, R. E., M. R. Rose, S. C. Simpson, and S. C. Tadler. 1991. Long-term experimental evolution in *Escherichia coli*. I. Adaptation and divergence during 2,000 generations. *American Naturalist* 138:1315-1341.
- Lenski, R. E., and M. Travisano. 1994. Dynamics of adaptation and diversification: a 10,000-generation experiment with bacterial populations. *Proceedings of the National Academy of Sciences of the USA*. 91:6808-6814.
- Levin, B. R. 1972. Coexistence of two asexual strains on a single resource. *Science* (Washington, D.C.) 175:1272-1274.
- Levin, B. R. 1981. Periodic selection, infectious gene exchange and the genetic structure of *E. coli* populations. *Genetics* 99:1-23.
- Levin, B. R. 1988. Frequency-dependent selection in bacterial populations. *Philosophical Transactions of the Royal Society of London B, Biological Sciences*. 319:459-472.
- Miller, J. H. 1992. A short course in bacterial genetics. Cold Spring Harbor Laboratory Press, Plainview, New York
- Muller, H. J. 1932. Some genetic aspects of sex. *American Naturalist* 8:118-138.
- Rainey P. B., and M. Travisano. 1998. Adaptive radiation in a heterogeneous environment. *Nature* (London) 394:69-72.
- Reznick D. N., Shaw F. H., Rodd F. H., and R. G. Shaw. 1997. Evaluation of the rate of evolution in natural populations of guppies (*Poecilia reticulata*). *Science* (Washington, D.C.) 275:1934-1937.

- Rosenzweig, R. F., R. R. Sharp, D. S. Treves, and J. Adams. 1994. Microbial evolution in a simple unstructured environment: genetic differentiation in *Escherichia coli*. *Genetics* 137:903-917.
- Schluter D. 1996. Ecological speciation in postglacial fishes. *Philosophical Transactions of the Royal Society of London B, Biological Sciences* 351:807-814.
- Sniegowski, P. D., P. J. Gerrish, and R. E. Lenski. 1997. Evolution of high mutation rates in experimental populations of *Escherichia coli*. *Nature (London)* 387:703-705.
- Tilman, D. 1982. *Resource Competition and Community Structure*. Princeton University Press, Princeton, NJ.
- Travisano, M., and R. E. Lenski. 1996. Long-term experimental evolution in *Escherichia coli*. IV. Targets of selection and the specificity of adaptation. *Genetics* 143:15-26.
- Travisano, M., F. M. Vasi, and R. E. Lenski. 1995. Long-term experimental evolution in *Escherichia coli*. III. Variation among replicate populations in correlated responses to novel environments. *Evolution* 49:189-200.
- Treves, D. S., S. Manning, and J. Adams. 1998. Repeated evolution of an acetate-crossfeeding polymorphism in long-term populations of *Escherichia coli*. *Molecular Biology and Evolution* 15(7):789-797
- Turner, P. E., V. Souza, and R. E. Lenski. 1996. Tests of ecological mechanisms promoting the stable coexistence of two bacterial genotypes. *Ecology* 77:2119-2129.
- Vasi, F., M. Travisano, and R. E. Lenski. 1994. Long-term experimental evolution in *Escherichia coli*. II. Changes in life-history traits during adaptation to a seasonal environment. *American Naturalist* 144:432-456.
- Velicer, G. J., L. Kroos, and R. E. Lenski. 1998. Loss of social behaviors by *Myxococcus xanthus* during evolution in an unstructured habitat. *Proceedings of the National Academy of Sciences of the USA* 95:12376-12380.
- Wilson D. S. 1998. Adaptive individual differences within single populations. *Philosophical Transactions of the Royal Society of London B, Biological Sciences* 353:199-205.

Chapter 2

THE PHYLOGENETIC HISTORY OF A BALANCED POLYMORPHISM IN A LONG-TERM LABORATORY POPULATION OF *ESCHERICHIA COLI*

Over the last decade, it has become clear that asexual microbial populations can evolve extensive polymorphism even under the most ecologically simple laboratory conditions (Helling et al. 1988; Rainey et al. 2000; Rainey and Travisano 1998; Rosenzweig et al. 1994; Rozen and Lenski 2000). Predictably, polymorphism can evolve in response to spatial heterogeneity (Rainey and Travisano 1998) or temporal heterogeneity in resource availability (Bell 1997; Stewart and Levin 1972). However, such variability need not be experimentally imposed but can be generated *in situ* by the evolving bacterial populations themselves. The evolutionary production of environmental complexity, particularly the generation of metabolizable resources, can then serve as the catalyst for further evolutionary diversification. This was first observed by Helling et al. (1988), who described a population of *E. coli* that had been evolving in chemostats supplemented with a single resource and which developed a polymorphism maintained through cross-feeding and negative frequency-dependent selection.

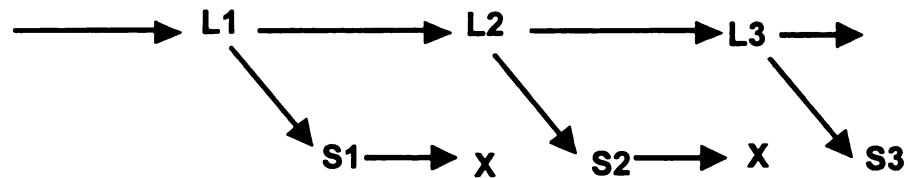
More recently, Rozen and Lenski (2000) described a polymorphism that evolved in a laboratory population of *E. coli* that has been serially propagated for more than 20,000 generations in a glucose-limited minimal medium (Lenski 1991; Cooper and Lenski

2000). The polymorphism consists of two morphs, called S and L, which coexist in a frequency-dependent manner. The relationship is maintained, despite the fact that the exponential growth rate of L is nearly 20% greater than S, because of two primary factors: 1) cross-feeding, where L cells secrete a resource upon which S cells can grow, and 2) a higher death rate of L during stationary phase. We found that S and L were phenotypically ancient (the two morphs have each existed for at least 12,000 generations), and that the relationship has been dynamic through time with their relative frequencies shifting repeatedly between about 10% and 90%. In this paper we use RFLP based on IS elements to examine the phylogenetic history of S and L. Further, we examine evidence for continued adaptation within S and L, with the aim of better understanding the factors that have influenced their dynamic coexistence.

Two alternative scenarios for the history of S and L are diagrammed in Figure 7; each scenario suggests distinct causes for the fluctuations in the relative frequencies of S and L through time. Consider, as starting conditions for both scenarios, that two hypothetical genotypes S1 and L1 coexist at some stable equilibrium based upon cross-feeding.

According to the scenario shown in Figure 7a, we first imagine a beneficial mutation, L2, that arises and replaces L1. L2 not only replaces L1 but is sufficiently better than S1 in competing for glucose that it also drives it to extinction. L2, however, still secretes useful metabolites which would enable the successful invasion of an S-like phenotype from within the L2 clade. These repeated bouts of extinction and re-evolution of "S" could cause dynamic fluctuations in the relative frequencies of both types over time.

(a)



(b)

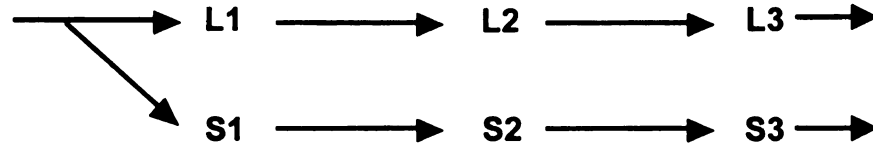


Figure 7. Two hypothetical models for the evolutionary history of the **S** and **L** morphs through time. Arrows represent genetic transitions and indicate ancestry. In (a) genotype **L1** gives rise to both **L2** and **S2** and then **L3**. This history is one of **S** extinction and re-evolution. (B) depicts phylogenetic continuity of **S** and **L**. This history is one of replacement only within type, e.g. **S** clones can only arise from earlier **S** clones. Accordingly, **S** and **L** both represent monophyletic clades.

In contrast, Figure 7b shows distinct monophyletic S and L clades. In this scenario, beneficial mutations that arose in L would not lead to the extinction of S. Beneficial mutations would also occur in S, but these would not cause the extinction of L. Accordingly, clones of each morph isolated from a single evolutionary time point would be more related to clones of the same morph from earlier time points than to the alternative morph from the same time point. If, during the period of coexistence, both the S and L clades continue evolving, it could generate the observed fluctuations in the relative frequencies of both morphs.

A unique aspect of this study is our ability to study genotypes from across a temporal series; this allowed us examine aspects of the S and L dynamic which could not be inferred from the phylogeny alone. Specifically, we sought evidence for continued adaptation within the putative S and L clades following their respective origins. To assess evidence for continued adaptation, we first examined the time course of genetic variation for both morphs. Two possible patterns were expected from this analysis. If S and L have continued to adapt, then genetic variation within each asexual clade would be periodically purged as beneficial mutations swept to fixation via the process of periodic selection (Atwood et al. 1951). Population-wide elimination of genetic variation is expected following each selective sweep since the genomes of strictly asexual organisms are linked to the sweeping beneficial mutation (Levin 1981). Alternatively, if adaptation did not continue, then we would expect genetic variation to increase, either indefinitely or

to some plateau, but not to show any significant decreases. Thus, an anticipated genetic signal of continued adaptation via periodic selection in an asexual population is the periodic reduction, and subsequent renewal, of genetic variation through time. In support of this analysis, we next sought direct evidence for continued adaptation within S and L. This was examined by direct competition assays between S and L clones that were isolated from distant evolutionary time points. If S or L has continued to adapt, then the fitnesses of genotypes from later evolutionary time points should exceed those from earlier evolutionary time points.

By using RFLP genetic fingerprinting with Insertion Sequences (IS) (Lawrence et al. 1989; Papadopoulos et al. 1999), we have examined the phylogenetic history of S and L. We have further assessed predictions based on IS data by conducting direct fitness comparisons between evolutionarily distant genotypes. Briefly, we conclude that the S/L polymorphism is genetically ancient and that both S and L are monophyletic clades. We also found evidence for continued adaptation within the S and L clades following their respective origins.

Materials and methods

Bacterial genotypes and culture conditions

The genotypes used in this study were derived from a single clone of *E. coli* B which was used to found a population that has been serially propagated for over 20,000

generations in glucose limited batch culture (see Lenski et al. 1991 for original strain description; Cooper and Lenski 2000). Two morphotypes, distinguishable on the basis of colony size and on time of appearance on tetrazolium-arabinose (TA) indicator agar, were observed in one of the evolving populations, Ara -2 (Rozen and Lenski 2000). Cells forming large colonies ~ 24 hours after plating, and colonies which remained small ~ 48 hours post plating, were designated L and S, respectively. Until this point, we have referred to the genotypes "S" or "L". Hereafter, S and L refer to specific clonal genotypes, and S and L refer to the groups of phenotypically similar clones which we will show are monophyletic clades.

During the course of this long-term experiment, aliquots from each population were collected every 500 generations and stored at -80°C (Lenski et al. 1991). Samples from these stocks that were taken before (1,000, 2,000, 3,000, 4,000, 5,000 and 6,000 generations) and after the phenotypic emergence of the S morph (6,500, 7,000, 9,000, 11,000, 13,000, 15,000, 17,000 generations) were plated on TA indicator plates to determine the frequency of S and L through time (Rozen and Lenski 2000). Ten random clones of each morph at each time point were tooth-picked from TA plates, grown overnight in LB broth and then frozen at -80°C in 15% glycerol. A total of 200 clones were isolated.

Competition experiments and fitness estimation

Competition assays were conducted to examine fitness changes within S and L over time.

Fitness differences were examined over the fairly long evolutionary interval between 13,000 and 17,000 generations. Because at both time points we have found evidence for genetic variation, fitness assays were conducted between S and L samples rather than single clones. To generate these samples, we isolated five randomly chosen S and L clones from both time points for a total of 20 clones. For each clone we selected a spontaneous Ara⁺ mutant by plating $\sim 10^8$ cells onto minimal-arabinose agar. Next, the total set of 40 clones was divided into eight samples of five clones each that corresponded to S or L, 13,000 or 17,000 generations, and Ara⁺ or Ara⁻ marker state.

All fitness assays were conducted between Ara⁺ and Ara⁻ samples which are distinguishable on TA plates. Ara⁺ colonies appear white on TA agar while Ara⁻ colonies are red. In previous work we determined that the Ara⁺ mutations are neutral with respect to fitness in DM25 (Rozen and Lenski 2000). For each fitness assay, both samples were grown for one full day in DM25 to ensure that they had each attained similar densities and physiological states. Following this acclimation period, equal densities of both samples were mixed and the change in their relative densities was measured over the course of six days. The mean relative fitness of both samples was calculated as the ratio of their Malthusian parameters, which for each sample was calculated as $m_i = (N_i[6] / N_i[0]) / 6$ d, where $N_i[0]$ and $N_i[6]$ are initial and final densities respectively.

DNA handling and analysis

DNA preparation, blotting, and hybridization

These methods have been described in detail elsewhere (Naas et al. 1994). Briefly, genomic DNA of each clone was isolated using Qiagen Genomic-tip kits according to manufacturer's specifications. DNA was then digested with *EcoRV* for ~ 3 hours at 37°C and electrophoresed overnight at 35V through 0.8% agarose gels. DNA was transferred to nylon membranes (Roche) using either capillary transfer (Sambrook et al. 1989) or vacuum transfer (Pharmacia Biotech Vacugene Pump).

DNA probes corresponding to the internal fragments of each of four IS elements were prepared as in Naas et al. (1994), except that the probes were labeled using the non-radioactive DIG kit (Roche) according to manufacturer's protocols. We probed for four IS elements: *IS1*, *IS3*, *IS150* and *IS186*. This set of four IS was chosen because they had been found to be phylogenetically informative in preliminary analyses and in a related study with these *E. coli* B populations (Papadopoulos et al. 1999). Southern blot hybridizations were conducted using the same kit used for probe labeling. Filters were probed with each IS element successively, and each hybridized probe was stripped prior to reprobing. All ambiguous IS positions were reexamined by co-migrating the corresponding DNA samples in adjacent lanes.

RFLP coding and analysis

RFLP fragments were scored as either present (1) or absent (0) for all clones to obtain a genotype-specific IS fingerprint. Shared IS fragments were assumed to be identical by

descent (i.e., homologous). This resulted in an IS fingerprint matrix of 200 clones by 128 total IS positions combined over the four IS elements. A distance matrix was computed that determined all pairwise distances between clones. To obtain the phylogeny of the 200 clones with respect to one another and their common ancestor, the 200 X 128 matrix was first examined by Neighbor Joining in PAUP* (Swofford 1998) with 5,000 bootstrap replicates. Next, 100 bootstrap replicates of both Parsimony and Neighbor Joining were conducted on genotypes collected just at specified evolutionary time points to more extensively examine the hypothesis that S and L are monophyletic clades. We examined genotypes collected from 6,500, 7,000, 9,000, 11,000, 13,000, 15,000, and 17,000. In all cases, the actual ancestral genotype was included as an outgroup.

L-specific PCR and the origin of L and S

Inverse PCR (Ochman et al. 1987) was used as in Schneider et al. (2000) to determine the genomic location of an L-specific IS3 mutation (D. Schneider, unpublished data). The resulting DNA sequence data were used to design PCR primers to specifically amplify the same region in putative L clones, and which would not amplify in either S or in the group that preceded S and L, called non-L/S. L-specific primers were RL130: 5' ctg tga ttg gga tca gcg gt 3' and RL131: 5' agc gtg ctg tgg ttt caa cc 3' which amplified an ~1,500 bp DNA fragment. All PCR were performed with Gibco taq polymerase according to the manufacturer's recommendations. Between 90-200 random clones were screened for the L-specific marker from each of six evolutionary time points. We examined samples

that were frozen after 2,000, 3,000, 4,000, 5,000, 6,000, and 7,000 generations of evolution.

In earlier work, we first observed **S** at 6,500 generations (see Rozen and Lenski 2000).

The phylogeny developed here provided compelling evidence that the genetic and phenotypic emergence of **S** coincided. These earlier data, based on phenotype, are reproduced for comparison with the new **L** data collected here.

Results

S and L are both monophyletic clades

The phylogenetic histories of **S** and **L** were determined through examination of RFLP from 200 clones. Specifically, we examined the hypotheses that **S** and **L** were both monophyletic clades. In Figure 8, we show the consensus Neighbor Joining topology for these clones based upon 5,000 bootstrap replicates in PAUP*. Three major groups can be delineated; **L**, **S**, and non-**L/S**. 1) **L** is a monophyletic clade that is first detected at 3,000 generations. 2) **S** is a monophyletic clade that is first seen at 6,500 generations. 3) Non-**L/S** is an essentially artificial grouping that contains the ancestral **IS** genotype and many, but not all, of its descendants. It is thus paraphyletic. The apparent monophyly of both **S** and **L** thus demonstrates that **S** and **L** fluctuations do not result from the continued extinction and then re-evolution of **S** genotypes, thereby rejecting the scenario depicted in Figure 7a.

Figure 8. Neighbor-joining topology for **S** and **L** rooted by using the actual ancestral genotype. Each taxon listed is a composite of generation, morph, and the fraction of clones from that time point and morph that had the identical IS genotype. For example, a taxon listed as "6.5K S (3/10)" represents a group of 3 identical **S** clones out of 10 **S** clones from 6,500 generations that were examined. Note that non-S/L clones are morphologically similar to **L** clones, and thus are indicated as **L** samples in this notation. Values on the main branches leading to **S** and **L** are derived from 5000 bootstrap replicates.

Despite (or perhaps as a result of) the large number of clones that we have examined, the Neighbor Joining bootstrap support for monophyly of **S** and **L**, 55 and 56 respectively, is only moderate. The bootstrap support may be influenced by the fact that our chronological data set includes genotypes from both terminal and internal nodes (i.e., the set of genotypes included extant and extinct varieties). To address this possibility we conducted analyses which removed the temporal sequence of genotypes and only considered the support for monophyly of **S** and **L** at specified evolutionary time points. In Table 1, we show that bootstrap support for the hypotheses that **S** and **L** are both monophyletic is substantially higher using this approach than when all 200 genotypes were examined together. This relationship is not dependent upon the specific phylogenetic method used. A second potential cause of reduced support for **S** and **L** monophyly is the possibility that some RFLP characters in **S** and **L** are convergent; this would occur when IS mutations in diverged clades caused the same alteration in RFLP pattern. Such events, which are assumed to be rare, causes divergent lineages to appear more related than they actually are and underestimates the number of changes since the lineages diverged (Bull et al. 1997). We have not yet identified the insertion sites for all IS differences between **S** and **L**, but of the several so far examined, one IS/ mutation appears to be convergent (D. Schneider, unpublished data).

Genetic and phenotypic origins of S and L

Until now, we had been unable to determine the time of origin for **L**, because it is phenotypically indistinguishable from the non-**L/S** group. In this work (see Figure 8), we

Table 1: Bootstrap support for monophyly of S and L using only the genotypes collected from each specified evolutionary generation. Except for the total data set which used 5,000 Neighbor Joining bootstrap replicated, all values are based upon 100 replicates. Parsimony bootstrap data were not obtained due to excessive computational time.

	Total		6,500		7,000		9,000		11,000		13,000		15,000		17,000	
	S	L	S	L	S	L	S	L	S	L	S	L	S	L	S	L
Parsimony	---	---	87	91	94	62	100	95	99	92	75	72	86	<50	80	50
Neighbor joining	55	56	91	90	99	72	99	93	94	96	65	86	81	88	81	86

discovered genetic evidence for a monophyletic **L** clade as early as 3,000 generations. Although 2,000 generations is the first point at which **L** was observed, it does not represent the time of the true genetic origin of the **L** clade; that is, the time at which **L** first invaded from a frequency of $1/N_e$ (or $\sim 1/3.3 \times 10^7$). To determine the time of genetic origin and the rate of invasion for **L**, we screened hundreds of randomly chosen clones with a PCR marker based on an *IS3* mutation that was diagnostic for the **L** clade. This both allowed us to determine the rate at which **L** invaded this evolving population (Dykhuizen 1990) and to extrapolate to the point of **L**'s genetic origin. In Figure 9 we show estimates of **L** frequency between 2,000 and 7,000 generations. We can estimate the time averaged fitness (Dykhuizen 1990) of **L** during its invasion from a frequency of ~ 0.005 at 2,000 generations to ~ 0.79 , 5,000 generations later. This estimate is 1.0014; which places the time of genetic origin of **L** at the earliest points in the history of this population (in fact, as early as the first day).

Figure 9 also shows the dynamics of invasion of the **S** clade, based on data from Rozen and Lenski (2000). These data are based on phenotype, but as we have shown here (Figure 8), the grouping of the phenotypically described **S** clones into a monophyletic clade is also well supported by the molecular genetic RFLP data. Estimates for the time of the genetic origin of **S**, based on the rate at which the **S** phenotype invaded, ranged from 2,000 to 6,000 generations depending on whether or not frequency-dependent fitness of **S** versus **L** was considered (Rozen and Lenski 2000). The early estimate assumed that **S** and **L** fitness at the time of **S**'s origin was not frequency-dependent, while

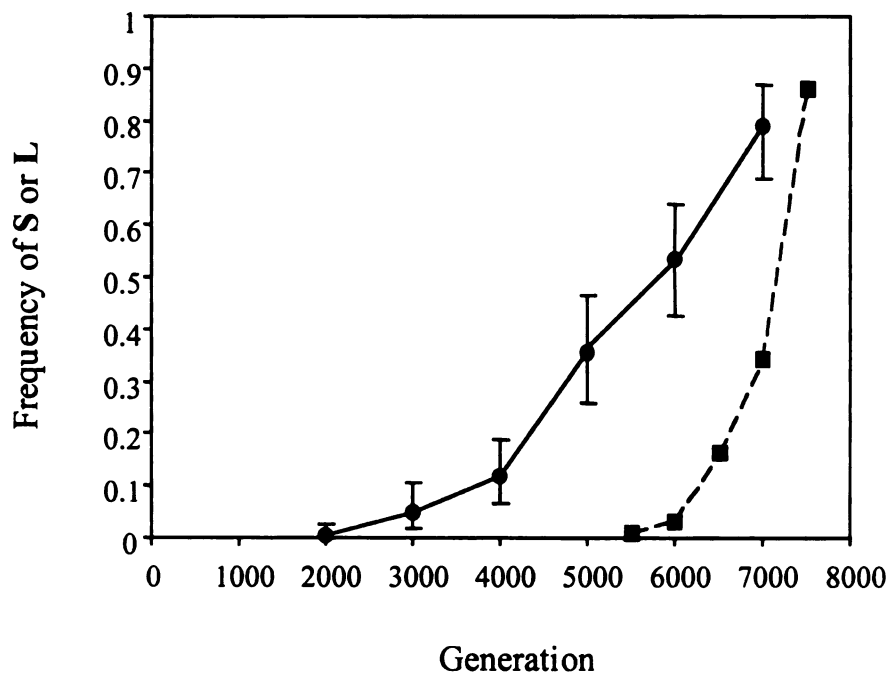


Figure 9. Trajectories of invasion of S and L. L data are based on IS frequencies (see text) while S data are based on phenotypic data from Rozen and Lenski (2000). L is represented by circles and S is shown as squares. Error bars on the data for L are confidence intervals based on the binomial distribution. L frequency estimates are for the portion of the population that was not S (i.e. L and non-L/S).

the latter estimate assumed that S fitness advantage at its first occurrence was approximately 10%, a value that was experimentally determined from competition assays between S and L clones (Rozen and Lenski 2000). The results presented here, that the first genetic evidence for S (at 6,500) coincides their first phenotypic appearance, suggests that S and L exhibited frequency-dependent fitness from the earliest points in their coexistence.

Adaptation following the independent origins of S and L

To assess whether S and L were continuing to adapt following their independent origins, we first examined the time course of genetic variation within each clade, as shown in Figure 10. Two possible patterns were expected from this analysis. If S and L were continuing to adapt then genetic variation would have been periodically purged as each new beneficial mutation rose in frequency and ultimately displaced the rest of the population. Alternatively, if there was no continued adaptation then we would expect genetic variation to have increased, either indefinitely or to some plateau, but not to show any significant decreases. Estimates for genetic variation were obtained by calculating the pair-wise distances of each clone to all others from the same morph and time point.

As evident in Figure 10, neither S nor L show monotonic increases in genetic variation through time. Instead, the time course of genetic variation for both clades is fluctuating and punctuated by periods of decline. Two clear declines between adjacent time points

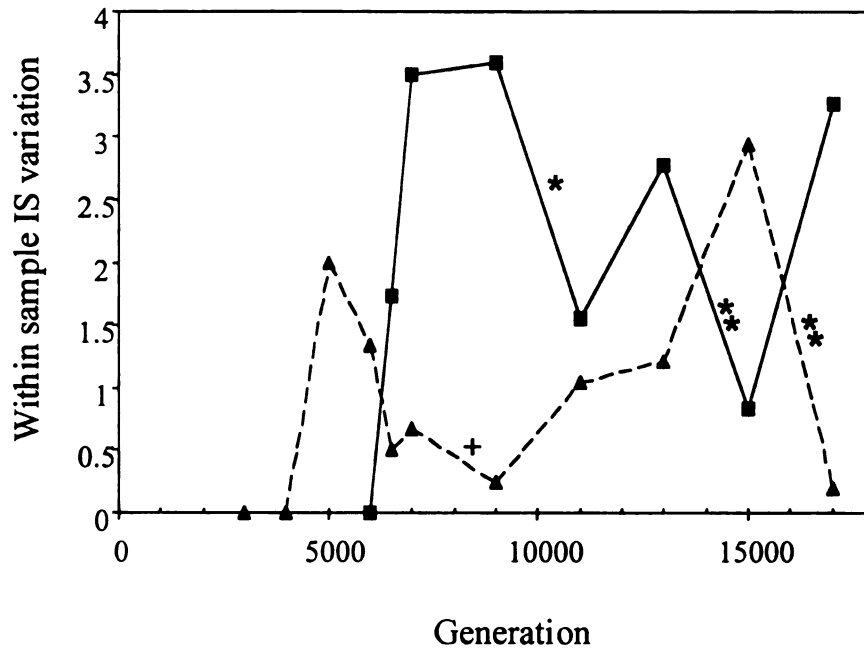


Figure 10. Time course of genetic variation in S and L, as calculated from pairwise genetic distances within samples. The S clade is represented by squares and solid lines and the L clade is represented by triangles and dashed lines. Significant declines in genetic variation over specific time intervals are shown by asterisks. **, $P < 0.01$. *, $0.01 < P < 0.05$. +, $0.05 < P < 0.1$.

are evident in both **S** and **L**--in **S** from 9,000-11,000 generations and from 13,000-15,000 generations, and in **L** from 7,000-9,000 generations and from 15,000-17,000 generations. To determine the statistical significance of the two declines in each clade, the change in mean pairwise distance from consecutive time points was compared using a two-tailed Mann-Whitney U test; p-values were adjusted with a sequential Bonferroni test (Rice 1989) to correct for the fact that comparisons were made across many temporally adjacent samples. Across three of the four noted intervals, we found evidence for significant reductions in genetic variation (summarized in Figure 10). The fourth interval, between 7,000 and 9,000 generations in **L**, was not quite significant ($0.05 < p < 0.1$) after correction for multiple comparisons. Because reductions of genetic variation are indicative of the substitution of beneficial mutations via periodic selection, we infer that both clades have continued to adapt following their initial appearance.

To more directly examine evidence for adaptation, we measured fitness change within **S** and **L** between 13,000 and 17,000 generations. Evidence shown in Figure 11 indicates that both clades have adapted during this interval. Two comparisons, corresponding to both Ara marker combinations, were conducted for each morph. Because we found no statistical influence of Ara marker on fitness for either morph (**S**: $F_{1,8} = 1.028$, $P = 0.34$; **L**: $F_{1,8} = 0.738$, $P = 0.415$), the data were combined. The mean fitness of **S** and **L** increased by $\sim 2\%$ and $\sim 1.5\%$ respectively, between 13,000 and 17,000 generations. The magnitude of fitness gains did not differ between **S** and **L** ($t_{18} = 0.889$, $P = 0.39$).

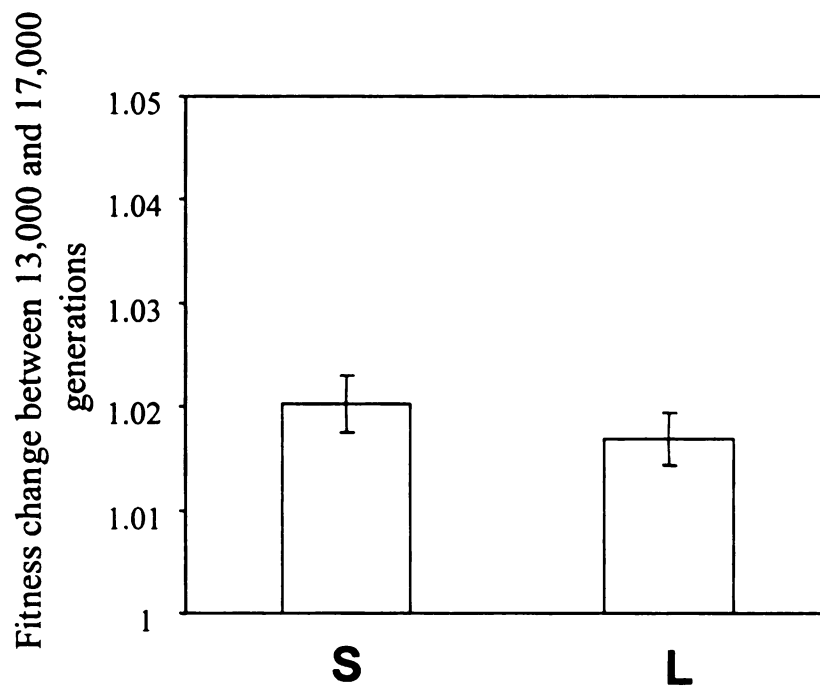


FIGURE 11. Changes in mean fitness between 13,000 and 17,000 generations within S and L. During this interval, the fitness of both S and L increased significantly. Error bars are 95% confidence intervals based on ten replicates.

Discussion

We examined previously the phenotypic history and mechanisms of persistence of two morphs, S and L, that evolved in a laboratory population of *E. coli* (Rozen and Lenski 2000). Here, we use Insertion Sequences to examine the phylogenetic history of S and L with the subsidiary aim of understanding the causes for the fluctuations in their relative frequencies through evolutionary time. As shown in Figure 7, we posed two scenarios for the history of S and L. Our data support the scenario shown in Figure 7b. That is, S and L are each monophyletic clades and therefore S and L fluctuations do not result from episodes of extinction followed by phenotypically convergent re-evolution of either morph. In addition, we provide two types of evidence that L and S have continued to evolve and adapt following their independent origins at ~ 2,000 and ~ 6,500 generations, respectively (Figure 9). It is their continued adaptation, which necessarily changes their mutual interaction, the presumably caused the fluctuations in the relative frequencies of S and L over evolutionary time.

Although a large number of genotypes was assayed in this study, the overall bootstrap confidence of the monophyly of S and L is only moderate (Figure 8 and Table 1). Two factors appear to have reduced the bootstrap confidence level: inclusion of a temporal sequence of genotypes and partially convergent RFLP patterns. In the first case, we have found that removal of the temporal sequence, by focusing sequentially only on contemporaneous samples, dramatically increases confidence in S and L monophyly

(Table 1). The importance of convergent RFLP is at present unclear. Schneider et al (2000) and Cooper (2000) have identified one convergent IS mutation that affected all 12 replicate populations from the long-term study of Lenski et al. (1990), although Schneider et al (2000) also found nine other IS-mediated mutations that were unique to either of two focal populations in that study. In the dimorphic Ara-2 population that is the focus of our study, we have also identified one putative case of convergence, involving an IS1 mediated mutation (D. Schneider, unpublished data). In future work we intend to identify the genomic location of each new IS position in this population to more fully examine the possibility of convergent IS-mediated events.

Through evolutionary time, the relative frequencies of S and L have oscillated between about 10% and 90%. Because this trend could result from continued adaptation within both clades following their origins, we sought evidence for such adaptive evolution. A signal of continued adaptation in asexual populations is the periodic reduction of genetic variation that results from new beneficial mutations that sweep to fixation. Figure 10 shows a series of such purging events, which provides evidence for continued adaptation within S and L. In addition, we show direct evidence in Figure 11 that fitness within both S and L increased between 13,000 and 17,000 generations. Although these changes in fitness are fairly small, they could easily drive substantial fluctuations in the frequencies of S and L over periods of thousands of generations, which is the scale at which these fluctuations are seen.

Perhaps the simplest explanation for fitness increases in both clades is that **S** and **L** have continued to find new genetic solutions to the problems posed by the laboratory environment--glucose, 37°C, pH, etc. Alternatively, **S** and **L** may be finding solutions to the problems of living with one another. That is, over the lengthy period of their coexistence, the particular features and products of **S** and **L** may have become the most important facets of the environment for one another. In that case, the evolutionary changes that we observed could be more accurately described as coevolutionary. While we do not explicitly examine evidence for coevolution here, this might be a future direction of this work.

Some of the fluctuations in IS genetic diversity shown in Figure 10 are associated with the fixation of a new IS mutation in either **S** or **L**. Especially for these cases, but generally for all new IS, it is tempting to speculate that the mutations are themselves causally associated with the changes in fitness that have occurred in this population. IS elements and transposons in *E. coli* and other organisms are known to be capable of causing beneficial mutations (Blot 1994; Blot et al. 1994). This benefit can occur either directly, as in the case of the bleomycin-resistance cassette of Tn5 (Blot et al. 1994), or indirectly through gene loss or polar effects that create novel promoters for genes downstream of the insertion (Hall 1999). Two examples are particularly relevant here since they both occurred in long-term evolving laboratory populations of *E. coli*. In one example, Cooper (2000) identified an IS150 mediated deletion of the *rbs* operon in all twelve replicate *E. coli* populations from the long-term study of Lenski et al. (1991).

This IS insertion eliminated the ability to catabolize ribose, and conferred an ~1.5% fitness benefit in the glucose minimal medium. In the second example, Treves et al. (1998) found repeated insertions of either IS30 or IS3 upstream of the *acs* locus, which caused increased expression of the gene and thereby enhanced the ability for mutated cells to scavenge acetate.

These examples make clear that benefits can be directly derived from IS-mediated events, but it is important to note that such mutations can also become fixed due to genetic drift or hitchhiking. While pure drift is unlikely to cause such rapid fixation of new mutations in these large populations, we cannot definitively distinguish between direct selection and hitchhiking due to selection acting on beneficial mutations elsewhere in the genome. To examine this we are in the process of constructing specific genotypes that alter the state of several IS mutations that were fixed in either S or L. We will thus be able to examine the direct phenotypic consequences of each new IS alone, in combination, and as a function of genetic background. Mutations that confer a fitness benefit must have achieved fixation owing to a direct selective advantage, while those IS mutations with deleterious or neutral effects could only have fixed via hitchhiking.

Genetic polymorphisms, often affecting ecologically relevant phenotypes are found in many populations. They can assume many forms and have a variety of causes. Although retrospective experiments can often determine the processes that give rise to and maintain specific polymorphisms, this is not always possible. Even in an apparently straight-

forward laboratory system, such as was examined here, the task is challenging and would not be possible except for the fact that *E. coli* clones and populations can be maintained in suspended animation. This feature has enabled us to determine the mechanisms of **S** and **L** persistence (Rozen et al. 2000) as well as, in this work, the phylogenetic and adaptive history of both clades. Measurement of adaptive change, in particular, required this extensive time frame as both the rate of initial invasion of **S** and **L** (Figure 9) and subsequent fitness change within both morphs (Figure 11) would have been too small to be detected by short term measurements. However, as indicated above, important steps remain in our efforts to understand the evolution of **S** and **L**.

Lituratione cited

- Atwood, K. C., L. K. Schneider, and F. J. Ryan. 1951. Periodic selection in *Escherichia coli*. Proceedings of the National Academy of Sciences of the USA 37:146-155.
- Bell, G. 1997. Selection: The Mechanism of Evolution. Chapman & Hall, New York, NY USA.
- Blot, M. 1994. Transposable elements and adaptation of host bacteria. Genetica 93:5-12.
- Blot, M., B. Hauer, and G. Monnet. 1994. The Tn5 Bleomycin resistance gene confers improved survival and grown advantage on *Escherichia coli*. Molecular and General Genetics 242:595-601.
- Bull, J. J., M. R. Badgett, H. A. Wichman, J. P. Huelsenbeck, D. M. Hillis, A. Gulati, C. Ho, and I. J. Molineux. 1997. Exceptional convergent evolution in a virus. Genetics 147:1497-1507.
- Cooper, V. S. 2000. Consequences of ecological specialization in long-term evolving populations of *Escherichia coli*. Ph.D. dissertation, Michigan State University, East Lansing, MI.
- Cooper, V. S., and R. E. Lenski. 2000. The population genetics of ecological specialization in evolving *E. coli* populations. Nature 407:736-739.
- Dykhuizen, D. E. 1990. Experimental studies of natural selection in bacteria. Annual Review of Ecology and Systematics 21:373-398.
- Hall, B. G. 1999. Transposable elements as activators of cryptic genes in *E. coli*. Genetica 107:181-187.
- Helling, R. B., C. N. Vargas, and J. Adams. 1988. Evolution of *Escherichia coli* during growth in a constant environment. Genetics 116:349-358.
- Lawrence, J. G., D. E. Dykhuizen, R. F. Dubose, and D. L. Hartl. 1989. Phylogenetic analysis using Insertion-Sequence fingerprinting in *Escherichia coli*. Molecular Biology and Evolution 6:1-14.
- Lenski, R. E., M. R. Rose, S. C. Simpson, and S. C. Tadler. 1991. Long-term experimental evolution in *Escherichia coli* .I. Adaptation and divergence during 2,000 generations. American Naturalist 138:1315-1341.
- Levin, B. R. 1981. Periodic selection, infectious gene exchange and the genetic structure

- of *E. coli* populations. *Genetics* 99:1-23.
- Naas, T., M. Blot, W. M. Fitch, and W. Arber. 1994. Insertion sequence-related genetic variation in resting *Escherichia coli* K-12. *Genetics* 136:721-730.
- Ochman, H., A. S. Gerber, and D. L. Hartl. 1987. Genetic applications of an inverse polymerase chain reaction. *Genetics* 120:621-623.
- Papadopoulos, D., D. Schneider, J. Meier-Eiss, W. Arber, R. E. Lenski, and M. Blot. 1999. Genomic evolution during a 10,000-generation experiment with bacteria. *Proceedings of the National Academy of Sciences, USA* 96:3807-3812.
- Rainey, P. B., A. Buckling, R. Kassen, and M. Travisano. 2000. The emergence and maintenance of diversity: insights from experimental bacterial populations. *Trends in Ecology and Evolution* 15:243-247.
- Rainey, P. B., and M. Travisano. 1998. Adaptive radiation in a heterogeneous environment. *Nature* 394:69-72.
- Rice, W. R. 1989. Analyzing tables of statistical tests. *Evolution* 43:223-225.
- Rosenzweig, R. F., R. R. Sharp, D. S. Treves, and J. Adams. 1994. Microbial evolution in a simple unstructured environment: genetic differentiation in *Escherichia coli*. *Genetics* 137:903-917.
- Rozen, D. E., and R. E. Lenski. 2000. Long-term experimental evolution in *Escherichia coli*. VIII. Dynamics of a balanced polymorphism. *American Naturalist* 155:24-35.
- Sambrook, J., E. F. Fitch, and T. Maniatis. 1989. *Molecular Cloning*. Cold Spring Harbor Laboratory Press, New York.
- Schneider, D., E. Duperchy, E. Coursange, R. E. Lenski, and M. Blot. 2000. Long-term experimental evolution in *Escherichia coli*. IX. Characterization of Insertion Sequence-mediated mutations and rearrangements. *Genetics* 156:477-488.
- Stewart, F. M., and B. R. Levin. 1972. Partitioning of resources and the outcome of interspecific competition: a model and some general considerations. *American Naturalist* 107:171-198.
- Swofford, D. L. 1998. *PAUP*: Phylogenetic analysis using parsimony*. Sinauer Associates, Sunderland, Mass.
- Treves, D. S., S. Manning, and J. Adams. 1998. Repeated evolution of an acetate-

crossfeeding polymorphism in long-term populations of *Escherichia coli*. *Molecular Biology and Evolution* 15:789-797.

Chapter 3

THE ROLE OF IS MUTATIONS IN THE EVOLUTION OF A BALANCED POLYMORPHISM IN A LABORATORY POPULATION OF *ESCHERICHIA COLI*

Introduction

Since their discovery, it has become clear that transposable elements are ubiquitous and can comprise a substantial proportion of their hosts' genomes (Kidwell and Evgen'ev 1999; Kidwell and Lisch 1997). Less clear, however, are the consequences of transposition for the evolutionary dynamics of their hosts. Like point mutations or insertions and deletions, the fitness effects of transposon induced mutations can range from deleterious to beneficial. Transposons differ from other types of mutations, however, in that their potential for horizontal transfer allows them to escape the damages that they cause (Charlesworth et al. 1984). For this reason, transposons have been thought of as genomic parasites (Doolittle and Sapienza 1980) and the possibility that they may be "selfish" has generated much discussion. If transposons were strictly selfish genomic parasites, high rates of horizontal transmission could offset their deleterious effects. However, it is also possible that mobile elements persist because they play an important role in the adaptation of their hosts (Blot 1994; Kidwell and Lisch 1997). In this work we examine the role of one class of mobile elements, Insertion Sequences (IS), in the adaptive changes that have occurred in an evolving population of *E. coli* that has achieved a selectively maintained balanced polymorphism.

Insertion Sequences (IS) are the predominant class of mobile elements in *E. coli*. They are extremely variable in copy number (Deonier 1996) and are responsible for generating a large fraction of new mutational variation (Hall 1999a; Rodriguez et al. 1992). The consequences of IS mutations can be manifest in a variety of ways (Mahillon and Chandler 1998). First, IS insertions can eliminate gene function through disruption of an open reading frame. Second, because IS carry promoters, they can cause polar effects whereby insertion into one gene alters or eliminates expression of adjacent genes (Hall 1999b; Mahillon and Chandler 1998). In a few cases, a direct benefit of IS mediated events has been demonstrated (Treves et al. 1998); however, these instances are limited. The work described in this paper is part of an ongoing program (see Papadopoulos et al. 1999; Schneider et al. 2000) to determine the role of IS elements in the adaptive evolution of laboratory populations of *E. coli*.

Lenski established twelve replicate populations of *E. coli* that have been maintained in serial batch culture for more than 20,000 generations (Cooper and Lenski 2000; Lenski et al. 1991). In one of the twelve populations, we identified a balanced polymorphism that has been both extremely dynamic and remarkably persistent (Rozen and Lenski 2000). Two clades, called L and S, were first observed at approximately 3,000 and 6,500 generations, respectively, and have persisted in a negative frequency-dependent fashion ever since (D. Rozen, unpublished). The frequency dependence results from three factors. First, L is able to invade a population of S cells because of an ~20% growth rate advantage. Second, S, despite its growth rate deficit, is able to invade a population of L

cells because it can metabolize one or more products that L cells secrete during growth.

Finally, L cells die during periods of starvation at a greater rate than S cells, an effect that is somehow exacerbated by the presence of S cells. When S and L are grown together for a short period of time they ultimately reach an equilibrium. However, over the more than 15,000 generations of coexistence their frequencies have oscillated between about 10% and 90% (Rozen and Lenski 2000).

In earlier work (D. Rozen, unpublished) that was directed towards elucidating the phylogenetic history of S and L, RFLP fingerprints (based on using IS as probes) were collected for 200 genotypes of both clades. In that work, IS were used as markers to provide information about the history of both groups. However, three features suggested that the IS mutations might themselves be causally linked to the adaptive changes that occurred in both clades. First, we found a series of IS mutations that were fixed within each clade. Second, the time of appearance of some IS mutations coincided closely with one another and with the first observation of the S clade. Finally, recent work in this (Cooper 2000) and other systems (Treves et al. 1998) has found evidence for IS mediated beneficial mutations. Consequently, we sought to determine the role of new IS mutations in the evolution of S and L.

New IS mediated mutations can become fixed through two routes, genetic hitchhiking or selection (Papadopoulos et al. 1999). In the case of hitchhiking, a beneficial mutation that was destined for fixation would have occurred on a genetic background that carried a

neutral or deleterious IS mutation. If we were able to isolate the direct consequence of this mutation, we would find that its effect would be either neutral or deleterious (though not deleterious enough to override the advantage of the beneficial mutation elsewhere in the genome). In the case of direct selection, the IS mediated mutation would itself confer a benefit.

In this work, we characterize five IS mutations that became fixed in either the **S** or **L** clade. We first identify the type and location of each mutation. Next we construct a set of isogenic genotypes that differ in the allelic state of two of the five mutations. Finally, we determine the direct fitness consequences of these mutations, alone, in combination, and as a function of genetic background. Together, this set of manipulated genotypes allowed us to determine: 1) if these two IS played a causal role in the adaptive changes that took place in the **S** clade; 2) if they interacted to influence **S** fitness; and 3) if the effects observed in **S** could be generalized to other genetic backgrounds.

Materials and Methods

Bacterial Strains and Plasmids

Table 2 summarizes of the bacterial strains used in this study. All genotypes were derived from a single clone of *E. coli* B (REL606) which was used to found twelve replicate populations that have been serially transferred in glucose limited batch culture for 20,000 generations (see Lenski et al. 1991 for original strain description). In one of

Table 2. Bacterial strains used in this study. A "/" mark is used to indicate experimental manipulations and a "+" is used to indicate wild-type.

Strain	Relevant strain properties	Notation in text/Figures
REL 606	Ancestral genotype of <i>Escherichia coli</i> B which is unable to utilize arabinose (Ara-) and is resistant to streptomycin	Anc
REL 607	Spontaneous Ara ⁺ mutant of REL 606	Anc/Ara ⁺
GBE 102	REL 606 with a deletion of a portion of the menaquinone operon	Anc/ Δ menC
GBE 107	REL 606 with a deletion in a portion of gene b2875	Anc/ Δ b2875
GBE 126	REL 606 with both mutations from GBE 102 and GBE 107	Anc/ Δ menC/ Δ b2875
REL 7409	Descendant of REL 606 that evolved for 18,000 generations (Small morph). It contains an IS186 insertion in <i>menC</i> and an IS150 insertion in b2875.	S
REL 7411	Spontaneous Ara ⁺ mutant of REL 7409	S/Ara ⁺
GBE 106	S with the menaquinone operon restored to wild-type	S/ <i>menC</i> ⁺
GBE 122	S with the b2875 gene restored to wild-type	S/b2875 ⁺
GBE 123	S with both changes from GBE 106 and GBE 122	S/ <i>menC</i> ⁺ /b2875 ⁺
REL 7410	Descendant of REL 606 that evolved for 18,000 generations (Large morph)	L
REL 7412	Spontaneous Ara ⁺ mutant of REL 7410	L/Ara ⁺
GBE 100	L with a deletion of a portion of the menaquinone operon	L/ Δ menC
GBE 108	L with a deletion in a portion of gene b2875	L/ Δ b2875
GBE 109	L with both mutations from GBE 100 and GBE 108	L/ Δ menC/ Δ b2875

the twelve populations, we identified a balanced polymorphism between two genotypes, S and L, which coexist via negative frequency dependence for fitness (Rozen and Lenski 2000). We used S and L to refer to the coexisting clades, S and L whereas "S" and "L" refer to specific clones isolated at 18,000 generations. In addition, we will refer to REL606 as Anc (for Ancestor). Anc, S, and L are unable to utilize arabinose (Ara-) and appear as red colonies when plated on tetrazolium-arabinose (TA) indicator plates. Spontaneous Ara⁺ revertants were isolated for REL606, S, and L by plating $\sim 10^8$ cells on minimal-arabinose agar. Ara⁺ genotypes make white colonies on TA plates. All genotypes were isolated as single clones and stored at -80°C in a 15% glycerol solution.

Host genotypes for cloning experiments were *E. coli* strains JM109 and SM10 λ pir. The plasmid used for gene cloning was pBC (Stratagene). For allelic replacement, the suicide plasmid pDS132 (D. Schneider, unpublished data) was used, which contains the following elements: 1) a chloramphenicol resistance gene, 2) the replication origin R6K (*oriR_{R6K}*), 3) the *sacB* gene that encodes levan sucrase, which is toxic to *E. coli* in the presence of sucrose, and 4) the *mob* region of plasmid RP4.

Growth Conditions and Media

Bacteria were cultured in Davis minimal medium supplemented with thiamine hydrochloride (at $2 \times 10^{-3} \mu\text{g mL}^{-1}$) and glucose at $25 \mu\text{g mL}^{-1}$, which supports a stationary phase cell density of $\sim 5 \times 10^7 \text{ mL}^{-1}$. This medium, hereafter called DM25, is

the same as was used during the long-term evolution experiment. To examine the role of menaquinone (also known as Vitamin K₂; Sharma et al. 1993) competition assays were conducted in DM25 with the addition of menaquinone (Sigma). For routine molecular work, we used Luria-Bertani broth (LB) with the addition of chloramphenicol (30 µg/ml) or streptomycin (50 µg/ml), where necessary. All cultures were grown in 10 mL of media in 50 mL Erlenmeyer flasks at 37°C and 120 rpm.

Competition Experiments and Fitness Estimation

Competition assays (Lenski et al. 1991) were performed to determine the fitness effects of IS mutations. All fitness assays were conducted between Ara⁺ and Ara⁻ genotypes, which are distinguishable on TA plates. In previous work, we determined that the Ara⁺ mutations are neutral with respect to fitness in DM25 in all three genetic backgrounds: Anc, S, and L (Lenski et al. 1991; Rozen and Lenski 2000). For each fitness assay, both competitors were grown for one full day in DM25 to ensure that they had each attained similar densities and physiological states. Following this acclimation period, competitors were mixed and the change in their relative densities was measured over the course of one day. For competitions between S and L, competitors were mixed at relative frequencies of either 1:9 or 9:1 to allow detection of frequency-dependent effects. In all other cases, we mixed equal volumes of the two competitors. Relative fitness of competing genotypes was calculated as the ratio of their Malthusian parameters, which for each genotype was calculated as $m_i = (N_i[1] / N_i[0]) / 1 \text{ d}$, where $N_i(0)$ and $N_i(1)$ are

initial and final densities, respectively.

Epistatic effects between mutations on fitness were estimated as in Bohannan et al. (1999). That is, the expected fitness of a double mutant was calculated using a multiplicative model of mutation interactions. If W_1 and W_2 are the fitness estimates for each single mutant, then the expected fitness of the double mutants is $W_{12} = W_1 \times W_2$. Expected values were calculated as the product of independent, replicated paired values for each single mutant. We then used a paired t-test to determine whether the set of observed fitness estimates for the double mutant differed significantly from the expectations based on the measurements from single mutants.

DNA Preparations, Blotting, and Hybridization

Genomic DNA and plasmid DNA were extracted using Qiagen Genomic-tip and plasmid purification midi-prep kits. For Southern hybridization, genomic DNA was digested with *EcoRV* for at least three hours at 37°C and electrophoresed overnight at 35V through 0.8% agarose gels. DNA was transferred to nylon membranes (Roche) using capillary transfer (Sambrook et al. 1989) and Southern hybridizations were performed using the non-radioactive DIG kit (Roche) under high-stringency conditions.

Determination of Genes Affected by IS Mediated Events

Mutation identity was determined by inverse PCR (Ochman et al. 1987), as described in Schneider et al. (2000). Briefly, restricted DNA was separated on 0.8% agarose gels.

Fragments that contained the IS of interest were cut from gels, purified, and self ligated with T4 DNA ligase at 5-10 µg/mL. Self ligated fragments were used as template in PCR experiments with primers directed out from the IS. Primers for each IS are as follows:

IS1: G3, 5'-GTCATCGGGCATTATCTGAAC-3' and G4, 5'-

AGAAGCCACTGGAGCACC-3'; IS150: G5, 5'-GATCCTGTAACCATCATCAG-3'

and G21, 5'-CATCCTGTTCTGCACTCTGA-3'; IS186: 5'-

CGGCATTACGTGCCGAAG-3' and G8, 5'-GGTGGCCATTCGTGGGAC-3'.

Amplified products were cloned using the PCR-script Cam cloning kit (Stratagene) and sequenced using the same primers as above. We attempted to determine the genomic location of each IS mutation by conducting a BLAST (Altschul et al. 1997) search against the *E. coli* K-12 sequence. Genomic location and the type of IS mediated mutations were confirmed using PCR products as Southern hybridization probes.

Allelic Replacements

Isogenic constructs were engineered to examine the fitness effects of two of the IS insertions found to have been fixed in S. Allelic replacement of wild type and mutant alleles was performed in three genetic backgrounds. First, in S, we used allelic replacement to restore the wild-type version of the mutant alleles. In Anc and L, we used allelic replacement to replace the wild-type allele with a mutant allele. The parent strains and resulting constructs are listed in Table 2.

Allelic replacement was conducted using the suicide vector pDS132. For all constructs, the allele to be manipulated (in either the mutant or wild-type form) was first cloned into pDS132 at *Sma*I cloning sites. Next, the suicide plasmid was transformed into SM10 λ *pir*, which was then used as a donor for subsequent plasmid transfer to Anc, S, or L.

For plasmid transfer and gene replacement, recipient cells (Anc, S, or L) were mated to SM10 λ *pir* carrying pDS132. Recipient transconjugants were selected on agar medium containing chloramphenicol and streptomycin, the latter of which counter-selected SM10 λ *pir* donor cells. The plasmid pDS132 is unable to initiate replication in recipient cells; thus, stable expression of chloramphenicol resistance required that pDS132 become recombined into the host chromosome at the site of the allele to be replaced. Next, chloramphenicol resistant cells were plated onto LB agar supplemented with sucrose, which is made lethal to *E. coli* by the product of *sacB*. Only cells from which the plasmid has been excised can grow under these conditions. Recombination during plasmid excision either restores the original allele or introduces the mutant allele. Thus, after *sacB* counterselection, several clones were screened by PCR and Southern hybridization to identify constructs that had incorporated the mutant allele.

Results

Gene locations

Five mutations that became fixed in either the S or L clade were examined (Table 3). We identified two simple insertions into open reading frames (ORF2), and three complex deletions involving completely uncharacterized genomic regions. For this latter set of mutations, no homologous regions within the *E. coli* K-12 genome could be identified.

In S, an IS186 inserted into *menC* (Sharma et al. 1993), one of the genes in an operon involved in the biosynthesis of menaquinone, which is a membrane bound component of the electron transport system (Meganathan 1996). Menaquinone is also known as Vitamin K₂, and is thought to be synthesized by a single pathway in *E. coli* (Meganathan 1996). Menaquinone biosynthesis is most active during anaerobic growth (Meganathan 1996).

Also in S, an IS150 inserted into the uncharacterized ORF designated b2875. We conducted a psi-BLAST (Altshul et al. 1997) search for homologues of b2875 in Genbank. Homologues were found in many other sequenced microbes, but in no case has the function of this gene been identified.

The third IS mediated mutation in S involved the deletion of a more than 10 kb fragment, containing about ten identified ORFs, between two IS1 fragments. One flanking IS1

Table 3: Genomic location of IS mutations that became fixed in S and L clades

	IS type	minute on <i>E. coli</i> K-12 chromosome	Gene(s) influenced
S	IS186	51.15	simple insertion into <i>menC</i>
	IS150	64.895	simple insertion into b2875
	IS1	~97.486-97.829	complex deletion between two IS1 elements
	IS3	no homology with K-12	complex deletion between two IS3 elements
L	IS3	no homology with K-12	complex deletion between two IS3 elements

element was in *fimB*, which encodes a recombinase involved in phase variation. This IS/ insertion in the *fimB* mutation is present in the ancestral strain (Schneider et al. 2000) and its location was not changed by this deletion event. The second flanking IS/ element is in an unidentified gene downstream from *sgcR*, which is part of a putative operon at 97.5 minutes on the *E. coli* K-12 chromosome. Although *sgcR* is uncharacterized, it is homologous to other transcriptional regulators in *E. coli*.

The final two mutations involved IS3 mediated deletions, one in S and one in L. Neither of these mutations could be identified because the genes affected do not show homology with any known *E. coli* K-12 genes. In both cases, however, we confirmed that the sequence was present in the ancestral *E. coli* B by using the inverse-PCR fragment as a Southern probe.

Rate of invasion for IS insertions into *menC* and b2875

Of the five IS mediated mutations that became fixed in S and L, two were chosen for further investigation. These two were chosen because they were generated by simple genetic events whose putative "knockout" effects could be closely approximated by creating precise deletion alleles and then using allelic replacement methods. Also, these two mutations invaded S early in its history and appeared to fix rapidly. In Figure 12, we show the dynamics of emergence and fixation of the IS mutations in *menC* and b2875 compared to several other IS/86 and IS/50 mutations that occurred in S during the same time period, but which did not become fixed. In contrast to the other IS mutations which

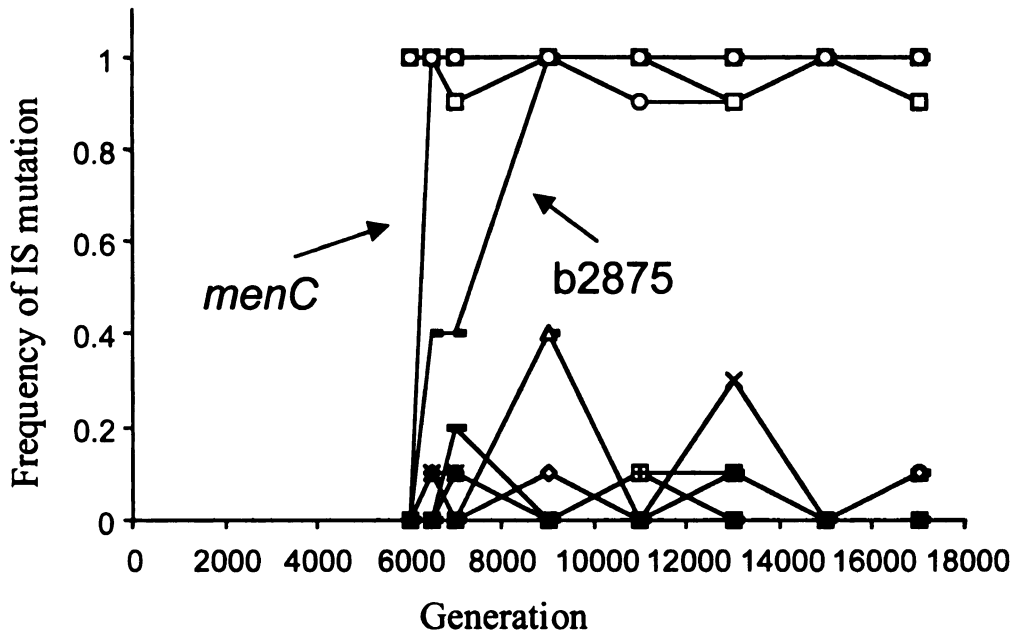


Figure 12: Frequency of new *IS186* and *IS150* mutations in the S clade, where each line represents a single distinct IS mutation. Data are based on RFLP from ten independent genotypes from each evolutionary time point. When the initial frequency is equal to 1, IS presence is ancestral and IS loss is derived. The reverse is true when initial frequency is equal to 0. Two mutations, *menC* and *b2875*, are highlighted. Other mutations, which do not fix, are detected at just one or two sampling periods.

were transient, *menC* and b2875 became fixed rapidly, suggesting that they might be directly beneficial to the cells in which they first arose.

Fitness consequences of IS mutations

The fitness effects of *menC* and b2875 were measured in three sets of isogenic constructs, as summarized in Table 2, using S, L, and Anc as genetic backgrounds. The reconstructed strains allowed us to determine the individual and combined effect of both mutations in all three backgrounds. Because these IS mediated events actually occurred in S during the course of the long term experiment, results from the S genetic background are most relevant for understanding the adaptive role of IS in this clade. Examination of IS effects in other backgrounds allowed us to determine whether the effects observed in S were dependent upon their genetic background (Anc vs. S vs. L).

Manipulations of menaquinone and *menC*

Two approaches were employed to study the role of *menC* in the evolution of S and L. In the first, fitness assays were conducted between S and L in the presence of exogenous menaquinone. Under normal experimental conditions, S and L show frequency-dependent fitness, where both clones exhibit a fitness advantage when rare. Figure 13 shows the results of competitions between S and L, both with and without exogenous menaquinone and at three initial S frequencies; the statistical summary is provided in Table 4. In contrast to normal conditions, where S cells exhibit an advantage versus L only when rare, the addition of menaquinone provides S with a fitness advantage versus

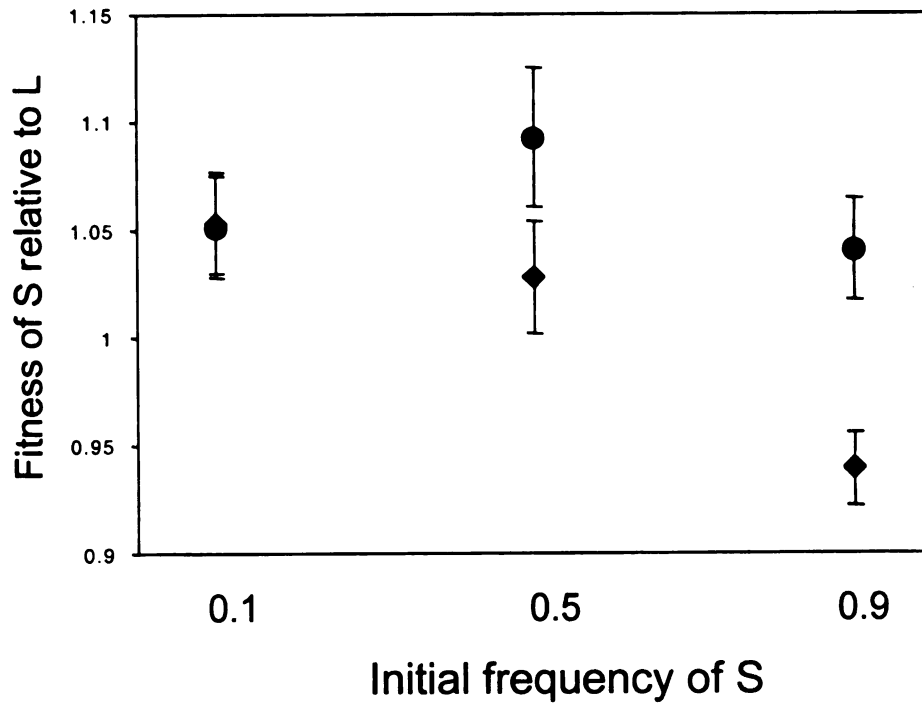


Figure 13: Frequency-dependent relative fitness of S versus L during one day competition assays. Diamonds show results from competition experiments conducted in DM25 medium. Squares show results of competition experiments conducted in DM25 with supplemented menaquinone. Error bars are standard errors based on fourteen replicates within each category.

Table 4: Analysis of covariance for fitness of S competed against L, with and without supplemented menaquinone, and at different initial frequencies.

Source	d.f.	MS	<i>F</i>	<i>P</i>
frequency	1	0.054	5.87	0.0186
menaquinone	1	0.007	0.74	0.3925
frequency*menaquinone	1	0.037	3.96	0.0501
error	80	0.009		

L at all frequencies. In other words, supplemental menaquinone appears to eliminate the S/L frequency-dependence. This shift is supported statistically by the significant effect of menaquinone supplementation ($P = 0.0186$) and by the suggestive interaction term between menaquinone treatment and initial frequency ($P = 0.0501$).

Because we cannot be certain that exogenous menaquinone mimics the relevant extracellular or intracellular concentrations of this molecule during S and L competition, we sought to evaluate the IS mutation in *menC* more directly by measuring the fitness of *menC* manipulated genotypes. As suggested earlier, the evolved *menC* knock out could either have been beneficial, in which case restoration to wild-type would be deleterious, or may have been neutral or even deleterious, in which restoration would be neutral or even beneficial. The first result would imply that this mutation was fixed by selection in S, while the alternative would suggest that this mutation fixed due to hitchhiking. We competed S with a restored *menC*⁺ allele against S/Ara⁺ and found that the functional *menC*⁺ allele reduced fitness by nearly 4% (Figure 14). The cost of allele restoration to wild type allows us to infer that the IS mutation in *menC* confers a direct benefit to S. In contrast to the fitness benefit observed in S, the *menC* deletion in Anc/ Δ *menC* or L/ Δ *menC* is neutral when either deletion construct is competed against a *menC*⁺ but otherwise isogenic strain. Using one-way analysis of variance, we find that the fitness effect of the *menC* state, however, is indeed significantly dependent on genetic background ($F_{2,37} = 5.76$, $P = 0.0066$).

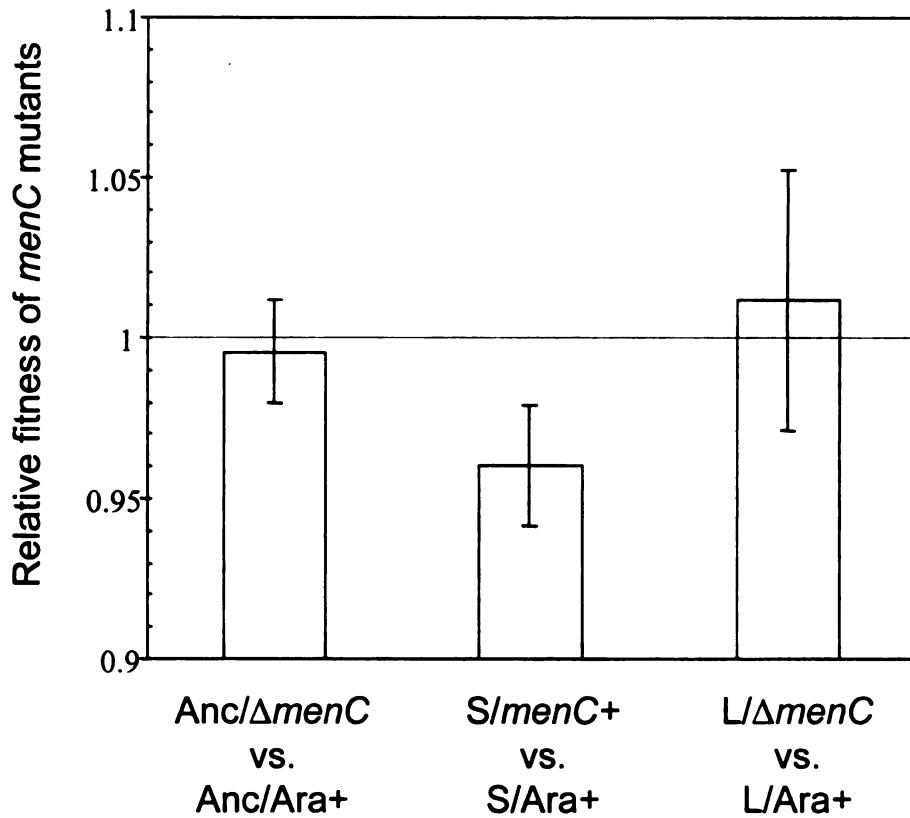


Figure 14: Fitness effects of *menC* allelic replacement in Anc, S, and L. For each competition, an Ara⁻ manipulated genotype was competed against an otherwise isogenic Ara⁺ strain. For both Anc/ $\Delta menC$ and L/ $\Delta menC$, the wild-type *menC* was replaced with the mutant deletion allele. In S/*menC*⁺, the wild-type allele was restored. Error bars represent 95% confidence intervals based on fifteen replicates for Anc and S, and ten replicates for L.

Restoration and deletion of b2875

The fitness effects of b2875 are shown in Figure 15. In contrast to *menC*, restoration to wild-type of b2875 in S is neutral, as it is in Anc. The b2875 deletion in L, however, is functionally lethal. Lethality in L/ Δ b2875 is only observed in DM25. In LB, cells grow to high densities, and on rich agar medium they form colonies indistinguishable from those of the unmanipulated genotype. We find that the fitness of S/b2875+ and Anc/ Δ b2875 do not differ from one another ($F_{1,28} = 2.31$, $P = 0.14$).

Restoration and deletion of both *menC* and b2875

Fitnesses of double mutants are shown in Figure 16. As was observed for each mutation alone, the fitness of double mutants was found to be highly dependent on genetic background ($F_{2,37} = 6.43105$, $P = 0.004$) based on one-way analysis of variance. Next, we examined evidence for epistatic interactions between *menC* and b2875 within each genetic background. The magnitude of epistatic effects between mutations was estimated by comparing the expected fitness of the double mutant (calculated by taking the product of the individual fitness measurements) to the observed fitness of the double mutant. As shown in Table 5, there is strong epistasis between *menC* and b2875 in L ($P < 0.0001$) that compensates for the lethality caused by in Δ *menC* in L. We do not observe any epistatic interaction between *menC* and b2875 in S or in Anc.

IS mutations and frequency dependence

The fitness effects of *menC* and b2875 on the frequency dependence between S and L are

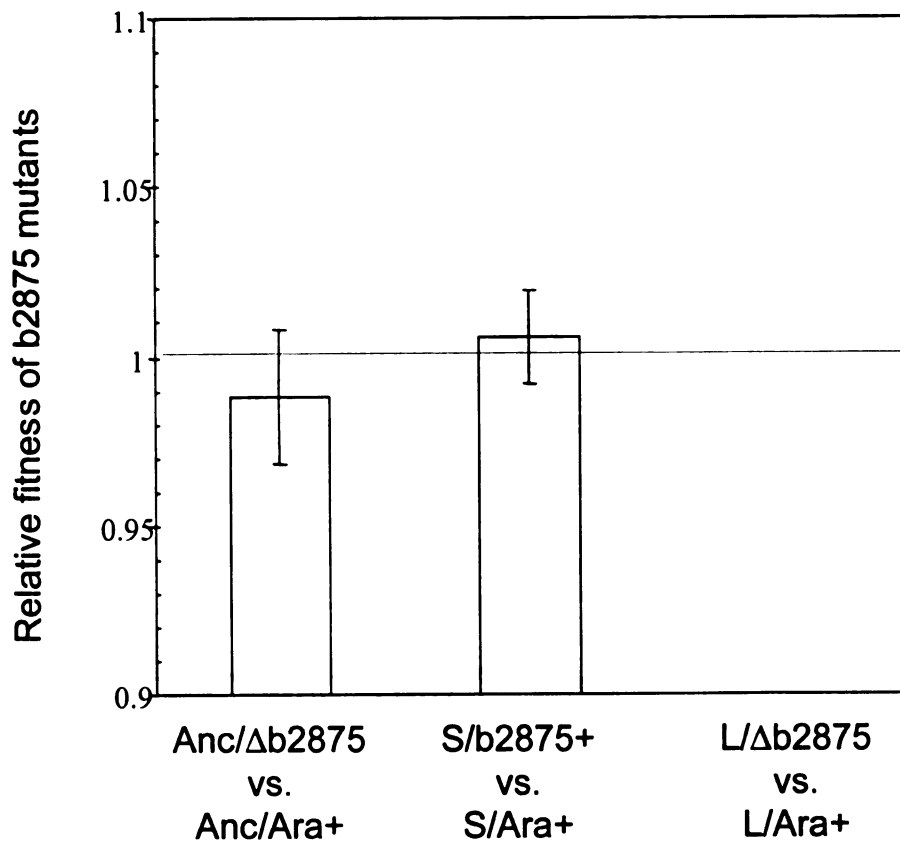


Figure 15: Fitness effects of b2875 allelic replacement in Anc, S, and L. For each competition, an Ara⁻ manipulated genotype was competed against an otherwise isogenic Ara⁺ strain. For both Anc/ Δ b2875 and L/ Δ b2875, wild-type b2875 was replaced with the mutant deletion allele. In S/b2875+, the wild-type allele was restored. Error bars represent 95% confidence intervals based on fifteen replicates for Anc and S. Fitness of L/ Δ b2875 could not be directly measured because this genotype did not grow in DM25; its fitness, therefore, is effectively 0.

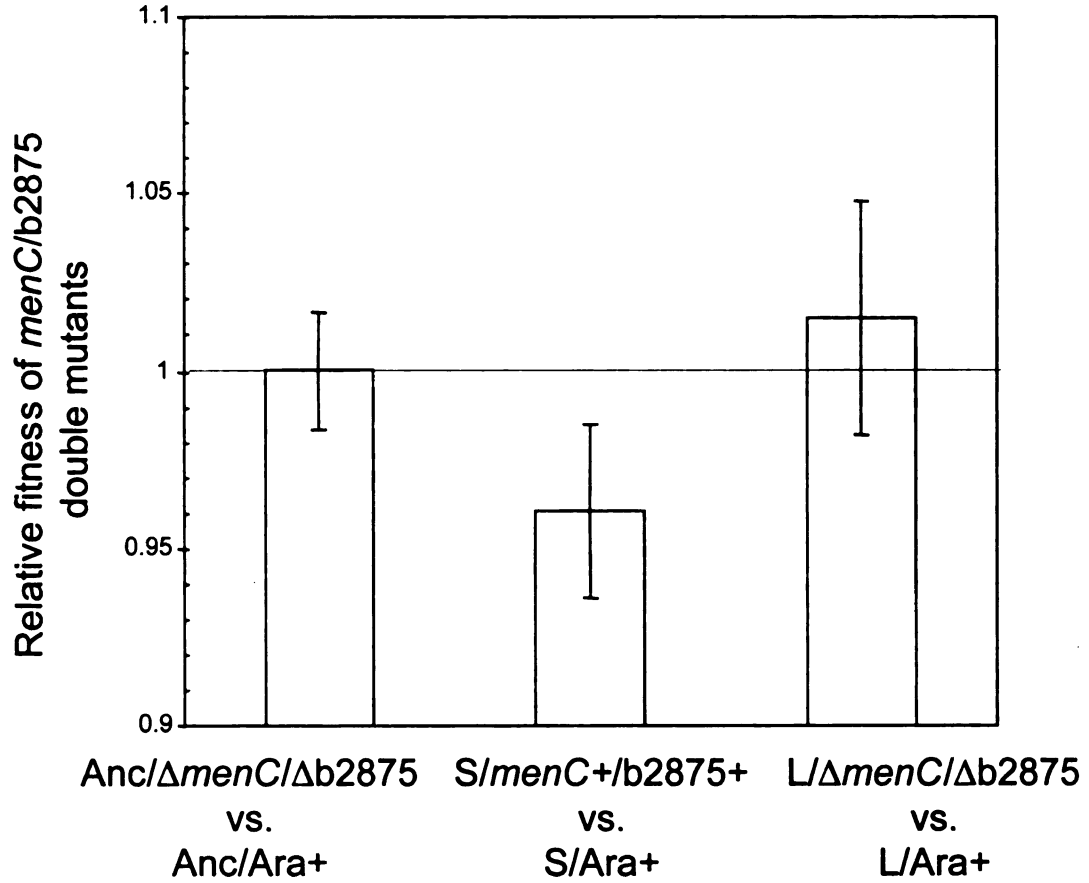


Figure 16: Fitness effects of double mutants containing both *menC* and *b2875* allelic replacements in Anc, S, and L. For each competition, an Ara⁻ manipulated genotype was competed versus an otherwise isogenic Ara⁺ strain. For both Anc/ $\Delta menC/\Delta b2875$ and L/ $\Delta menC/\Delta b2875$, the wild-type *menC* and *b2875* alleles were replaced with the mutant deletion alleles. In S/ $menC^+/b2875^+$, the wild-type alleles for both genes were restored. Error bars represent 95% confidence intervals based on fifteen replicates for Anc and S, and ten replicates for L.

Table 5. Comparison between observed and expected fitness of double mutants, the latter assuming a multiplicative model of gene interaction. Expected values were generated by calculating the product of the individual observations for each mutation and comparing them, using a paired t-test, to observed values for the double mutant. In the case of $\Delta b2875$ in L, we conservatively used a value of 0.01 for fitness rather than 0.

Genetic Background	Number of Paired Comparisons	Mean Expected Fitness	Mean Observed Fitness	Mean Difference Between Expected and Observed	Standard Error of Difference	<i>P</i>
S	15	0.966	0.9606	0.005	0.0277	0.7298
L	10	0.010	1.0152	1.005	0.0148	< 0.0001
Anc	15	0.984	1.0003	0.016	0.0235	0.1953

shown in Figure 17 and in analyzed statistically Table 6. Recall that during paired competition assays, the fitness of S versus L is frequency-dependent. For these competition experiments, S, or S that had either or both IS mutations restored to wild-type, was competed versus L. A three-way ANOVA was conducted which examined the influence of *menC* and b2875 restoration on S:L frequency-dependence. [Because of unbalanced data within categories, these analyses were conducted using ProcMixed in SAS which is robust to unbalanced designs. We observed no qualitative differences between these results and those obtained by three-way ANOVA. The latter results are thus presented for ease of viewing.] S and the reconstructed single and double mutants all have higher fitness versus L when initially rare, and lower fitness when initially common. These data indicate that neither the *menC* nor the b2875 IS insertion was sufficient to have caused the frequency-dependent relationship between S and L. In addition, there was a significant effects main effect of b2875 restoration, and a marginally significant interaction between *menC* and frequency (Table 6). The effect of b2875 restoration was positive in the presence of L, indicating that the evolved b2875::IS150 mutation in S was detrimental. This contrasts with the neutrality of b2875 restoration, also in S, but when L was not in the environment (Figure 15). The interaction of *menC* and frequency is such that the frequency-dependence is actually strong with the restored wild-type *menC*⁺ then with the evolved *menC*::IS186. Hence, the frequency-dependent interaction between S and L, even partially, by either IS mediated mutation in S.

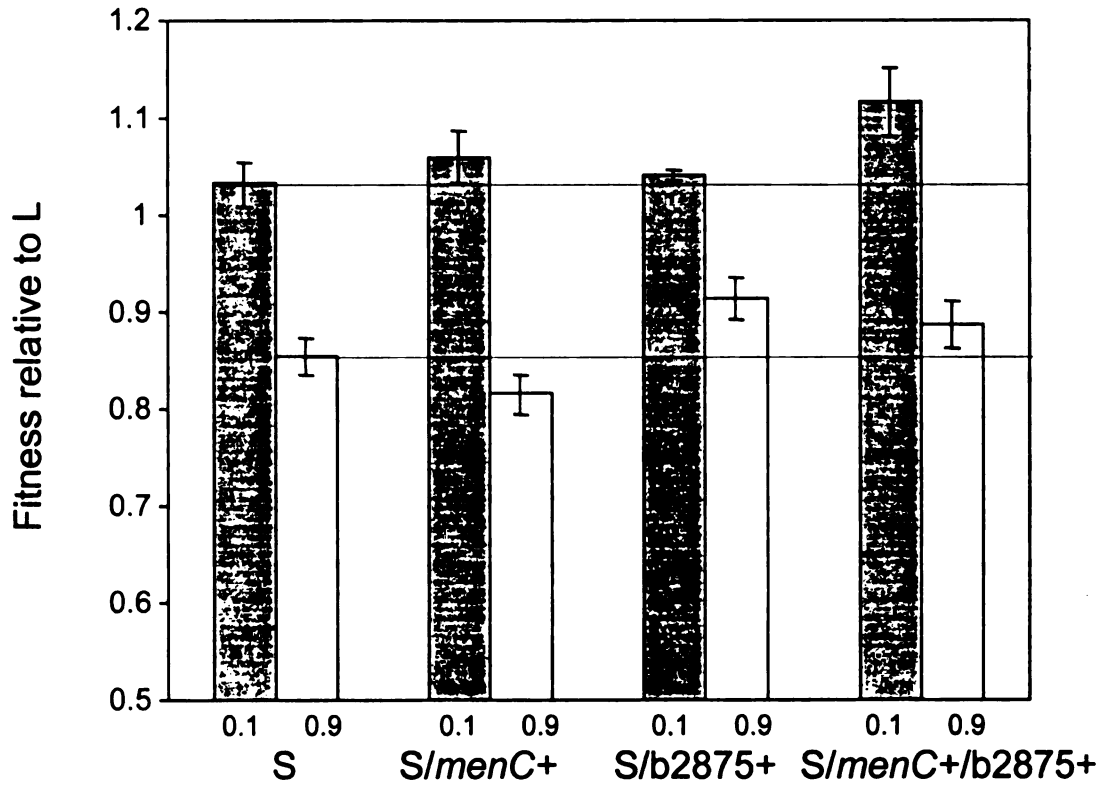


Figure 17: Frequency-dependent relative fitness of S, S/*menC*+, S/b2875, and S/*menC*+/b2875+, each competed against L. Competition assays were run at two initial frequencies, 1:9 and 9:1, for each of the four genotypes. Shaded bars show fitness when the initial frequency of the S genetic background was 10%, and clear bars show fitness when the initial frequency of the S genetic background was 90%. Error bars represent standard errors.

Table 6: Three way analysis of variance of fitness of S and S-derived mutants when competed against L. We examined the influence of *menC* and b2875, as well as their interaction across environments. The number of replicates within treatments is unequal. Data were thus first analyzed with Proc-Mixed in SAS, which is robust to unbalanced designs. The results and conclusions from this analysis did not differ from the more traditional three-way ANOVA, which are presented here.

Source	d.f	MS	<i>F</i>	<i>P</i>
frequency	1	0.4093	122.32	<0.0001
<i>menC</i>	1	<0.0001	0	0.9498
b2875	1	0.0329	9.83	0.0034
<i>menC</i> *b2875	1	0.0004	0.14	0.7087
frequency* <i>menC</i>	1	0.0115	3.44	0.0715
frequency*b2875	1	0.0005	0.16	0.6874
frequency* <i>menC</i> *b2875	1	<0.0001	0	0.9568
error	37	0.0033		

Discussion

We have previously documented the evolution of a balanced polymorphism, maintained by frequency-dependent selection, in a laboratory population of *E. coli* (Rozen and Lenski 2000). Two clades, **S** and **L**, have coexisted dynamically for more than 10,000 generations. In earlier work, we used RFLP analysis with IS elements as probes to study the phylogenetic history of **S** and **L**, and we identified a series of mutations that became fixed in either the **S** or **L** clade (D. Rozen, unpublished). We proposed that these fixed mutations were good candidates for being causally involved in the evolutionary dynamics of **S** and **L**. In this work, we genetically characterized five IS mediated mutations, and we examined the fitness consequences of two of these mutations alone and in combination, as well as in three different genetic backgrounds: Anc, **S**, and **L**.

Of the five IS mediated mutations that we characterized, two were simple insertions while three involved complex deletions (Table 3). Both mutations invaded rapidly in **S** (Figure 12) and became fixed. Because they invaded rapidly in **S**, and because their effects could be mimicked using standard genetic techniques, these two mutations were chosen for further analysis. The other mutations could not be constructed because they involved large deletions, could not be located in the *E. coli* genome, or both. However, it is important to note that these other IS mediated mutations may still have been important for the adaptation of **S** or **L**.

We identified, in S, an IS186 insertion into *menC*, one gene of an operon involved in the biosynthesis of menaquinone (Sharma et al. 1993). While it is likely that this insertion led to the inactivation of *menC*, the gene might still be transcribed because it shares an upstream promoter with *menB* that can initiate transcription of *menC* (Sharma et al. 1993). The reported functions of menaquinone in *E. coli* are said to be restricted to electron transport during anaerobic conditions (Meganathan 1996). Anaerobic conditions do not exist in the experimental regime used in the evolution experiment and competition assays. Thus, the loss of menaquinone production might be beneficial if expression during aerobic conditions is costly.

Two methods were employed to examine the potential role of *menC* in S and L. First, the effect of exogenously added menaquinone on the fitness of S versus L was examined. Under normal conditions, S and L show frequency-dependent fitness with S having an advantage when rare that depends on its ability to utilize product(s) secreted in the medium by L (Rozen and Lenski 2000). Contrary to the currently understood exclusively anaerobic importance of menaquinone, we hypothesized that menaquinone might influence the frequency-dependent relationship between S and L if it were a key metabolite. By adding menaquinone, S's fitness, might be decoupled from the frequency of L. The results shown in Figure 13 support this hypothesis. The fitness of S cells was significantly increased, especially when S was common, by the presence of menaquinone in the culture medium. The results of our study therefore indicate that menaquinone plays a physiological role even under aerobic conditions and moreover, a *menC* mutation

can be complemented by exogenous menaquinone. However, further work is required to determine whether, under normal conditions, this molecule itself or an intermediate product of its biosynthesis, is supplied by L to S. Recently, menaquinone, or one of the products generated during its biosynthesis, was also shown to be involved in a cross-feeding interaction in *Shewanella putrefaciens* under aerobic conditions (Newman and Kolter 2000).

The second approach taken to understanding the role of *menC* in the evolution of S and L was to measure the fitness consequences of mutations at this locus. We found that restoration to the wild-type allele in the S clone, S/*menC*⁺, caused a nearly 4% reduction in fitness when measured against S/Ara⁺ (Figure 14). Thus, the original IS mutation in S appears to have been directly beneficial. We also saw a marginally significant effect of *menC* on S and L frequency-dependence (Figure 17 and Table 6), but this effect acts to hinder, rather than promote to the coexistence of S and L. Hence, the fitness of the *menC*⁺ remained strongly frequency-dependent when competed against L. These findings indicate that the IS insertion in *menC* became fixed in the S clade owing to a general fitness advantage of this mutation, rather than a frequency-dependent benefit. These findings therefore also support the hypothesis that menaquinone production is costly, at least in the S background. However, *menC* is apparently neutral in the ancestral and L backgrounds (Figure 14), which might suggest that S has become deficient in the regulation as well as expression of *menC*.

At present, it remains unclear why the genetic and environmental manipulations produced different results with regard to the role of *menC* and menaquinone in the frequency-dependent interaction between S and L. One possibility is that exogenous menaquinone, while added at apparently physiologically relevant levels, exceeded that normally provided by L. A second possible explanation is that exogenous menaquinone is toxic to L cells; this would allow an advantage to S at all L frequencies. Finally, the possibility exists that the original IS insertion in *menC* in S caused polar effects on up or downstream genes, and that these effects were not reproduced by our genetic manipulations.

Because b2875 is as yet an uncharacterized ORF in *E. coli* and other species with homologues, it was not possible to generate a functional hypothesis concerning the effects of its alteration by mutation. However, as suggested with *menC*, if b2875 production is costly in our experimental regime, then its loss via an IS insertion might be beneficial. As shown in Figure 15, the fitness effect of the b2875 IS mutation was neutral when in competition with S/b2875⁺. Furthermore, its restoration to wild-type does not significantly alter S and L frequency-dependence (Figure 17 and Table 6), although b2875 restoration significantly improved the fitness of S/b2875⁺ when competed against L. As with *menC*, these effects are in a direction that would work to hinder, rather than contribute to, S invasion. Thus the b2875::IS150 mutation in S was either neutral or deleterious in the two contexts in which it was tested. Therefore, b2875 likely became fixed due to genetic hitchhiking with some (unknown) beneficial mutation.

While the $\Delta b2875$ deletion in *Anc*/ $\Delta b2875$ was neutral with respect to fitness, its effect in *L*/ $\Delta b2875$ was lethal (Figure 15). This lethality in *L* is environment dependent and is only observed in minimal, not rich medium. Thus $\Delta b2875$ behaves as an auxotrophic mutation in the *L* background, but not in *Anc* or *S*. It remains unclear, however, what essential function is not being served in *L* cells. Because we do not observe lethality of $\Delta b2875$ mutations in either *Anc* or *S*, we infer that $b2875$ interacts epistatically with one or more of the mutations that became fixed during the evolution of the *L* clade. We hope to identify these mutations to determine the cause of this unexpected lethal phenotype. We also acknowledge that while all efforts have been made to confirm that the mutation in *L* was cleanly generated, the possibility remains that an unrelated mutation arose elsewhere in the genome during construction of *L*/ $\Delta b2875$ that have caused the observed lethality. This possibility can be excluded by generating an independent $\Delta b2875$ mutation in the *L* background.

Epistatic effects between mutations are recognized when the combined effect of mutations is different from what is expected based on their individual effects. Epistatic interactions between deleterious mutations have been widely studied in evolutionary biology (Bohannan et al. 1999; deVisser et al. 1997; Elena and Lenski 1997) because they may be important for the origin and maintenance of sexual reproduction (Kondrashov 1998; Barton and Charlesworth 1998) and for the origin and maintenance of reproductive boundaries between species (Orr and Presgraves 2000). We examined the

interaction between *menC* and b2875 in three genetic backgrounds, and only in L do we observe evidence for epistasis (Table 5). Here the lethal effect of b2875 is fully compensated by the neutral mutation in *menC*, although the cause of this interaction remains obscure. In neither S nor Anc do we find evidence for epistatic interactions between the two mutations. We infer that epistatic interactions between *menC* and b2875 were not critical for the fixation of these mutations in S.

Epistasis is also observed when the effect of a single mutation varies as a function of the genetic background in which a mutation is expressed. This form of epistasis has been frequently observed in plant breeding (Doebley et al. 1995; Lukens and Doebley 1999). In these cases, however, large genomic regions have been introgressed between genotypes and it has not been possible to determine if background dependent epistasis resulted from the single agronomically important locus or from genes linked to it (Lukens and Doebley 1999). In the work described here, background effects are observed that are specific to individual loci. On the one hand, this is surprising given that S, L, and Anc are so recently diverged. Alternatively, these effects may have been anticipated because of the highly integrated nature of biochemical pathways (Neidhardt and Savageau 1996). If found to be general, background effects of this sort would indicate that the effects of mutations are highly contingent.

We identified and manipulated two IS insertion mutations that became fixed in one of two clades involved in a balanced polymorphism. Both mutations had demonstrable

effects on fitness, but only in specific contexts of the other mutation, a particular genetic background, a certain competitor, or some complex combination thereof. Given these complications, and the difficulty of reconstructing the exact circumstances during their initial invasion, we cannot unequivocally determine their modes of invasion. However, it appears most likely, given the preponderance of evidence, that *menC*::IS186 invaded clade S owing to the direct action of selection, whereas b2875::IS150 seems to have been fixed in the S clade by hitchhiking with some other, unknown, mutation.

Literature cited

- Altshul, S., F. Stephen, T. L. Madden, A. A. Shaffer, J. Zhang, Z. Zhang, W. Miller, and L. D. J. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* 25:3389-3402.
- Barton, N. H., and B. Charlesworth. 1998. Why sex and recombination? *Science* 281:1986-1990.
- Blot, M. 1994. Transposable elements and adaptation of host bacteria. *Genetica* 93:5-12.
- Bohannon, B. J. M., M. Travisano, and R. E. Lenski. 1999. Epistatic interactions can lower the cost of resistance to multiple consumers. *Evolution* 53:292-295.
- Charlesworth, B., P. Sniegowski, and W. Stephan. 1984. The evolutionary dynamics of repetitive DNA in eukaryotes. *Nature* 371:215-220.
- Cooper, V. S. 2000. Consequences of ecological specialization in long-term evolving populations of *Escherichia coli*. Ph.D. dissertation. Michigan State University, East Lansing, MI.
- Cooper, V. S., and R. E. Lenski. 2000. The population genetics of ecological specialization in evolving *E. coli* populations. *Nature* 407:736-739.
- Deonier, R. C. 1996. Native Insertion Sequence Elements: Locations, Distributions, and Sequence Relationships. Pp. 2000-2011 in F. C. Neidhardt, R. Curtiss III, J. L. Ingraham, E. C. C. Lin, K. Brooks Low, B. Magasanik, W. S. Reznikoff, M. Riley, M. Schaechter and H. E. Umbarger, eds. *Escherichia coli* and *Salmonella*: Cellular and Molecular Biology. ASM Press, Washington, D. C.
- deVisser, J. A. G. M., R. F. Hoekstra, and H. vandenEnde. 1997. Test of interaction between genetic markers that affect fitness in *Aspergillus niger*. *Evolution* 51:1499-1505.
- Doebley, J., A. Stec, and C. Gustus. 1995. *teosinte branched 1* and the origin of maize: evidence for epistasis and the evolution of dominance. *Genetics* 141:333-346.
- Doolittle, W. F., and C. Sapienza. 1980. Selfish genes, the phenotype paradigm and genome evolution. *Nature* 284:601-603.
- Elena, S. F., and R. E. Lenski. 1997. Test of synergistic interactions among deleterious

- mutations in bacteria. *Nature* 390:395-398.
- Hall, B. G. 1999a. Spectra of spontaneous growth-dependent and adaptive mutations at *ebgR*. *Journal of Bacteriology* 181:1149-1155.
- Hall, B. G. 1999b. Transposable elements as activators of cryptic genes in *E. coli*. *Genetica* 107:181-187.
- Kidwell, M. G., and M. B. Evgen'ev. 1999. How valuable are model organisms for transposable element studies? *Genetica* 107:103-111.
- Kidwell, M. G., and D. Lisch. 1997. Transposable elements as sources of variation in animals and plants. *Proceedings of the National Academy of Sciences, USA* 94:7704-7711.
- Kondrashov, A. S. 1998. Measuring spontaneous deleterious mutation process. *Genetica* 103:183-197.
- Lenski, R. E., M. R. Rose, S. C. Simpson, and S. C. Tadler. 1991. Long-term experimental evolution in *Escherichia coli* .I. Adaptation and divergence during 2,000 generations. *American Naturalist* 138:1315-1341.
- Lukens, L. N., and J. Doebley. 1999. Epistatic and environmental interactions for quantitative trait loci involved in maize evolution. *Genetical Research, Cambridge* 74:291-302.
- Mahillon, J., and M. Chandler. 1998. Insertion Sequences. *Microbiology and Molecular Biology Reviews* 62:725-774.
- Meganathan, R. 1996. Biosynthesis of the isoprenoid quinones menaquinone (Vitamin K₂) and ubiquinone (Coenzyme Q). Pp. 642-656 in F. C. Neidhardt, R. Curtiss III, J. L. Ingraham, E. C. C. Lin, K. Brooks Low, B. Magasanik, W. S. Reznikoff, M. Riley, M. Schaechter and H. E. Umbarger, eds. *Escherichia coli* and *Salmonella*: Cellular and Molecular Biology. ASM Press, Washington, D. C.
- Neidhardt, F. C., and M. A. Savageau. 1996. Regulation beyond the operon. Pp. 1310-1324 in F. C. Neidhardt, R. Curtiss III, J. L. Ingraham, E. C. C. Lin, K. Brooks Low, B. Magasanik, W. S. Reznikoff, M. Riley, M. Schaechter and H. E. Umbarger, eds. *Escherichia coli* and *Salmonella*: Cellular and Molecular Biology. ASM Press, Washington, D. C.
- Newman, D. K., and R. Kolter. 2000. A role for excreted quinones in extracellular electron transfer. *Nature* 405:94-96.

- Ochman, H., A. S. Gerber, and D. L. Hartl. 1987. Genetic applications of an inverse polymerase chain reaction. *Genetics* 120:621-623.
- Orr, H. A., and D. C. Presgraves. 2000. Speciation by postzygotic isolation: forces, genes and molecules. *Bioessays* 22:1085-1094.
- Papadopoulos, D., D. Schneider, J. Meier-Eiss, W. Arber, R. E. Lenski, and M. Blot. 1999. Genomic evolution during a 10,000-generation experiment with bacteria. *Proceedings of the National Academy of Sciences, USA* 96:3807-3812.
- Rodriguez, H., E. T. Snow, U. Bhat, and E. L. Loechler. 1992. An *Escherichia coli* plasmid-based, mutational system in which *supF* mutants are selectable: insertion elements dominate the spontaneous spectra. *Mutation Research* 270:219-231.
- Rozen, D. E., and R. E. Lenski. 2000. Long-term experimental evolution in *Escherichia coli*. VIII. Dynamics of a balanced polymorphism. *American Naturalist* 155:24-35.
- Sambrook, J., E. F. Fitch, and T. Maniatis. 1989. *Molecular Cloning*. Cold Spring Harbor Laboratory Press, New York.
- Schneider, D., E. Duperchy, E. Coursange, R. E. Lenski, and M. Blot. 2000. Long-term experimental evolution in *Escherichia coli*. IX. Characterization of Insertion Sequence-mediated mutations and rearrangements. *Genetics* 156:477-488.
- Sharma, V., R. Maganathan, and M. E. S. Hudspeth. 1993. Menaquinone (Vitamin K₂) biosynthesis: Cloning, nucleotide sequence, and expression of the *menC* gene from *Escherichia coli*. *Journal of Bacteriology* 175:4917-4921.
- Treves, D. S., S. Manning, and J. Adams. 1998. Repeated evolution of an acetate-crossfeeding polymorphism in long-term populations of *Escherichia coli*. *Molecular Biology and Evolution* 15:789-797.

Chapter 4

EXPLORING THE UTILITY OF MICROARRAYS FOR IDENTIFYING CAUSES OF ADAPTIVE DIFFERENCES BETWEEN S AND L

Introduction

An important goal of evolutionary biology is to identify the genetic factors which underlie the adaptive phenotypic differences between populations (Futuyma 1998; Rose and Lauder 1996). Microarrays, which allow the expression of every gene in a genome to be simultaneously measured, offer a promising new method of addressing this goal. Though microarrays have been primarily used to discover the function and regulation of newly identified genes (Arfin et al. 2000; Chu et al. 1988; deRisi et al. 1997; Duggan et al. 1999; Richmond et al. 1999; Tao et al. 1999), they may also be used by evolutionary biologists to gain insight into the mechanistic basis of evolution (Ferea et al. 1999). Using this approach, gene expression can be monitored and compared across genotypes with distinct evolutionary histories and fitness levels and in different environments of interest. Genes whose expression is increased or decreased across genotypes are genes whose products may be causally associated with fitness differences and are candidates for further manipulation. In addition, by examining suites of co-regulated genes it may be possible to trace expression changes to mutations at upstream regulatory loci.

In this work we use DNA microarrays to examine global patterns of gene expression from genotypes isolated from a long term laboratory population of *Escherichia coli* (Lenski et al. 1991; Lenski and Travisano 1994). Specifically, we examine the global

gene expression patterns of two clones, S and L, that were sampled from clades that have coexisted as a balanced polymorphism for more than 10,000 generations (Rozen and Lenski 2000). Our aims in this work are: 1) to identify candidate genes and pathways that may be causally involved in the evolution of the S and L clades, and 2) to assess more generally the utility of these methods for identifying the mechanisms of adaptation.

Lenski established twelve replicate populations of *E. coli* that have been maintained in serial batch culture for more than 20,000 generations (Cooper and Lenski 2000; Lenski et al. 1991; Lenski and Travisano 1994). In one of the twelve populations, we observed the origin of a balanced polymorphism (Rozen and Lenski 2000). Two clades, called S and L, had emerged by 3,000 and 6,500 generations, respectively, and have coexisted ever since. Over the course of many thousands of generations, S and L frequencies fluctuate between about 10% and 90%, but when individual S and L clones are grown together for a few hundred generations of time they reach a frequency-dependent equilibrium. The interaction results from three factors. First, L is able to invade a population of S cells owing to a ~20% maximum growth rate advantage on the limiting glucose. Second, despite its growth rate deficit, S is able to invade a population of L cells as a result of an ability to metabolize one or more products that L cells secrete during growth. Third, L cells die during periods of starvation at a greater rate than S cells, an effect that is somehow increased by the presence of S cells.

While the ecological and dynamical mechanisms which enable S and L to coexist in the

short term are known, the genetic and physiological bases of the persistence are not. In this work, we conducted three paired comparisons with gene arrays in order to begin to understand the genetic and phenotypic bases of S and L coexistence. Each comparison corresponds to one of the three factors outlined above. In Experiment 1, we compared the expression profiles of S and L during exponential growth on glucose. In Experiment 2, we compared the expression profiles of S cells exponentially growing alone and in the presence of L secretions (L conditioned media). Finally, in Experiment 3, we compared gene expression of S and L during starvation conditions, where L's rate of mortality exceeds S's by nearly 2% per hour.

Materials and Methods

Strains, Media and Growth Conditions

The two genotypes used in this study, S and L, were isolated from a single population of *E. coli* B that had been serially propagated in glucose limited batch culture for 18,000 generations of evolution (see Lenski et al. 1991 and Rozen and Lenski 2000, for further strain descriptions). S and L were isolated because of different physical appearance and colony growth on tetrazolium-arabinose (TA) indicator agar. They were subsequently found to exhibit frequency-dependence for fitness.

Unless otherwise noted, bacteria were cultured in Davis minimal medium supplemented

with thiamine hydrochloride (at $2 \times 10^{-3} \mu\text{g mL}^{-1}$) and glucose at $25 \mu\text{g mL}^{-1}$ (hereafter DM25), which supports a stationary phase cell density of $\sim 5 \times 10^7 \text{ mL}^{-1}$. Bacteria were grown in 10 mL culture tubes inoculated with 0.1 mL from a stationary phase culture and grown at 37°C to either mid-exponential or stationary phase (24 hours after inoculation). For all experiments, a number of replicate cultures were grown and combined prior to RNA extraction. This was done, rather than using a single larger culture volume, so that we could most faithfully reproduce the conditions of the environment in which the S and L clades evolved. Each experimental treatment was replicated twice. The same cultures and data were used for the identical treatments in Experiments 1 and 2 (S growing exponentially in DM25). Thus, there were a total of 10 expression analyses: three experiments, each with two treatments and replicated twice, with the one treatment overlap noted.

To obtain media that was conditioned by L cells for Experiment 2, L cells were grown for 24 hours and then vacuum filtered through $0.45 \mu\text{m}$ filters (Nalgene). This procedure removed all L cells from the media but retained L secretions. Following filtering, conditioned media was reconstituted with glucose to $25 \mu\text{g mL}^{-1}$ and then inoculated with S cells.

RNA Extraction, Probe Preparation, and Hybridization

For Experiments 1 and 2, cells were harvested at mid-log growth, and for Experiment 3,

cells were harvested at stationary phase. Otherwise, methods were identical for all treatments. Cells were vacuum filtered through 0.45µm filters (Nalgene), harvested by washing with TE buffer, and resuspended in 1.4 mL of Tris-EDTA buffer and RNAlater™ (Ambion) at 1:1. Cells were then pelleted by centrifugation for 2 minutes and resuspended in 100 µl TE buffer. Immediately following this step, RNA was purified using the RNeasy mini-column extraction kit (Qiagen) according to the manufacturer's recommendations. DNase treatment was performed directly on the RNeasy column using the Rnase-free Dnase kit (Qiagen). Extracted RNA was stored at -80°C until use.

For each sample, cDNA was prepared using the Panorama *E. coli* cDNA Labeling and Hybridization Kit (Sigma-Genosys). 4 µg total RNA was labeled with ³³P dCTP using a set of *E. coli* specific primers. The primer set did not include primers for ribosomal RNA; thus this abundant cellular RNA did not become reverse-transcribed.

Unincorporated nucleotides were removed by filtration through Sephadex G-25 gel-filtration spin columns.

Labeled cDNA was hybridized to Panorama *E. coli* gene arrays (Sigma-Genosys), which each contain 4,290 *E. coli* specific open reading frames (ORFs) spotted in duplicate. Hybridization was carried out in roller tubes as specified by the manufacturer. Briefly, arrays were pre-hybridized for 1 hour at 65°C in Hybridization Solution (Sigma-

Genosys). Next, labeled cDNA was incubated with the arrays at 65°C for at least 15 hours of hybridization. After washing, the arrays were wrapped in clear plastic wrap and exposed to PhosphorImager Screens (Molecular Dynamics) for 24-48 hours. Following exposure, labeled cDNA was stripped from the arrays by boiling for 20 minutes in a 10 mM Tris, 1mM EDTA, and 1% SDS solution. After stripping, arrays were either stored at -20°C, or prepared for additional hybridization experiments.

Analysis

Exposed PhosphorImager Screens were scanned on a Molecular Dynamics Storm Imager 860 at a resolution of 50 μm . The resulting image was analyzed using ArrayVision software (Imaging Research Inc.) and downloaded into a Microsoft Excel (1998) spreadsheet for further manipulation. Data from each array were normalized by adjusting the average (of duplicate spots) expression intensity of each ORF to the total image intensity. The \log_{10} of each value was used to allow convenient comparison between arrays. Relative expression was calculated as the \log_2 ratio of normalized values. Functional categories were assigned according to the annotated database for the *E. coli* K-12 MG1655 sequence (Blattner et al. 1997; Riley 1988).

In most microarray studies, expression data are collected either from one genotype exposed to distinct environmental conditions, or from a few genotypes exposed to the same environment. Relative expression values for every gene are then calculated to

determine those genes for which expression has changed, and thus which may be important for the phenomenon under investigation. An important difficulty has been to determine what magnitude of expression change "matters" biologically and which differences can be considered real amongst the mass of data. Arbitrarily, the threshold above which a difference in gene expression has been considered "significant" is 2-fold (Cavalieri et al. 2000; deRisi et al. 1997; Richmond et al. 1999; Tao et al. 1999).

However, this cut-off has more to do with the perceived reproducibility of microarrays (and with an interest in making the mass of data more tractable) than with anything biologically, or even statistically, meaningful. Here, we have used t-tests based on replicate expression values to determine whether treatments significantly differed from one another.

To determine the statistical significance of expression differences for each gene, we conducted t-tests between replicated treatments using $p < 0.05$ as the significance criterion (Arfin et al. 2000). Tests were performed on log-transformed normalized values. Given the very large number of tests conducted, some significant expression differences will result from chance alone. Because the primary function of this work was to identify possible biological trends, and because of the severity of correcting for 4,287 tests, we did not perform corrections for multiple comparisons. However, we also report which expression differences survive more stringent statistical criteria of $p < 0.01$ or $p < 0.001$.

Results

Total expression differences were examined in three experiments: 1) S versus L during exponential growth in DM25; 2) S growing in DM25 versus S in DM25 that was conditioned by L cells; 3) and S versus L during stationary phase in DM25. Within each experiment, the two conditions are referred to as treatments.

We found 867 genes whose expression differed significantly between S and L in Experiment 1, 894 differences in Experiment 2 and 200 differences in Experiment 3. By chance alone, we expect roughly 5%, or 215 genes, to show significant expression differences at the $p < 0.05$ level. Except for the comparison between S and L at stationary phase, we observed an ~ 4-fold excess of expression differences, which suggests that only a small fraction of the statistically significant expression differences are the result of false positive errors.

In Figures 18-20 we show scatter plots of expression values, and plots of the same data expressed as relative gene expression. Data are presented first for all genes, and then with increasing stringency of p -values from $p < 0.05$ to $p < 0.001$. As expected, the number of significant gene expression differences decreases with increasing statistical stringency. However, for experiments 1 and 2 (Figures 1 and 2) the number of significant differences remains several times higher than would be expected from chance alone, even at $p < 0.001$. This excess of statistically significant expression differences

Figure 18: Scatter plots of expression values, and histograms of relative expression for S and L exponentially growing in DM25. Data are presented first for all genes, and then with increasing stringency of p -values from $p < 0.05$ to $p < 0.001$. Note that many more genes exceed these statistical criteria than would be expected by chance.

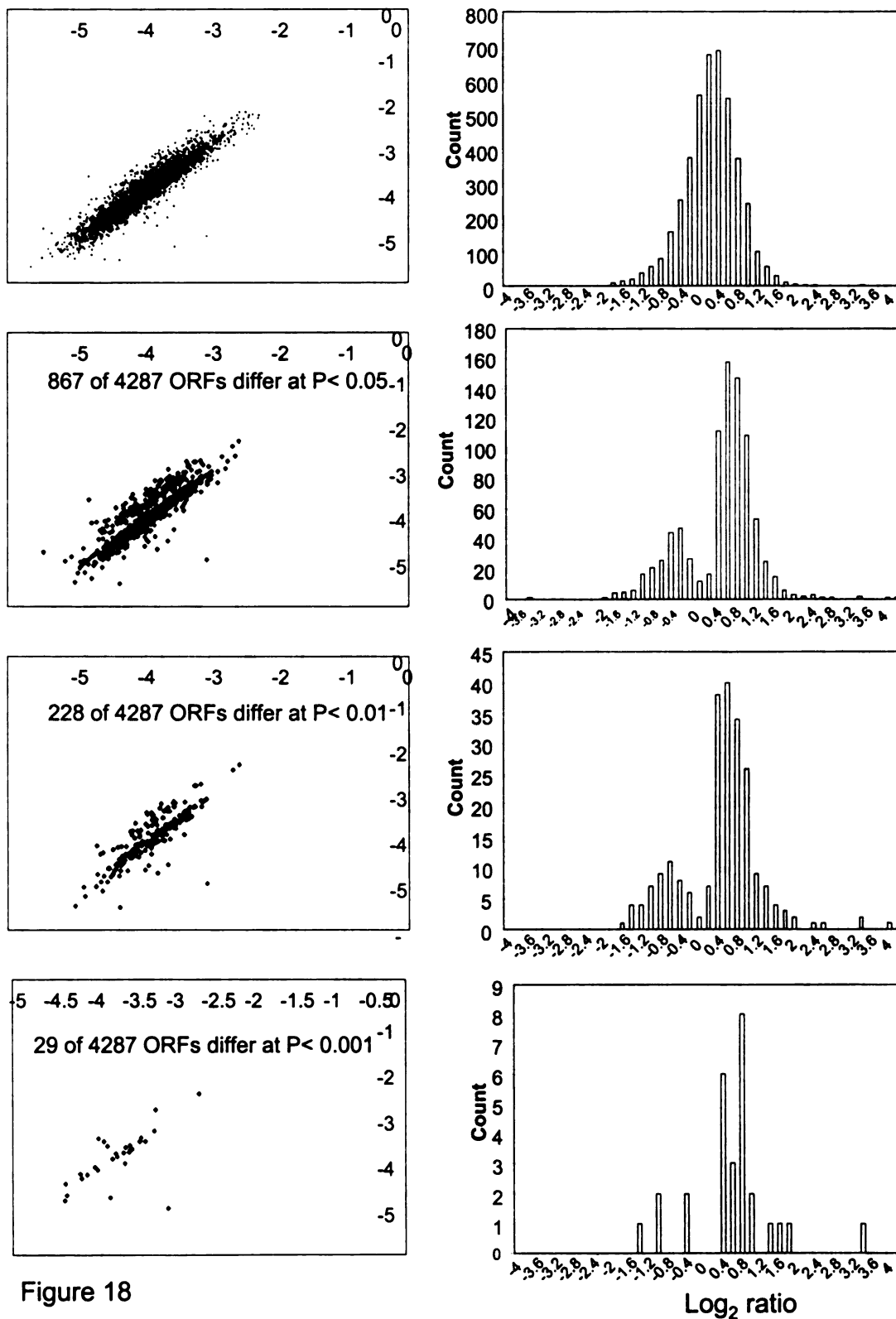


Figure 18

Figure 19: Scatter plots of expression values, and histograms of relative expression for S grown in DM25 and S grown in DM25 that has been conditioned by L cells. Data are presented first for all genes, and then with increasing stringency of p -values from $p < 0.05$ to $p < 0.001$. Note that many more genes exceed these statistical criteria than would be expected by chance.

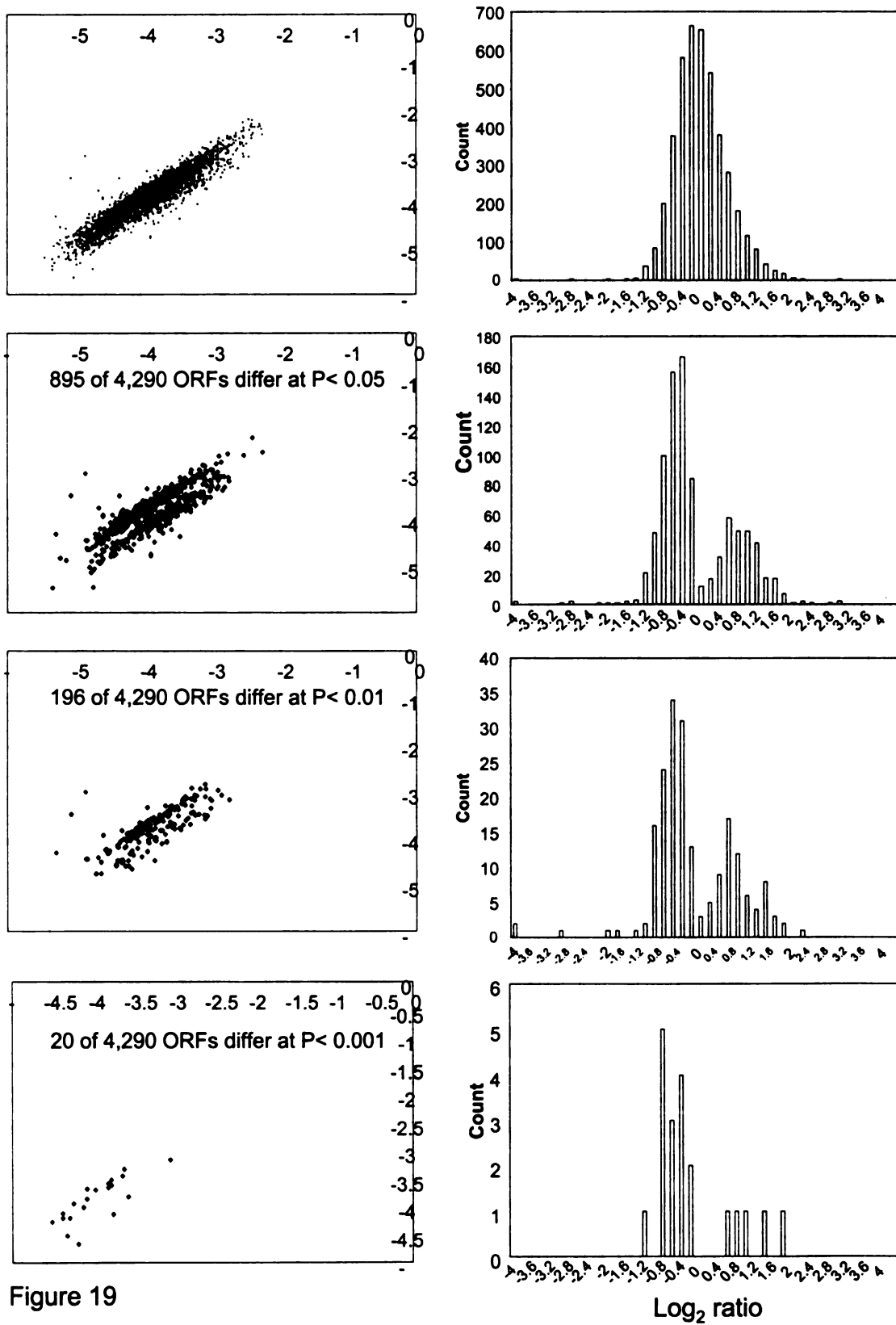


Figure 19

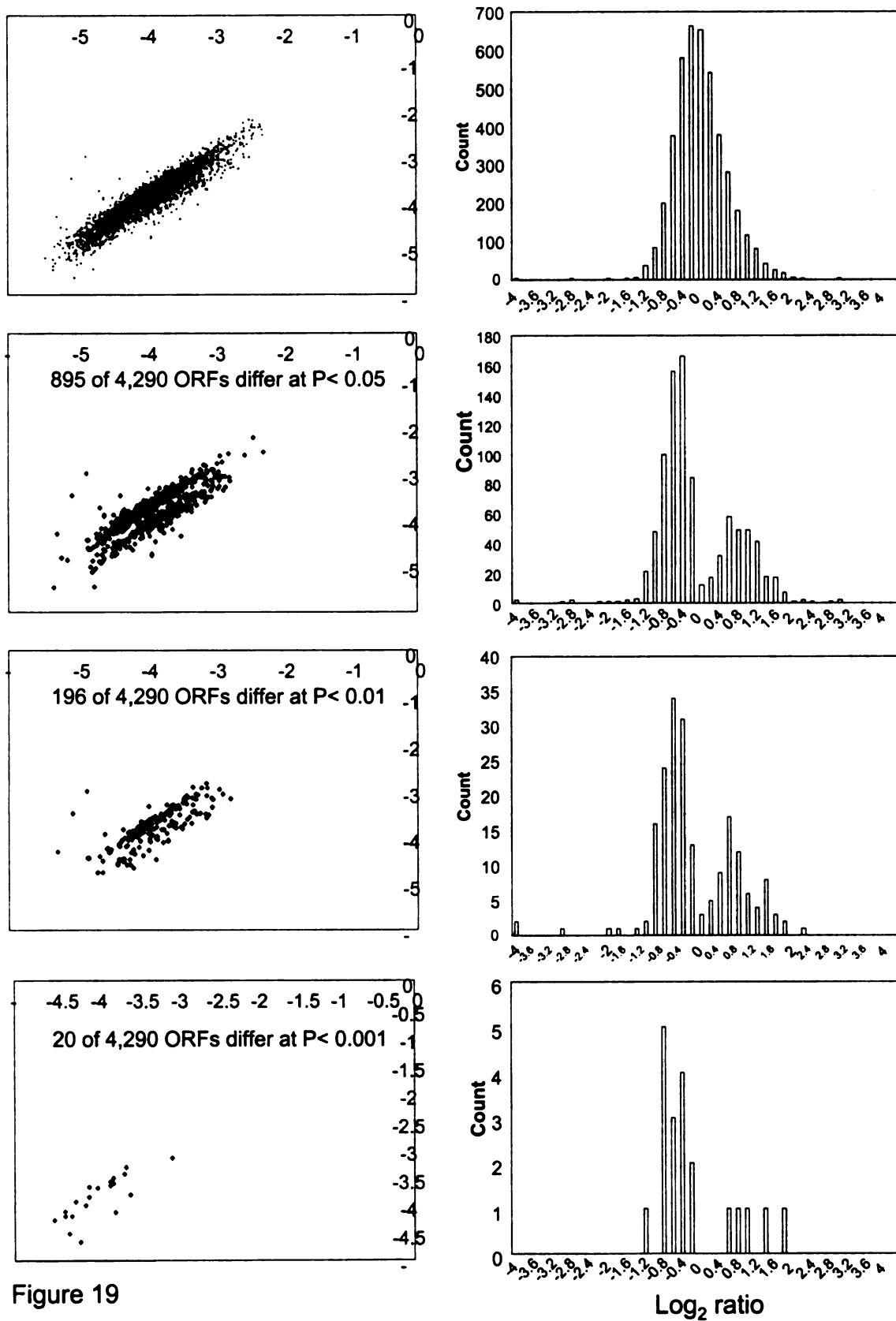


Figure 19

Figure 20: Scatter plots of expression values, and histograms of relative expression for S and L during stationary phase in DM25. Data are presented first for all genes, and then with increasing stringency of p -values from $p < 0.05$ to $p < 0.01$. Notice that there is no excess of significant deviations, unlike in the two earlier experiments.

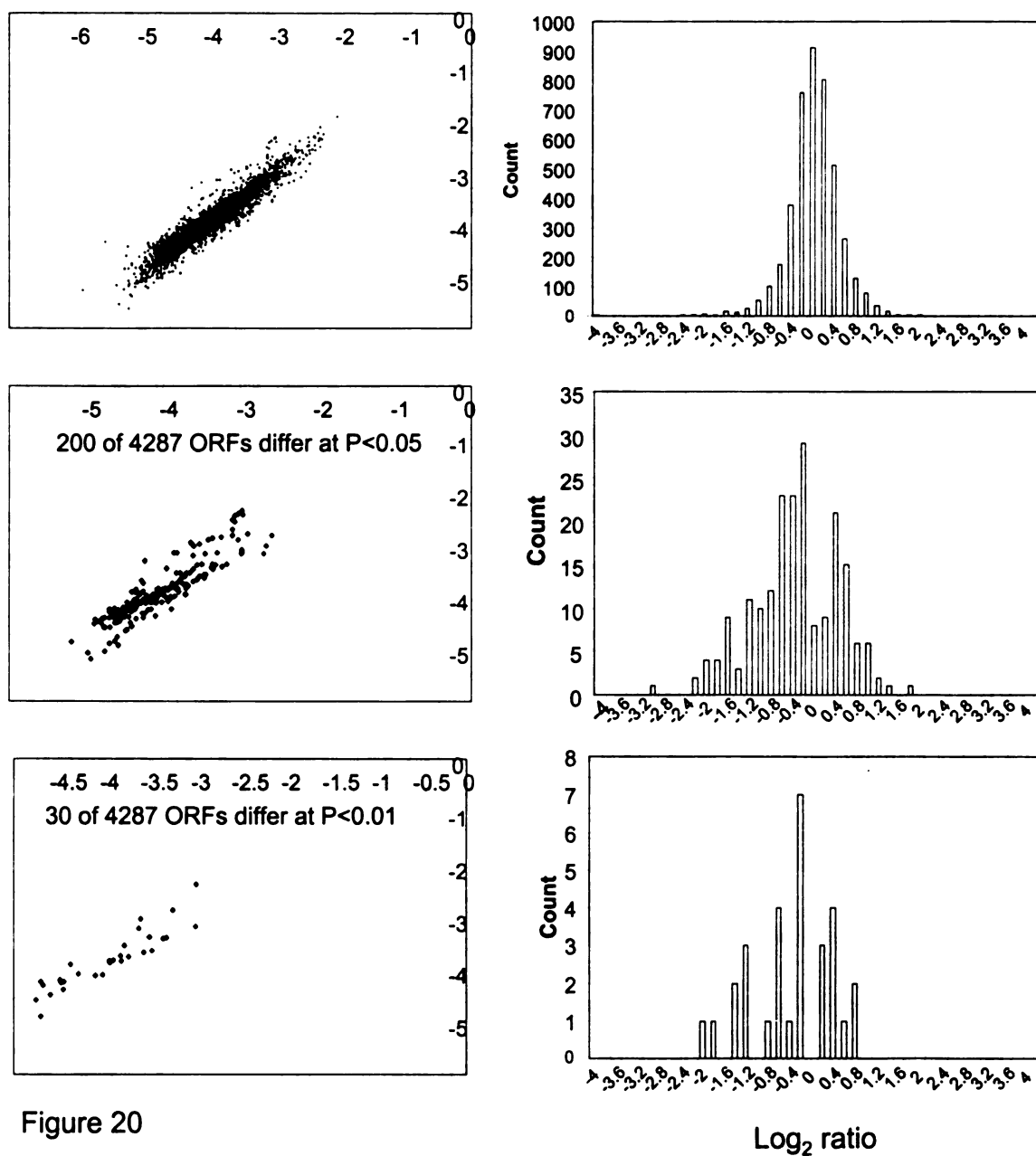


Figure 20

further indicates that most differences are not the result of false positives.

While the relative gene expression for some genes in Figures 18-20 (right panels) exceeds a 2-fold difference, most significant differences are much smaller. (Note that the data in Figures 18-20 are \log_2 transformed; thus, a relative expression value of 1 is equal to a 2-fold expression difference of non-transformed values). In Experiment 1, the relative expression of only 168 out of 867 statistically significant expression differences exceed a 2-fold difference. And for Experiment 2, only 172 out of 895 differences exceeds a 2-fold cut-off. By current methods, most of the gene expression differences identified as being statistically significant would not have been identified.

Functional Categorization of Gene Expression Differences

We find extensive gene expression differences between certain treatments. As a first step to understanding these differences at a functional level, and to better decipher the evolutionary differences between S and L, all statistically distinguishable genes were subdivided into functional classes. The 4,287 *E. coli* ORFs assayed here were divided into 22 broad functional categories, each of which contains a large number of individual genes and operons (Blattner et al. 1997; Riley 1988). This functional grouping allowed us to examine whether expression differences between treatments are essentially random with respect to biologically meaningful groupings, or are concentrated in specific functional classes.

In Tables 7-9, we show gene expression differences in each experiment, according to functional category. For each table, two statistical analyses are presented, each of which employs a binomial test (using the freeware Binomial Test from Bill Engels at the University of Wisconsin, Madison). First, we tested whether, within each functional category, the number of significant expression differences exceeded the 5% that would be expected by chance alone. Second, we tested whether one or the other treatment exhibited a greater number of more highly expressed genes in each functional class. In other words, we evaluated whether the *relative* expression of genes for each function (and for genes that were already found to differ between treatments) is distributed equally among treatments. For example, in Table 7, the first functional category is "Amino acid biosynthesis and metabolism", which contains 131 genes. Of these 131 genes, 27 (20.61%) show a significant difference in expression between S and L during exponential growth. Then, of these 27 genes, 8 show higher relative expression in S, while 19 show higher relative expression in L. This distribution also differs significantly from the null expectation of an equal distribution of relative expression differences between S and L.

In both Experiments 1 and 2 (Tables 7 and 8), most functional categories show significantly different relative expression across treatments. In contrast, in Experiment 3 (Table 9), only one functional group displays an excess of statistically significant expression differences. For nearly all functional categories in Experiments 1 and 2, one treatment or the other shows a disproportionate number of genes with higher relative expression. Such differences could result either from increased expression in one

Table 7. Comparison of expression differences between S and L during exponential growth in DM25. Genes are arranged according to function (see text). In column 2, I show the total number of genes that significantly differ between S and L. Text shown in bold in column 3 indicates an excess of significant differences beyond the null expectation of 5%. In columns 4 and 5, total differences are divided according to treatment. Bold text in columns 4 and 5 indicates that one treatment contained significantly more cases of elevated expression than the other. Both tests are based on the binomial distribution, and bold values denote deviations at $p < 0.05$.

	Total genes per functional category	Total differences between S and L during exponential growth in DM25	% Total genes showing significant differences	Genes Higher in S	Genes Higher in L
Total	4287	867	20.22	656	211
Amino acid biosynthesis and metabolism	131	27	20.61	8	19
Biosynthesis of cofactors, prosthetic groups and carriers	103	19	18.45	13	6
Carbon compound catabolism	130	27	20.77	25	2
Cell processes (incl. adaptation, protection)	188	58	30.85	58	0
Cell structure	182	39	21.43	37	2
Central intermediary metabolism	188	43	22.87	36	7
DNA replication, recombination, modification and repair	115	28	24.35	18	10
Energy metabolism	243	73	30.04	73	0
Fatty acid and phospholipid metabolism	48	8	16.67	1	7
Hypothetical, unclassified, unknown	1631	185	11.34	76	109
Membrane proteins	13	7	53.85	7	0
Nucleotide biosynthesis and metabolism	58	5	8.62	1	4
Other known genes	26	2	7.69	0	2
Phage, transposon, or plasmid	87	36	41.38	35	1
Putative chaperones	9	1	11.11	1	0
Putative enzymes	251	60	23.90	52	8
Putative regulatory proteins	133	57	42.86	57	0
Putative transport proteins	146	52	35.62	52	0
Regulatory function	45	9	20.00	9	0
Structural proteins	42	22	52.38	22	0
Transcription, RNA processing and degradation	55	10	18.18	4	6
Translation, post-translational modification	182	32	17.58	6	26
Transport and binding proteins	281	67	23.84	65	2

Table 8. Comparison of expression differences between S growing in DM25 and S growing in DM25 conditioned by L cells. Genes are arranged according to function (see text). In column 2, I show the total number of genes that significantly differ between the two treatments. Text shown in bold in column 3 indicates an excess of significant differences beyond the null expectation of 5%. In columns 4 and 5, total differences are divided according to treatment. Bold text in columns 4 and 5 indicates that one treatment contained significantly more cases of elevated expression than the other. Both tests are based on the binomial distribution, and bold values denote deviations at $p < 0.05$.

	Total genes per functional category	Total differences between S growing in DM25 and S growing in DM25 plus L conditioned media	% Total genes showing significant differences	Genes Higher in S in DM25	Genes Higher in S in L- conditioned media
Total	4287	894	20.84	600	294
Amino acid biosynthesis and metabolism	131	24	18.32	15	9
Biosynthesis of cofactors, prosthetic groups and carriers	103	22	21.36	16	6
Carbon compound catabolism	130	26	20.00	19	7
Cell processes (incl. adaptation, protection)	188	58	30.85	53	5
Cell structure	182	42	23.08	35	7
Central intermediary metabolism	188	42	22.34	33	9
DNA replication, recombination, modification and repair	115	33	28.70	21	12
Energy metabolism	243	68	27.98	66	2
Fatty acid and phospholipid metabolism	48	4	8.33	1	3
Hypothetical, unclassified, unknown	1631	255	15.62	76	179
Membrane proteins	13	5	38.46	5	0
Nucleotide biosynthesis and metabolism	58	9	15.52	7	2
Other known genes	26	6	23.08	1	5
Phage, transposon, or plasmid	87	24	27.59	22	2
Putative chaperones	9	2	22.22	1	1
Putative enzymes	251	62	24.70	49	13
Putative regulatory proteins	133	45	33.83	43	2
Putative transport proteins	146	29	19.86	29	0
Regulatory function	45	12	26.67	10	2
Structural proteins	42	20	47.62	20	0
Transcription, RNA processing and degradation	55	8	14.55	3	5
Translation, post-translational modification	182	25	13.74	12	13
Transport and binding proteins	281	73	25.98	63	10

Table 9. Comparison of expression differences between S and L during stationary phase. Genes are arranged according to function (see text). In column 2, I show the total number of genes that significantly differ between the two treatments. Text shown in bold in column 3 indicates an excess of significant differences beyond the null expectation of 5%. In columns 4 and 5, total differences are divided according to treatment. Bold text in columns 4 and 5 indicates that one treatment contained significantly more cases of elevated expression than the other. Both tests are based on the binomial distribution, and bold values denote deviations at $p < 0.05$.

	Total genes per functional category	Total differences between S and L during stationary phase	% Total genes showing significant differences	Genes Higher in L	Genes Higher in S
Total	4287	200	4.67	139	61
Amino acid biosynthesis and metabolism	131	9	6.87	8	1
Biosynthesis of cofactors, prosthetic groups and carriers	103	10	9.71	9	1
Carbon compound catabolism	130	3	2.31	3	0
Cell processes (incl. adaptation, protection)	188	5	2.66	4	1
Cell structure	182	9	4.95	6	3
Central intermediary metabolism	188	10	5.32	9	1
DNA replication, recombination, modification and repair	115	5	4.35	2	3
Energy metabolism	243	12	4.94	9	3
Fatty acid and phospholipid metabolism	48	1	2.08	0	1
Hypothetical, unclassified, unknown	1631	55	3.37	21	34
Membrane proteins	13	0	0	0	0
Nucleotide biosynthesis and metabolism	58	3	5.17	1	2
Other known genes	26	4	15.38	4	0
Phage, transposon, or plasmid	87	16	18.39	16	0
Putative chaperones	9	0	0	0	0
Putative enzymes	251	8	3.19	5	3
Putative regulatory proteins	133	12	9.02	10	2
Putative transport proteins	146	12	8.22	10	2
Regulatory function	45	1	2.22	1	0
Structural proteins	42	4	9.52	4	0
Transcription, RNA processing and degradation	55	2	3.64	1	1
Translation, post-translational modification	182	7	3.85	6	1
Transport and binding proteins	281	12	4.27	10	2

treatment or a decrease in the other. At present, we cannot distinguish between these possibilities.

Expression Differences in S Compared with L During Exponential Growth in DM25

We observe an excess of genes with higher relative expression in L than S for genes that are related to translation (Table 7). Within this class, 21 of 36 genes encode ribosomal proteins and 6 encode aminoacyl tRNA synthases. In addition, we find a greater number genes involved in amino acid biosynthesis with higher relative expression in L than in S.

For most other functional categories, most genes' relative expression is higher for S than for L. Because S is known to cross-feed one or more L secreted products, it is especially interesting that many genes involved in transport and metabolism show higher relative expression in S. Table 10 lists the 65 specific genes involved in transport of amino acids, carbohydrates, and small molecules that have higher relative expression in S than L, as well as the two genes showing the opposite direction. We also observe higher relative expression in S for genes involved in carbon compound degradation, and throughout central metabolism and energy metabolism for both anaerobic and aerobic respiration. Within the broad categories of central and energy metabolism, we see expression increases in fermentation, glycolysis, and the TCA cycle.

Several global regulatory genes are more highly expressed in S, most notably *cyaA*, which encodes adenylate cyclase that converts AMP to cAMP. cAMP regulates many

Table 10: Expression differences between S and L during log growth in DM25 for genes in the broad functional category "Transport and Binding Proteins".

Functional Class	Gene	Log ₂ (S/L)	P value	Gene Function
Higher in S				
Not classified	b0709	0.8035	*	
	b0829	0.5661	*	
	mdlA	0.4464	*	
	yejF	0.8375	*	hypothetical ABC transporter in bcr 5'region
	b0830	0.5753	**	
	<i>frvR</i>	0.3584	*	putative frv operon regulatory protein
	<i>mdlB</i>	0.7501	*	
	<i>b0831</i>	0.3007	*	
	<i>ybaL</i>	0.5111	*	hypothetical protein in gsk 3'region
	<i>cirA</i>	0.6885	*	colicin I receptor precursor
Outer membrane constituents				
Protein, peptide secretion	<i>secD</i>	0.8371	*	protein-export membrane protein
	<i>ffh</i>	0.1918	*	signal recognition particle protein
	<i>secF</i>	1.5740	*	protein-export membrane protein
	<i>oppB</i>	0.8137	*	oligopeptide transport system permease protein
	<i>dppD</i>	0.3842	*	dipeptide transport ATP-binding protein
Transport of Amino acids, amines	<i>msyB</i>	1.2595	**	multicopy suppressor of SecY
	<i>gltK</i>	1.6667	*	glutamate/aspartate transport system permease protein
	<i>potB</i>	0.4086	*	spermidine/putrescine transport system permease protein
	<i>gltJ</i>	0.6809	***	glutamate/aspartate transport system permease protein
	<i>cadB</i>	1.1057	*	probable cadaverine/lysine antiporter
	<i>gltP</i>	0.4834	*	proton glutamate symport protein
	<i>potE</i>	1.5332	*	putrescine-ornithine antiporter
	<i>aroP</i>	0.5291	**	aromatic amino acid transport protein
	<i>tdcC</i>	0.7899	*	threonine-serine permease
	<i>yaaJ</i>	0.4626	*	hypothetical 51.7 kD protein
Transport of Anions	<i>glnQ</i>	0.6456	*	glutamine transport ATP-binding protein
	<i>mtr</i>	0.6700	**	tryptophan-specific permease
	<i>narK</i>	0.8305	*	nitrite extrusion protein
	<i>pitB</i>	0.6268	*	probable low-affinity inorganic phosphate transporter

Table 10 (continued):

Transport of Carbohydrates, organic acids, alcohols	<i>modE</i>	0.6374	*	molybdenum transport ATP-binding protein
	<i>cynX</i>	0.8052	*	
	<i>modC</i>	1.1008	*	
	<i>gntU_1</i>	0.2034	*	
	<i>rbsB</i>	0.3614	*	periplasmic ribose-binding protein precursor
	<i>melB</i>	0.6110	*	thiomethylgalactoside permease II
	<i>fruB</i>	0.4820	*	pts system, fructose-specific IIA/FPR component
	<i>celB</i>	0.6738	*	glycerol-3-phosphatase transporter
		0.6824	*	glucuronide permease
	<i>rbsD</i>	0.1566	*	high affinity ribose transport protein
	<i>xylE</i>	0.8226	**	xylose-proton symport
	<i>gntT</i>	1.0605	*	high-affinity gluconate transporter
	<i>gntU_2</i>	0.9875	*	
	<i>ascF</i>	0.5584	**	phosphotransferase enzyme IABC- Asc
Transport of Cations	<i>nhaA</i>	0.9796	*	Na(+)/H(+) antiporter 1
	<i>fhuD</i>	1.0427	*	ferrichrome-binding periplasmic protein precursor
	<i>fecD</i>	0.3973	*	iron(III) dicitrate transport system permease protein
	<i>nhaR</i>	1.1699	*	transcriptional activator protein
	<i>fepA</i>	1.3763	*	ferrienterobactin receptor precursor
	<i>corA</i>	0.9709	*	
	<i>fes</i>	1.1672	*	enterochelin esterase
	<i>lnt</i>	1.1627	*	apolipoprotein N-acyltransferase
	<i>kefC</i>	0.7722	*	glutathione-regulated potassium-efflux system protein
	<i>chaA</i>	0.3132	***	putative calcium/proton antiporter
	<i>fepD</i>	0.6226	*	ferric enterobactin transport protein
	<i>kdpC</i>	1.3472	*	potassium-transporting ATPase C chain
	<i>mgtA</i>	1.4042	*	Mg(2+) transport ATPase, P-type 1
	<i>trkG</i>	0.3147	*	trk system potassium uptake protein
	<i>fepB</i>	0.8837	*	ferrienterobactin-binding periplasmic protein precursor
	<i>fhuE</i>	0.3681	*	outer-membrane receptor for Fe(III)- coprogen, Fe(III)-ferrioxamine B and Fe(III)-rhodotrucic acid precursor
	<i>nikE</i>	0.5424	*	

Table 10 (continued):

Transport of Nucleosides, purines, pyrimidines	<i>nupC</i>	2.2742	**	nucleoside permease
	<i>codB</i>	0.4789	*	cytosine permease
Transport of small molecules: Other	<i>cydC</i>	0.6299	***	transport ATP-binding protein
	<i>msbA</i>	0.6818	*	probable transport ATP-binding protein
Higher in L				
Not classified	b1485	-0.2082	*	
Transport of Anions	<i>cysP</i>	-0.7658	*	thiosulfate-binding protein precursor

0.5 > P > 0.1, *. 0.01 > P > 0.001, **. 0.001 > P > 0.0001, ***.

hundreds of genes in *E. coli* (Botsford and Harman 1992; Saier et al. 1996). To examine the possible consequences of *cyaA* up-regulation in S, we looked at whether known cAMP-regulated genes showed coordinate expression differences between S and L. Table 11 lists the subset of genes that are regulated by *cyaA* (as determined through expression studies and by the presence of a CRP binding site) (Salgado et al. 2000) which display significant differences between S and L. Among these are many of the transport and metabolic genes listed in Table 10. Of the 53 genes shown, only 5 have higher expression in L, while the remainder are elevated in S. Because not all of the direct targets of cAMP in *E. coli* are known, Table 11 is not exhaustive. In addition, Table 11 does not include genes and operons that are indirectly influenced by cAMP. Thus it is possible that many more than 53 of the 867 significant expression differences trace to expression differences at *cyaA*.

Of the 867 genes whose gene expression significantly differs between S and L, 185 are in genes that, as yet, have no recognized function. For this functional class, we find more genes that show higher relative expression in L than in S. However, this excess in L may be an artifact of the normalization process, whereby total expression in each treatment must sum to 1. Given that most functional categories are more highly expressed in S, this "catch all" category may artificially appear to be more highly expressed in L.

S Grown in DM25 versus S Grown in DM25 Conditioned by L Secretions

When grown in DM25 conditioned by L secretions, the growth rate of S cells increases

Table 11: Genes that are putatively regulated by cAMP that differ between S and L during exponential growth in DM25.

Functional Class	Gene	log ₂ (S/L)	p-value	Gene Description
Higher in S				
Amino acid biosynthesis and metabolism	<i>dadA</i>	0.3423	**	D-amino acid dehydrogenase
Central intermediary metabolism	<i>gntK</i>	1.8023	*	thermoresistant glucokinase
	<i>speE</i>	0.3668	*	spermidine synthase
	<i>speC</i>	0.6952	*	
	<i>speF</i>	0.1577	*	ornithine decarboxylase
	<i>glnK</i>	0.7768	**	nitrogen regulatory protein P-II
	<i>galF</i>	1.4811	***	UTP-glucose-1-phosphate uridylyltransferase
Cell processes	<i>caiC</i>	1.1941	*	probable crotonobetaine/carnitine- CoA ligase
	<i>caiA</i>	1.2769	***	probable carnitine operon oxidoreductase
	<i>cheR</i>	0.6889	*	chemotaxis protein methyltransferase
	<i>cheA</i>	0.5638	*	chemotaxis protein
	<i>treA</i>	0.4374	*	periplasmic trehalase precursor
	<i>proX</i>	0.1932	**	glycine betaine-binding periplasmic protein precursor
Carbon compound catabolism	<i>melA</i>	0.3275	*	alpha-galactosidase
	<i>bglB</i>	0.7059	*	phospho-beta-glucosidase
	<i>lacZ</i>	0.6241	*	beta-galactosidase
	<i>bglG</i>	0.4840	*	positive regulatory protein;
	<i>treC</i>	0.3673	***	trehalose-6-phosphate hydrolase
	<i>srlR</i>	0.4690	*	
	<i>galK</i>	0.6895	**	galactokinase
	<i>fucI</i>	0.6094	**	L-fuculose isomerase
	<i>malT</i>	0.6398	*	
	<i>rhaB</i>	0.3772	*	rhamnulokinase
Energy metabolism	<i>glpG</i>	0.6227	*	
	<i>fixA</i>	1.0726	*	
	<i>fixB</i>	1.4057	*	
	<i>gltA</i>	1.8288	**	
	<i>sucD</i>	1.5075	*	succinyl-coA synthetase alpha chain

Table 11 (continued):

	<i>sdhB</i>	0.8907	*	succinate dehydrogenase iron-sulfur protein
	<i>sucA</i>	1.4047	*	2-oxoglutarate dehydrogenase E1 component
Regulatory function	<i>cyaA</i>	0.5066	*	adenylate cyclase
Cell structure	<i>glgP</i>	0.4910	*	alpha-glucan phosphorylase
	<i>glgX</i>	0.7528	*	glycogen operon protein
	<i>cirA</i>	0.6885	*	colicin I receptor precursor
Transport and binding proteins	<i>gltK</i>	1.6667	*	glutamate/aspartate transport system permease protein
	<i>gltJ</i>	0.6809	***	glutamate/aspartate transport system permease protein
	<i>gltP</i>	0.4834	*	proton glutamate symport protein (glutamate- aspartate carrier
	<i>tdcC</i>	0.7899	*	threonine-serine permease
	<i>glnQ</i>	0.6456	*	glutamine transport ATP-binding protein
	<i>gntU_1</i>	0.2034	*	
	<i>rbsB</i>	0.3614	*	periplasmic ribose-binding protein precursor
	<i>melB</i>	0.6110	*	thiomethylgalactoside permease II
	<i>glpT</i>	0.5674	*	glycerol-3-phosphatase transporter
	<i>rbsD</i>	0.1566	*	high affinity ribose transport protein
	<i>gntT</i>	1.0605	*	high-affinity gluconate transporter
	<i>gntU_2</i>	0.9875	*	
Translation, post-translational modification	<i>nupC</i>	2.2742	**	nucleoside permease NupC
	<i>ppiC</i>	0.7876	*	peptidyl-prolyl cis-trans isomerase C
Higher in L				
Amino acid biosynthesis and metabolism	<i>ilvL</i>	-0.8747	**	ilvGMEDA operon leader peptide
Amino acid biosynthesis and metabolism	<i>ilvM</i>	-1.8169	*	acetohydroxy acid synthase II, small subunit
Central intermediary metabolism	<i>speG</i>	-0.9321	*	spermidine N1-acetyltransferase
Carbon compound catabolism	<i>araC</i>	-0.5661	*	arabinose operon regulatory protein
Fatty acid and phospholipid metabolism	<i>fadR</i>	-0.7410	*	fatty acid--fatty acyl responsive DNA-binding protein

Table 11 (continued):

Data taken from Botsford and Harman (1992) and Saier et al (1996) and Salgado et al (2000). $0.5 > P > 0.1$, *. $0.01 > P > 0.001$, **. $0.001 > P > 0.0001$, ***.

significantly (Rozen and Lenski 2000). It is this alteration in growth rate that is most critical for the ability for S cells to invade a population of L cells. An important difference between this expression Experiment 2 and Experiment 1 is that no mutations differentiate the treatments. All expression differences result solely from the environmental shift. Surprisingly, gene expression profiles from S growing in DM25 and from S growing in L-conditioned medium differ from one another (Figure 19) by approximately the same number of genes as S and L (Figure 18). One possible interpretation is that L secretions are tremendously complex. Alternatively, the extensive differences may result from influences that route through a smaller number of effector regulatory genes, as suggested above for *cyaA*.

For nearly all functional categories, the relative expression of S grown in DM25 exceeds that of S grown in L conditioned media (Table 8). This effect is seen for genes involved in aerobic and anaerobic respiration, fermentation and glycolysis. Because of many apparently coordinated expression increases, the cause of many of these differences may result from change in expression of global regulatory genes. Two such critical factors show elevated expression in S in DM25: 1) *cyaA* and 2) *rpoS*, which is involved in the regulation of stationary phase specific genes (Hengge-Aronis 1996; Huisman et al. 1996; Loewen et al. 1998). In Tables 12 and 13 we show relative expression in genes that are putatively regulated by *cyaA* and *rpoS*, respectively. Forty-six genes that are regulated by *cyaA* show significant expression differences between S grown in DM25 and S grown in DM25 conditioned by L. Of these 46, 35 show higher relative expression in S grown

Table 12: Genes that are putatively regulated by cAMP (cyaA) that differ between S grown in DM25 and S grown in DM25 conditioned by L cells.

Functional Class	Gene	Log ₂ (S in L filtrate/S)	P value	Gene Function
Higher in S grown in DM25				
Amino acid biosynthesis and metabolism	<i>ilvH</i>	-0.6524	*	acetolactate synthase isozyme III small subunit
	<i>dadX</i>	-0.3757	**	alanine racemase, catabolic precursor
Biosynthesis of cofactors, prosthetic groups and carriers Carbon compound catabolism	<i>gor</i>	-0.4573	*	glutathione oxidoreductase
	<i>araJ</i>	-0.4601	*	AraJ protein precursor
	<i>araA</i>	-0.3607	*	L-arabinose isomerase
	<i>lacA</i>	-0.4424	*	
	<i>bglB</i>	-0.5742	*	phospho-beta-glucosidase
	<i>lacZ</i>	-0.6183	*	beta-galactosidase
	<i>treC</i>	-0.2119	**	trehalose-6-phosphate hydrolase
	<i>bglA</i>	-0.5342	*	
	<i>galK</i>	-0.5721	*	galactokinase
	<i>aldB</i>	-0.1633	*	aldehyde dehydrogenase b
	<i>fucI</i>	-0.8019	**	L-fucose isomerase
	<i>galT</i>	-0.3482	*	galactose-1-phosphate uridylyltransferase
Cell processes	<i>ftsW</i>	-0.8623	**	cell division protein
	<i>caiC</i>	-1.0138	*	probable crotonobetaine/carnitine-CoA ligase
Cell structure	<i>glgP</i>	-0.5429	*	alpha-glucan phosphorylase
	<i>glgA</i>	-0.6882	*	glycogen synthase
	<i>glgC</i>	-0.3009	*	glucose-1-phosphate adenylyltransferase
Central intermediary metabolism	<i>speC</i>	-0.8816	*	
	<i>glpQ</i>	-0.6984	*	glycerophosphoryl diester phosphodiesterase

Table 12 (continued):

Energy metabolism	<i>glpD</i>	-0.6672	**	aerobic glycerol-3-phosphate dehydrogenase
	<i>pflB</i>	-0.5369	*	formate acetyltransferase I
	<i>fixA</i>	-0.9662	*	FixA protein
	<i>aldA</i>	-0.7912	*	lactaldehyde dehydrogenase A
Regulatory function	<i>cyaA</i>	-0.5448	***	adenylate cyclase
Transcription, RNA processing and degradation	<i>rpoS</i>	-0.5372	**	RNA polymerase sigma subunit RpoS (sigma-38)
Translation, post-translational modification	<i>ppiC</i>	-0.9548	*	peptidyl-prolyl cis-trans isomerase C
Transport and binding proteins	<i>cirA</i>	-0.9882	*	colicin I receptor precursor
	<i>gltK</i>	-1.3709	*	glutamate/aspartate transport system
	<i>gltP</i>	-0.7222	*	proton glutamate symport protein
	<i>araG</i>	-0.6633	*	L-arabinose transport ATP-binding protein
	<i>lamB</i>	-0.8200	*	phage lambda receptor protein
	<i>glpT</i>	-0.6148	*	glycerol-3-phosphatase transporter
	<i>mglA</i>	-0.9162	*	galactoside transport ATP-binding protein mglA
Higher in S grown in L conditioned media				
Amino acid biosynthesis and metabolism	<i>ilvM</i>	1.4284	*	acetohydroxy acid synthase II, small subunit
	<i>tnaL</i>	1.9599	*	tna operon leader peptide
Carbon compound catabolism	<i>focA</i>	0.4535	*	probable formate transporter
	<i>araB</i>	0.8155	*	L-ribulokinase
	<i>araC</i>	1.1558	*	arabinose operon regulatory protein
	<i>rhaD</i>	0.2040	*	rhamnulose-1-phosphate aldolase

Table 12 (continued):

Cell processes	<i>treR</i>	0.4553	*	trehalose operon repressor
Central intermediary metabolism	<i>gntK</i>	1.4287	*	thermoresistant glucokinase
Energy metabolism	<i>frdC</i>	0.1727	*	fumarate reductase, membrane anchor polypeptide
Nucleotide biosynthesis and metabolism	<i>udp</i>	0.7377	*	uridine phosphorylase
Transport and binding proteins	<i>lacY</i>	0.7268	*	lactose permease

Data taken from Botsford and Harman (1992), Saier et al (1996), and Salgado et al (2000).
0.5 > P > 0.1, *. 0.01 > P > 0.001, **. 0.001 > P > 0.0001, ***.

Table 13: Genes that are putatively regulated by rpoS that differ between S grown in DM25 and S grown in DM25 conditioned by L cells.

Functional Class	Gene	Log2 (S in L filtrate/S)	P value	Gene Function
Higher in S grown in DM25				
Carbon compound catabolism	<i>lacZ</i>	-0.6183	*	beta-galactosidase
	<i>treC</i>	-0.2119	**	trehalose-6-phosphate hydrolase
	<i>galK</i>	-0.5721	*	galactokinase
	<i>aldB</i>	-0.1633	*	aldehyde dehydrogenase b
	<i>galT</i>	-0.3482	*	galactose-1-phosphate uridylyltransferase
Cell processes	<i>katE</i>	-0.7032	**	catalase HP11
	<i>otsA</i>	-0.4293	**	alpha trehalose phosphate synthase
Cell structure	<i>glgA</i>	-0.6882	*	glycogen synthase
Energy metabolism	<i>glpD</i>	-0.6672	**	aerobic glycerol-3-phosphate dehydrogenase
	<i>aldA</i>	-0.7912	*	lactaldehyde dehydrogenase A
	<i>hyaB</i>	-0.3722	*	hydrogenase-1 large chain
	<i>hyaE</i>	-0.8181	***	hydrogenase-1 operon
	<i>hmpA</i>	-1.2063	*	flavoheмоprotein
Transport and binding proteins	<i>mglA</i>	-0.9162	*	galactoside transport ATP-binding protein mglA
Higher in S grown in L conditioned media				
Cell structure	<i>csgD</i>	0.7927	**	putative regulatory protein
	<i>csgE</i>	1.0275	*	assembly /transport component in curli production
DNA replication, recombination, modification and repair	<i>himA</i>	1.5544	*	integration host factor alpha-subunit
	<i>mutH</i>	1.5979	*	

Data taken from Hengge-Aronis (1996), Huisman et al (1996), and Loewen et al (1998).
0.5 > P > 0.1, *. 0.01 > P > 0.001, **. 0.001 > P > 0.0001, ***.

in DM25. Eighteen genes that are putatively regulated by *rpoS* show significantly different expression between S grown in DM25 and S grown in DM25 conditioned by L cells, 14 of which are higher in S from DM25. Again, these lists may not include all genes that are directly regulated by these global regulators nor do they include the many genes that are indirectly influenced by their expression.

Of the 294 genes that show higher relative expression in S grown in L-conditioned medium, 179 are in genes that have no identified function. This catch-all category is the only one to show an excess of elevated expression in the conditioned medium. Again, this difference may be a artifact of the normalization procedure.

S versus L in DM25 during Stationary Phase

A higher death rate of L than of S during stationary phase contributes to the coexistence of S and L by partially offsetting L's growth rate advantage (Rozen and Lenski 2000). Despite this demographic difference between S and L during stationary phase, global expression patterns differ at only 181 genes. Because about 5% = 215 genes of 4,287 genes should differ significantly by chance alone, we do not have sufficient confidence in this level of differentiation to warrant detailed consideration of any single category or gene.

Discussion

We examined global gene expression patterns in two clones, S and L, using microarrays. We sought to determine the extent of expression differences between the clones in evolutionarily relevant scenarios, and more generally, to ascertain the utility of this approach for understanding the mechanistic bases of adaptive changes that have occurred in the evolving populations of *E. coli* that have been generated in the Lenski lab. Three main comparisons were conducted, each corresponding to factors that contribute to the coexistence of S and L. In Experiment 1, we compared the expression profiles of S and L during exponential growth. In Experiment 2, we compared the expression profiles of S cells growing alone and in the presence of L secretions (L-conditioned media). Finally, in Experiment 3, we compared gene expression of S and L during stationary phase. For all but Experiment 3, we identified an enormous number of genes whose expression differed significantly between treatments. Because of the large number of expression differences observed, microarray analysis must be considered only a first step in the search for the mechanistic bases of adaptive differences in these *E. coli* populations.

S and L differ in maximum growth rate by nearly 20%. This difference largely explains how L cells can invade a population of S cells. However, this growth rate difference presents a potential difficulty in this analysis because growth rate in *E. coli* is known to influence the expression of some genes (Bremer and Dennis 1996). In particular, ribosomal content scales with growth rate (Bremer and Dennis 1996), and this effect may

explain the increased expression in L of functions related to translation and amino acid biosynthesis (Table 7) . At present it is unknown whether these observed differences were causally involved in L's faster growth rate, or only a consequence of it. Further examination using chemostats to control growth rate may help resolve this issue.

However, such a study will be itself compromised with respect to evolutionary inferences because it will dramatically differ from the batch culture environment in which S and L evolved. For this reason, we believe that the culture regimes used in this study are warranted because they most faithfully reproduce the relevant environments for these genotypes.

At present, we do not know the number of beneficial or neutral mutations that differentiate the S and L clones. We also do not know the proportion of these mutational differences that cause any alteration in transcription level. Thus, it was difficult to predict how many expression differences would be uncovered using microarrays. Nevertheless, given the relatively recent divergence of S and L, and because of the extensive expression differences observed in Experiment 2 (where no mutational differences exist), it is likely that some, if not most, of the expression differences between our treatments result from changes in regulatory genes having widespread pleiotropic effects. Two candidate genes that could be responsible for such extensive pleiotropic effects in our experimental treatments are *cyaA* and *rpoS*, both of which are global regulatory genes and both of which show higher relative expression in S than in L (Experiment 1).

The gene *cyaA* encodes adenylate cyclase, the enzyme that produces intracellular cAMP. This gene, in part, regulates one of the best studied metabolic regulons in *E. coli* (Saier et al. 1996). Intracellular cAMP concentrations are responsive to concentrations of metabolizable carbohydrates, among other things. The best known cAMP regulating substrate is glucose, although other substrates, are also known to influence cAMP levels (Postma et al. 1996). By repressing the transcription of adenylate cyclase, a high concentration of glucose results in the repression of alternative catabolic functions. When glucose concentration decreases, cAMP concentration rises, thereby de-repressing many cAMP responsive genes and allowing alternative substrates to be exploited and then metabolized through the various pathways of central metabolism (Botsford and Harman 1992; Saier et al. 1996). *rpoS* encodes the stationary phase sigma factor, σ^S , and regulates dozens of genes that are thought to be involved in starvation survival and preparation for growth when suitable substrate becomes available (Loewen et al. 1998). As might be expected from increased expression of these regulatory loci, Tables 11-13 show that many genes that are regulated by *cyaA* and *rpoS* show differential expression across our treatments.

How might changes at these regulatory loci influence the ecological interaction between S and L? One possibility is that global de-repression of genes for transport and catabolism may allow S to simultaneously utilize glucose plus the metabolites present in L secretions. That is, these changes may enable cross-feeding. However, while this

hypothesis is intuitively appealing, it is inconsistent with our expression data. In particular, when we compare the global expression pattern of S growing in DM25 with S growing in medium conditioned by L cells, we found that most cAMP and *rpoS* regulated genes were more highly expressed in the former environment. These data are the reverse of the expectation that growing S in medium conditioned by L secretions would stimulate the transcription of genes that might enable cross-feeding. This counter-intuitive result cannot be readily explained *unless* these regulons have been "short-circuited" in the S genotype. That is, we know from Experiment 1 that S over-expresses many of the genes that one would expect to support a cross-feeding life-style in the conditioned medium. However, S evolved for more than 10,000 generations in a medium that did contain L, so S may have mutated to avoid--or even invert--the counter-effective repression due to cAMP and σ^S .

Another possible explanation derives from the fact that glucose is not the sole regulator of *cyaA* (Postma et al. 1996). It is influenced by other products of the PTS transport, by other carbohydrates, such as lactose, glutamate, and glucose-6-phosphate, and by perhaps other gene products (Postma et al. 1996). If S cells growing in L secretions are provided with substrate for one or more of the *cyaA* derepressed genes, then this might in turn decrease expression of adenylate cyclase and thus many of the *cyaA* regulated downstream loci. Similar effects may influence the global influence of *rpoS* regulated gene expression. These are merely hypotheses, but ones that might be examined further

by studying expression changes, and phenotypic consequences, in genotypes that are experimentally mutated in *cyaA* or *rpoS*.

In contrast to Experiments 1 and 2, we saw no compelling differences in Experiment 3 which compared gene expression between S and L cells during stationary phase (Figure 20, Table 9). L cells die at a higher rate than S cells during stationary phase (Rozen and Lenski 2000), which suggests that the two types have important physiological differences during this time period. There are several explanations for this negative result. One possibility is that the relevant physiological differences do not depend on gene expression at the level of transcription. Another possibility is that the survival difference does not depend on new expression at all, but instead reflect differences in cellular constituents that were produced prior to stationary phase, during cell growth. Finally, the higher cell death rate of L is exacerbated by the presence of S cells (Rozen and Lenski 2000), a factor that was not reflected in the design of Experiment 3. It would be interesting, therefore, to compare gene expression in L cells during stationary phase in media with and without conditioning by S cells, much as Experiment 2 examine the effect of L secretions of S during exponential growth.

A final interesting result is the observation that expression differences were often found for genes whose functions are, as yet, unknown. In all three experiments, we found that the treatment that showed an more increases in this class was lower in all other classes summed together (Tables 7-9). The explanation for this may well lie with the

normalization procedure we used to convert raw expression values for each ORF into a fraction of the total expression over all ORFs. Although this normalization was necessary to allow meaningful comparisons across arrays, it might also have introduced artifacts. Specifically, by forcing total expression for each array to sum to 1, large expression increases in certain genes would cause apparent decreases in other genes. We do not know how pervasive this artifact is. However, the possibility that it might be important suggests, again, the necessity for caution with, and even independent confirmation of, all microarray data.

We have examined the potential utility of microarrays for studying the physiological mechanisms of genetic adaptation in S and L. While this new approach provides tremendous amounts of data on the genes and pathways that became altered through adaptive evolution, identifying individual mutations with this approach is problematic. In addition, because the number of genes that show altered expression is enormous as a result of pleiotropy, and because of possible artifacts due to normalization, all microarray data need to be confirmed using additional methods and assays.

Literature Cited

- Arfin, S. M., A. D. Long, E. T. Ito, L. Toller, M. M. Riehle, E. S. Paegle, and G. W. Hatfield. 2000. Global gene expression profiling in *Escherichia coli* K12: The effects of integration host factor. *Journal of Biological Chemistry* 275:29672-29684.
- Blattner, F. R., G. Plunkett III, 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* 277:1453-1474.
- Botsford, J. L., and J. G. Harman. 1992. Cyclic AMP in prokaryotes. *Microbiological Reviews* 56:100-122.
- Bremer, H., and P. P. Dennis. 1996. Modulation of chemical composition and other parameters of the cell by growth rate. Pp. 1553-1569 in F. C. Neidhardt, R. Curtiss III, J. L. Ingraham, E. C. C. Lin, K. Brooks Low, B. Magasanik, W. S. Reznikoff, M. Riley, M. Schaechter and H. E. Umbarger, eds. *Escherichia coli* and *Salmonella*: Cellular and Molecular Biology. ASM Press, Washington, D. C.
- Cavalieri, D., J. P. Townsend, and D. L. Hartl. 2000. Manifold anomalies in gene expression in a vineyard isolate of *Saccharomyces cerevisiae* revealed by DNA microarray analysis. *Proceedings of the National Academy of Sciences, USA* 97:12369-12374.
- Chu, S., J. DeRisi, M. Eisen, J. MulHolland, D. Botstein, P. O. Brown, and I. Herskowitz. 1988. The transcriptional program of sporulation in budding yeast. *Science* 282:699-705.
- Cooper, V. S., and R. E. Lenski. 2000. The population genetics of ecological specialization in evolving *E. coli* populations. *Nature* 407:736-739.
- deRisi, J. L., V. R. Iyer, and P. O. Brown. 1997. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* 278:680-686.
- Duggan, D. J., M. Bittner, Y. Chen, P. Meltzer, and J. M. Trent. 1999. Expression profiling using cDNA microarrays. *Nature Genetics* (suppl.) 21:10-14.
- Ferea, T. L., D. Botstein, P. O. Brown, and R. F. Rosenzweig. 1999. Systematic changes in gene yeast expression patterns following adaptive evolution in yeast. *Proceedings of the National Academy of Sciences, USA* 96:9721-9726.

- Futuyma, D. J. 1998. *Evolutionary Biology*. Sinauer Associates, Sunderland, Mass.
- Hengge-Aronis, R. 1996. Regulation of gene expression during entry into stationary phase. Pp. 1497-1512 in F. C. Neidhardt, R. Curtiss III, J. L. Ingraham, E. C. C. Lin, K. Brooks Low, B. Magasanik, W. S. Reznikoff, M. Riley, M. Schaechter and H. E. Umbarger, eds. *Escherichia coli* and *Salmonella*: Cellular and Molecular Biology. ASM Press, Washington, D. C.
- Huisman, G. W., D. A. Siegele, M. M. Zambrano, and R. Kolter. 1996. Morphological and physiological changes during stationary phase. Pp. 1672-1682 in F. C. Neidhardt, R. Curtiss III, J. L. Ingraham, E. C. C. Lin, K. Brooks Low, B. Magasanik, W. S. Reznikoff, M. Riley, M. Schaechter and H. E. Umbarger, eds. *Escherichia coli* and *Salmonella*: Cellular and Molecular Biology. ASM Press, Washington, D. C.
- Lenski, R. E., M. R. Rose, S. C. Simpson, and S. C. Tadler. 1991. Long-Term Experimental Evolution in *Escherichia-Coli* .1. Adaptation and Divergence During 2,000 Generations. *American Naturalist* 138:1315-1341.
- Lenski, R. E., and M. Travisano. 1994. Dynamics of Adaptation and Diversification - a 10,000- Generation Experiment With Bacterial-Populations. *Proceedings of the National Academy of Sciences, USA* 91:6808-6814.
- Loewen, P. C., B. Hu, J. Strutinsky, and R. Sparling. 1998. Regulation in the *rpoS* regulon of *Escherichia coli*. *Canadian Journal of Microbiology* 44:707-717.
- Postma, P. W., L. J. W., and G. R. Jacobson. 1996. Phosphoenolpyruvate:Carbohydrate phosphotransferase systems. Pp. 1149-1174 in F. C. Neidhardt, R. Curtiss III, J. L. Ingraham, E. C. C. Lin, K. Brooks Low, B. Magasanik, W. S. Reznikoff, M. Riley, M. Schaechter and H. E. Umbarger, eds. *Escherichia coli* and *Salmonella*: Cellular and Molecular Biology. ASM Press, Washington, D. C.
- Richmond, C. S., J. D. Glasner, R. Mau, H. Jin, and F. R. Blattner. 1999. Genome-wide expression profiling in *Escherichia coli* K-12. *Nucleic Acids Research* 27:3821-3835.
- Riley, M. 1988. Genes and proteins in *Escherichia coli* K-12. *Nucleic Acids Research* 26:54.
- Rose, M. R., and G. V. Lauder. 1996. *Adaptation*. Pp. 511. Academic Press, San Diego, California.
- Rozen, D. E., and R. E. Lenski. 2000. Long-term experimental evolution in *Escherichia*

coli. VIII. Dynamics of a balanced polymorphism. *American Naturalist* 155:24-35.

Saier Jr., M. H., T. M. Ramseier, and J. Reizer. 1996. Regulation of carbon utilization. Pp. 1325-1343 in F. C. Neidhardt, R. Curtiss III, J. L. Ingraham, E. C. C. Lin, K. Brooks Low, B. Magasanik, W. S. Reznikoff, M. Riley, M. Schaechter and H. E. Umbarger, eds. *Escherichia coli* and *Salmonella*: Cellular and Molecular Biology. ASM Press, Washington, D. C.

Salgado, H., A. Santos-Zavaleta, S. Gama-Castro, D. Millan-Zarate, F. R. Blattner, and J. Collado-Vides. 2000. RegulonDB (version 3.0): transcriptional regulation and operon organization in *Escherichia coli* K-12. *Nucleic Acids Research* 28:65-67.

Tao, H., C. Bausch, C. Richmond, F. R. Blattner, and C. Conway. 1999. Functional genomics: Expression analysis of *Escherichia coli* growing on minimal and rich media. *Journal of Bacteriology* 181:6425-6440.

MICHIGAN STATE UNIVERSITY LIBRARIES



3 1293 02328 7711