

University of Montana

ScholarWorks at University of Montana

Graduate Student Theses, Dissertations, &
Professional Papers

Graduate School

2009

Adaptive Variation in Microbes: Insights from Wild and Experimental Populations of *Escherichia coli*

Margaret Ann Kinnersley

Follow this and additional works at: <https://scholarworks.umt.edu/etd>

Let us know how access to this document benefits you.

Recommended Citation

Kinnersley, Margaret Ann, "Adaptive Variation in Microbes: Insights from Wild and Experimental Populations of *Escherichia coli*" (2009). *Graduate Student Theses, Dissertations, & Professional Papers*. 10837.

<https://scholarworks.umt.edu/etd/10837>

This Dissertation is brought to you for free and open access by the Graduate School at ScholarWorks at University of Montana. It has been accepted for inclusion in Graduate Student Theses, Dissertations, & Professional Papers by an authorized administrator of ScholarWorks at University of Montana. For more information, please contact scholarworks@mso.umt.edu.

ADAPTIVE VARIATION IN MICROBES: INSIGHTS FROM WILD AND

EXPERIMENTAL POPULATIONS OF *Escherichia coli*

By

MARGARET ANN KINNERSLEY

Bachelors of Science, The University of Texas at Austin, 1998

Dissertation

presented in partial fulfillment of the requirements
for the degree of

Doctorate of Philosophy in
in Microbiology and Biochemistry, Molecular Biology
The University of Montana
Missoula, MT

May 2009

Approved by:

Perry Brown, Associate Provost for Graduate Education
Graduate School


William Holben, PhD Co-Chair
Biological Sciences

Frank Rosenzweig, PhD Co-Chair
Biological Sciences

Steve Lodmell, PhD
Biological Sciences

Scott Miller, PhD
Biological Sciences

Mark Pershouse, PhD
Biomedical and Pharmaceutical Sciences



Abstract

Intraspecific differences in genome composition and gene regulation are widespread in both natural and artificial prokaryotic systems. Understanding the molecular basis, population dynamics and fitness consequences of these differences can provide useful insight into many aspects of microbial ecology and evolution. The work presented here is a study of molecular variation in both natural and experimental populations of *E. coli*, conducted with the ultimate goal of gaining a better understanding of niche adaptation and the nature of molecular variation in microbes. In Chapter 2, the mechanistic basis of adaptation and diversification in a polymorphic experimental population of *E. coli* that spontaneously arose after ~700 generations of glucose limitation in chemostats was explored. The results highlight the importance of mutations in both global and gene-specific regulators in maintaining the stable co-existence of clones, and the profound effect that founder genotype can have on evolutionary outcome. Chapter 3 examines the extant variation in genome composition at the gene level between natural isolates *E. coli* from different mammalian host species to address the basic question of how genetic measures of diversity are correlated with habitat variation. Our work shows that genome content is a more reliable indicator of host affiliation than a number of fingerprinting methods commonly used to distinguish host source, and that human-derived strains show patterns of gene presence/absence consistent with elevated genome recombination and convergence compared to isolates from other animals. The work in Chapter 4 extends these observations to include differences in gene transcription and suggests that mutations affecting the regulation of certain genes have occurred in parallel between unrelated isolates from the same host source. Finally, in Chapter 5, I describe a classroom inquiry developed during my year with the ECOS program at UM designed to introduce students to the nitrogen cycle from both a microbial and plant perspective. The broader significance and future directions of Chapters 2-5 are detailed in Chapter 6.

Abstract	1
Introduction	1
Figure and Tables	1
Chapter 2: Microarray comparative genomic hybridization of <i>Escherichia coli</i> from humans and animal hosts	1
Abstract	1
Introduction	1
Materials and Methods	1
Results	1
Discussion	1
Conclusions/Summary	1
Literature Cited	1
Figure and Tables	1

Chapter 4: Transcriptional profiles of *Escherichia coli* from different niches
 and the influence of gene expression evolution on niche adaptation 158

Table of Contents

Chapter 1- Introduction	1
Overview	1
Natural History	1
Genetic diversity and niche adaptation in laboratory populations of <i>E. coli</i>	3
Genetic diversity and niche adaptation in natural populations of <i>E. coli</i>	4
Study Rationale	6
Research questions by chapter	7
Literature Cited	9
Chapter 2- Genomic analysis of an evolved polymorphism in <i>E. coli</i>	12
Abstract	12
Introduction	14
Materials and Methods	18
Results	24
Discussion	44
Conclusion/Summary	55
Acknowledgements	56
Literature Cited	56
Figures and Tables	63
Chapter 3- Microarray comparative genomic hybridization of <i>Escherichia coli</i> from human and animal hosts	108
Abstract	108
Introduction	109
Materials and Methods	112
Results	117
Discussion	126
Conclusion/Summary	134
Literature Cited	134
Figures and Tables	139

Chapter 4- Transcriptional profiling of <i>Escherichia coli</i> from different mammalian hosts shows convergence in gene expression consistent with niche adaptation	153
Abstract	153
Introduction	154
Materials and Methods	156
Results	159
Discussion	164
Conclusion.....	168
Literature Cited	170
Figures and Tables	172
Chapter 5-An exploration in nitrogen cycling and plant growth	185
Abstract	185
Introduction	185
Learning Goals for Students.....	186
Before the Experiment	187
Student Preparation	188
Procedure Overview	191
Phase 1- Making Compost	189
Phase 2- Tracking the Nitrogen Cycle	191
Phase 3- Growing Plants	192
Assessment	193
Conclusion.....	194
Literature Cited	194
Figures and Tables	195
Chapter 6- Synthesis	206
Literature Cited	213

List of Tables

Chapter 2

Table 1. Bacterial Strains.....	74
Table 2. Expression levels of selected genes from 1-class and 4-class SAM analyses.....	75
Table 3. Sequenced Genes.....	77
Supplementary Table 1. Failed and low concentration PCR reactions.....	85
Supplementary Table 2. Top 91 significant genes by 1-class SAM for evolved isolates grown individually.....	91
Supplementary Table 3. Top 93 significant genes by 4-class SAM for evolved isolates grown individually.....	97
Supplementary Table 4. Primers used for qRT-PCR.....	105
Supplementary Table 5. Sequencing primers.....	106

Chapter 3

Table 1. Bacterial Strains.....	146
Table 2. Genes absent in all isolates.....	147
Table 3. Amplified genes.....	148
Supplementary Table 1. Genes absent in all isolates.....	150
Supplementary Table 2. Amplified genes.....	152

Chapter 4

Supplementary Table 1. Biolog results for all twelve wild isolates and <i>E. coli</i> K-12.....	177
Supplementary Table 2. Genes with significantly different expression between host groups by 4-class SAM.....	180
Supplementary Table 3. List of global regulators for 4-class SAM significant genes..	182

Chapter 5

Table 1. Materials to be purchased.....	198
Figure 4. Genes that were considered diagnostic for at least one out of eight of the host, dog or wild <i>B. meli</i>	184
Supplementary Figure 4. PCR with primers specific for <i>B. meli</i> and <i>B. meli</i> on the dog and wild <i>B. meli</i> hybridization results.....	185

List of Figures

Chapter 2

- Figure 1.** Array Comparative Genomic Hybridization (a-CGH) of each adaptive clone versus their common ancestor, JA12263
- Figure 2.** 1-class SAM analysis for terminal isolates grown in chemostat monoculture64
- Figure 3.** Top 93 significant genes by 4-class SAM for evolved isolates grown in chemostat monoculture66
- Figure 4.** Expression profile SAM analysis of strains in co-culture reflects many, but not all regulatory changes observed when strains are grown in monoculture68
- Figure 5.** Some genes differ markedly between the monoculture and consortium expression profiles.70
- Figure 6.** Cladogram depicting the likely evolutionary relationship between CV101, CV103, CV116 and CV115.72
- Supplementary Figure 1.** rep-PCR and PFGE fingerprints of chemostat isolates. ...78
- Supplementary Figure 2.** Global transcriptional response of evolved clones.79
- Supplementary Figure 3.** Overview of Central Metabolic Transcriptional Response.81
- Supplementary Figure 4.** qRT-PCR results for lamB, flgB, and acs.83

Chapter 3

- Figure 1.** Clustering of rep-PCR and PFGE fingerprints139
- Figure 2.** Whole genome “fingerprints” show better clustering of the human, cow and deer isolates than fingerprints generated by rep-PCR and PFGE.....141
- Figure 3.** Hierarchical clustering of genes that are diagnostic for at least two out of the three human strains.142
- Figure 4.** Genes that were considered diagnostic for at least two out of three of the bear, deer or cow *E. coli*.144
- Supplementary Figure 1.** PCR with primers specific for the *fec* and *ins* loci confirm the comparative genome hybridization results149

Chapter 4

- Figure 1.** Carbon source utilization profiles of all twelve wild isolates for the 38 compounds that showed variable growth.....172
- Figure 2.** Heatmap showing the 86 genes that had significantly different expression patterns among the four host groups.....173
- Figure 3.** Pie chart depicting the distribution of significant genes from the 4-class SAM analysis by functional group175
- Figure 4.** Comparison of dendograms generated using the expression pattern of the 86 genes differentially expressed between host groups176

Chapter 5

- Figure 1.** A simplified version of the nitrogen cycle.....197
- Supplementary Figure 1.** Example student lab notebook199

CHAPTER 1

Introduction

Overview

Escherichia coli has been used as a model organism for genetic, biochemical, physiological and evolutionary studies for over six decades. More is understood about its genetics and biochemistry than any other prokaryote, and molecular biological techniques that have been developed and optimized using *E. coli* have built a foundation onto which the study of other microorganisms can be based. However, despite its near ubiquity in the microbiology laboratory, there is still more to be learned about the forces that shape the natural history, physiological variation and population genetics of *E. coli*.

Natural history

E. coli is a facultatively anaerobic commensal inhabitant of the intestinal tract of all warm-blooded mammals, some birds and some reptiles. In most mammals, one or a few dominant strains of *E. coli* that persist in an individual for decades coexist with several transient strains that may exit the intestine in as little as 26 hours (Caugant, Levin et al. 1981). Some researchers believe that new strains can immigrate into the intestinal tract from ingested fecal material, food or water, while others believe that it is nearly impossible for exogenous strains of *E. coli* to invade past established gut flora (Caugant, Levin et al. 1981; Freter, Brickner et al. 1983; Winfield and Groisman 2003). Intestinal doubling times for *E. coli* are thought to be between 5 and 12 hours in the colon and perhaps faster in the small intestine where the contents of the ileum empty into the large

intestine approximately every hour. Under these conditions *E. coli* is forced to divide rapidly or risk being flushed out (Levin 1981; Schaechter 2001; Winfield and Groisman 2003).

Where *E. coli* resides in the mammalian intestinal tract appears to depend largely on the taxonomic affiliation and gastrointestinal physiology of the host. Mammals can be roughly divided into two groups based on digestive morphology: foregut fermenters and hindgut fermenters. Foregut fermenters (also known as ruminants) such as cattle and deer are herbivorous and have highly compartmentalized stomachs in which the breakdown of food particles occurs prior to nutrient absorption in the small intestine. *E. coli* is also an occupant of the rumen in these animals and may thus have access to ingested food before host enzymes do. By contrast, in hindgut fermenters such as humans and bears, microbial contact with ingesta is restricted by the physiology of the intestines; here host enzymatic digestion and nutrient absorption occurs in the stomach and upper small intestine while most microbes, including *E. coli*, colonize the lower small intestine, cecum and colon where they encounter only digesta and undigested food particles (Stevens 1988). These basic host physiological differences can have a large effect on the types and quantities of metabolic substrates available to *E. coli*, intestinal retention times, microbial population densities and microbial community composition. Thus, it is clear that the biochemical and physiological environments that *E. coli* experiences are different depending on the taxonomic affiliation and physiology of their host, and it is reasonable to speculate that differences in *E. coli* adaptive physiology would correlate well with host intestinal physiology.

Genetic diversity and niche adaptation in laboratory populations of *E. coli*

Short generation times and ease of cultivation have made *E. coli* the organism of choice for the study of natural selection in the laboratory. Such “experimental evolution” studies have vastly expanded our knowledge of myriad aspects of evolutionary biology including the molecular bases of adaptation, the importance of evolutionary trade-offs and the origin and maintenance of diversity (Elena and Lenski 2003; Zeyl 2006). Early microbial evolution experiments led to two key observations regarding the maintenance of variation in large, asexual populations. First, variation that arises through mutation propagates via “periodic selection” events in which fitter genotypes displace less fit competitors (Muller 1932; Novick and Szilard 1950; Atwood, Schneider et al. 1951). Second, competition for the same limiting resource reduces variation, an observation that led to the development of the competitive exclusion principle which asserts that competitors cannot simultaneously occupy the same ecological niche (Gause 1934; Hardin 1960).

One noteworthy example of laboratory evolution in action that appears to violate these principles and is of particular relevance to the work presented here involves the metabolic diversification of *E. coli* propagated under glucose limitation in chemostat culture (Helling, Vargas et al. 1987; Rosenzweig, Sharp et al. 1994). In this simple, unstructured environment, large populations of *E. coli* founded by a single clone evolve into multiple clones that coexist for scores, if not hundreds of generations. This repeatable phenomenon appears to be a special case of niche adaptation in which each clone occupies a metabolic niche created by the incomplete catabolism of the limiting nutrient, glucose.

The precise genetic basis of this stable polymorphism remains obscure. However, changes in the frequency of a neutral marker over the course of the experiment indicated that the total number of accrued mutations was small (Helling, Vargas et al. 1987). Thus, genetic diversity in this system is likely to be quite low despite the high degree of adaptive diversification. The rapid rate at which *E. coli* can adapt to novel environments in the absence of recombination even under simple conditions raises a number of important questions about how *E. coli* responds to selective pressure in its more complex natural environment.

Genetic diversity and niche adaptation in natural populations of *E. coli*

Most of what is currently known about *E. coli* population genetics in its natural habitat has been gleaned from analyses of how molecular markers vary through space and time. One of the simplest ways to detect whether or not environmental differences influence *E. coli* population structure is to determine what proportion of naturally occurring variation can be attributed to host taxonomic affiliation. Estimates of *E. coli* population genetics and phenotypic diversity have been calculated using a variety of methods and for a number of different culture collections. Unfortunately, these estimates do not always agree. Two decades of work on the *E. coli* reference (ECOR) collection by Milkman and others have indicated that *E. coli* populations are primarily clonal in nature and have a relatively low level of genetic diversity ($H=0.343$) (Ochman 1983). This observation led to the development of the clonality hypothesis which postulates that natural bacterial populations are the descendants of a single ancestral clone and genetic recombination plays little (if any) role in the evolution of extant lineages (Selander 1987).

Critics have argued that because the ECOR collection consists mainly of isolates from humans and zoo animals living in close proximity to one another, that low diversity in this group is the result of sampling bias. Further, sequencing of several bacterial genomes including *E. coli* K12 has demonstrated that recombination occurs much more frequently in prokaryotes than previously believed. Opponents of the clonality hypothesis such as Maynard Smith (1991) have suggested that bacterial species are more likely organized into an ecotypic structure in which each ecotype is adapted to a different environmental condition (i.e. niche). Under the ecotype hypothesis, variation can be purged from a single ecotype by periodic selection events but maintained at the species level. Thus, the observed genetic variation for *E. coli* over a large geographical area could be high, but adaptation to the host gut environment may still be important at the local scale.

Efforts to characterize diversity in *E. coli* isolated from wild mammals have highlighted the partitioning of this variation by host taxonomic group suggesting that at least part the *E. coli* niche is defined at the level of host species. Souza et al. (1999) typed over 200 strains from 81 different animal species from South America and Australia using multilocus enzyme electrophoresis (MLEE) and found a diversity index of $H = 0.682$ -the highest reported value to date. Roughly 7.5% of this diversity was correlated with host order and 2.5% with host diet. Gordon and Lee (1999) reported a relatively low MLEE mean genetic diversity of 0.27 for *E. coli* isolated from 16 different mammalian families in Australia, but found that 5.5% was attributable to host order. Analysis of non-nucleic acid based measures of diversity such as carbon utilization profiles, antibiotic resistance profiles and plasmid content have suggested that geography

and adaptation to the gut environment might be significant selective forces behind observed phenotypic variation in the Souza dataset (Souza V. 1999). Differences in thermal tolerance profiles of 21 strains of *E. coli* from 11 mammalian genera analyzed by Okada and Gordon (2001) also showed a significant association to host taxonomic group.

Study Rationale

No study to date has attempted to address what aspects of molecular variation in natural *E. coli* populations are adaptive for the unique physiological and biochemical environments created by the digestive systems of animals, or which molecular/biochemical markers are best suited for detecting such variation. The central dogma of molecular biology states that the flow of genetic information in a cell proceeds in a linear fashion from DNA through RNA into protein. Most of the work that has been done on variation in natural populations of *E. coli* has focused on the genome and very little has attempted to determine the impact of adaptation on the transcriptome (i.e the RNA complement of the cell). Although the transcriptome is difficult to study *in vivo*, a number of microbial laboratory evolution experiments have shown that the most significant mutations fixed in prokaryotic genomes during adaptation to different environmental conditions have global or single-gene regulatory effects. If physiological differences between strains of *E. coli* from the intestinal environment of different animal hosts are the result of a modest number of regulatory mutations, such differences are unlikely to be detected by coarse genomic methods if they are the result of a point mutation, small insertion, deletion or inversion. Cooper et al. (Cooper 2003; Elena and Lenski 2003) have demonstrated that using microarrays to identify parallel changes in

gene expression among replicate populations of *E. coli* evolved under laboratory selective conditions can lead investigators to candidate genes in which regulatory mutations occur. This should also be true for natural *E. coli* populations.

In contrast to the many technical challenges that must be overcome to adequately address the causes and consequences of adaptive evolution in complex natural communities of *E. coli*, the study of laboratory populations is relatively straightforward thanks to advances in genomic technologies such as comparative genome hybridization (Ochman and Santos 2005), expression profiling (Cooper, Remold et al. 2008), and very high-throughput sequencing (Herring, Raghunathan et al. 2006). With a combination of these techniques, experimentally evolved microbial “populations” can be dissected into their constituent parts and the genetic basis of phenotypic adaptive variation can be accurately determined. These types of analyses have the potential to provide valuable insight not only into the rules that govern evolution in a controlled setting but have the potential to provide valuable insight into the same processes in natural populations.

Research questions by chapter

In Chapter 2, I explore the molecular basis for the evolution of a balanced polymorphism in a laboratory population of *E. coli*. This study is a direct continuation of a series of landmark experiments conducted by Julian Adams and colleagues (Adams, Kinney et al. 1979; Helling, Vargas et al. 1987; Rosenzweig, Sharp et al. 1994; Treves, Manning et al. 1998). Previous work has shown that glucose-limited chemostat cultures initiated with a single clone of *E. coli* K12 repeatedly evolved into a consortium of at least three genetically distinct ecotypes maintained by metabolic cross-feeding.

However, the genetic basis of this cross-feeding interaction remains largely uncharacterized. Using a combination of targeted gene sequencing and microarray transcriptional profiling, I specifically test the hypothesis that a limited number of mutations in global regulatory genes are responsible for the enhanced acquisition and assimilation of the primary limiting resource, glucose, but that specialization on secondary resources by the subdominant clones is the result of mutations at key structural loci.

In Chapter 3, I present the results of a study designed to pinpoint differences in genome composition between strains of wild *E. coli* collected from the feces of four different mammalian hosts. Here, I am primarily concerned with identifying which patterns of gene presence/absence are most useful for determining the animal origin of individual *E. coli* isolates. I specifically address the questions “to what extent do traditional methods of measuring genetic diversity in *E. coli* accurately reflect genomic content?” and “to what degree is genetic variation influenced by host species affiliation?”. The answers to these questions have a direct impact on the ability wastewater managers to rapidly and accurately monitor water quality, and the applicability of my results to the development of molecular markers for tracking the source of fecal water contamination is discussed.

In Chapter 4, I build upon the work presented in Chapter 3 by addressing the extent to which adaptation to the selective environment of the mammalian intestine might result in convergent patterns of gene expression in *E. coli*. My interest in this question stems directly from the results of Chapter 2 and the increasingly large number of reports that implicate mutations in global regulators as the driving force behind adaptation of *E.*

coli novel laboratory conditions (Kurlandzka, Rosenzweig et al. 1991; Turner, Souza et al. 1996; Notley-McRobb, King et al. 2002; Pelosi, Kuhn et al. 2006; Cooper, Remold et al. 2008). I explicitly test the hypothesis that differences in gene expression between *E. coli* populations will reflect differences in digestive system morphology and physiology of the host from which they were derived. This study is one of the first of its kind to use microarray transcriptional profiling of natural isolates grown in a "common garden" to explore larger issues of adaptive evolution in complex environments.

Finally, Chapter 5 pertains to the year I spent as an ecologist in residence with the "Ecologists, Educators and Schools" (ECOS) program at the University of Montana. My participation in this program was instrumental to my development as an educator and in this chapter I present an original curriculum piece designed give students in grades 5-8 fun, hands-on microbiology experience while still meeting the appropriate National Science Education Content Standards for Science as an Inquiry and Life Science. The five-week lesson presented combines traditional "lecture" style pedagogy with inquiry based investigation into the microbial process of composting, biochemistry of the nitrogen cycle and the effects of incomplete cycling on plant growth.

Literature Cited

- Adams, J., T. Kinney, et al. (1979). "Frequency-Dependent Selection for Plasmid-Containing Cells of *Escherichia coli*." *Genetics* **91**(4): 627-637.
- Atwood, K. C., L. K. Schneider, et al. (1951). "Periodic Selection in *Escherichia coli*." *Proceedings of the National Academy of Sciences of the United States of America* **37**(3): 146-155.
- Caugant, D. A., B. R. Levin, et al. (1981). "Genetic diversity and temporal variation in the *E. coli* population of a human host." *Genetics* **98**(3): 467-90.
- Cooper, T., Rozen D., and R.E. Lenski (2003). "Parallel Changes in Gene Expression after 20,000 Generations of Evolution in *Escherichia coli*." *PNAS* **100**(3): 1072-1077.

- Cooper, T. F., S. K. Remold, et al. (2008). "Expression profiles reveal parallel evolution of epistatic interactions involving the CRP regulon in *Escherichia coli*." PLoS Genet **4**(2): e35.
- Elena, S. F. and R. E. Lenski (2003). "Evolution experiments with microorganisms: the dynamics and genetic bases of adaptation." Nat Rev Genet **4**(6): 457-69.
- Freter, R., H. Brickner, et al. (1983). "Survival and implantation of *Escherichia coli* in the intestinal tract." Infect Immun **39**(2): 686-703.
- Gause, G. F. (1934). The Struggle for Existence. New York, Dover.
- Hardin, G. (1960). "The competitive exclusion principle." Science **131**: 1292-7.
- Helling, R. B., C. N. Vargas, et al. (1987). "Evolution of *Escherichia coli* during growth in a constant environment." Genetics **116**(3): 349-58.
- Herring, C. D., A. Raghunathan, et al. (2006). "Comparative genome sequencing of *Escherichia coli* allows observation of bacterial evolution on a laboratory timescale." Nat Genet **38**(12): 1406-12.
- Kurlandzka, A., R. F. Rosenzweig, et al. (1991). "Identification of adaptive changes in an evolving population of *Escherichia coli*: the role of changes with regulatory and highly pleiotropic effects." Mol Biol Evol **8**(3): 261-81.
- Levin, B. R. (1981). "Periodic selection, infectious gene exchange and the genetic structure of *E. coli* populations." Genetics **99**(1): 1-23.
- Muller, H. J. (1932). "Some Genetic Aspects of Sex." The American Naturalist **66**(703): 118-138.
- Notley-McRobb, L., T. King, et al. (2002). "rpoS mutations and loss of general stress resistance in *Escherichia coli* populations as a consequence of conflict between competing stress responses." J Bacteriol **184**(3): 806-11.
- Novick, A. and L. Szilard (1950). "Experiments with the Chemostat on Spontaneous Mutations of Bacteria." Proceedings of the National Academy of Sciences of the United States of America **36**(12): 708-719.
- Ochman, H. and S. R. Santos (2005). "Exploring microbial microevolution with microarrays." Infect Genet Evol **5**(2): 103-8.
- Ochman, H., Wilson, R.A., Whittam, T.S., and R.K. Selander (1983). "Enzyme Polymorphism and Genetic Population Structure in *Escherichia coli* and *Shigella*." Journal of General Microbiology **129**: 2715-2726.
- Pelosi, L., L. Kuhn, et al. (2006). "Parallel changes in global protein profiles during long-term experimental evolution in *Escherichia coli*." Genetics **173**(4): 1851-69.
- Rosenzweig, R. F., R. R. Sharp, et al. (1994). "Microbial evolution in a simple unstructured environment: genetic differentiation in *Escherichia coli*." Genetics **137**(4): 903-17.
- Schaechter, M. (2001). "*Escherichia coli* and *Salmonella* 2000: the view from here." Microbiol Mol Biol Rev **65**(1): 119-30.
- Souza V., R., M., Valera, A., and L.E. Eguiarte (1999). "Genetic Structure of Natural Populations of *Escherichia coli* in Wild Hosts on Different Continents." Applied and Environmental Microbiology **65**(8): 3373-3385.
- Stevens, C. E. (1988). Comparative physiology of the vertebrate digestive system. New York, Cambridge University Press.

- Treves, D. S., S. Manning, et al. (1998). "Repeated evolution of an acetate-crossfeeding polymorphism in long-term populations of *Escherichia coli*." Mol Biol Evol **15**(7): 789-97.
- Turner, P. E., V. Souza, et al. (1996). "Tests of Ecological Mechanisms Promoting the Stable Coexistence of Two Bacterial Genotypes." Ecology **77**(7): 2119-2129.
- Winfield, M. D. and E. A. Groisman (2003). "Role of nonhost environments in the lifestyles of *Salmonella* and *Escherichia coli*." Appl Environ Microbiol **69**(7): 3687-94.
- Zeyl, C. (2006). "Experimental evolution with yeast." FEMS Yeast Res **6**(5): 685-91.

Supplement

Microbial populations evolved by a single clone and experimentally selected under glucose limitation developed diverse polymorphs. We sought to discover genetic mechanisms underlying the emergence and persistence of a polymorphic *Escherichia coli* population that arose under long-term glucose limitation. Adaptive changes in metabolic genes, as well as genetic changes in protein architecture, cell cycle, and other genes were observed. However, as described in the transcriptional profile of evolved clones are markedly differentiated. Many of the expression changes are consistent with our understanding of *E. coli*'s long-term growth adaptation to glucose limitation. All adaptive clones exhibit reduced activity of the stationary-phase sigma factor σ^{S} and up-regulation of glucose transport genes, including glycoprotein LacII and the inducible transporter MjxABC. Other σ -protein differences (e.g., an 8-fold up-regulation of $\sigma^{Hsp}Caf$ expression) are clone-specific and correlate precisely with the unique gene profiles in the system. Transcriptional profiling of evolved isolates in chemically defined media reveals a third class of genes whose expression in the chemostat clone differs from that observed when the clone is cultured alone. Many of these genes are part of the CpxR-mediated stress response. CpxR activation is constitutive, likely results from overexpression of acetate that is normally a waste-scavenging signal in bio-cultures. Targeted sequencing of genes involved

Genomic analysis of an experimentally evolved polymorphism in *Escherichia coli*

Kinnersley, M., Holben, W., and F. Rosenzweig

Abstract

Microbial populations founded by a single clone and experimentally evolved under resource limitation sometimes become polymorphic. We sought to discover genetic mechanisms underlying the emergence and persistence of a polymorphic *Escherichia coli* population that arose under long-term glucose limitation. Aside from a 29 kb deletion in the dominant clone, no large-scale changes in genome architecture distinguish evolved clones from their common ancestor. However, in chemostat monoculture the transcriptional profiles of evolved clones are markedly differentiated. Many of the expression changes are consistent with our understanding of *E. coli*'s long-term genetic adaptations to glucose limitation. All adaptive clones exhibit reduced activity of the stationary-phase sigma factor σ^S and up-regulation of glucose transport genes, including glycoporin LamB and the galactose transporter MglABC. Other expression differences (e.g., an 8-fold up-regulation of acetyl-CoA synthetase) are clone-specific and confirm previous reports of acetate cross-feeding in this system. Transcriptional profiling of evolved isolates in chemostat co-culture reveals a third class of genes whose expression in the dominant clone differs from that observed when the clone is cultured alone. Many of these genes are part of the CpxR-mediated stress response. CpxR activation in monoculture likely results from extracellular accumulation of acetate that is removed by acetate-scavenging strains in co-culture. Targeted sequencing of genes previously

implicated in clonal diversification shows that limiting glucose conditions initially favored a glucose-scavenging strain from which all evolved isolates ultimately arose. Global regulatory mutations in σ^S as well as small-scale regulatory mutations affecting the maltose and acetyl CoA synthetase operons contribute to the evolution of cross-feeding. Finally we identified two mutations in the ancestor that likely pre-disposed the experimental population to develop specialists that feed upon overflow metabolites. Subsequent mutations in subpopulations leading to specialization emphasize the importance of compensatory rather than gain-of-function mutations in this system. Our observations that polymorphism is quickly established in an asexual population, that adaptive mutants arise without large-scale change in genome architecture and that morphs have both common and unique patterns of gene expression influenced by whether they are cultured separately or together underscore the importance of regulatory change, founder genotype and the biotic environment in the adaptive evolution of microbes.

Introduction

For over half a century evolutionary biologists have sought to elucidate mechanisms by which adaptive variation arises and persists. Laboratory selection experiments on extant and induced genetic variation in model organisms transformed the study of these mechanisms from a retrospective to a prospective endeavor. Early on, Dobzhansky and colleagues demonstrated that genetically diverse populations of *Drosophila* rapidly adapt to laboratory selection for temperature tolerance (Wright and Dobzhansky 1946; Dobzhansky 1947; Dobzhansky and Spassky 1947). Later studies using *Drosophila* showed that other complex traits including aspects of life-history (Rose 1984), behavior (Ricker and Hirsch 1985; Ricker and Hirsch 1988; Ricker and Hirsch 1988), physiology (Gefen, Marlon et al. 2006) and development (Prasad, Shakarad et al. 2001) all respond to laboratory selection (Huey and Rosenzweig, 2009). Experimental evolution has now been fruitfully applied to multiple Eukaryotic systems, illuminating in each how history, phenotypic plasticity, genetic architecture and development interact to constrain evolutionary trajectories (Stuber, Moll et al. 1980; van Oortmerssen and Bakker 1981; Crabbe, Kosobud et al. 1985; Baer and Lynch 2003; Denver, Morris et al. 2005).

Still, long generation times and practical limits on lab population size make higher eukaryotes imperfectly suited to experimentally investigating the tempo, trajectory and molecular mechanisms by which evolutionary change occurs. Both difficulties are easily overcome by using microbes such as phage, bacteria and yeast (Elena and Lenski 2003; Zeyl 2006). The earliest studies to employ microbial systems led to two generalizations concerning the maintenance of variation in large, asexual populations. First, over ecological time and in the absence of spatial structure and differential predation, competition for the same limiting resource selects for one fittest

variant, an insight that came to be known as the “competitive exclusion principle” (Gause 1934; Hardin 1960). Second, over evolutionary time variation that arises by mutation is subject to “periodic selection” leading to a succession of genotypes each more fit than its immediate predecessor (Muller 1932; Novick and Szilard 1950; Atwood, Schneider et al. 1951).

Notwithstanding these generalizations, experiments have shown that even simple laboratory environments can support evolution of consortia consisting of stably-coexisting microbial genotypes. This phenomenon has been demonstrated in spatially and temporally unstructured chemostats (Helling, Vargas et al. 1987; Rosenzweig, Sharp et al. 1994), in temporally-structured batch cultures (Turner, Souza et al. 1996; Rozen and Lenski 2000; Friesen, Saxer et al. 2004; Spencer, Bertrand et al. 2007; Le Gac, Brazas et al. 2008), and in spatially-structured microcosms (Rainey and Travisano 1998). In each setting the emergence and persistence of polymorphism in the absence of sexual recombination would seem to require cohabitants to exploit alternative ecological opportunities (i.e., unoccupied niche space), and/or to accept trade-offs between being a specialist and a generalist (as reviewed in (Rainey, Buckling et al. 2000), also see (Zhong, Khodursky et al. 2004). The particular adaptive strategy that evolves likely depends on the mode of selection. In serial dilution batch culture, where available resources vary cyclically, different phases of growth are likely to come under selection leading to clones that have either reduced lag time, increased maximum specific growth rate or enhanced capacity to grow or to survive at high cell densities in the presence of low nutrients. Cyclical environments may therefore bring balancing selection to bear on these different phenotypes, especially if antagonistic pleiotropy precludes evolution of one fittest genotype having all of these advantageous traits. Similarly, in spatially

structured environments mutants that can successfully colonize novel microhabitats may be at a selective advantage. By contrast, in continuous nutrient-limited environments (e.g., chemostats), selection is likely to favor clones that evolve a low K_m for the limiting resource or greater efficiency in converting that resource to progeny. Ultimately, in each of these environments the evolutionary outcome is determined by founder strain(s)' genotype, the genetic pathways leading to each adaptive strategy, and the propensity of key steps along those pathways to undergo mutation and act pleiotropically.

Only recently have we begun to understand how balanced polymorphisms arise in asexual populations. In the case of serial dilution batch culture, Rozen et al. (Rozen, Philippe et al. 2009) recently demonstrated that differences in the activity of the global regulator RpoS underlie co-existence of two *E. coli* isolates that have different propensities to survive extended stationary phase. However, the precise genetic basis for these activity differences remains obscure. Investigating polymorphism in a spatially structured microcosm, Bantinaki and co-workers demonstrated that a mat-forming variant of *Pseudomonas fluorescens* colonizes the air-broth interface owing to a structural mutation in a methylesterase that modulates expression of a cellulosic polymer (Bantinaki, Kassen et al. 2007). Finally, in the case of a polymorphic *E. coli* population first described by Helling et al. (Helling, Vargas et al. 1987), small-scale regulatory mutations affecting expression of a single operon (*acs-actP-yjcH*) partly explain repeated evolution of acetate cross-feeding under continuous glucose limitation (Treves et al. 1998). When individual clones from such populations are grown in monoculture, however, 2D-PAGE reveals strain-specific differences in ca. 20% of identifiable proteins expressed, suggesting the presence of other mutations with highly pleiotropic effects (Kurlandzka, Rosenzweig

et al. 1991). Thus, regardless of the experimental system, considerable uncertainty remains as to whether mutations at regulatory or at structural gene loci consistently deliver greater fitness increments, which category of mutation better explains the maintenance of diversity, and whether one type is more likely to precede the other in an evolutionary sequence leading to balanced polymorphism.

We sought to address these questions by investigating the experimental population first described by Helling et al. (Helling, Vargas et al. 1987). We specifically tested the hypothesis that enhanced uptake and assimilation of the primary resource, glucose, results from one (or few) mutations in global regulators, but that specialization on secondary resources arises from mutations at key structural loci. We predicted that major global regulatory changes would precede structural changes required for specialization. Lastly, we anticipated that in comparing the consortium's expression profile to that of individual members grown in monoculture we would discover emergent properties of the system not visible using a purely reductionist approach.

Aside from a 29 Kb deletion in the dominant clone, the evolved clones and their ancestor are virtually indistinguishable by rep-PCR and array comparative genome hybridization. However, gene expression profiling of each strain in monoculture indicates that the ancestral strain significantly differs from each evolutionary-adapted strain at ~180 loci. These observations are broadly consistent with the report (Kurlandzka, Rosenzweig et al. 1991) that expressed protein levels differ between ancestral and evolved strains at ~160 out of approximately 700 spots resolvable by 2-D PAGE of ³⁵S-labelled cells. Significance Analysis of Microarrays (SAM) indicates that 21 out of the top 91 significant genes similarly expressed in all clones are up-regulated and are primarily involved in metabolism and transport. The

remaining 70 are down-regulated and belong to a variety of functional categories. In addition, nearly all of these are part of the σ^S regulon. Both sets of expression differences were ultimately tied to shared mutations that affected the activity of σ^S and the regulation of the maltose operon. SAM analysis across clones reveals that most expression differences that distinguish one isolate from another are either related to motility or have unknown function; a majority of these are regulated by CRP and the global stress regulator CpxR. The “community” expression profile of clones grown in co-culture is strikingly similar to the profiles of three of the four clones grown in monoculture, suggesting that the CpxR effect may be related to biochemical interactions between strains; indeed, the dominant clone, which over-secretes acetate in monoculture, is apparently relieved of acetate feedback on gene expression by the presence of an acetate-scavenging subpopulation in the consortium. Finally, the discovery of previously unrecognized ancestral regulatory mutations in loci required for acetate and glycerol catabolism demonstrates how ancestral genotype critically influences evolutionary outcomes, even in simple model systems. Overall, our results show that global regulatory change followed by small scale regulatory change promotes rapid adaptive evolution in a simple, unstructured, resource-limited environment, and that founder genotype and chemical interactions among clones not only facilitate co-evolution, but also strongly impact their respective patterns of gene expression.

Materials and Methods

Strains, media and culture conditions

Escherichia coli JA122, CV101, CV103 CV115 and CV116 were stored at -80°C in 20% glycerol (See Table 1). Davis minimal media was used for all liquid cultures

with 0.025% glucose added for batch cultures and 0.0125% for chemostats (Helling, Kinney et al. 1981). Inocula for chemostat cultures were prepared by growing isolated colonies from TA plates in liquid media for 16-20 hours at 30°C, pelleting the cells at 3,000 rpm and resuspending the pellet in fresh media. A portion of this suspension was used to inoculate chemostats to a density that approximated the expected steady-state density. Chemostats contained Davis minimal media with 0.0125% glucose and were maintained at 30°C at a dilution rate of $\approx 0.2/\text{hr}$ for 70 hours (~ 15 generations). A_{600} readings and spread plate cell counts were taken at regular intervals to monitor growth. Cell densities at the end of 70 hours were between 1.5 and 2.5×10^8 cells mL^{-1} . At the end of each chemostat run, three aliquots of 40 mL of culture were immediately filtered onto 0.2 μm nylon membranes, flash-frozen in liquid nitrogen and stored at -80°C for RNA extraction.

For transcriptional profiling, each strain was grown in triplicate on three different occasions with independently prepared batches of media. To reduce the effect of variation in media preparation, cultures of JA122 were grown concomitantly such that each experimental chemostat had a corresponding reference fed off of the same media reservoir.

Nucleic acid extraction

Genomic DNA was extracted from cells grown in batch culture using a modification of methods described by Syn and Swarup (Syn and Swarup 2000). Subsequent to DNA precipitation, spun pellets were re-suspended in TE pH 8.0 containing 50 $\mu\text{g}/\text{mL}$ DNase-free RNase A and incubated at 37°C for 30 minutes. Samples were re-extracted once with phenol:chloroform (3:1), once with

phenol:chloroform (1:1) and twice with chloroform. Following re-precipitation the DNA was resuspended in TE pH 8.0.

Total RNA was extracted using an SDS lysis/ hot phenol method developed by the Dunham lab <http://www.genomics.princeton.edu/dunham/MDyeastRNA.htm>. Frozen filters were mixed with 4 mL lysis solution (10 mM EDTA, 0.5% SDS, 10 mM Tris pH 7.4) and vortexed to remove cells. An equal volume of acid phenol (pH 4.5) was added and the mixture was incubated at 65°C for 1 hour with frequent mixing. The entire extraction was transferred to a phase-lock gel tube (5Prime Inc., Gaithersburg, MD) and centrifuged according to the manufacturer's instructions. The aqueous layer was extracted twice more with chloroform: isoamyl alcohol (24:1) and precipitated with ethanol. Pellets were dried and resuspended in RNase free water, treated with 0.1U/ μ l RQ1 RNase-free DNase at 37°C for 1 hour (Promega, Madison WI), then further purified using the Qiagen RNeasy Mini kit. RNA quality was assessed on agarose denaturing gels as well as using a Bioanalyzer (Agilent Technologies) and quantified spectrophotometrically.

Array design

Microarrays were fabricated using full-length open reading frame PCR products generated using the Sigma-Genosys ORFmers primer set and reaction conditions and cycling parameters recommended by the manufacturer (Sigma-Genosys, The Woodlands, TX). This set contains primer pairs for all 4290 known and hypothetical ORFs in *E. coli* K12 MG1655. PCR reactions were repeated and pooled as necessary to obtain at least 3 μ g of DNA. Pooled reactions were ethanol precipitated, resuspended and further purified using a Qiagen MinElute96 UF PCR purification kit. 5-10 μ l of clean product was run on agarose gels for quantification

and to verify that the product was the correct size. 192 genes were excluded because they were either the wrong size, produced multiple products or failed to amplify after repeated attempts. An additional 19 genes amplified poorly and consequently were spotted at a lower concentration but were retained in the analyses (see Supplementary Table 1). Products were standardized to a concentration of 2 μ g, dried and resuspended in 10 μ l 3X SSC for printing. Arrays were printed in Corning Gaps II aminosilane coated slides using a 48-pin Stanford-UCSF style arrayer at the Stanford Functional Genomics Facility (Stanford, CA).

Array-based Comparative Genome Hybridization (a-CGH) and expression profiling

Microarray Expression Profiling and Comparative Genome Hybridization

were performed using protocols developed at the J. Craig Venter Institute (<http://pfgrc.tigr.org/protocols/protocols.shtml>) with the following modifications. For a-CGH 5 μ g of genomic DNA was sonicated to an average fragment length of 2-5 kb using a Branson Digital Sonifier at 11% amplitude for 1.1 seconds and a final concentration of 0.5 mM, 1:1 aa-dUTP:dTTP labeling mixture was used in the Klenow reaction. For expression profiling, 20 μ g of total RNA was reverse transcribed using 9 μ g of random hexamer and 0.83 mM 1:1 aa-dUTP:dTTP labeling mixture. Slides were blocked in 5X SSC, 0.1% SDS, 1% Roche Blocking Reagent prior to hybridization

(<http://www.genomics.princeton.edu/dunham/MDhomemadeDNA.pdf>) (Roche Applied Science, Mannheim, Germany). Hybridized arrays were scanned using an Axon 4000B scanner (Molecular Devices, Sunnyvale, CA).

qRT-PCR

Quantitative RT-PCR was performed using the Step-One Plus Real-Time PCR System (Applied Biosystems (ABI), Foster City, CA). Primers and probes were designed using the default parameters with Primer Express 3.0 and purchased from Integrated DNA Technologies (IDT, Coralville, IA). 2 μ g of total RNA was treated with RNase-free DNase to remove residual DNA and subsequently reverse transcribed using the ABI High Capacity cDNA Reverse Transcription Kit. 1 μ l of cDNA was added to 1X TaqMan Gene Expression Master Mix containing 900 nM each primer and 250 nM probe and cycled using the universal cycling program for the StepOne system. Relative amounts of each transcript were calculated using the $\Delta\Delta C_t$ method using *mdaB* as an endogenous control. The sequences of the primers and probes used are shown in Table 2.

Image processing and statistical methods

a-CGH images were processed using a combination of GenePix Pro 6.0, the TIGR TM4 software suite available at (ref) and Microsoft Excel. Image analysis and spot filtering was done in GenePix. a-CGH spots were considered acceptable if they: (1) passed the default flag conditions imposed by the software during spot finding; (2) had an intensity : background ratio > 1.5 and overall intensity > 350 in the reference channel; and (3) had an intensity:background ratio of > 1.0 in the experimental channel. GenePix files were converted to TIGR MEV format using Express Converter. Ratios were normalized using total intensity normalization and replicate spots were averaged using TIGR MIDAS. Results were viewed using Caryoscope 3.0.9. One a-CGH comparison was performed for each experimental isolate using the ancestor JA122 as the reference genome.

For transcriptional profiling, spots were considered acceptable if the regression R^2 was >0.6 or the sum of the median intensities for each channel minus the median background was >500 . Spots that contained saturated pixels in both channels were excluded from the analysis but spots that were saturated in only one channel were flagged and retained. GenePix results were converted to TIGR MEV format using Express Converter. Ratios were normalized by total intensity normalization and replicate spots were averaged using TIGR MIDAS. Results were viewed and analyzed using TIGR MeV. Three comparisons including one dye-flip pair were performed for each biological replicate for a total of nine comparisons for each strain. Genes that did not have acceptable spots for 2 out of the 3 biological replicates were excluded from the downstream analysis. For each biological replicate, reference RNA was prepared from independent JA122 cultures that were grown at the same time off of the same media reservoir.

Significance Analysis of Microarrays (Tusher, Tibshirani et al. 2001) (SAM) was used to examine expression differences between strains using a multi-class comparison consisting of four groups. Similarities among strains were identified using one-class SAM and differences between the strains were examined using a 4-class SAM. δ cutoffs were assigned either by eye, (in which case the median false discovery rate (FDR) was equal to 0%), or set at the 0% FDR threshold. In all cases these settings resulted in q-values of 0. The default settings for all other parameters were retained. The average (mean) \log_2 ratios for biological and technical replicates were calculated after SAM analysis using Microsoft Excel.

Pair-wise Pearson correlation coefficients between array and qRT-PCR expression data were calculated as in Larkin et al., (2005) using Microsoft Excel.

Regulon Comparisons

Transcription unit, regulon and operon information was collated from the EcoCyc Database at <http://www.ecocyc.org> (Karp, Keseler et al. 2007). Predicted regulatory binding site information was obtained via TractorDB (<http://www.tractor.lncc.br>) (Gonzalez, Espinosa et al. 2005).

Data archiving

Data are available through the NIH GEO database.

Results

Bacterial Strains

Table 1 summarizes phenotypic data on the Helling et al. strains, much of which has been previously published (Helling et al. 1987; Rosenzweig et al., 1994). Certain features of these strains' physiology merit review, as they provide context for interpreting the results of our monoculture and community expression analyses. CV101, CV103, CV115, and CV116 were isolated at 770 generations from a glucose-limited chemostat operated under aerobic conditions at $D=0.2 \text{ h}^{-1}$. CV103 was the numerically dominant clone, comprising greater than two-thirds the population, followed in relative abundances by CV116, CV101 and CV115.

When cultured in glucose minimal media under nutrient *non*-limiting conditions the order of μ_{\max} , clones' maximum specific growth rates, was CV116>CV115>CV101>CV103. Likewise, the order of relative growth yield in batch culture was CV116>CV115=CV101>CV103. Interestingly, the dominant clone, CV103, was unique among all evolved clones in that it grew more slowly in batch culture and produced fewer cells than the common ancestor, JA122. This observation

is consistent with incomplete metabolism of the limiting growth substrate, glucose. None of the evolved clones demonstrated enhanced uptake of labeled 2-deoxyglucose, a non-metabolizable analogue assimilated via the II^{Man} glucose transport pathway in *E. coli*. However, relative to their common ancestor, all of the evolved clones showed enhanced uptake of the glucose analogue ^{14}C - α -methylglucoside (αMG) which is assimilated via the $\text{IIB}^{\text{Glc}}/\text{III}^{\text{Glc}}$ pathway. Moreover, the dominant clone, CV103, accumulated significantly more αMG intracellularly than the other evolved strains (Helling, Vargas et al. 1987). Not surprisingly, at steady state in glucose-limited chemostats the residual substrate concentration, K_s , was found to be an order of magnitude less in CV103 than was observed for CV101, and less than half what was seen for CV116. On the other hand, unlike CV101 and CV116, CV103 was found to release into both batch and chemostat media appreciable amounts of acetate, creating a niche favorable for the evolution of cross-feeding. CV101 filled that niche, and scavenges this substrate to the detection limit of spectrophotometric assay (Rosenzweig, Sharp et al. 1994).

Genomic Characterization

To assess the level of large-scale genetic variation between the ancestor and the evolved clones, we performed rep-PCR fingerprinting and array-CGH. BoxAIR rep-PCR fingerprints were indistinguishable (see Supplementary Figure 1). However, a-CGH revealed an approximately 30 Kb deletion in CV103 (Figure 1). A total of 27 genes were affected by the deletion, 12 of which have no known function. Of the remaining 15, 3 have a predicted function based on homology to previously characterized genes and 12 are involved in a variety of cellular processes including

transcription, arginine biosynthesis, anaerobic respiration, nitrogen metabolism and glycoprotein biosynthesis.

Transcriptional profiling reveals changes in gene expression common to all adaptive clones, relative to their common ancestor

We used DNA microarrays to assess the global transcriptional response of each evolved strain to growth under glucose limitation in chemostat monoculture. In these experiments, evolved clones were grown to steady state in chemostat monoculture under conditions identical to those under which they evolved. In each case, steady state transcripts levels were estimated in relation to a common reference: the ancestral strain, JA122, grown in parallel under identical conditions. On average, the expression of 6.8% (or approximately 279 genes) of the measurable transcriptome is at least 2-fold up or down regulated in the evolved isolates versus JA122 (Supplementary Figure 2). This number compares reasonably well with an early proteomic analysis report on the Helling et al. strains grown in monoculture. Within the limits of their resolution (~700 proteins) Kurlandzka et al. (1991) found ~160 protein level differences between evolved clones and their common ancestor JA122.

1-class SAM identified 91 genes whose expression was significantly up- or down-regulated in all clones when each was grown in chemostat monoculture (Figure 2, Supplementary Table 2). The 21 up-regulated genes, representing 9 unique transcription units, were primarily involved in carbon catabolism while the remaining 70 down-regulated genes from 59 transcription units belonged to a variety of MultiFun classes including carbon metabolism, building block/macromolecule biosynthesis, transport and adaptation to osmotic stress.

Genes up-regulated in all evolved strains

Four of the nine transcription units up-regulated in all evolved isolates (*lamB*, *mglBAC*, *galS* and *rhaBAD*) are involved in carbon metabolism and are positively regulated by CRP, a major global regulator of catabolite-sensitive operons (Figure 2, Table 2) (Zheng, Constantinidou et al. 2004; Perrenoud and Sauer 2005). *LamB* and the *mgl* operon are also regulated by the stationary phase sigma factor RpoS, and along with *galS* (*mglD*) have all previously been shown to be targets of selection during long-term adaptation to glucose limitation (Notley-McRobb and Ferenci 1999; Notley-McRobb and Ferenci 1999). Interestingly, transcript abundance of LrhA, a LysR-family transcriptional dual regulator, is also increased in all evolved strains. LrhA is thought to be indirectly involved in the degradation of RpoS during log phase as well as directly responsible for the repression of the flagellar gene master regulator FlhDC (Gibson and Silhavy 1999; Lehnen, Blumer et al. 2002). FlhDC is required for the transcription of the flagellar regulon and has also been identified in a microarray analysis as a repressor of the *mglBAC* genes (Liu and Matsumura 1994; Pruss, Liu et al. 2001). Finally, expression of the IS5 insertion element transposase, *insH* is elevated in all isolates. Although the overexpression of *acs* in CV101 described by Treves et al. (Treves, Manning et al. 1998) is the result of IS30 movement, significantly upregulation of *insH* in this context is intriguing, especially considering the extent to which insertion element movement has influenced adaptation in other experimental evolution studies (as reviewed in (Schneider and Lenski 2004)).

Genes down-regulated in all evolved strains

A number of genes that are down-regulated in the 1-class SAM analysis (12) are also involved in central metabolism (see Figure 2 and Supplementary Figure 3). Most notably, the expression of two components of the glucose-specific PTS permease, *ptsG* (EIIB/C^{Glc}) and *crr*, as well as the non-specific PTS component *ptsH* (HPr) are all significantly lower, which is surprising given the low level of glucose present in the chemostat and the demonstrated improvement in glucose uptake exhibited by all of the evolved isolates (Table 2) (Rosenzweig, Sharp et al. 1994; Rahman, Hasan et al. 2006).

In glycolysis and the pentose phosphate pathway, 5 genes also show decreased transcript levels in the evolved isolates. These include two enzymes responsible for converting fructose-6-phosphate into glyceraldehyde-3-phosphate (*tpiA* and *fbaB*) and enolase (*eno*), which catalyzes the final step in the conversion of 2-phosphoglycerate to phosphoenolpyruvate. FbaB is typically not transcribed during aerobic growth on glucose and may consequently be the primary aldolase for gluconeogenesis, as it is only turned on during growth on gluconeogenic substrates such as glycerol (Scamuffa and Caprioli 1980). Curiously, enolase has a secondary role as part of the degradosome in *E. coli* that is responsible for the rapid degradation of *ptsG* mRNA in response to high levels of glucose-6-phosphate and fructose-6-phosphate (Morita, Kawamoto et al. 2004). Transketolase B and transaldolase A (*tktB* and *talA*), which act in the non-oxidative branch of the pentose phosphate pathway, also show decreased expression; however, both are variants of a more active isoenzyme.

We also note diminished expression of 7 genes which play a role in mixed acid fermentation: pyruvate oxidase (*poxB*), pyruvate formate-lyase (*pflB*), acetaldehyde dehydrogenase (*adhE*), both ethanol and alcohol dehydrogenase (*adhP*

and *adhE*) and D-lactate dehydrogenase (*ldhA*). While lower transcript levels of these enzymes do not necessarily mean that corresponding enzyme levels are insufficient to convert pyruvate into fermentation products under glucose limitation, the pattern of down-regulation suggests that the conversion of pyruvate into acetyl-CoA most likely occurs via the pyruvate dehydrogenase complex.

Finally, transcripts needed for the manufacture of motility and attachment structures, in particular the flagellin (*flhC*) and curlin (*csgA*) genes, also show decreased expression, an observation that is perhaps not surprising considering that the chemostat environment is well-mixed, and that attachment and motility may be of limited utility therein (Table 2).

Regulation of genes similarly expressed in all evolved isolates

Given that Helling et al. reported that relatively few adaptive sweeps in the ~700 generations leading to the establishment of the polymorphic population, it is reasonable to consider the possibility that many of the observed changes in gene expression are coordinately regulated. Indeed, we found that many changes are attributable to two global regulators, σ^S and CRP. A striking number (33%) of the 91 up- and down-regulated genes are part of the RpoS-mediated stress response (Figure 2). This is particularly noteworthy given that deleterious mutations in *rpoS* are frequently encountered in both wild and experimental *E. coli* populations as a response to prolonged low nutrient conditions (Ferenci 2001; Notley-McRobb, King et al. 2002; Ferenci 2003). Nineteen genes (21%) are regulated by CRP and an additional 13 (14%) have predicted CRP binding sites (Figure 2).

Transcriptional profiling also reveals changes in gene expression that distinguish adaptive clones from one another

To ascertain how the transcriptional profiles of evolved clones differ from one another we implemented a 4-class SAM analysis (Figure 3A, Table 2). Aside from the anticipated overexpression of *acs-yjcHG* in CV101, the transcription patterns of CV101, CV115 and CV116 appear remarkably similar. By contrast, CV103 differs from the other three at a number of loci, and accounts for the great majority (~94%) of significant differences that distinguish adaptive clones. At a δ value of 0.27, a total of 93 genes from 66 transcription units significantly differ in steady state expression levels in at least one isolate. These genes tend to fall into three MultiFun classes: metabolism, cell structure and transport. Under the category of metabolism, forty-four genes from twenty-seven transcription units vary in their relative expression patterns. The metabolism-building block biosynthesis subclass contained the most independent transcription units (8/27), including *acs-yjcHG* (acetyl CoA synthetase), which is up-regulated in CV101. It is noteworthy that 12 out of 66 transcription units in this group have been shown to be regulated by the extracytoplasmic stress response regulator CpxR, and that 16 are regulated by CRP or have predicted CRP binding sites (Gonzalez, Espinosa et al. 2005; Karp, Keseler et al. 2007).

Genes down-regulated in CV103 but up-regulated in CV101, CV115 and CV116

Relative to its ancestor JA122 grown in monoculture, 27 genes from 12 transcription units show diminished expression in CV103, but increased expression in the other evolved clones. Of these, 4 are thought to be up-regulated by CRP directly and 2 have predicted CRP binding sites in their promoter regions. Especially noteworthy note are the flagellar motor complex and flagellar hook gene transcripts

which are conspicuously down-regulated in CV103, but up-regulated in CV101, CV115 and CV116. Up-regulation of flagellar genes has been previously observed both under glucose limitation and during growth on secondary carbon sources such as acetate (Oh, Rohlin et al. 2002; Polen, Rittmann et al. 2003; Franchini and Egli 2006; Zhao, Liu et al. 2007). Thus, in certain respects CV103 appears to exhibit a transcriptional response inconsistent with adaptation to nutrient-poor conditions. The flagellar master switch, FlhDC can be induced by CRP but may also be repressed by phosphorylated OmpR, a transcriptional shift that would be expected to lead to down-regulation of the majority of flagellar transcripts (Liu and Matsumura 1994). Interestingly, *fliC*, the gene that encodes flagellin, the flagellar structural subunit, is down-regulated in *all* isolates in the 1-class SAM analysis suggesting that despite differences in motor complex and hook gene transcript levels, all four evolved strains are unable to make an intact flagellum. These results are supported by the observation only the ancestor displays movement when grown in motility agar (data not shown).

Multiple CRP-induced transport-related gene transcripts also show diminished relative abundance in CV103. Both the galactitol-PTS-permease operon (part of the tagatose-6-phosphate pathway) as well as the gene for the OmpF outer membrane porin are repressed in CV103 (Table 2). Moreover, our observation that expression of *ompF* mRNA is diminished in CV103 is strikingly consistent with previous observations that *ompF* protein expression is greatly diminished in this strain relative to other members of the consortium and their common ancestor (Kurlandzk, Rosenzweig et al. 1991). OmpF expression has been studied extensively in relation to culture under glucose-limitation (Liu and Ferenci 1998; Zhang and Ferenci 1999; Liu and Ferenci 2001; Maharjan, Seeto et al. 2006). Typically, aerobic glucose limitation

leads to increased *ompF* expression as part of a general strategy by the cell to increase membrane permeability. While the regulation of this response is complex and involves multiple factors, it is important to note that high intracellular acetyl phosphate levels may down-regulate *ompF* expression by phosphorylating OmpR, a negative regulator of *ompF* transcription (Pratt, Hsing et al. 1996; Liu and Ferenci 2001).

Finally, CV103 shows a relative decrease in transcript levels of cytochrome *bo* oxidase, a terminal respiratory chain oxidase used during aerobic growth, and *eutD*, a predicted acetyl transferase that remains largely uncharacterized in *E. coli*, but has been shown to be required for efficient acetate excretion in *Salmonella* (Starai, Garrity et al. 2005).

Genes up-regulated in CV103 but unchanged or down-regulated in CV101, CV115 and CV116

Forty genes representing 35 transcription units were significantly up-regulated in CV103 but were unchanged or down-regulated in the other evolved strains (Figure 3A). Considering these in the light of available regulatory information, the most important effector appears to be CpxR, which controls 8 transcription units. Of the remaining 27, three are regulated by CRP and four have predicted CRP binding sites. Several genes in this group function to mitigate cellular stress, perhaps most notably the heat-shock sigma factor RpoH, which is controlled by both CpxR and CRP (Table 2) (Zheng, Constantinidou et al. 2004; Zahrl, Wagner et al. 2006). Aside from mediating the cellular response to high temperature, RpoH is also transcribed during carbon starvation and exposure to hyperosmotic conditions (VanBogelen, Kelley et al. 1987), Jenkins 1991). RpoH is the sigma factor for five other transcription units that

are up-regulated in CV103, including those for ExoX nuclease and MutL, both of which are involved in DNA mismatch repair, and *raiA*, a translation elongation inhibitor that interacts directly with the ribosomal A site to prevent binding of aminoacyl-tRNAs during stationary phase (Agafonov, Kolb et al. 2001). Although not part of the RpoH regulon, two functionally-related genes also show increased transcript abundance: CpxP, a CpxR-mediated extra-cytoplasmic stress response regulator and potential chaperone, and DegP, a high-temperature protease/chaperone (Table 2). DegP is normally induced by CpxR and is responsible for degrading misfolded proteins at elevated temperatures. At lower temperatures (between 28° C and 37° C) DegP loses protease activity and instead assists in proper folding of the maltose operon regulator, MalS (Spiess, Beil et al. 1999). CpxP, which negatively regulates CpxR, is itself positively regulated at the transcriptional level by CpxR and is therefore responsible for modulating the Cpx response. CpxP accomplishes this by interacting with the histidine kinase for CpxR, CpxA. Thus, elevated levels of CpxP should ultimately result in repression of the CpxR-mediated stress response as a result of CpxA being unable to phosphorylate CpxR. In fact, our data suggest the opposite.

Several transport related genes are up-regulated only in CV103: the binding subunit of a glycerol-3-phosphate transporter, UgpB, a divalent metal cation transporter FieF, and the high-affinity molybdenum transporter ModCBA. While expression of the glycerol-3-phosphate transporter UgpABCE does not enable cells to grow on that substrate as a sole carbon source, it does contribute significantly to cellular phosphate economy. The *ugp* operon is typically induced under conditions of phosphate limitation as part of the PHO regulon and is negatively affected by σ^S . Conditions that result in the accumulation of acetyl phosphate (such as growth on pyruvate or inactivation of acetate kinase) induce the PHO regulon (Wanner 1992),

and could therefore be expected to induce the transcription of *ugpB*. Moreover, null mutations in *rpoS* result in significantly higher expression of *ugpB* under phosphorus starvation conditions (Taschner, Yagil et al. 2004). Interestingly, *ugp* genes have a high degree of homology to members of the maltose operon; MalK and UgpC can functionally substitute for one another (Overduin, Boos et al. 1988).

The second transporter gene uniquely up-regulated in CV103 is *fieF*, a divalent metal cation transporter which has no known regulatory interactions with CRP, σ^S or CpxR, but which is physically located immediately downstream from *cpxP*. The transcriptional terminator of the *cpxP* is rho-independent, raising the interesting possibility that transcription of *fieF* results from a disrupted terminator, leaving *fieF* under the de facto regulatory control of CpxR (Danese and Silhavy 1998).

Finally, we note in CV103 increased expression of the ModCBA transporter. This transcriptional unit is positively controlled by CRP and is needed to secure molybdenum as a cofactor for the catalytic function of various molybdoenzymes, most of which are expressed under anaerobic conditions. In *E. coli* there are only a few known aerobic molybdoenzymes- biotin sulfoxide reductase (BisC), nitrate reductase (NRZ), and formate dehydrogenase (FDH-O) (Kozmin, Pavlov et al. 2000). Transcription of the NRZ operon (*narZYWV*) is positively controlled by σ^S , and the entire operon is deleted in CV103 (as shown in Figure 1) (Chang, Wei et al. 1999).

Genes whose expression is unchanged in CV103 but altered in CV101, CV115 and CV116

Five genes from three transcription units show ancestral levels of expression in CV103 but are clearly down-regulated in CV101, CV115 and CV116. *yrbL* and *yqjCDE* have no known function, but both are positively controlled by PhoP under

low Mg²⁺ conditions (Minagawa, Ogasawara et al. 2003; Zwir, Shin et al. 2005). *otsA*, a trehalose-6-phosphate synthase, is repressed in all of the isolates, but has a slightly lower transcription level in CV103. The expression of *otsA* is normally stimulated by σ^S in stationary phase as well as in response to cold and heat shock (Kandror, DeLeon et al. 2002).

As noted above, the most obvious difference between CV101 and the other evolved strains is overexpression of *acs* (acetyl CoA synthetase) and *actP* (acetate/glycolate permease) (Figure 3 and Table 2). Aside from these, only a few genes distinguish CV101 from CV115 and CV116. For example, CV115 has increased expression of *livF* and *livG*, part of the leucine ABC transporter and branched-chain amino acids transporter, while CV116 displays a higher transcript level *fieF*, a putative siderophore outer membrane receptor (Table 2).

Our global gene expression analyses are in overall agreement with previously published proteomic, biochemical and genetic data for these same isolates (Helling et al 1987; Kurlandzka et al. 1991; Rosenzweig et al 1994; Treves et al. 1998). *acs* overexpression by CV101 has now been confirmed by multiple lines of investigation. Also, both mRNA and protein profiling indicate that relative to the common ancestor, up-regulation of *lamB* occurs at steady state under glucose limitation in all evolved isolates. Likewise, down-regulation of *ompF* in CV103 and its concomitant up-regulation in the other three strains is confirmed by both techniques. Lastly, although increased expression of *rpoH* (Figure 3) was not observed on 2-D gels, Kurlandzka et al. did observe increased expression of σ^{32} -dependent proteins such as GroES and GroEL.

Transcriptional profiling of the evolved consortium

Reconstruction experiments demonstrated that three of the evolved strains could stably coexist in continuous culture as a consortium, and that their coexistence was made stable by cross-feeding (Rosenzweig et al., 1994). When limited on 0.0125% glucose, the consortium reproducibly apportioned as ~70% CV103, 20% CV116 and 10% CV101 at steady state. To better understand the mechanism underlying stable coexistence we interrogated the consortium transcriptome using DNA microarrays. In general, we observe that genes significantly up or down in the 1-class SAM monoculture analysis behave similarly when clones are co-cultured (Figure 4). Furthermore, consortium profiling extends the results of the monoculture analyses to include other members of operons previously identified by 1-class SAM. For example, *malK* and *malM* (which are co-transcribed with *lamB*), as well as *malF*, *G* and *S* form two separate, but similarly regulated transcription units. Each shows increased expression when cells are cultured as a consortium (Figure 4D).

A number of transcripts which were not scored as significant in the monoculture 1-class SAM using the stringent “by eye” FDR cutoff, were scored as significant using 0% FDR. A subset of these were also found to be up-regulated in the consortium, including genes for a second glycerol-3-phosphate transporter/phosphodiesterase, *glpTQ*, part of the G3P-dehydrogenase, *glpA*, as well as the fumarase genes, *fumA* and *fumC* (Figure 4D, Supplementary Figure 3).

Comparison of the consortium transcriptional profile to the 4-class monoculture SAM led to the surprising observation that a relatively small number of genes were significant in both analyses. From the expression levels of the acetate transporter *actP* (which is co-transcribed with *acs*) in both analyses, it is clear that consortium transcript levels are a reasonable approximation to the predicted “average”

of the monoculture data. However, the consortium profile for these common genes most closely matched the CV101, CV115 and CV116 profiles, despite the fact that CV103 is the numerically dominant member of the equilibrium chemostat population (see Figure 5). To ascertain whether this phenomenon was a general feature of the dataset, we looked at transcript levels across all samples for genes that were either (A) significant in the consortium analysis but not in the monoculture experiments or (B) significant in the monoculture experiments but not in the consortium profile. For this comparison, the more stringent "by-eye" significance cutoff was used. In both cases, the vast majority of genes that were differentially regulated in CV103 monoculture (and thus distinguished this isolate from the other clones) again had transcript levels that closely matched CV101, CV115 and CV116. While this analysis is undoubtedly limited by the fact that the individual contributions of isolates cannot be dissected from the consortium RNA pool, the sheer number of gene transcripts that follow this trend strongly suggests that CV103 has a different gene expression profile in the shared metabolic environment of the consortium than it has when grown in isolation.

Confirming expression changes for select genes by RT-qPCR

Three genes (*lamB*, *acs* and *flgB*) with different relative expression levels were selected for q-RT-PCR. In all three cases the PCR results closely matched the array results with correlation coefficients ranging from 0.78-0.99 (see Supplementary Figure 3).

Sequence analysis of candidate genes

To place our results in the context of previously published work and to potentially identify new mutations that contribute to the transcriptional profiles of the

evolved isolates, 13 candidate genes and their corresponding regulatory elements were sequenced (Table 3). Our selection of candidate genes was motivated by previous observation that members of the evolved polymorphism had differentiated from one other and their common ancestor with respect to glucose, acetate and glycerol metabolism.

Glucose transport and assimilation – Mutations that enhance the ability of *E. coli* to move glucose across the inner and outer membranes are commonly observed during adaptation to glucose limitation. Glucose can cross the outer membrane by passing through either the general porins OmpC and OmpF or the maltodextrin porin LamB, which is part of the *mal* regulon. Transcriptional changes relative to the ancestor were observed for both *ompF* and *lamB* as well as a number of the other *mal* regulon genes.

As the regulation of OmpF is complex and involves a number of different regulators (any of which might be a mutational target), sequencing efforts were focused on the LamB structural gene, the *mal* transcriptional activator MalT and the *mal* repressor Mlc. Other groups have reported adaptive mutations in the first ~360 amino acids of MalT eliminate the need for maltotriose inducer and thus allow continuous induction of the *mal* genes (Dardonville and Raibaud 1990; Notley-McRobb and Ferenci 1999; Schlegel, Danot et al. 2002). Likewise, mutations in Mlc that abolish repressor activity and lead to increased transcription of MalT are also common under glucose limitation (Notley-McRobb and Ferenci 1999). Despite the fact that upregulation of the *malEFG*, *malK-lamB-malM* and *malS* transcription units in the monoculture and consortium SAM analyses strongly pointed to increased transcription of the entire *mal* regulon, we were surprised to find that there were no mutations in *mlc* for any of the isolates or the promoter region/structural gene for

malt in CV103 or JA122. Similarly, the LamB gene itself was also unchanged across all isolates. However, when we sequenced the same portion of *malt* for the remaining strains, we identified an A→E substitution at aa 53 present in CV101, CV115 and CV116. Despite the large number of mutations in MalT that have been observed by others during chemostat adaptation to glucose limitation, none reside in the same structural motif as aa 53 (Table 3). This region of the protein (from aa44-55) forms a helix that is positioned between two ATP binding motifs and is part of a larger and widely-recognized nucleotide-binding P-loop NTPase domain (Leipe, Koonin et al. 2004). Whereas most members of this family of NTPases typically have a non-polar residue at this position, CV101, CV115 and CV116 have acquired a polar substitution. The mechanistic significance of this substitution is currently unknown.

From the *E. coli* periplasm, glucose can cross the inner membrane via the phosphotransferase system, the glucose/galactose transporter MglBAC and/or the galactose MFS transporter GalP. As no report of mutations in GalP under glucose limitation have been reported in the literature, we focused our sequencing efforts on the *ptsG* structural gene known upstream regulatory region, the *mgl* transcriptional repressor *mglD* and the *mgl* operator. In our 1-class SAM analyses, the PTS enzyme II^{glc} (PtsG/Crr) and the MglBAC transporter were differentially transcribed, with II^{glc} being repressed and MglBAC upregulated. While mutations in *ptsG* do confer a moderate fitness advantage in glucose-limited chemostat culture, upregulation of MglBAC either by inactivating its repressor, MglD, or eliminating the repressor binding site in the *mgl* operator exerts a much greater effect on glucose transport (Manche, Notley-McRobb et al. 1999; Notley-McRobb and Ferenci 1999; Maharjan, Seeto et al. 2007). As might be expected from their relative activities, no mutations were found in the *ptsG* structural gene or its upstream regulatory sequence for any of

the evolved isolates, but they all shared the same mutation in the *mgl* operator: a single G→T transversion located 3 base-pairs from the end of *mglD* (**Table 3**). This substitution is identical to one previously reported and lies within the repressor binding site thus allowing semi-constitutive transcription of *mglBAC* and increased transport of glucose into the cytoplasm (Notley-McRobb and Ferenci 1999).

Acetate uptake and secretion – The basis of the acetate-scavenging behavior of CV101 was previously identified to be IS element-mediated constitutive over-expression of acetyl-CoA synthetase (Treves, Manning et al. 1998). Re-sequencing of the *acs* gene and promoter region for the common ancestor, JA122, highlighted the importance of the ancestral promoter composition relative to the fully sequenced *E. coli* K-12 strain MG1655: JA122, CV103, CV115 and CV116 all share an A→T substitution at position -93 relative to the *acs* start site (**Table 3**). No additional changes were found in either the promoter region or the *acs* gene for any of the isolates with the caveat that a ~50 base pair segment of the *acs* sequence of JA122, CV115 and CV116 (between nucleotides 391 and 441) was not sequenced due to a technical failure. However, it is unlikely that this region contains a mutation as the sequence for CV103 is identical to that of the reference strain MG1655.

Thus, while the ancestral sequence of the *acs* promoter region was accurately determined by Treves et al. (1998), at the time of publication, JA122 was considered the “wild-type” condition when in fact the opposite is true: the A at position -93 is conserved across the *E. coli* clade of the *Enterobacteriaceae*. This phylogenetically-related group contains genera (*Citrobacter*, *Shigella*, *Salmonella* and *Escherichia*) that live almost exclusively in the gastrointestinal tract of warm-blooded mammals, an environment in which extracellular acetate is an important source of carbon (Wolfe 2005). The base-pair in question lies in the first of two CRP binding sites for the

proximal *acs* promoter P2. This CRP binding site is required for induction of acetyl CoA synthetase; directed point mutations in this region cause a 40-80% decrease in transcription (Beatty, Browning et al. 2003). In the absence of a constitutive mutation such as the IS-element insertion in CV101, transcriptional control of this locus is thought to occur primarily via induction. This induction is sensitive to the level of cAMP in the cell, i.e. higher cAMP concentrations (in conjunction with CRP) appear to stimulate *acs* expression (Kumari, Beatty et al. 2000). At the 0% FDR threshold cutoff, the average level of CRP transcript compared to the ancestor is slightly lower in CV103 versus the other isolates (Figure 3B). Taken together, these data strongly suggest that the ancestor, as well as CV103, CV115 and CV116 exhibit *less* than wild-type expression of *acs* and that the induction of this operon in CV103 may be inhibited by higher glucose consumption and/or lower levels of CRP. Restoring base-pair -93 to the wild-type state is all that is required to generate an acetate scavenging strain that can stably co-exist with CV103 (Treves, Manning et al. 1998).

E. coli will excrete acetate via the phosphotransacetylase/acetate kinase pathway if the glycolytic flux is such that not all of the acetyl CoA generated can be efficiently utilized by the TCA cycle. Given that CV103 scavenges more glucose and accumulates more acetate in batch and chemostat monoculture than either its ancestor or CV116, and given that the kinetics of acetate kinase are comparable between all of the isolates in chemostat culture (Rosenzweig, Sharp et al. 1994), we sequenced the gene for phosphotransacetylase (*pta*). However, no mutations were found among any of the strains with the caveat that we were not able to capture the first 17 bp of the gene.

Glycerol and Glycerol-3-phosphate metabolism – Rosenzweig et al. (1994)

presented enzyme kinetic data suggesting differential metabolism of glycerol by CV116 relative to CV101 and CV103. Glycerol amendment of the media used to feed the evolved consortium altered clone frequencies as predicted by these data. We speculated that a mutation in glycerol kinase (*glpK*) could explain these observations. Sequencing of *glpK* did uncover a single point-mutation in CV116; however, as this was a silent substitution (glycine → glycine, amino acid 225) we cannot argue that the mutation has adaptive significance (Table 3). Mutations in the glycerol-3-phosphate regulon repressor, GlpR, could also account for enhanced glycerol metabolism by CV116. Sequencing of this gene revealed a glycine → alanine substitution at amino acid 55 in all of the isolates, including the ancestor. After re-examining the ancestry of JA122, we found that the *glpR* mutation could be traced back to its progenitor *E. coli* K12 strain, C600. While this is a fairly modest mutation, it occurs at a highly conserved position and has been previously reported to result in constitutive expression of genes involved in glycerol utilization (Koch, Hayashi et al. 1964; Elvin, Hardy et al. 1985; Holtman, Thurlkill et al. 2001). Despite this mutation, the regulon is still subject to glucose-mediated catabolite repression at the transcriptional level, as well as post-translational inhibition by IIA^{glc} (*crr*, downregulated in all evolved strains) and fructose-1,6-bisphosphate (Koch, Hayashi et al. 1964; Holtman, Pawlyk et al. 2001). In regard to the behavior of our isolates, the relative activity levels of glycerol kinase (*glpK*) and glycerol-3-phosphate dehydrogenase (*glpD*) are lower in CV103 relative to the other strains (Rosenzweig, Sharp et al. 1994). However, no significant differences in *glpK*, *F* (the glycerol facilitator) or *D* expression were detected between the parent and evolved strains on our arrays, as would be expected if the operon is constitutively active across all isolates.

Conversely, *glpT* (the glycerol-3-phosphate transporter), *glp Q*, and *glpA* (a subunit of anaerobic glycerol-3-phosphate dehydrogenase) are all significantly upregulated in the 1-class community analyses as well as at the 0%-FDR level in monoculture. Considering that the ancestor already has constitutive expression of *glpT*, this increase is surprising but entirely consistent with the observation that either glycerol or glycerol-3-phosphate cross-feeding maintains the CV103/CV116 equilibrium (Rosenzweig, Sharp et al. 1994). While both sets of genes are positively regulated by CRP and repressed by GlpR, *glpTQ* and *A* are additionally regulated by Fis, FlhDC and Fnr. The status of Fis and Fnr cannot be reliably deduced from our array data, but judging by the aforementioned up-regulation of the flagellar genes in CV101, CV115 and CV116, FlhDC, the flagellar “master-switch” would appear to be active in these strains and not in CV103. Regulation of glycerol utilization genes appears to be complex; our results suggest that the natural state of JA122 and all of its descendants is one in which the glycerol regulon is constitutively transcribed, but increased glucose consumption in CV103 mitigates enzyme activity post-transcriptionally.

Global regulators of carbon metabolism – Considering the large number of coordinately transcribed genes whose expression levels differed significantly in the evolved strains relative to their ancestor, we strongly suspected alterations in global regulatory pathways had occurred during the course of the evolution experiment. As 39% of the down-regulated genes in the 1-class analysis are part of the σ^S regulon, and mutations in *rpoS* have been repeatedly observed in glucose-limited chemostat cultures, we sequenced this gene (Notley-McRobb, King et al. 2002; King, Ishihama et al. 2004). We found that all evolved isolates shared a C→T transition at nucleotide 97 that resulted in an amino acid 33 Q→amber mutation, suggesting this

mutation arose early in the experiment. Given the severe nature of the resulting truncation it is likely that this mutation negatively affects σ^S activity. Interestingly, the *rpoSAm* mutation at this position has been observed in a number of other *E. coli* strains (Atlung, Nielsen et al. 2002). Further, our experimental strains carry the supE44 amber suppressor. In suppressor-free strains that carry the *rpoSAm* mutation, translation of a truncated $\Delta 1-53$ σ^S can proceed from a downstream secondary translation initiation region (Subbarayan and Sarkar 2004). This shortened RpoS, while not fully functional, retains partial activity and appears to have a preference for supercoiled promoters (Rajkumari and Gowrishankar 2002; Gowrishankar, Yamamoto et al. 2003). To test whether or not our evolved isolates retained any σ^S activity, we performed catalase and glycogen staining assays as described elsewhere. Compared to their common ancestor, all four of the evolved strains showed reduced catalase activity (weak bubbling after more than 5 seconds) as well as an impaired ability to accumulate glycogen (little to no staining with iodine) indicating that σ^S activity was indeed diminished. These observations are in concordance with our expression profiling results in general and with the reduced expression of *katE* (catalase HP11) in particular (see Figure 2).

Many genes from our expression analyses are also known to be regulated by cAMP-CRP. However, we observed no mutations either in the promoter regions or structural genes for CRP or adenylate cyclase.

Discussion

We have used a combination of microarray-based comparative genome hybridization, transcriptional profiling, and gene-specific sequencing to identify the genetic bases that support the evolution and persistence of a single-species consortium in a

spatially-unstructured, chemostat environment (Helling, Vargas et al. 1987; Kurlandzka, Rosenzweig et al. 1991; Rosenzweig, Sharp et al. 1994; Treves, Manning et al. 1998). Previous communications have shown that coexistence arises from cross-feeding interactions in which the limiting resource, glucose, is incompletely metabolized by the dominant clone, leaving residual metabolites in the media which support growth by other clones. Protein profiling of evolved clones in monoculture indicates that global regulatory mutations were at least partly responsible for adaptive phenotypes (Kurlandzka, Rosenzweig et al. 1991). The genetic basis of CV101's ability to scavenge acetate has been identified as a regulatory mutation that specifically affects the transcription of the acetyl Co-A synthetase operon (Treves, Manning et al. 1998). Still unknown, however, are the genetic mechanisms that explain why all adaptive clones are significantly better at assimilating glucose than their common ancestor, why the dominant clone, CV103, does not re-assimilate residual metabolites, as well as how CV103 and CV116 can stably coexist.

Adaptation to glucose limitation-strategies and mutations shared by all evolved clones

Our results show that all four evolved clones share a common regulatory response to long-term glucose limitation: In general, genes involved in the phosphotransferase system, glycolysis, the pentose-phosphate pathway and mixed acid fermentation are down-regulated whereas TCA cycle genes are up-regulated (Supplementary Figure 2). Strikingly similar changes in central metabolic gene expression have been reported for *E.coli* in batch culture as well as Baker's yeast following adaptive evolution in long-term glucose-limited chemostat culture (Ferea, Botstein et al. 1999; Jansen, Diderich et al. 2005; Le Gac, Brazas et al. 2008). The

repeatability of this phenomenon across replicate experiments within the same species as well as across Domains strongly argues that microbes may have limited options for increasing fitness in glucose-limited environments. Our 1-class microarray analysis and sequencing results indicate that the changes in levels of the stationary-phase sigma factor, σ^S , could account for many of the genes significantly down-regulated in all strains. Recent microarray analysis of the effects of an *rpoS* knockout on metabolism in rich medium batch culture clearly shows that all central metabolic pathways (including the phosphotransferase system and the TCA cycle) were down-regulated during early stationary phase relative to wild type (Rahman, Hasan et al. 2006). Our array data are broadly consistent with this result with the notable exception of the TCA cycle, which showed higher relative expression in our strains. However, transcriptional profiling of the same *rpoS* knockout during exponential phase shows the TCA cycle very strongly up-regulated (Rahman, Hasan et al. 2006). Under continuous nutrient limitation *E. coli* populations achieve a steady-state that approximates late exponential/early stationary phase growth in batch culture. It is tempting to speculate that the pattern of expression we observe for genes in central metabolism is what might be observed if an *rpoS* knockout were grown under our experimental conditions. Alternatively, upregulation of TCA cycle genes in our strains may be the result of altered activity exhibited by incomplete suppression of the *rpoSAm* mutation, translation of truncated σ^S , or simply be the effect of yet-to-be identified regulatory mutation(s). While it may seem that reduced expression of glycolytic genes should be disadvantageous under glucose limitation, minimizing the concentration of catabolic enzymes needed is likely to be energetically favorable provided the concentration does not fall below the minimum necessary to metabolize available substrate.

In addition to shared global expression patterns for central metabolic genes, our microarray results show that evolved isolates also up-regulate genes involved in movement of glucose across the outer and inner membranes. Increased transcription of the inner membrane Mgl galactose ABC-transporter (which also transports glucose) appears to be a common response to continuous glucose limitation, and in this regard our experimental system is no exception. This regulatory adjustment is easily accounted for by a mutation present in all of the evolved isolates in the *mgl* operator sequence, a mutation that presumably interferes with GalS-mediated suppression of *mgl* transcription (Notley-McRobb and Ferenci 1999; Notley-McRobb, Seeto et al. 2003). Similarly, increased movement of glucose into the periplasm in the evolved isolates is undoubtedly due in part to the overexpression of the LamB glycoporin, another hallmark feature of *E. coli* adaptation to glucose limitation (Notley-McRobb and Ferenci 1999; Hua, Yang et al. 2004).

In Ferenci and colleagues' experiments, adaptive overexpression of LamB (which is part of the *mal* regulon) results from mutations in the *mal* repressor Mlc and/or its activator MalT (Notley-McRobb and Ferenci 1999; Notley-McRobb, Seeto et al. 2003). Sequencing of *mlc* and its associated regulatory region failed to identify mutations in any of our evolved clones. We did find a mutation in the gene encoding MalT, but its distribution was limited to CV101, CV115 and CV116 and its location was unique relative to other mutations in MalT that have been characterized as constitutive. It is surprising that this mutation does not occur in CV103 considering that on average, CV103 has 3-6 fold higher transcript levels of *lamB* and other MalT responsive genes than the other three strains (significant in a between-subjects t-test, $p=0.0007$). The superior glucose scavenging ability of CV103 may be attributable to the fact that it produces incrementally more LamB than its cohabitants, but this

increase can be sufficiently explained either by inactivation of Mlc or by constitutive mutation in MalT, as neither occurs. Additional mutations (such as those that affect ptsG) or transcriptional differences (such as an increase in OmpF) that could contribute to enhanced glucose uptake and have been observed in other chemostat experiments could not be found (Maharjan, Seeto et al. 2006). While increased LamB expression is likely due, in part, to defective *rpoS*, this mutation is shared by all evolved isolates and therefore cannot account for the differences observed between strains (Notley-McRobb, King et al. 2002). An alternative explanation may lie in the structure of MalT from CV101, CV115 and CV116. Based on the distribution of mutations in *rpoS*, *gals*, *acs* and *glpK*, it is highly probable that the *malT* mutation occurred in the common ancestor of CV101, CV115 and CV116 prior to specialization of CV101 on acetate but after the divergence of CV103. If this mutation results in a constitutive activator (and therefore increased LamB expression) in CV101, CV115 and CV116, the still higher transcript levels of LamB and superior glucose uptake of CV103 remain unexplained. If the shared increase in LamB expression is primarily due to a different mutation that is present in all of the isolates (such as the defect in σ^S), then the adaptive significance of the *malT* mutation in CV101, CV115 and CV116 is unclear. All other reported mutations in the N-terminal portion of MalT that occur during glucose-limited chemostat culture result in constitutive expression of the MalT protein (Notley-McRobb and Ferenci 1999). However, despite the relatively large number of these that have been characterized (at least 16), none is in the same position or motif as the one we report here (Notley-McRobb and Ferenci 1999; Leipe, Koonin et al. 2004).

Strangely, adaptation to long-term glucose limitation in a batch culture frequently selects for mutations that partially or fully inactivate MalT, one of which

does occur in the same helix as our mutation (Pelosi, Kuhn et al. 2006). If our mutation results in a weakened activator (and consequently less LamB) there exists the intriguing (although speculative) possibility that in this particular environment, surrendering a portion of the limiting nutrient in order to acquire more overflow metabolite could provide an advantage to minority clones that specialize on excess excreted carbon. Clearly, additional experiments will be required to establish the precise effect of the *malt* mutation in CV101, CV115 and CV116. Alternatively, CV103 may simply have higher levels of endogenous maltotriose inducer or may harbor as yet unidentified changes that affect *lamB* transcription and/or glucose uptake via another route. In either case, the mechanism (whether physiological or genetic) promises to be a unique and interesting one and will be the subject of future investigations.

The evolution of cross-feeding between CV101 and CV103

The constitutive overexpression of acetyl-CoA synthetase that enables CV101 to capture overflow acetate from the dominant clone has a clearly documented mutational basis which has been re-confirmed by our microarray and sequencing results. What has been heretofore unresolved was why CV103 is unable to efficiently recover its own acetate. This characteristic is measurable both as high equilibrium acetate concentration in chemostat culture as well as lack of detectable acetyl CoA synthetase activity under low-glucose batch cultivation. Given that the ancestor, JA122, has a weakened *acs* promoter and acetate scavenging at low concentration almost exclusively occurs via the *acs* pathway, CV103's inability to recover acetate when consuming glucose could be explained by this genetic predisposition compounded by increased catabolite repression of *acs* as a consequence of increased

glucose transport. The rate of glucose uptake, equilibrium acetate concentration and acetyl CoA synthetase measurements of CV116 under glucose limitation support this contention: all are intermediate between JA122 and CV103. However, in the presence of acetate and glycerol, *acs* activity in CV103 is negligible while CV116 exhibits ancestral levels, indicating that in CV103 *acs* is neither appropriately activated nor repressed in this isolate. The regulation of *acs* expression is quite complex and at the very least appears to involve the integration of signals from the TCA cycle, glyoxylate bypass and phosphotransacetylase/acetate kinase (*pta/ackA*) acetate dissimilation pathway (Kumari, Beatty et al. 2000; Wolfe 2005; Veit, Polen et al. 2007). Changes in the regulation or structure of acetate kinase were previously ruled out based on enzyme activity and K_m measurements (Rosenzweig, Sharp et al. 1994). In the present study, we sequenced the promoter and full structural gene for *acs* as well as the other enzyme in the dissimilation pathway, *pta*. With the exception of the first 17 base-pairs of *pta* (which were not sequenced) we found no mutations. Thus, the genetic basis for the loss of *acs* activity in CV103 remains obscure.

The evolution of cross-feeding between CV101 and CV103

Increased glycerol uptake coupled with the observation that addition of glycerol increases the equilibrium frequency of CV116 co-cultured with CV103 led to the conclusion that CV116's success in the chemostat was due, at least in part, to glycerol cross-feeding (Rosenzweig, Sharp et al. 1994). Sequencing of the glycerol kinase gene (the rate limiting step in the metabolism of extracellular glycerol) identified a mutation in CV116 that was not present in the other isolates. However, given that this was a silent substitution resulting in a codon change from an abundant to a rare tRNA, and given that the surrounding sequence bears little homology to a

glycerol repressor (*glpR*) binding site, it is difficult to argue that that this mutation has adaptive significance. We therefore next targeted *glpR* and were surprised to find a mutation that was not only present in the ancestor but was present in the *E. coli* progenitor strain from which JA122 was derived. This mutation has been characterized and results in constitutive expression of the glycerol regulon. Many *glpR*-regulated genes did not show appreciable expression differences on our microarrays, as would be expected if they were also upregulated in the ancestor. However, three genes that did show increased transcript level *only* in CV101, CV115 and CV116 were the glycerol-3-phosphate transporter (*glpT*), the glycerophosphoryl diester phosphodiesterase (*glpQ*) and the anaerobic glycerol-3-phosphate dehydrogenase (*glpA*). These genes are also controlled by *GlpR*, but they have additional regulators not shared by other genes in the regulon. Based on these observations, it appears likely that CV116 is able to recover and metabolize extracellular glycerol-3-phosphate better than CV103 and JA122 by upregulating the expression of *glpT*. CV101, though it shares the increased expression of these genes, may not be able to effectively transport glycerol due to molecular feedback arising from *acs* overexpression.

The consortium expression profile does not recapitulate monoculture profiles.

Transcriptional profiling of the consortium RNA pool led to the unexpected observation that in monoculture, CV103 has a different pattern of gene expression than it has when co-cultured with CV101 and CV116. The genes primarily affected are those that distinguish CV103 from the other clones in the 4-class SAM analysis, suggesting that a global regulatory mechanism is responsible for the shift in expression. Two global regulators dominate the 4-class SAM analysis: CRP and

CpxR together explain expression patterns for nearly half the transcription units which distinguish CV103. CRP is known or predicted to influence the expression of 24% of T.U.s, though none of these are under the exclusive control of CRP. While CpxR controls a smaller proportion of CV103-specific T.U.s, (19%), most of these are *solely* regulated by CpxR. Thus, CpxR regulation underlies much of CV103's expression pattern in monoculture; this effect is reversed when CV103 is co-cultured with the subdominant clones.

One dramatic environmental difference between the glucose-limited CV103 monoculture environment and the consortium environment is the concentration of extracellular acetate. When CV101 is present acetate is scavenged and cannot accumulate. CpxR in its phosphorylated form mediates a global response to extracytoplasmic stressors such as high osmolarity, misfolded outer membrane protein or alkaline pH (as reviewed in (Ruiz and Silhavy 2005) but there have been no reports of a direct connection between extracellular acetate concentration and CpxR activation. However, CpxR can be activated by acetyl-P, the high-energy intermediate of the *pta/ackA* pathway that accumulates during exponential phase growth on glucose (Fredericks, Shibata et al. 2006; Klein, Shulla et al. 2007; Keating, Shulla et al. 2008). Recent epistatic analysis has suggested that CpxR phosphorylation might be inhibited by an unidentified signal that is dependent upon normal function of the Pta-AckA pathway ("substance Y") (Wolfe, Parikh et al. 2008). Regardless of the precise molecular nature of the interaction, it seems clear that CpxR activation is intimately connected to acetate dissimilation. We previously reported that the K_m for acetate kinase in CV103 and CV116 was lower than that of JA122 and CV101 (Rosenzweig, Sharp et al. 1994). Given the low equilibrium acetate concentration in the chemostat, it was concluded that this decrease in K_m

should not significantly affect acetate uptake or secretion. However, alterations in acetate kinase activity, increased acetate secretion or reduced acetate uptake could conceivably affect the overall performance of the Pta-AckA pathway and thus influence intracellular levels of acetyl-P and/or "substance Y." Such interactions could be reasonably postulated to elicit a CpxR-mediated transcriptional response when extracellular acetate concentrations increase (as in CV103 monoculture).

Ancestral genotype constrains possible evolutionary trajectories

Shared mutations in *rpoS* and *mglD* strongly support the hypothesis that competition for the limiting nutrient, glucose, was the primary selective force operating in the chemostat prior to metabolic divergence of CV101 and CV116 (Rosenzweig, Sharp et al. 1994). Increased glucose consumption coupled with acetate and glycerol secretion by CV103 created a favorable environment for the evolution of clones that could efficiently consume these two overflow metabolites. While screening for mutations that contributed to the emergence of cross-feeding populations, we unexpectedly encountered ancestral regulatory mutations in both the acetate and glycerol metabolic pathways that affect the induction of acetyl CoA synthetase (the primary acetate scavenging pathway) and the glycerol regulon repressor GlpR. As a result, it appears that the ancestor is unable to efficiently recover excreted acetate and constitutively overexpresses the glycerol regulon. We believe that these two mutations in the ancestor profoundly influenced the evolutionary outcome of these experiments (as well as the replicate evolution experiments reported in Treves et al. 1998, which showed similar qualitative results). Impaired acetate scavenging by the progenitor of CV103 undoubtedly accelerated or predisposed the evolution of a strain that could efficiently utilize this substrate. We

cannot argue that acetate scavenging clones would not have eventually arisen from a purely "wild-type" inoculum, but the repeatability of their emergence as well as the precise way in which they were invariably generated (activation of *acs* by reversion of the ancestral mutation or IS element insertion) suggests that there was strong selective pressure for changes at the *acs* locus. The influence of the ancestral GlpR mutation is less clear: overexpression of the glycerol dissimilation pathway could affect the excretion of glycerol-3-phosphate by CV103 or enhance the ability CV116 to recover it. In either case it seems unlikely that the presence of the GlpR mutation is mere coincidence.

The founder effect is generally disregarded in microbial evolution experiments because immense population sizes enable a pool of variants to be rapidly generated by mutation and also buffer against severe genetic bottlenecks. The results presented here suggest that experimental evolution studies *are* influenced by the founding genotype and such constraints can underlie evolution of stable polymorphisms. At least one mutation instrumental in the evolution and maintenance of cross-feeding was compensatory rather than neomorphic. Thus, the exploration of new biochemical opportunities required recovery of old functions in addition to the development of novel traits. These observations may not be confined strictly to experimental systems as many natural microbial populations (such as those that cause nosocomial or chronic infections) are also founded by clones.

Transcriptional profiling and targeted gene sequencing expanded and confirmed certain aspects of our understanding of the mechanisms that drive adaptation and diversification. All identified nonsynonymous mutations were regulatory in nature, but not strictly confined to global regulators. Initial selection in the chemostat favored mutations that enhance competitive acquisition of the limiting

resource (such as those in *rpoS* and *mgl*), but ancestral regulatory mutations like those in *acs* and perhaps *glpR* explain much of the unique behavior of this system. The transcriptional effect of some adaptations were apparent even when consortium members were grown in isolation, while the expression of others appeared to be depend on the metabolic activity of sibling clones. Finally, even under strong selection, at least one of the most beneficial mutations served to restore lost function thereby creating a stable cross-feeding interaction.

Conclusion/Summary

The advantages of *E. coli* as a model organism for experimental evolution lie in its ease of cultivation, large population sizes, rich history of investigation and perceived simplicity of adaptive response. An attempt to understand the process of adaptation of *E. coli* to a single environmental stressor led to the unexpected discovery that biological diversity can evolve and endure even under the simplest of conditions.

Out of necessity, previous efforts to characterize the nature of our microbiological consortium relied upon the assumption that the sum of the individual units was mechanistically equal to the behavior of the whole. And indeed, detailed analysis of each member in isolation provided useful information about both their shared evolutionary history and individual adaptive strategies. However, treating the intact consortium as a single unit revealed a transcriptomic behavior that was clearly different from a simple aggregation of the "atomized" parts (sensu Gould and Lewontin, 1979). Future experiments which rely on advances in genome sequencing, cell labeling and sorting will enable us to dissect the consortium into its individual components prior to analysis, and precisely identify the characteristics that define

each clone's adaptive strategy. The challenge of deconvoluting individual metabolic responses in this system underscores the complexity of even a simple three-membered "community." And our finding that that community's sum does not strictly equal its parts makes clear that experimental microbial evolution is a powerful tool to study the evolution of emergent properties in complex biological systems.

Acknowledgements

This work was supported by the University of Montana Office of Sponsored Programs, and subcontract E4406-117101 of NIH grant GM63800-01 to FR. The authors gratefully acknowledge the critical commentary of Evgueny Kroll, Steve Lodmell, Scott Miller, Mark Pershouse, Alan Wolfe and three anonymous reviewers.

Literature Cited

- Adams, J., T. Kinney, et al. (1979). "Frequency-Dependent Selection for Plasmid-Containing Cells of *Escherichia coli*." Genetics **91**(4): 627-637.
- Agafonov, D. E., V. A. Kolb, et al. (2001). "Ribosome-associated protein that inhibits translation at the aminoacyl-tRNA binding stage." EMBO Rep **2**(5): 399-402.
- Atlung, T., H. V. Nielsen, et al. (2002). "Characterisation of the allelic variation in the rpoS gene in thirteen K12 and six other non-pathogenic *Escherichia coli* strains." Mol Genet Genomics **266**(5): 873-81.
- Atwood, K. C., L. K. Schneider, et al. (1951). "Periodic Selection in *Escherichia coli*." Proceedings of the National Academy of Sciences of the United States of America **37**(3): 146-155.
- Baer, C. F. and M. Lynch (2003). "Correlated evolution of life-history with size at maturity in *Daphnia pulex*: patterns within and between populations." Genet Res **81**(2): 123-32.
- Bantinaki, E., R. Kassen, et al. (2007). "Adaptive divergence in experimental populations of *Pseudomonas fluorescens*. III. Mutational origins of wrinkly spreader diversity." Genetics **176**(1): 441-53.
- Beatty, C. M., D. F. Browning, et al. (2003). "Cyclic AMP receptor protein-dependent activation of the *Escherichia coli* acsP2 promoter by a synergistic class III mechanism." J Bacteriol **185**(17): 5148-57.

- Chang, L., L. I. Wei, et al. (1999). "Expression of the Escherichia coli NRZ nitrate reductase is highly growth phase dependent and is controlled by RpoS, the alternative vegetative sigma factor." *Mol Microbiol* **34**(4): 756-66.
- Crabbe, J. C., A. Kosobud, et al. (1985). "Bidirectional selection for susceptibility to ethanol withdrawal seizures in *Mus musculus*." *Behav Genet* **15**(6): 521-36.
- Danese, P. N. and T. J. Silhavy (1998). "CpxP, a stress-combative member of the Cpx regulon." *J Bacteriol* **180**(4): 831-9.
- Dardonville, B. and O. Raibaud (1990). "Characterization of malT mutants that constitutively activate the maltose regulon of Escherichia coli." *J Bacteriol* **172**(4): 1846-52.
- Denver, D. R., K. Morris, et al. (2005). "The transcriptional consequences of mutation and natural selection in *Caenorhabditis elegans*." *Nat Genet* **37**(5): 544-8.
- Dobzhansky, T. (1947). "Adaptive Changes Induced by Natural Selection in Wild Populations of *Drosophila*." *Evolution* **1**(1/2): 1-16.
- Dobzhansky, T. and B. Spassky (1947). "Evolutionary Changes in Laboratory Cultures of *Drosophila pseudoobscura*." *Evolution* **1**(3): 191-216.
- Elena, S. F. and R. E. Lenski (2003). "Evolution experiments with microorganisms: the dynamics and genetic bases of adaptation." *Nat Rev Genet* **4**(6): 457-69.
- Elvin, C. M., C. M. Hardy, et al. (1985). "Pi exchange mediated by the GlpT-dependent sn-glycerol-3-phosphate transport system in Escherichia coli." *J Bacteriol* **161**(3): 1054-8.
- Ferea, T. L., D. Botstein, et al. (1999). "Systematic changes in gene expression patterns following adaptive evolution in yeast." *Proc Natl Acad Sci U S A* **96**(17): 9721-6.
- Ferenci, T. (2001). "Hungry bacteria--definition and properties of a nutritional state." *Environ Microbiol* **3**(10): 605-11.
- Ferenci, T. (2003). "What is driving the acquisition of mutS and rpoS polymorphisms in Escherichia coli?" *Trends Microbiol* **11**(10): 457-61.
- Franchini, A. G. and T. Egli (2006). "Global gene expression in Escherichia coli K-12 during short-term and long-term adaptation to glucose-limited continuous culture conditions." *Microbiology* **152**(Pt 7): 2111-27.
- Fredericks, C. E., S. Shibata, et al. (2006). "Acetyl phosphate-sensitive regulation of flagellar biogenesis and capsular biosynthesis depends on the Rcs phosphorelay." *Mol Microbiol* **61**(3): 734-47.
- Friesen, M. L., G. Saxer, et al. (2004). "Experimental evidence for sympatric ecological diversification due to frequency-dependent competition in Escherichia coli." *Evolution* **58**(2): 245-60.
- Gause, G. F. (1934). *The Struggle for Existence*. New York, Dover.
- Gefen, E., A. J. Marlon, et al. (2006). "Selection for desiccation resistance in adult *Drosophila melanogaster* affects larval development and metabolite accumulation." *J Exp Biol* **209**(Pt 17): 3293-300.
- Gibson, K. E. and T. J. Silhavy (1999). "The LysR homolog LrhA promotes RpoS degradation by modulating activity of the response regulator sprE." *J Bacteriol* **181**(2): 563-71.
- Gonzalez, A. D., V. Espinosa, et al. (2005). "TRACTOR_DB: a database of regulatory networks in gamma-proteobacterial genomes." *Nucleic Acids Res* **33**(Database issue): D98-102.
- Gowrishankar, J., K. Yamamoto, et al. (2003). "In vitro properties of RpoS (sigma(S)) mutants of Escherichia coli with postulated N-terminal subregion 1.1 or C-terminal region 4 deleted." *J Bacteriol* **185**(8): 2673-9.

- Hardin, G. (1960). "The competitive exclusion principle." *Science* **131**: 1292-7.
- Helling, R. B., T. Kinney, et al. (1981). "The maintenance of Plasmid-containing organisms in populations of *Escherichia coli*." *J Gen Microbiol* **123**(1): 129-41.
- Helling, R. B., C. N. Vargas, et al. (1987). "Evolution of *Escherichia coli* during growth in a constant environment." *Genetics* **116**(3): 349-58.
- Holtman, C. K., A. C. Pawlyk, et al. (2001). "Reverse genetics of *Escherichia coli* glycerol kinase allosteric regulation and glucose control of glycerol utilization in vivo." *J Bacteriol* **183**(11): 3336-44.
- Holtman, C. K., R. Thurlkill, et al. (2001). "Unexpected presence of defective *glpR* alleles in various strains of *Escherichia coli*." *J Bacteriol* **183**(4): 1459-61.
- Hua, Q., C. Yang, et al. (2004). "Analysis of gene expression in *Escherichia coli* in response to changes of growth-limiting nutrient in chemostat cultures." *Appl Environ Microbiol* **70**(4): 2354-66.
- Jansen, M. L., J. A. Diderich, et al. (2005). "Prolonged selection in aerobic, glucose-limited chemostat cultures of *Saccharomyces cerevisiae* causes a partial loss of glycolytic capacity." *Microbiology* **151**(Pt 5): 1657-69.
- Kandror, O., A. DeLeon, et al. (2002). "Trehalose synthesis is induced upon exposure of *Escherichia coli* to cold and is essential for viability at low temperatures." *Proc Natl Acad Sci U S A* **99**(15): 9727-32.
- Karp, P. D., I. M. Keseler, et al. (2007). "Multidimensional annotation of the *Escherichia coli* K-12 genome." *Nucleic Acids Res* **35**(22): 7577-90.
- Keating, D. H., A. Shulla, et al. (2008). "Optimized two-dimensional thin layer chromatography to monitor the intracellular concentration of acetyl phosphate and other small phosphorylated molecules." *Biol Proced Online* **10**: 36-46.
- King, T., A. Ishihama, et al. (2004). "A regulatory trade-off as a source of strain variation in the species *Escherichia coli*." *J Bacteriol* **186**(17): 5614-20.
- Klein, A. H., A. Shulla, et al. (2007). "The intracellular concentration of acetyl phosphate in *Escherichia coli* is sufficient for direct phosphorylation of two-component response regulators." *J Bacteriol* **189**(15): 5574-81.
- Koch, J. P., S. Hayashi, et al. (1964). "The Control of Dissimilation of Glycerol and L-Alpha-Glycerophosphate in *Escherichia Coli*." *J Biol Chem* **239**: 3106-8.
- Kozmin, S. G., Y. I. Pavlov, et al. (2000). "Hypersensitivity of *Escherichia coli* Delta(*uvrB*-bio) mutants to 6-hydroxylaminopurine and other base analogs is due to a defect in molybdenum cofactor biosynthesis." *J Bacteriol* **182**(12): 3361-7.
- Kumari, S., C. M. Beatty, et al. (2000). "Regulation of acetyl coenzyme A synthetase in *Escherichia coli*." *J Bacteriol* **182**(15): 4173-9.
- Kurlandzka, A., R. F. Rosenzweig, et al. (1991). "Identification of adaptive changes in an evolving population of *Escherichia coli*: the role of changes with regulatory and highly pleiotropic effects." *Mol Biol Evol* **8**(3): 261-81.
- Larkin, J. E., B. C. Frank, et al. (2005). "Independence and reproducibility across microarray platforms." *Nat Methods* **2**(5): 337-44.
- Le Gac, M., M. D. Brazas, et al. (2008). "Metabolic changes associated with adaptive diversification in *Escherichia coli*." *Genetics* **178**(2): 1049-60.
- Lehnen, D., C. Blumer, et al. (2002). "LrhA as a new transcriptional key regulator of flagella, motility and chemotaxis genes in *Escherichia coli*." *Mol Microbiol* **45**(2): 521-32.
- Leipe, D. D., E. V. Koonin, et al. (2004). "STAND, a class of P-loop NTPases including animal and plant regulators of programmed cell death: multiple,

- complex domain architectures, unusual phyletic patterns, and evolution by horizontal gene transfer." *J Mol Biol* **343**(1): 1-28.
- Liu, X. and T. Ferenci (1998). "Regulation of porin-mediated outer membrane permeability by nutrient limitation in *Escherichia coli*." *J Bacteriol* **180**(15): 3917-22.
- Liu, X. and T. Ferenci (2001). "An analysis of multifactorial influences on the transcriptional control of *ompF* and *ompC* porin expression under nutrient limitation." *Microbiology* **147**(Pt 11): 2981-9.
- Liu, X. and P. Matsumura (1994). "The FlhD/FlhC complex, a transcriptional activator of the *Escherichia coli* flagellar class II operons." *J Bacteriol* **176**(23): 7345-51.
- Maharjan, R., S. Seeto, et al. (2006). "Clonal adaptive radiation in a constant environment." *Science* **313**(5786): 514-7.
- Maharjan, R. P., S. Seeto, et al. (2007). "Divergence and redundancy of transport and metabolic rate-yield strategies in a single *Escherichia coli* population." *J Bacteriol* **189**(6): 2350-8.
- Manche, K., L. Notley-McRobb, et al. (1999). "Mutational adaptation of *Escherichia coli* to glucose limitation involves distinct evolutionary pathways in aerobic and oxygen-limited environments." *Genetics* **153**(1): 5-12.
- Minagawa, S., H. Ogasawara, et al. (2003). "Identification and molecular characterization of the Mg²⁺ stimulon of *Escherichia coli*." *J Bacteriol* **185**(13): 3696-702.
- Morita, T., H. Kawamoto, et al. (2004). "Enolase in the RNA degradosome plays a crucial role in the rapid decay of glucose transporter mRNA in the response to phosphosugar stress in *Escherichia coli*." *Mol Microbiol* **54**(4): 1063-75.
- Muller, H. J. (1932). "Some Genetic Aspects of Sex." *The American Naturalist* **66**(703): 118-138.
- Notley-McRobb, L. and T. Ferenci (1999). "Adaptive *mgl*-regulatory mutations and genetic diversity evolving in glucose-limited *Escherichia coli* populations." *Environ Microbiol* **1**(1): 33-43.
- Notley-McRobb, L. and T. Ferenci (1999). "The generation of multiple co-existing *mgl*-regulatory mutations through polygenic evolution in glucose-limited populations of *Escherichia coli*." *Environ Microbiol* **1**(1): 45-52.
- Notley-McRobb, L., T. King, et al. (2002). "*rpoS* mutations and loss of general stress resistance in *Escherichia coli* populations as a consequence of conflict between competing stress responses." *J Bacteriol* **184**(3): 806-11.
- Notley-McRobb, L., S. Seeto, et al. (2003). "The influence of cellular physiology on the initiation of mutational pathways in *Escherichia coli* populations." *Proc Biol Sci* **270**(1517): 843-8.
- Novick, A. and L. Szilard (1950). "Experiments with the Chemostat on Spontaneous Mutations of Bacteria." *Proceedings of the National Academy of Sciences of the United States of America* **36**(12): 708-719.
- Oh, M. K., L. Rohlin, et al. (2002). "Global expression profiling of acetate-grown *Escherichia coli*." *J Biol Chem* **277**(15): 13175-83.
- Overduin, P., W. Boos, et al. (1988). "Nucleotide sequence of the *ugp* genes of *Escherichia coli* K-12: homology to the maltose system." *Mol Microbiol* **2**(6): 767-75.
- Pelosi, L., L. Kuhn, et al. (2006). "Parallel changes in global protein profiles during long-term experimental evolution in *Escherichia coli*." *Genetics* **173**(4): 1851-69.

- Perrenoud, A. and U. Sauer (2005). "Impact of global transcriptional regulation by ArcA, ArcB, Cra, Crp, Cya, Fnr, and Mlc on glucose catabolism in *Escherichia coli*." *J Bacteriol* **187**(9): 3171-9.
- Polen, T., D. Rittmann, et al. (2003). "DNA microarray analyses of the long-term adaptive response of *Escherichia coli* to acetate and propionate." *Appl Environ Microbiol* **69**(3): 1759-74.
- Prasad, N. G., M. Shakarad, et al. (2001). "Correlated responses to selection for faster development and early reproduction in *Drosophila*: the evolution of larval traits." *Evolution Int J Org Evolution* **55**(7): 1363-72.
- Pratt, L. A., W. Hsing, et al. (1996). "From acids to osmZ: multiple factors influence synthesis of the OmpF and OmpC porins in *Escherichia coli*." *Mol Microbiol* **20**(5): 911-7.
- Pruss, B. M., X. Liu, et al. (2001). "FlhD/FlhC-regulated promoters analyzed by gene array and lacZ gene fusions." *FEMS Microbiol Lett* **197**(1): 91-7.
- Rahman, M., M. R. Hasan, et al. (2006). "Effect of rpoS gene knockout on the metabolism of *Escherichia coli* during exponential growth phase and early stationary phase based on gene expressions, enzyme activities and intracellular metabolite concentrations." *Biotechnol Bioeng* **94**(3): 585-95.
- Rainey, P. B., A. Buckling, et al. (2000). "The emergence and maintenance of diversity: insights from experimental bacterial populations." *Trends Ecol Evol* **15**(6): 243-247.
- Rainey, P. B. and M. Travisano (1998). "Adaptive radiation in a heterogeneous environment." *Nature* **394**(6688): 69-72.
- Rajkumari, K. and J. Gowrishankar (2002). "An N-terminally truncated RpoS (σ (S)) protein in *Escherichia coli* is active in vivo and exhibits normal environmental regulation even in the absence of rpoS transcriptional and translational control signals." *J Bacteriol* **184**(12): 3167-75.
- Ricker, J. P. and J. Hirsch (1985). "Evolution of an instinct under long-term divergent selection for geotaxis in domesticated populations of *Drosophila melanogaster*." *J Comp Psychol* **99**(4): 380-90.
- Ricker, J. P. and J. Hirsch (1988). "Genetic changes occurring over 500 generations in lines of *Drosophila melanogaster* selected divergently for geotaxis." *Behav Genet* **18**(1): 13-25.
- Ricker, J. P. and J. Hirsch (1988). "Reversal of genetic homeostasis in laboratory populations of *Drosophila melanogaster* under long-term selection for geotaxis and estimates of gene correlates: evolution of behavior-genetic systems." *J Comp Psychol* **102**(3): 203-14.
- Rose, M. R. (1984). "Laboratory Evolution of Postponed Senescence in *Drosophila Melanogaster*." *Evolution* **38**(5): 1004-1010.
- Rosenzweig, R. F., R. R. Sharp, et al. (1994). "Microbial evolution in a simple unstructured environment: genetic differentiation in *Escherichia coli*." *Genetics* **137**(4): 903-17.
- Rozen, D. E. and R. E. Lenski (2000). "Long-Term Experimental Evolution in *Escherichia coli*. VIII. Dynamics of a Balanced Polymorphism." *Am Nat* **155**(1): 24-35.
- Rozen, D. E., N. Philippe, et al. (2009). "Death and cannibalism in a seasonal environment facilitate bacterial coexistence." *Ecol Lett* **12**(1): 34-44.
- Ruiz, N. and T. J. Silhavy (2005). "Sensing external stress: watchdogs of the *Escherichia coli* cell envelope." *Curr Opin Microbiol* **8**(2): 122-6.

- Scamuffa, M. D. and R. M. Caprioli (1980). "Comparison of the mechanisms of two distinct aldolases from *Escherichia coli* grown on gluconeogenic substrates." Biochim Biophys Acta **614**(2): 583-90.
- Schlegel, A., O. Danot, et al. (2002). "The N terminus of the *Escherichia coli* transcription activator MalT is the domain of interaction with MalY." J Bacteriol **184**(11): 3069-77.
- Schneider, D. and R. E. Lenski (2004). "Dynamics of insertion sequence elements during experimental evolution of bacteria." Res Microbiol **155**(5): 319-27.
- Spencer, C. C., M. Bertrand, et al. (2007). "Adaptive diversification in genes that regulate resource use in *Escherichia coli*." PLoS Genet **3**(1): e15.
- Spiess, C., A. Beil, et al. (1999). "A temperature-dependent switch from chaperone to protease in a widely conserved heat shock protein." Cell **97**(3): 339-47.
- Starai, V. J., J. Garrity, et al. (2005). "Acetate excretion during growth of *Salmonella enterica* on ethanolamine requires phosphotransacetylase (EutD) activity, and acetate recapture requires acetyl-CoA synthetase (Acs) and phosphotransacetylase (Pta) activities." Microbiology **151**(Pt 11): 3793-801.
- Stuber, C. W., R. H. Moll, et al. (1980). "Allozyme Frequency Changes Associated with Selection for Increased Grain Yield in Maize (*ZEA MAYS* L.)." Genetics **95**(1): 225-236.
- Subbarayan, P. R. and M. Sarkar (2004). "A stop codon-dependent internal secondary translation initiation region in *Escherichia coli* rpoS." RNA **10**(9): 1359-65.
- Syn, C. K. and S. Swarup (2000). "A scalable protocol for the isolation of large-sized genomic DNA within an hour from several bacteria." Anal Biochem **278**(1): 86-90.
- Taschner, N. P., E. Yagil, et al. (2004). "A differential effect of sigmaS on the expression of the PHO regulon genes of *Escherichia coli*." Microbiology **150**(Pt 9): 2985-92.
- Treves, D. S., S. Manning, et al. (1998). "Repeated evolution of an acetate-crossfeeding polymorphism in long-term populations of *Escherichia coli*." Mol Biol Evol **15**(7): 789-97.
- Turner, P. E., V. Souza, et al. (1996). "Tests of Ecological Mechanisms Promoting the Stable Coexistence of Two Bacterial Genotypes." Ecology **77**(7): 2119-2129.
- Tusher, V. G., R. Tibshirani, et al. (2001). "Significance analysis of microarrays applied to the ionizing radiation response." Proc Natl Acad Sci U S A **98**(9): 5116-21.
- van Oortmerssen, G. A. and T. C. Bakker (1981). "Artificial selection for short and long attack latencies in wild *Mus musculus domesticus*." Behav Genet **11**(2): 115-26.
- VanBogelen, R. A., P. M. Kelley, et al. (1987). "Differential induction of heat shock, SOS, and oxidation stress regulons and accumulation of nucleotides in *Escherichia coli*." J Bacteriol **169**(1): 26-32.
- Veit, A., T. Polen, et al. (2007). "Global gene expression analysis of glucose overflow metabolism in *Escherichia coli* and reduction of aerobic acetate formation." Appl Microbiol Biotechnol **74**(2): 406-21.
- Wanner, B. L. (1992). "Is cross regulation by phosphorylation of two-component response regulator proteins important in bacteria?" J Bacteriol **174**(7): 2053-8.
- Wolfe, A. J. (2005). "The acetate switch." Microbiol Mol Biol Rev **69**(1): 12-50.
- Wolfe, A. J., N. Parikh, et al. (2008). "Signal integration by the two-component signal transduction response regulator CpxR." J Bacteriol **190**(7): 2314-22.

- Wright, S. and T. Dobzhansky (1946). "Genetics of Natural Populations. Xii. Experimental Reproduction of Some of the Changes Caused by Natural Selection in Certain Populations of *Drosophila Pseudoobscura*." Genetics **31**(2): 125-56.
- Zahrl, D., M. Wagner, et al. (2006). "Expression and assembly of a functional type IV secretion system elicit extracytoplasmic and cytoplasmic stress responses in *Escherichia coli*." J Bacteriol **188**(18): 6611-21.
- Zeyl, C. (2006). "Experimental evolution with yeast." FEMS Yeast Res **6**(5): 685-91.
- Zhang, E. and T. Ferenci (1999). "OmpF changes and the complexity of *Escherichia coli* adaptation to prolonged lactose limitation." FEMS Microbiol Lett **176**(2): 395-401.
- Zhao, K., M. Liu, et al. (2007). "Adaptation in bacterial flagellar and motility systems: from regulon members to 'foraging'-like behavior in *E. coli*." Nucleic Acids Res **35**(13): 4441-52.
- Zheng, D., C. Constantinidou, et al. (2004). "Identification of the CRP regulon using in vitro and in vivo transcriptional profiling." Nucleic Acids Res **32**(19): 5874-93.
- Zhong, S., A. Khodursky, et al. (2004). "Evolutionary genomics of ecological specialization." Proc Natl Acad Sci U S A **101**(32): 11719-24.
- Zwir, I., D. Shin, et al. (2005). "Dissecting the PhoP regulatory network of *Escherichia coli* and *Salmonella enterica*." Proc Natl Acad Sci U S A **102**(8): 2862-7.

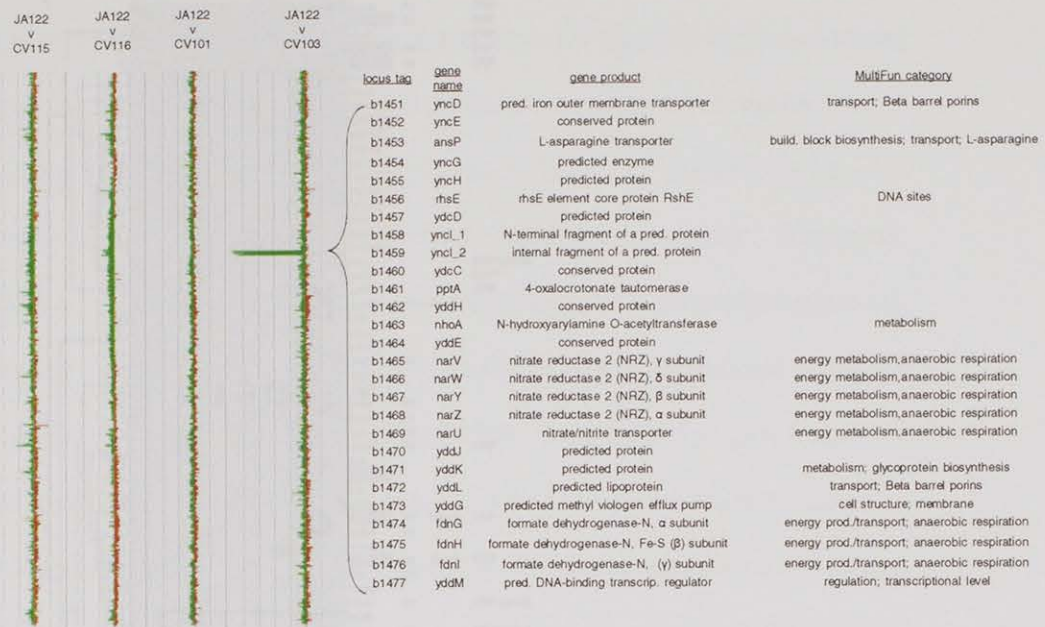


Figure 1. array Comparative Genomic Hybridization (a-CGH) of each adaptive clone versus their common ancestor, JA122. CV103 has sustained an approximately 30 Kb deletion relative to JA122 comprising a total of 27 genes of either unknown function or involved in transcription, arginine biosynthesis, anaerobic respiration, nitrogen metabolism and glycoprotein biosynthesis. Cy-5 labeled genomic DNA from each evolvant (red bars) was hybridized against Cy-3 labeled genomic DNA from JA122 (green bars). The \log_2 ratio of hybridization intensities is depicted along a linear map of the *E. coli* K-12 MG1655 chromosome with genes closest to the origin at the top. Grey lines denote a 2-fold difference in target hybridization. The deleted portion of the CV103 chromosome shown as an excess of hybridization in the reference channel encompasses the 27 genes detailed in the table to the right.

Figure 2. Heatmap showing the expression levels of genes in response to...

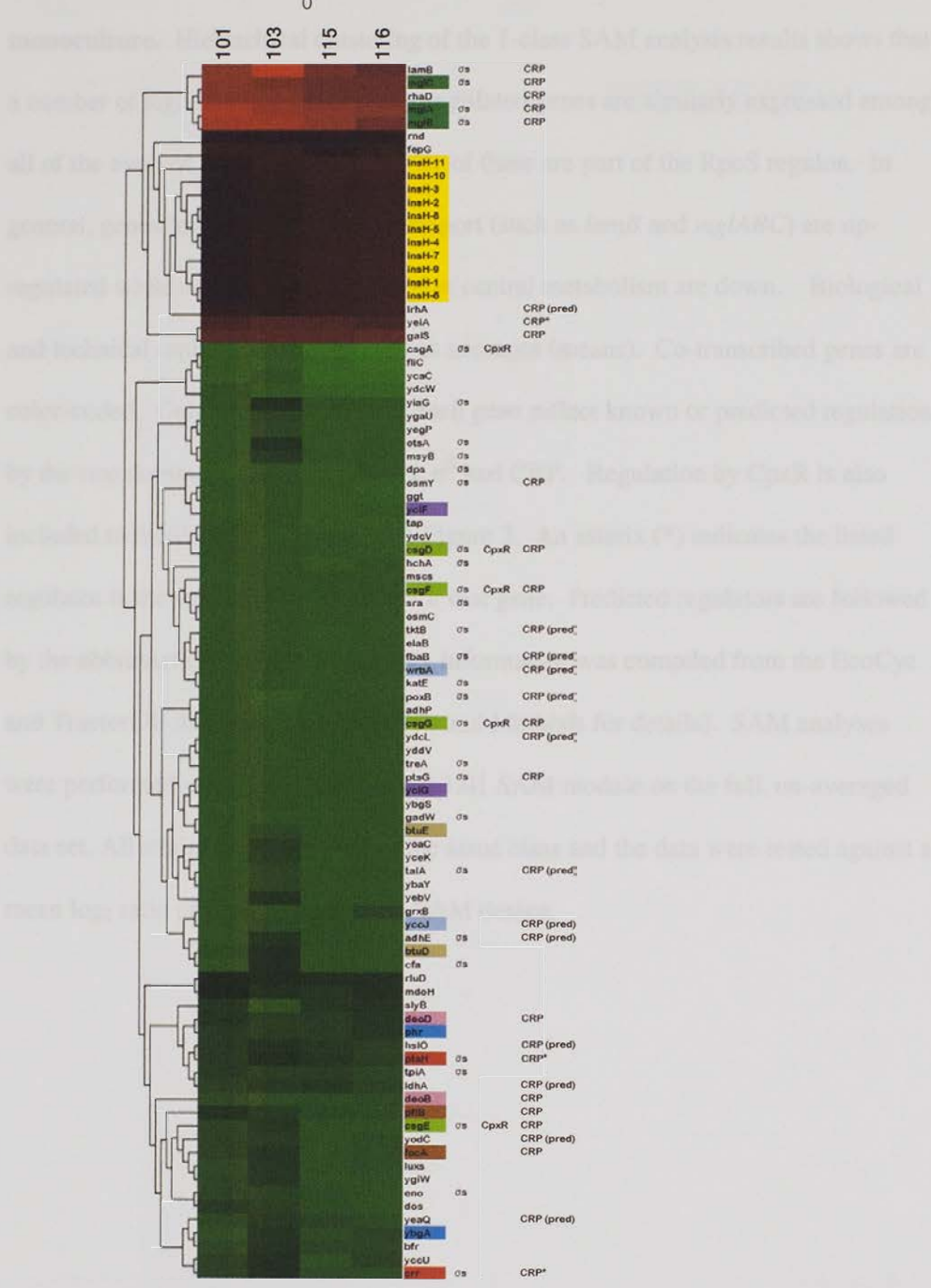


Figure 2. 1-class SAM analysis for terminal isolates grown in chemostat

monoculture. Hierarchical clustering of the 1-class SAM analysis results shows that a number of significantly up- or down-regulated genes are similarly expressed among all of the evolved isolates. The majority of these are part of the RpoS regulon. In general, genes involved in glucose transport (such as *lamB* and *mglABC*) are up-regulated while several genes involved in central metabolism are down. Biological and technical replicates are displayed as averages (means). Co-transcribed genes are color-coded. Columns to the right of each gene reflect known or predicted regulation by the two dominant global regulators, σ^S and CRP. Regulation by CpxR is also included to facilitate comparison with Figure 3. An asterisk (*) indicates the listed regulator is the sole known regulator for that gene. Predicted regulators are followed by the abbreviation "pred." Regulatory information was compiled from the EcoCyc and TractorDB databases (see Materials and Methods for details). SAM analyses were performed using the TIGR MeV 4.1.01 SAM module on the full, un-averaged data set. All strains were assigned to the same class and the data were tested against a mean \log_2 ratio of 0 using the 1-class SAM design.

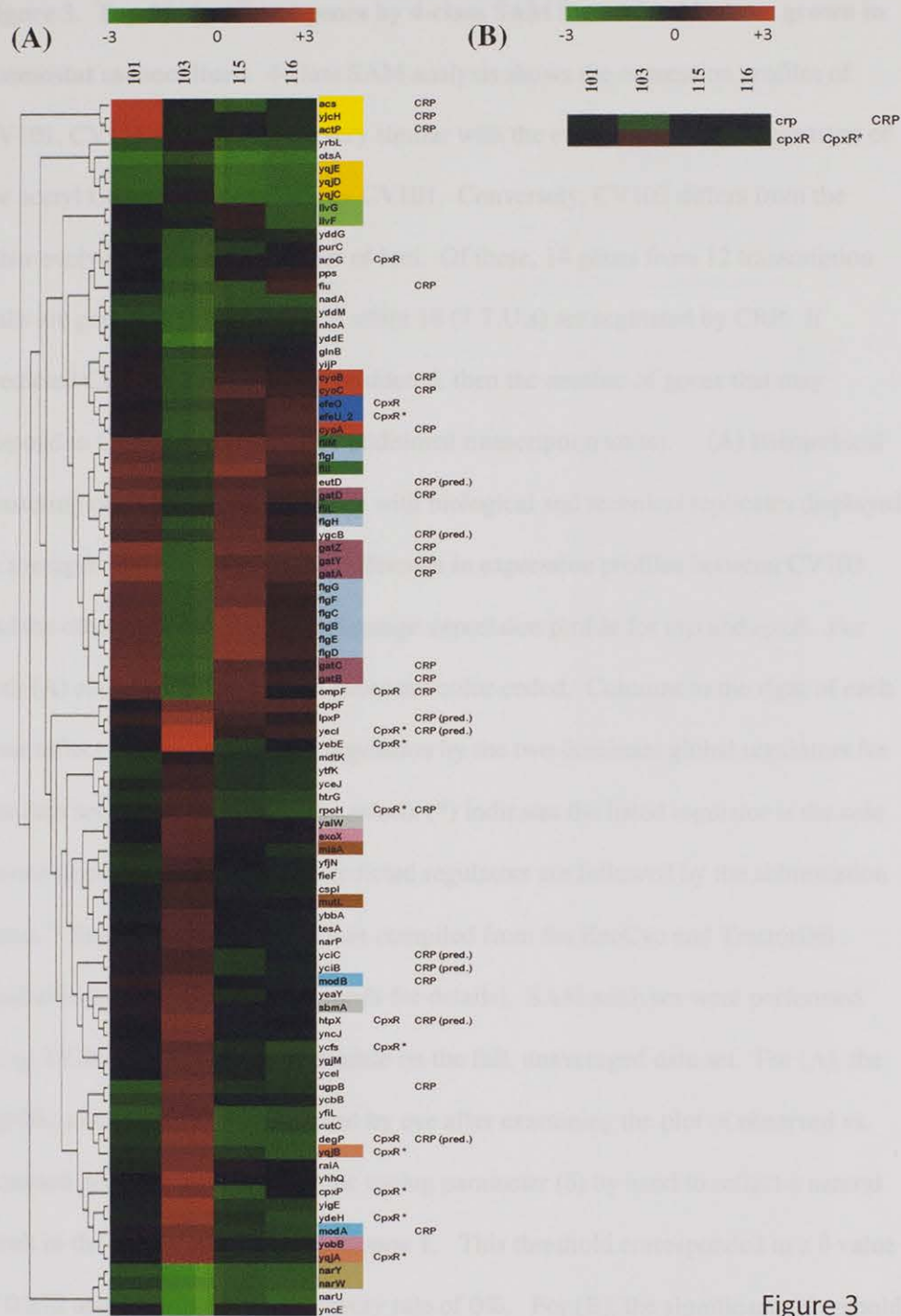


Figure 3

Figure 3. Top 93 significant genes by 4-class SAM for evolved isolates grown in chemostat monoculture. 4-class SAM analysis shows the expression profiles of CV101, CV115 and CV116 are very similar with the exception of over-expression of the acetyl CoA synthase operon in CV101. Conversely, CV103 differs from the other evolved isolates at a number of loci. Of these, 14 genes from 12 transcription units are part of the CpxR regulon while 18 (7 T.U.s) are regulated by CRP. If predicted CRP binding sites are considered, then the number of genes that may respond to CRP increases to 26 (8 additional transcription units). **(A)** Hierarchical clustering of all 93 significant genes with biological and technical replicates displayed as averages (means) showing the difference in expression profiles between CV103 and the other three strains. **(B)** Average expression profile for *crp* and *cpxR*. For both (A) and (B), co-transcribed genes are color-coded. Columns to the right of each gene reflect known or predicted regulation by the two dominant global regulators for this data set, CpxR and CRP. An asterisk (*) indicates the listed regulator is the sole known regulator for that gene. Predicted regulators are followed by the abbreviation "pred." Regulatory information was compiled from the EcoCyc and TractorDB databases (see Materials and Methods for details). SAM analyses were performed using TIGR MeV 4.1.01 SAM module on the full, unaveraged data set. For (A), the significance threshold was assigned by eye after examining the plot of observed vs. expected d-values and adjusting the tuning parameter (δ) by hand to reflect a natural break in the data from a line with slope= 1. This threshold corresponded to a δ value of 0.272 and a median false-discovery rate of 0%. For (B), the significance threshold was assigned using the highest δ value that gave a median false discovery rate of 0%, an analysis that returned a total of 303 significant genes, only two of which are displayed.

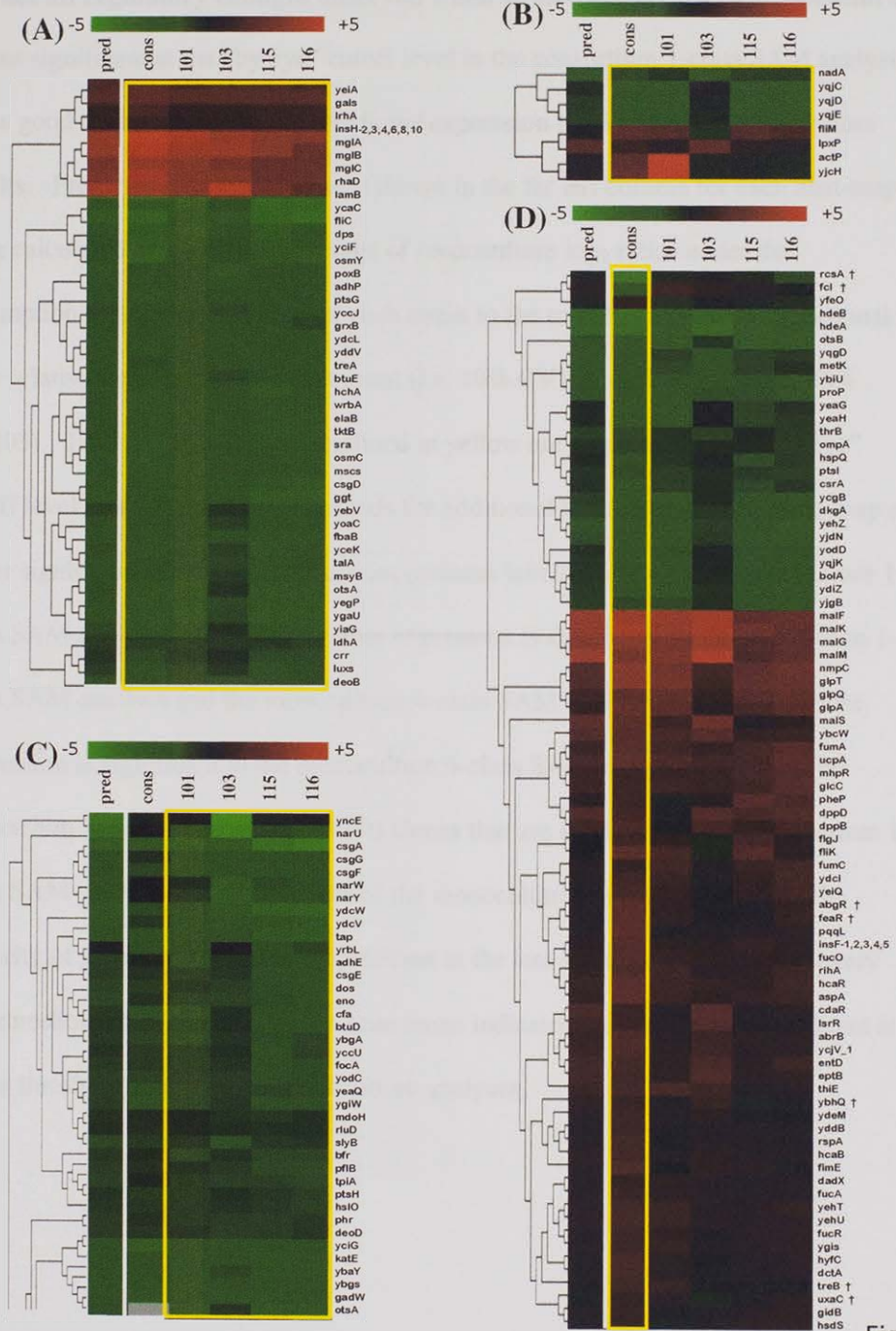


Figure 4

Figure 4. Expression profile SAM analysis of strains in co-culture reflects many, but not all regulatory changes observed when strains are grown in monoculture. Genes significant at the “by eye” cutoff level in the consortium 1-class SAM analysis show good agreement with their predicted expression levels based on monoculture results. Predicted expression levels (shown in the far left column for each heat-map) were calculated as a weighted average of monoculture \log_2 ratios under the assumption that the contribution of each strain to the total RNA pool is proportional to their relative frequency in the chemostat (i.e. 10% CV101, 20% CV116 and 70% CV103). For each panel, genes outlined in yellow are significant at the “by eye” cutoff level (see Materials and Methods for additional information). **(A)** Heat-map of genes significant in both the consortium (column labeled “cons”) and monoculture 1-class SAM analyses. **(B)** Genes whose expression is significant in the consortium 1-class SAM analysis and the monoculture 4-class SAM analysis. **(C)** Genes whose expression is significant in the monoculture 4-class SAM analysis but not in the consortium 1-class SAM analysis. **(D)** Genes that are significant in the consortium 1-class SAM analysis but not in either of the monoculture analyses. However, the majority of genes in panel D *are* significant at the less stringent 0% false discovery rate threshold. † to the right of the gene name indicates the gene is **not** significant at either threshold in any of the monoculture analyses.

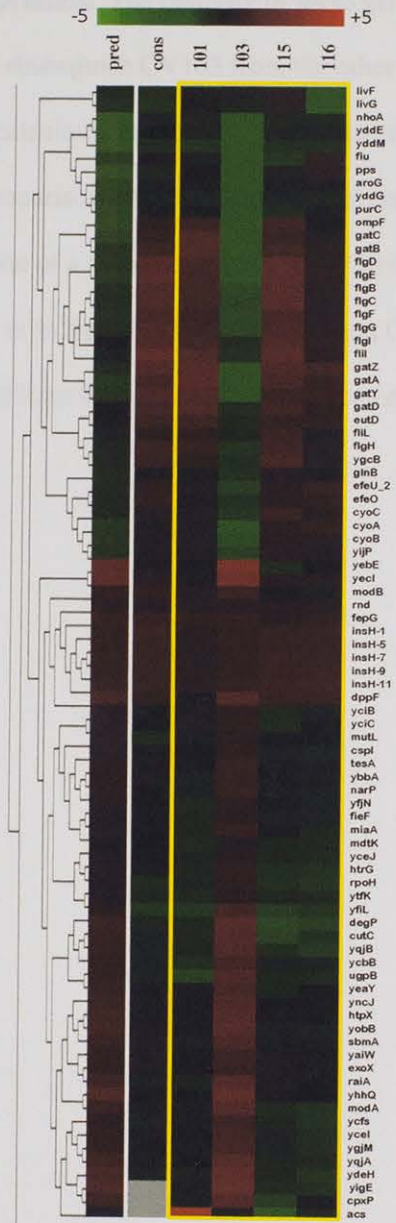


Figure 5

Figure 5. Some genes differ markedly between the monoculture and consortium expression profiles. The majority of these genes are those from the 4-class SAM analysis that distinguish CV103 from the other evolved isolates. Predicted expression levels were calculated as for Figure 4 and are shown in the far left column marked “pred”. Comparison of the consortium and predicted transcriptional profiles suggests that expression of a number of genes in CV103 changes depending on whether it is grown alone or in the presence of CV101 and CV116. Grey boxes indicate the gene was excluded from the analysis due to a lack of high-quality signal on the array.

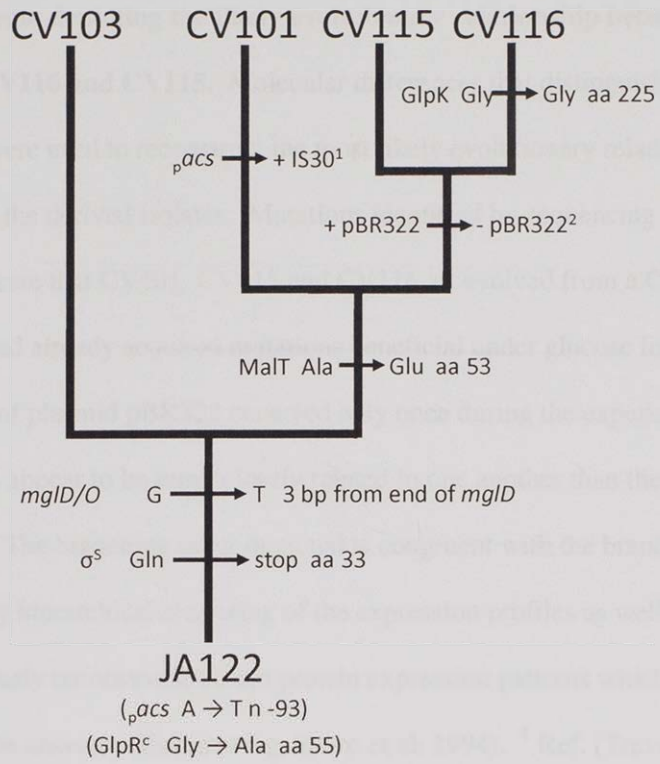


Figure 6. Cladogram depicting the likely evolutionary relationship between CV101, CV103, CV116 and CV115. Molecular differences that distinguish clones from one another were used to reconstruct the most likely evolutionary relationships between all four of the derived isolates. Mutations identified by sequencing of targeted genes indicate that CV101, CV115 and CV116 all evolved from a CV103-like ancestor that had already acquired mutations beneficial under glucose limitation. Assuming the loss of plasmid pBR322 occurred only once during the experiment, CV115 and CV116 appear to be more closely related to one another than they are to CV101 or CV103. The branching order depicted is congruent with the branching order determined by hierarchical clustering of the expression profiles as well as phylogenies previously reconstructed from protein expression patterns which place CV103 closest to the ancestor (Rosenzweig, Sharp et al. 1994). ¹ Ref. (Treves, Manning et al. 1998). ² Ref. (Helling, Vargas et al. 1987).

Table 1. Bacterial Strains

Strain	Relevant Characteristics ²	Specific growth rate (hr ⁻¹) ₂	Relative growth yield ³	Rate of glucose uptake (μmol αMG/min/gm) ³	Equilibrium [glucose] (nmol/mL) ³	Equilibrium [acetate] (nmol/mL) ³
RH201 ¹	CGSC 5346 <i>F- thi 1leu6 thi1 lacY1 tonA21 supE44 hss1 glpR200</i>					
JA104	Derivative of RH 201 <i>F- thi 1 lacY1 araD139gdh supE44 hss1</i> ; lysogenic for λ.					
JA122	As JA104 but contains plasmid pBR322Δ5	0.44 ± 0.01	1.14 ± 0.02	1.19 ± 0.09	1.84 ± 0.48	194 ± 20
CV101	Derivative of JA122; isolated after 773 generations, Amp ^R	0.50 ± 0.02	1.11 ± 0.02	1.66 ± 0.06	0.88 ± 0.31	0 ± 0
CV103	As CV101 but independent isolate which forms small colonies on T, Amp ^R	0.40 ± 0.01	0.81 ± 0.04	2.46 ± 0.16	0.07 ± 0.03	252 ± 70
CV115	Derivative of JA122, isolated after 773 generations, lacks plasmid	0.55 ± 0.02	1.11 ± 0.02	ND	ND	ND
CV116	As CV115 but forms small colonies on TA	0.60 ± 0.01	1.20 ± 0.03	1.61 ± 0.11	0.19 ± 0.05	40 ± 25

¹ Adams, Kinney et al. (1979) (Adams, Kinney et al. 1979)

² Data from Helling, Vargus and Adams (1987), Table 1(Helling, Vargas et al. 1987)

³ Data from Rosenzweig et al. (1994), Table 2 (Rosenzweig, Sharp et al. 1994)

TABLE 2. Expression levels of selected genes from 1-class and 4-class SAM analyses.

ID	gene	SAM class	mean log ₂ CV101 /JA122	mean log ₂ CV103 /JA122	mean log ₂ CV115 /JA122	mean log ₂ CV116 /JA122	gene product	Transcription Unit	MultiFun Category
b4069	<i>acs</i>	4-class	3.9	0.0	-1.5	-0.3	acetyl-CoA synthetase	acs-yjcHG	Metabolism; Building Block Biosynthesis; Acetate utilization; Central intermediary metabolism;
b4484	<i>cpxP</i>	4-class	0.0	1.7	-0.8	0.0	reg. of Cpx response	cpxP	Cell processes; Adaptations; Regulation; 2-component regulatory system
b2417	<i>crr</i>	1-class	-0.9	-0.5	-1.1	-1.1	glucose-specific enzyme IIA component of PTS	ptsHI-crr (ptsHp1)	Metabolism; carbon utilization; The PTS Fructose-Mannitol (Fru) Family, Transport; substrate; D-glucose/trehalose
b1073	<i>flgB</i>	4-class	1.3	-1.1	2.0	0.7	flagellar component of basal-body rod	flgBCDEFGHIJ	Metabolism; Macromolecule Biosynthesis; Flagellum; Motility (incl. chemotaxis, energy taxis, aerotaxis, redox taxis), cell structure;
b1923	<i>fliC</i>	1-class	-2.6	-2.7	-3.5	-3.8	flagellar filament structural protein (flagellin)	fliC	Metabolism; Macromolecule Biosynthesis flagella
b2151	<i>galS</i>	1-class	1.9	2.0	2.0	2.2	DNA-binding transcriptional repressor	galS	Metabolism; Carbon utilization; Regulation; Transcriptional repressor
b1732	<i>katE</i>	1-class	-1.9	-1.5	-2.2	-2.1	hydroperoxidase HPII(III) (catalase)	katE	Cell processes; Protection; Detoxification (xenobiotic metabolism)
b4036	<i>lamB</i>	1-class	3.6	5.1	2.9	2.2	maltose outer membrane porin	malK-lamB-malM (malKp)	Transport; (The Outer Membrane Porin (OMP) Functional Superfamily); The Sugar Porin (SP) Family
b3454	<i>livF</i>	4-class	0.1	-0.1	0.6	-1.5	leucine/isoleucine/valine transporter subunit	livKHMGF	Primary Active Transporters; (isoleucine/valine/leucine); amino acid transport/metabolism; ABC superfamily

b2149	<i>mglA</i>	1-class	4.5	3.9	3.6	3.3	methyl-galactoside transporter	mglBAC (mglBp)	Metabolism; Carbon utilization; The ATP-binding Cassette (ABC) Superfamily
b0929	<i>ompF</i>	4-class	1.2	-1.7	1.3	0.0	outer membrane porin 1a (1a;b;F)	ompF	Transport; β -barrel porins (Outer Membrane Porin (OMP) Functional Superfamily)
b1101	<i>ptsG</i>	1-class	-1.8	-1.4	-1.7	-1.4	PTS system glucose-specific IICB component	ptsG	Metabolism; Carbon utilization; Regulation; Posttranscriptional; Transport Information transfer; Transcriptional Regulation; σ factors, anti- σ factors; adaptation to stress; temperature extremes
b3461	<i>rpoH</i>	4-class	-0.2	0.8	-0.5	-0.7	RNA polymerase, $\sigma 32$ (σH) factor	rpoH	

Table 3. Sequenced Genes

Locus	Gene product	MG1655 position (gene length)	transcriptional start (relative to translational start)	sequenced region relative to translational start site	mutations
<i>acs</i>	acetyl-CoA synthetase (AMP-forming)	4,283,436 ← 4,285,394 (1959 bp)	-224	CV103: -439 → end +14 ; JA122, CV101, CV115 and CV116: -439 → +391, +441 → end+14	A→T, position -93. Shared by JA122, CV101, CV103, CV115 and CV116. CV101 also has an IS 30 element insertion in the promoter as previously reported.
<i>crp</i>	CRP transcriptional dual regulator	3,484,142 → 3,484,774 (633 bp)	-167	-163 → <u>547</u>	none
<i>cya</i>	adenylate cyclase	3,989,176 → 3,991,722 (2547 bp)	-379	-428 → end +56	none. CV115 not sequenced.
<i>glpK</i>	glycerol kinase	4,113,737 ← 4,115,245 (1509 bp)	gene internal to mRNA start	+18 → end +9	Gly → Gly at aa 225 in CV116. JA122, CV101, CV103 and CV115 unchanged.
<i>glpR</i>	sn-Glycerol-3-phosphate repressor	3,557,870 ← 3,558,628 (759 bp)	-286	-25 → end +23	Gly → Ala, aa 55 in JA122, CV101, CV103, CV115 and CV116.
<i>lamB</i>	maltose high-affinity receptor	4,245,994 → 4,247,334 (1341 bp)	gene internal to mRNA start	-16 → end + 241	none
<i>malT</i>	maltose operon transcriptional regulator	3,551,107 →3,553,812 (2706 bp)	-61	-541 → <u>+1125</u>	Ala→Glu, aa 53 in CV101, CV115 and CV116. JA122 and CV103 unchanged.
<i>mgID</i>	GalS transcriptional dual regulator	2,238,650 ← 2,239,690 (1041 bp)	-42	-158 → end + 503	G→T transversion located 3 base-pairs from the end of mgID. Shared by CV101, CV103, CV115 and CV116. Absent in JA122
<i>mlc</i>	DgsA transcriptional repressor	1,665,368 ←1,666,588 (1221 bp)	-39	-75 → end +41	none
<i>pta</i>	phosphate acetyltransferase	2,412,769 → 2,414,913 (2145 bp)	gene internal to mRNA start	JA122: +17 → <u>+1865</u> ; CV101, CV103, CV116: +17 → end +50	none. CV115 not sequenced.
<i>ptsG</i>	enzyme II _{glc}	1,157,092 →1,158,525 (1434 bp)	-243	-297 → end +37	none
<i>rpoS</i>	RNA polymerase, sigma S (sigma 38) factor	2,864,581 ← 2,865,573 (993 bp)	-567	-185 → end +48	Gln→stop aa 33 in CV101, CV103, CV115 and CV116. Unchanged in JA122.
<i>spoT</i>	GDP diphosphokinase / guanosine-3',5'-bis(diphosphate) 3'-diphosphate	3,820,423 → 3,822,531 (2109 bp)	unknown	-48 → <u>2105</u>	none

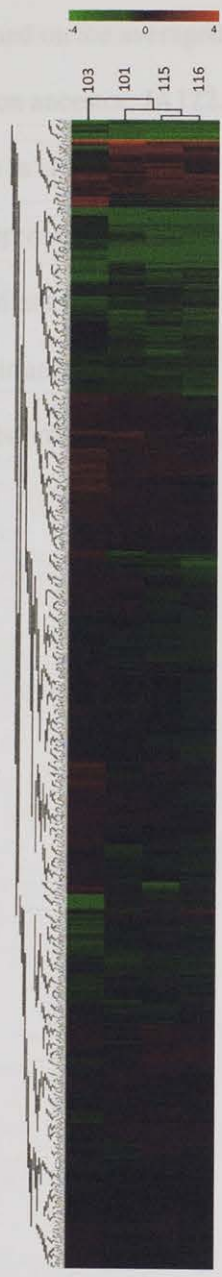


Supplementary Figure 2

Supplementary Figure 1. REP-PCR and PFGE fingerprints of chemostat isolates. (A) Box A1R fingerprints of the terminal chemostat isolates are indistinguishable from those of the ancestor, JA122

Supplementary Figure 1. Global transcriptional response of evolved strains.

Microarray technology was performed on 116 evolved strains. Hierarchical clustering was performed on the data to identify clusters of genes that are co-expressed in the adapted strains relative to the common ancestor. The heatmap shows a change with each row representing a gene. The color scale ranges from -4 (27% greater of the transcriptional density in the adapted strains relative to the common ancestor) to 4 (46% decrease in transcript abundance versus H1 genes).



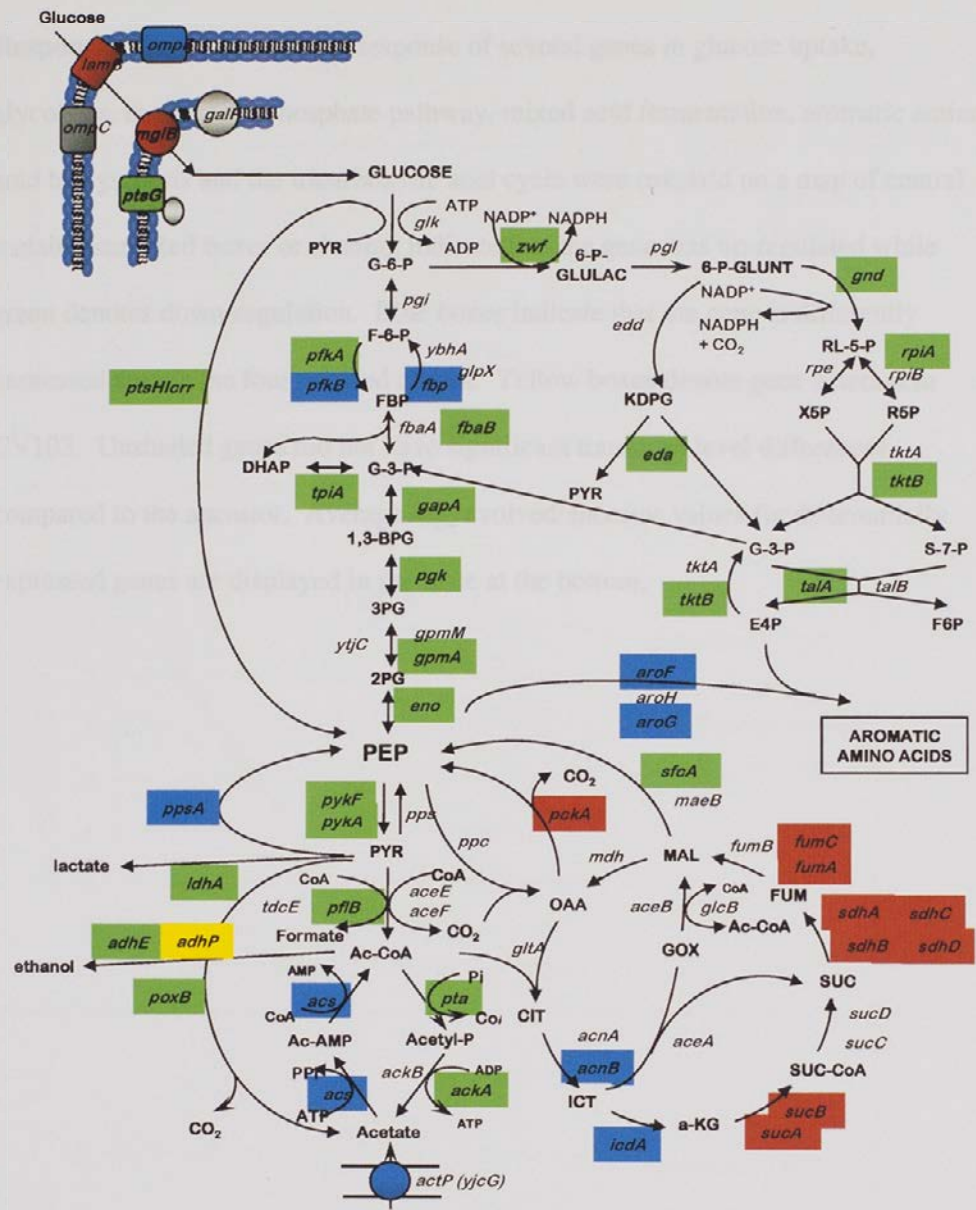
Supplementary Figure 2

Supplementary Figure 2. Global transcriptional response of evolved clones.

Hierarchical clustering was performed on the averaged transcriptional profiles for each adaptive clone relative to its common ancestor, JA122. Adaptive clones and their ancestor were grown to steady state in chemostat monoculture. Evolved clones are shown as columns with each row representing a single gene. On average, about 93% (279 genes) of the transcriptome did not show a two-fold or greater expression change in the adapted clones relative to their ancestor. Of the 7% that did exhibit this degree of change, decreases in transcript abundance were observed more often than increases (168 versus 111 genes).



Supplementary Figure 3. Overview of Central Metabolic Transportations

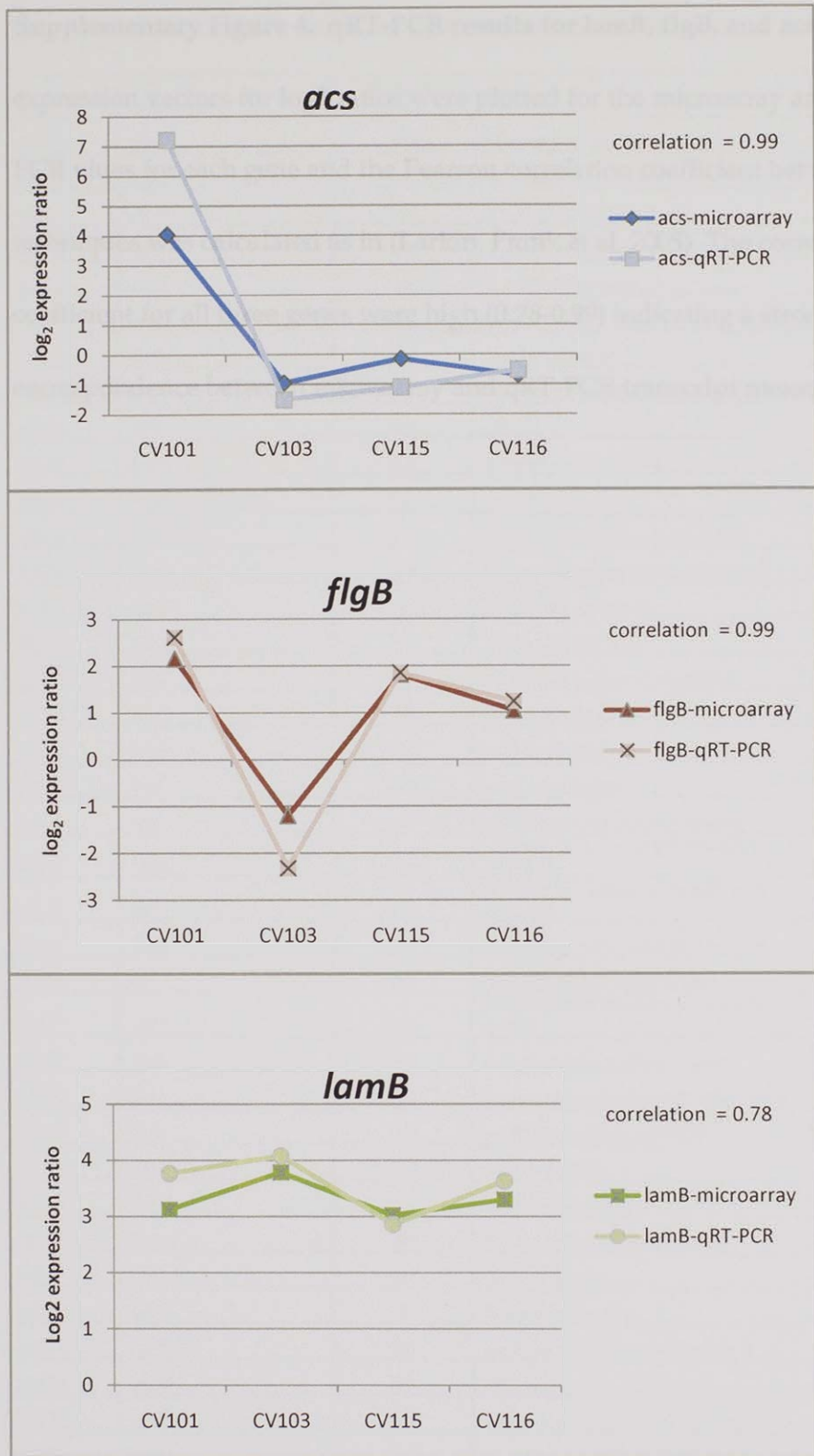


	101	103	115	116
<i>acnB</i>	1.13	-0.20	1.00	0.87
<i>ppsA</i>	0.02	-1.84	0.29	0.43
<i>aroF</i>	-0.56	0.86	-0.52	-0.63
<i>aroG</i>	0.14	-1.48	0.17	0.15
<i>icdA</i>	0.67	-0.42	0.92	0.81
<i>pfkB</i>	-1.47	-0.22	-1.20	-1.09
<i>fbp</i>	0.16	-0.67	-0.10	0.14

Supplementary Figure 3. Overview of Central Metabolic Transcriptional

Response. The transcriptional response of several genes in glucose uptake, glycolysis, the pentose phosphate pathway, mixed acid fermentation, aromatic amino acid biosynthesis and the tricarboxylic acid cycle were overlaid on a map of central metabolism. Red boxes or shading indicate that the gene was up-regulated while green denotes down-regulation. Blue boxes indicate that the gene is differently expressed among the four evolved clones. Yellow boxes denote gene deletion in CV103. Unshaded genes did not have significant transcript level differences compared to the ancestor. Average \log_2 evolved/ancestor values for differentially expressed genes are displayed in the table at the bottom.





Supplementary Table 1. Failed and low concentration PCR reactions

ID	gene name	PCR reaction status	gene product
b0012	htgA	bad	heat shock protein HtgA
b0024	yaaY	bad	predicted protein
b0031	dapB	bad	dihydrodipicolinate reductase
b0037	caiC	bad	probable crotonobetaine/carnitine-CoA ligase
b0062	araA	bad	L-arabinose isomerase
b0075	leuL	bad	leu operon leader peptide
b0080	fruR	bad	fructose repressor
b0083	ftsL	bad	cell division protein FtsL
b0089	ftsW	bad	cell division protein FtsW
b0131	panD	bad	aspartate 1-decarboxylase
b0144	yadB	bad	hypothetical protein
b0240	crI	bad	curlin genes transcriptional activator
b0269	yagF	bad	CP4-6 prophage; predicted dehydratase
b0270	yagG	bad	YagG GPH Transporter
b0271	yagH	bad	putative β -xylosidase
b0276	yagJ	bad	CP4-6 prophage; predicted protein
b0304	ykgC	bad	predicted oxidoreductase
b0319	yahE	bad	predicted protein
b0324	yahJ	bad	predicted deaminase
b0335	prpE	bad	predicted propionyl-CoA synthetase
b0349	mhpC	bad	2-hydroxy-6-ketono-2,4-dienedioate hydrolase
b0375	yaiV	bad	predicted DNA-binding transcriptional regulator
b0400	phoR	bad	phosphate regulon sensor protein PhoR
b0406	tgt	bad	tRNA-guanine transglycosylase
b0437	clpP	bad	ATP-dependent clp protease proteolytic subunit
b0457	ylaB	bad	conserved inner membrane protein
b0460	hha	bad	haemolysin expression modulating protein
b0462	acrB	bad	acriflavin resistance protein B
b0465	o1120	bad	predicted protein
b0466	ybaM	bad	hypothetical protein
b0497	rhdD	bad	RhdD protein precursor
b0517	f349	bad	predicted protein
b0521	ybcF	bad	hypothetical protein
b0525	ppiB	bad	peptidyl-prolyl cis-trans isomerase B
b0564	appY	bad	M5 polypeptide
b0575	ybdE	bad	hypothetical protein
b0586	entF	bad	enterobactin synthetase component F
b0659	ybeY	bad	conserved protein
b0663	b0663	bad	predicted ORF

b0700	rhcC	bad	RhcC protein precursor
b0703	ybfO	bad	conserved protein, rhs-like
b0717	ybgP	bad	putative chaperone
b0779	uvrB	bad	excision nuclease ABC subunit B
b0841	ybjG	bad	undecaprenyl pyrophosphate phosphatase
b0890	ftsK	bad	cell division protein FtsK
b0924	mukB	bad	cell division protein
b0939	ycbR	bad	predicted periplasmic pilin chaperone
b1031	b1031	bad	predicted ORF
b1084	rne	bad	ribonuclease E
b1102	fhuE	bad	outer-membrane receptor for Fe(III)-coprogen
b1117	lolD	bad	outer membrane-specific lipoprotein transporter subunit
b1194	ycgR	bad	protein involved regulation of flagellar motility
b1207	prsA	bad	ribose-phosphate pyrophosphokinase
b1229	tpr	bad	protamine-like protein
b1242	ycheE	bad	hypothetical protein
b1250	kch	bad	putative potassium channel protein
b1252	tonB	bad	TonB protein
b1265	trpL	bad	trp operon leader peptide
b1270	btuR	bad	COB(I) alamin adenosyltransferase
b1378	ydbK	bad	putative pyruvate synthase
b1387	maoC	bad	putative ring-cleavage enzyme of phenylacetate degradation
b1409	ynbB	bad	predicted CDP-diglyceride synthase
b1423	ydcJ	bad	conserved protein
b1432	b1432	bad	putative virulence protein
b1461	ydcE	bad	hypothetical protein
b1487	ddpA	bad	subunit of YddO/YddP/YddQ/YddR/YddS ABC transporter
b1489	dos	bad	cAMP phosphodiesterase, heme-regulated
b1495	yddb	bad	predicted ORF
b1496	ydda	bad	hypothetical ABC transporter in gadB 5' region
b1510	ydeK	bad	hypothetical protein in hipA 5' region
b1513	lsrA	bad	fused A12 transporter subunits of ABC superfamily
b1514	lsrC	bad	LsrC, subunit of LsrA/LsrC/LsrD/LsrB ABC transporter
b1595	ynfL	bad	predicted DNA-binding transcriptional regulator
b1602	pntB	bad	pyridine nucleotide transhydrogenase subunit-beta
b1617	uidA	bad	beta-D-glucuronidase
b1619	hdhA	bad	7-alpha-hydroxysteroid dehydrogenase
b1701	ydiD	bad	short chain acyl-CoA synthetase monomer
b1712	himA	bad	integration host factor alpha-subunit
b1715	pheM	bad	phenylalanyl-tRNA synthetase operon leader peptide
b1770	b1770	bad	predicted DNA-binding transcriptional regulator

b1786	yeaI	bad	predicted diguanylate cyclase
b1815	yoaD	bad	predicted phosphodiesterase
b1816	yoaE	bad	predicted inner membrane protein
b1831	yebJ	bad	predicted structural transport element
b1837	yebW	bad	predicted protein
b1845	ptrB	bad	protease II
b1859	yebI	bad	hypothetical protein
b1877	yecF	bad	predicted protein
b1903	b1903	bad	phantom gene
b1908	yecA	bad	conserved metal binding protein
b1928	yedD	bad	predicted protein
b1934	yedN	bad	predicted protein, C-ter fragment
b1942	fliJ	bad	flagellar FliJ protein
b1963	yedR	bad	predicted inner membrane protein
b1966	yedS_3	bad	predicted protein, C-ter fragment (pseudogene)
b1976	mtfA	bad	conserved protein
b1978	yeeJ	bad	adhesin
b1997	insC-3	bad	IS2 element protein InsA
b2018	hisL	bad	his operon leader peptide
b2118	yehI	bad	hypothetical protein
b2318	truA	bad	pseudouridylyl synthase I
b2360	yfdQ	bad	CPS-53 (KpLE1) prophage; predicted protein
b2420	yfeS	bad	conserved protein
b2432	yfeY	bad	predicted protein
b2457	cchA	bad	predicted protein
b2459	eufT	bad	predicted cobalamine adenosyltransferase
b2463	maeB	bad	NADP-linked malic enzyme
b2500	purN	bad	phosphoribosylglycinamide myltransferase
b2508	guaB	bad	inosine-5'-monophosphate dehydrogenase
b2520	yfhM	bad	conserved protein
b2535	csiE	bad	stationary phase inducible protein CsiE
b2543	yphA	bad	predicted inner membrane protein
b2557	purL	bad	phosphoribosylformylglycineamide synthetase
b2569	lepA	bad	GTP-binding protein LepA
b2606	rplS	bad	50S ribosomal subunit protein L19
b2647	ypjA	bad	adhesin-like autotransporter
b2751	cysN	bad	ATP sulfurylase (ATP:sulfate adenylyltransferase) subunit
b2752	cysD	bad	ATP sulfurylase (ATP:sulfate adenylyltransferase)
b2761	ygcB	bad	hypothetical protein in cysH 3' region
b2765	ygcM	bad	6-pyruvoyl tetrahydropterin synthase
b2837	galR	bad	galactose operon repressor
b2843	kduI	bad	5-keto-4-deoxyuronate isomerase
b2852	ygeH	bad	predicted transcriptional regulator

b2869	ygeV	bad	putative transcriptional regulator
b2920	ygfH	bad	propionyl-CoA:succinate CoA transferase
b3021	ygiT	bad	predicted DNA-binding transcriptional regulator
b3033	yqiB	bad	predicted dehydrogenase
b3038	ygiC	bad	predicted enzyme
b3050	yqiJ	bad	putative oxidoreductase
b3052	rfaE	bad	fused heptose 7-phosphate kinase/heptose 1-phosphate adenylyltransferase
b3073	ygiG	bad	probable ornithine aminotransferase
b3102	yqiG	bad	predicted S-transferase
b3125	yhaE	bad	tartronate semialdehyde reductase
b3137	agaY	bad	tagatose-bisphosphate aldolase agaY
b3166	truB	bad	tRNA pseudouridine 55 synthase
b3168	infB	bad	protein chain initiation factor 2
b3181	greA	bad	transcription elongation factor
b3212	gltB	bad	glutamate synthase (NADPH) large chain precursor
b3213	gltD	bad	glutamate synthase (NADPH) small chain
b3220	yhcG	bad	conserved protein
b3230	rpsI	bad	30S ribosomal subunit protein S9
b3297	rpsK	bad	30S ribosomal subunit protein S11
b3310	rplN	bad	50S ribosomal subunit protein L14
b3312	rpmC	bad	50S ribosomal subunit protein L29
b3331	yheI	bad	putative general secretion pathway protein j precursor
b3338	yheB	bad	endochitinase
b3378	yhfU	bad	predicted protein
b3384	trpS	bad	tryptophanyl tRNA synthetase
b3449	ugpQ	bad	glycerophosphoryl diester phosphodiesterase
b3482	rhsB	bad	RhsB core protein with unique extension
b3488	yhiJ	bad	predicted protein
b3489	yhiK	bad	predicted protein
b3521	yhjC	bad	predicted DNA-binding transcriptional regulator
b3593	rhsA	bad	rhsA protein precursor
b3606	yibK	bad	predicted rRNA methylase
b3612	yibO	bad	putative 2,3-bisphosphoglycerate-independent phosphoglycerate
b3643	rph	bad	RNase PH
b3765	yifB	bad	predicted ATP-dependent protease
b3768	ilvG_2	bad	acetolactate synthase II, large subunit, C-ter fragment (pseudogene)
b3772	ilvA	bad	threonine deaminase; threonine dehydratase biosynthetic
b3793	rfft	bad	4-alpha-l-fucosyltransferase
b3826	yigL	bad	sugar phosphatase
b3835	ubiB	bad	2-octaprenylphenol hydroxylase
b3885	yihX	bad	α -D-glucose-1-phosphatase

b3942	katG	bad	catalase hydroperoxidase I
b3955	yjJP	bad	conserved inner membrane protein
b4005	purD	bad	phosphoribosylglycineamide synthetase
b4006	purH	bad	phosphoribosylaminoimidazolecarboxamide formyltransferase
b4034	malE	bad	periplasmic maltose-binding protein
b4066	yjcF	bad	conserved protein
b4083	yjcS	bad	predicted alkyl sulfatase
b4099	phnI	bad	phnI protein
b4114	yjdB	bad	predicted metal-dependent hydrolase
b4138	dcuA	bad	anaerobic c4-dicarboxylate transporter dcuA
b4148	sugE	bad	SugES
b4156	yjeM	bad	YjeM APC transporter
b4177	purA	bad	adenylosuccinate synthetase
b4179	vacB	bad	VacB protein
b4194	yjIT	bad	subunit of L-ascorbate transporting phosphotransferase system
b4199	yjFY	bad	predicted protein
b4200	rpsF	bad	30S ribosomal subunit protein S6
b4201	priB	bad	primosomal replication protein n
b4208	cycA	bad	d-serine/d-alanine/glycine transporter
b4211	ytfG	bad	NAD(P)H:quinone oxidoreductase
b4215	ytfL	bad	predicted protein
b4221	ytfN	bad	conserved protein
b4222	ytfP	bad	conserved protein
b4242	mgtA	bad	Mg(2+) transport ATPase, P-type 1
b4249	yjI	bad	predicted oxidoreductase
b4256	yjgM	bad	predicted acetyltransferase
b4307	yjhQ	bad	KpLE2 phage-like element; predicted acetyltransferase
b4348	hsdS	bad	type I restriction enzyme <i>ecoli</i> specificity protein (s protei
b0015	dnaJ	low concentration	DnaJ protein
b0495	ybbA	low concentration	hypothetical ABC transporter
b1468	narZ	low concentration	respiratory nitrate reductase 2 alpha chain
b1687	ydiJ	low concentration	predicted FAD-linked oxidoreductase
b1732	katE	low concentration	catalase HPiI
b1823	cspC	low concentration	cold shock-like protein CspC
b1916	sdiA	low concentration	sdiA regulatory protein
b2145	yeiS	low concentration	predicted inner membrane protein
b2392	mntH	low concentration	MntH manganese ion NRAMP transporter
b2666	yqaE	low concentration	predicted membrane protein
b2791	truC	low concentration	tRNA pseudouridine 65 synthase
b2866	xdhA	low concentration	xanthine dehydrogenase, molybdenum binding subunit

b3016	ygiQ	low concentration	conserved protein
b4038	yjbl	low concentration	conserved protein
b4058	uvrA	low concentration	excision nuclease
b4068	yjcH	low concentration	conserved inner membrane protein
b4121	yjdF	low concentration	conserved inner membrane protein
b4193	yjfS	low concentration	predicted protein
b4206	ytfB	low concentration	predicted cell envelope opacity-associated protein

Supplementary Table 2. Top 91 significant genes by 1-class SAM for evolved isolates grown individually.

ID	gene	mean log ₂ CV101 /JA122	mean log ₂ CV103 /JA122	mean log ₂ CV115 /JA122	mean log ₂ CV116 /JA122	gene product	Transcription Unit	MultiFun category
b4036	<i>lamB</i>	3.6	5.1	2.9	2.2	maltose outer membrane porin (maltoporin)	malK-lamB-malM (malKp)	transport; (The Outer Membrane Porin (OMP) Functional Superfamily); The Sugar Porin (SP) Family
b2148	<i>mglC</i>	3.5	3.1	2.6	2.5	membrane component of an ABC superfamily methyl-galactoside transporter	mglBAC (mglBp)	metabolism; carbon utilization; transport; The ATP-binding Cassette (ABC) Superfamily
b3902	<i>rhaD</i>	3.4	3.4	2.7	2.7	rhamnose-1-phosphate aldolase	rhaBAD (rhaBp)	Metabolism; carbon utilization;
b2149	<i>mglA</i>	4.5	3.9	3.6	3.3	fused methyl-galactoside transporter subunits of ABC superfamily; ATP-binding components	mglBAC (mglBp)	metabolism; carbon utilization; The ATP-binding Cassette (ABC) Superfamily
b2150	<i>mglB</i>	4.1	4.1	3.6	2.7	periplasmic-binding component of an ABC superfamily methyl-galactoside transporter	mglBAC (mglBp)	Metabolism; carbon utilization; chaperoning, folding, transport; Primary Active Transporters; The ATP-binding Cassette (ABC) Superfamily
b1804	<i>rnd</i>	0.6	0.6	0.6	0.7	ribonuclease D	rnd	metabolism; macromolecule degradation; information transfer; RNA modification
b0589	<i>fepG</i>	1.0	1.0	0.7	1.0	membrane component of an ABC superfamily iron-enterobactin transporter	fepDGC (fepDp1)	transport; The ATP-binding Cassette (ABC) Superfamily
b3505	<i>insH-11</i>	0.7	1.0	1.2	1.2	IS5 transposase	insH-11	Extrachromosomal; transposon related
b3218	<i>insH-10</i>	0.7	0.9	1.1	1.1	IS5 transposase	insH-10	Extrachromosomal; transposon related
b0656	<i>insH-3</i>	0.6	0.9	1.0	1.0	IS5 transposase	insH-3	Extrachromosomal; transposon related
b0552	<i>insH-2</i>	0.7	0.8	1.0	1.1	IS5 transposase	insH-2	Extrachromosomal; prophage genes; transposon related
b2192	<i>insH-8</i>	0.6	0.9	1.0	1.1	IS5 transposase	insH-8	Extrachromosomal; transposon related
b1370	<i>insH-5</i>	0.7	0.9	1.0	1.1	IS5 transposase	insH-5	Extrachromosomal; prophage genes; transposon related
b1331	<i>insH-4</i>	0.6	0.9	1.0	1.1	IS5 transposase	insH-4	Extrachromosomal; transposon related
b2030	<i>insH-7</i>	0.6	0.9	1.1	1.1	IS5 transposase	insH-7	Extrachromosomal; transposon related
b2982	<i>insH-9</i>	0.6	0.9	1.1	1.2	IS5 transposase	insH-9	Extrachromosomal; transposon related

b0259	<i>insH-1</i>	0.7	0.9	1.1	1.1	IS5 transposase	insH-1	Extrachromosomal; prophage genes; transposon related
b1994	<i>insH-6</i>	0.6	0.9	1.1	1.1	IS5 transposase	insH-6	Extrachromosomal; transposon related
b2289	<i>lrhA</i>	0.9	0.9	1.1	1.1	DNA-binding transcriptional repressor of flagellar, motility and chemotaxis genes	lrhA	metabolism; energy metabolism, aerobic respiration, regulation; transcriptional level; repressor
b2147	<i>yeiA</i>	1.4	1.7	1.6	1.4	predicted oxidoreductase	yeiAT (yeiTp)	metabolism; central intermediary metabolism; unassigned reversible reactions
b2151	<i>galS</i>	1.9	2.0	2.0	2.2	DNA-binding transcriptional repressor	galS	metabolism; carbon utilization; regulation; transcriptional level; repressor
b1042	<i>csgA</i>	-2.9	-2.7	-4.0	-4.3	cryptic curlin major subunit	csgBAC (csgBp)	metabolism; macromolecules (cellular constituent) biosynthesis; glycoprotein, cell structure; pilus,
b1923	<i>fliC</i>	-2.6	-2.7	-3.5	-3.8	flagellar filament structural protein (flagellin)	fliC	metabolism; macromolecules (cellular constituent) biosynthesis; flagella
b0897	<i>ycaC</i>	-2.9	-1.8	-3.6	-3.6	predicted hydrolase	ycaC	Metabolism; carbon utilization; amino acids
b1444	<i>ydcW</i>	-1.4	-2.6	-3.3	-3.3	medium chain aldehyde dehydrogenase	ydcW	
b3555	<i>yiaG</i>	-2.0	-0.6	-2.4	-2.4	predicted transcriptional regulator	yiaG	regulation; transcriptional level
b2665	<i>ygaU</i>	-2.5	-1.0	-2.7	-2.6	predicted protein	ygaU	
b2080	<i>yegP</i>	-2.4	-1.1	-2.6	-2.4	predicted protein	yegP	
b1896	<i>otsA</i>	-2.4	-0.7	-2.6	-2.4	trehalose-6-phosphate synthase	otsBA (otsBp)	metabolism; central intermediary metabolism; glucose metabolism,
b1051	<i>msyB</i>	-2.3	-0.8	-2.4	-2.3	predicted protein	msyB	transport; substrate; protein
b0812	<i>dps</i>	-2.8	-2.2	-2.9	-2.4	Fe-binding and storage protein	dps	Information transfer; protein related; cell processes; adaptation to stress; starvation response
b4376	<i>osmY</i>	-2.3	-1.8	-2.8	-3.0	periplasmic protein	osmY	Cell processes; adaptation to stress; osmotic pressure
b3447	<i>ggt</i>	-2.2	-1.9	-2.9	-2.9	gamma-glutamyltranspeptidase	ggt	metabolism; macromolecules (cellular constituent) biosynthesis; thioredoxin, glutaredoxin
b1258	<i>yciF</i>	-2.6	-1.5	-2.8	-2.6	conserved protein	yciGFE (yciGp)	
b1885	<i>tap</i>	-1.5	-2.0	-2.8	-2.9	methyl-accepting protein IV	tar-tap-cheRBYZ (tarp)	regulation; posttranscriptional; cell processes; motility (incl.
b1443	<i>ydcV</i>	-1.4	-2.4	-2.5	-2.8	membrane component of an ABC superfamily	ydcSTUV	transport; The ATP-binding Cassette (ABC) Superfamily

						predicted spermidine/putrescine transporter		
b1040	<i>csgD</i>	-1.7	-1.8	-2.3	-2.6	DNA-binding transcriptional regulator of adhesion determinants	csgDEFG (csgDp2)	Information transfer; RNA related; transcription related, activator
b1967	<i>hchA</i>	-1.9	-1.9	-2.8	-2.0	Hsp31 molecular chaperone	hchA	
b2924	<i>mscS</i>	-2.0	-1.7	-1.9	-2.1	component of the MscS mechanosensitive channel	mscS	transport; Channel-type Transporters; The Small Conductance Mechanosensitive Ion Channel (MscS) Family
b1038	<i>csgF</i>	-1.8	-1.8	-2.1	-2.0	predicted transport protein	csgDEFG (csgDp2)	transport; Putative uncharacterized transport protein, cell structure; pilus, curli subunit
b1480	<i>sra</i>	-2.0	-1.7	-2.3	-2.3	30S ribosomal subunit protein S22	sraA (sraAp)	Information transfer; protein related; ribosomal proteins
b1482	<i>osmC</i>	-2.1	-1.9	-2.2	-2.1	osmotically inducible, stress-inducible membrane protein	osmC	Cell processes; adaptation to stress; osmotic pressure
b2465	<i>tktB</i>	-1.9	-1.3	-2.5	-2.3	transketolase 2, thiamin-binding	tktB	metabolism; carbon utilization; central intermediary metabolism; pentose phosphate shunt, non-oxidative branch; nucleotide and nucleoside conversions
b2266	<i>elaB</i>	-1.8	-1.3	-2.3	-2.2	conserved protein	elaB	
b2097	<i>fbaB</i>	-1.8	-1.2	-2.3	-2.1	fructose-bisphosphate aldolase class I	fbaB	metabolism; energy metabolism; carbon; glycolysis
b1004	<i>wrbA</i>	-2.0	-1.4	-2.3	-2.0	predicted flavoprotein in Trp regulation	wrbA-yccJ (wrbAp)	Metabolism; building block biosynthesis; amino acids; tryptophan
b1732	<i>katE</i>	-1.9	-1.5	-2.2	-2.1	hydroperoxidase HP11(III) (catalase)	katE	Cell processes; protection; detoxification (xenobiotic metabolism)
b0871	<i>poxB</i>	-1.6	-2.7	-2.0	-1.9	pyruvate dehydrogenase (pyruvate oxidase), thiamin-dependent, FAD-binding	ltaE-poxB-ybjI	metabolism; carbon utilization; central intermediary metabolism; pyruvate oxidation
b1478	<i>adhP</i>	-1.3	-2.2	-1.8	-1.7	alcohol dehydrogenase, 1-propanol preferring	adhP	metabolism; energy metabolism; carbon; anaerobic respiration
b1037	<i>csgG</i>	-1.6	-1.9	-1.7	-1.4	outer membrane channel lipoprotein	csgDEFG (csgDp2)	transport; Putative uncharacterized transport protein, cell structure; pilus, curli subunit
b1431	<i>ydcL</i>	-1.5	-1.5	-1.6	-1.8	predicted lipoprotein	ydcL	
b1490	<i>yddV</i>	-1.5	-1.5	-1.5	-1.8	predicted diguanylate cyclase	yddV-dos	

b1197	<i>treA</i>	-1.6	-1.3	-1.8	-1.9	periplasmic trehalase	treA	metabolism; central intermediary metabolism; adaptation to stress; osmotic pressure
b1101	<i>ptsG</i>	-1.8	-1.4	-1.7	-1.4	PTS system glucose-specific IICB component	ptsG	Metabolism; carbon utilization; regulation; type of regulation; posttranscriptional; transport
b1259	<i>yciG</i>	-1.7	-1.5	-1.8	-1.5	predicted protein	yciGFE (yciGp)	
b0753	<i>ybgS</i>	-1.9	-1.2	-1.9	-1.7	conserved protein	ybgS	
b3515	<i>gadW</i>	-1.6	-1.3	-2.1	-1.7	DNA-binding transcriptional activator	gadXW (gadXp)	Information transfer; RNA related; transcription related; regulation; activator
b1710	<i>btuE</i>	-1.5	-1.1	-2.0	-1.8	predicted glutathione peroxidase	btuCED	Metabolism; building block biosynthesis; cobalamin (Vitamin B12), transport; Primary Active Transporters; The ATP-binding Cassette (ABC) Superfamily
b1810	<i>yoaC</i>	-1.8	-0.9	-1.8	-2.1	predicted protein	yoaC	
b1050	<i>yceK</i>	-1.9	-0.9	-2.4	-1.8	predicted lipoprotein	yceK	
b2464	<i>talA</i>	-1.9	-1.0	-2.1	-1.8	transaldolase A	talA	metabolism; central intermediary metabolism; pentose phosphate shunt, non-oxidative branch
b0453	<i>ybaY</i>	-1.9	-1.0	-2.1	-2.0	predicted outer membrane lipoprotein	ybaY	metabolism; macromolecules biosynthesis; lipoprotein; glycoprotein
b1836	<i>yebV</i>	-1.4	-0.7	-2.2	-1.9	predicted protein	yebV	
b1064	<i>grxB</i>	-1.6	-1.3	-1.6	-1.0	glutaredoxin 2 (Grx2)	grxB	metabolism; macromolecules (cellular constituent) biosynthesis; large molecule carriers; thioredoxin, glutaredoxin
b1003	<i>yccJ</i>	-1.3	-1.1	-1.7	-1.4	predicted protein	wrbA-yccJ (wrbAp)	
b1241	<i>adhE</i>	-1.4	-0.7	-1.2	-1.6	fused acetaldehyde-CoA dehydrogenase and iron-dependent alcohol dehydrogenase and pyruvate-formate lyase deactivase	adhE	metabolism; energy metabolism, carbon; fermentation
b1709	<i>btuD</i>	-1.1	-0.6	-1.7	-1.3	vitamin B12 transporter subunit : ATP-binding component of ABC superfamily	btuCED	Metabolism; building block biosynthesis; cobalamin (Vitamin B12), transport; The ATP-binding Cassette (ABC) Superfamily
b1661	<i>cfa</i>	-1.5	-0.6	-1.6	-1.3	cyclopropane fatty acyl phospholipid	cfa	Metabolism; building block

						synthase		biosynthesis; fatty acid and phosphatidic acid
b2594	<i>rluD</i>	-0.4	-0.5	-0.6	-0.6	23S rRNA pseudouridine synthase	rluD-yfiH	Information transfer; RNA related; RNA modification
b1049	<i>mdoH</i>	-0.4	-0.7	-0.8	-0.6	glycosyl transferase	mdoGH (mdoGp1)	Cell processes; adaptation to stress; osmotic pressure
b1641	<i>slyB</i>	-0.8	-1.3	-0.8	-0.7	outer membrane lipoprotein	slyB	Cell structure; membrane
b4384	<i>deoD</i>	-0.7	-0.9	-0.9	-0.7	purine-nucleoside phosphorylase	deoCABD (deoCp1)	metabolism; central intermediary metabolism; nucleotide and nucleoside conversions
b0708	<i>phr</i>	-0.8	-0.9	-0.9	-0.8	deoxyribodipyrimidine photolyase	ybgA-phr (ybgAp1)	Metabolism; building block biosynthesis; riboflavin (Vitamin B2), FAD, FMN, information transfer; DNA related; DNA repair
b3401	<i>hslO</i>	-0.8	-0.7	-0.8	-0.9	heat shock protein Hsp33	hslR-hslO (hslRp)	Information transfer; protein related; chaperoning, folding
b2415	<i>ptsH</i>	-0.9	-0.6	-0.8	-0.9	phosphohistidinophoretin-hexose phosphotransferase component of PTS system (Hpr)	ptsHI-crr (ptsHp1)	metabolism; carbon utilization; Group Translocators; The Phosphotransferase System HPr (HPr) Family; transport; substrate; sugar
b3919	<i>tpiA</i>	-0.9	-0.9	-1.0	-1.1	triosephosphate isomerase	tpiA	metabolism; energy metabolism, carbon; glycolysis
b1380	<i>ldhA</i>	-1.0	-0.8	-0.8	-1.0	fermentative D-lactate dehydrogenase, NAD-dependent	ldhA	metabolism; energy metabolism, carbon; fermentation
b4383	<i>deoB</i>	-1.1	-1.2	-1.2	-1.1	phosphopentomutase	deoCABD (deoCp1)	metabolism; central intermediary metabolism; nucleotide and nucleoside conversions
b0903	<i>pflB</i>	-0.6	-0.9	-1.1	-1.1	pyruvate formate lyase I	focA-pflB (focAp1)	metabolism; energy metabolism, carbon; anaerobic respiration, carbon utilization; central intermediary metabolism; threonine catabolism
b1039	<i>csgE</i>	-1.0	-0.8	-1.5	-1.4	predicted transport protein	csgDEFG (csgDp2)	transport; Transporters of Unknown Classification; Putative uncharacterized transport protein, cell structure; pilus, curli
b1957	<i>yodC</i>	-1.0	-0.9	-1.4	-1.2	predicted protein	yodC	
b0904	<i>focA</i>	-1.0	-0.7	-1.4	-1.2	formate transporter	focA-pflB (focAp1)	metabolism; carbon utilization; transport; The Formate-Nitrite

								Transporter (FNT) Family
b2687	<i>luxS</i>	-1.1	-0.8	-1.3	-1.2	S-ribosylhomocysteine lyase	luxS	regulation; transcriptional level; complex regulation; quorum sensing
b3024	<i>ygiW</i>	-1.1	-0.8	-1.3	-1.2	conserved protein	ygiW	
b2779	<i>eno</i>	-1.0	-1.0	-1.2	-1.4	enolase	eno-pyrG (pyrGp)	metabolism; energy metabolism, carbon; glycolysis; anaerobic respiration; gluconeogenesis
b1489	<i>dos</i>	-0.8	-0.9	-1.2	-1.3	cAMP phosphodiesterase	yddV-dos	
b1795	<i>yeaQ</i>	-1.1	-0.9	-1.1	-1.1	conserved inner membrane protein	yeaQ	
b0707	<i>ybgA</i>	-1.2	-0.7	-1.2	-1.0	conserved protein	ybgA-phr (ybgAp1)	
b3336	<i>bfr</i>	-1.1	-0.7	-1.1	-1.0	bacterioferritin, iron storage and detoxification protein	bfd-bfr	Cell processes; adaptation to stress; Fe acquisition
b0965	<i>yccU</i>	-0.9	-0.6	-1.3	-1.0	predicted CoA-binding protein	yccU	
b2417	<i>crr</i>	-0.9	-0.5	-1.1	-1.1	glucose-specific enzyme IIA component of PTS	ptsHI-crr (ptsHp1)	metabolism; carbon utilization; The PTS Fructose-Mannitol (Fru) Family, transport; substrate; D-glucose/trehalose

Supplementary Table 3. Top 93 significant genes by 4-class SAM for evolved isolates grown individually.

ID	Gene	mean log ₂ CV101 /JA122	mean log ₂ CV101 /JA122	mean log ₂ CV101 /JA122	mean log ₂ CV101 /JA122	Gene product	Transcription unit	MultiFun description
b4484	<i>cpxP</i>	0.0	1.7	-0.8	0.0	regulator of Cpx response; possible chaperone involved in extra-cytoplasmic stress resistance	<i>cpxP</i>	Cell processes; adaptations; regulation; 2-component regulatory system
b1874	<i>cutC</i>	-0.4	1.8	-1.0	-0.6	copper homeostasis protein	<i>cutC</i>	Cell processes; protection; detoxification (xenobiotic metabolism)
b1552	<i>cspI</i>	0.0	0.8	-0.1	-0.2	Qin prophage; cold shock protein	<i>cspI</i>	Prophage genes and phage related functions; extrachromosomal
b2630	<i>yfjN</i>	-0.4	0.8	0.0	-0.1	CP4-57 prophage; RNase LS	<i>mlA-yfjO</i>	Prophage genes and phage related functions; extrachromosomal
b4170	<i>mutL</i>	0.0	0.8	0.1	0.0	methyl-directed mismatch repair protein	<i>yjeFE-amiB-mutL-miaA-hfq-hflXKC</i>	Information transfer; DNA repair
b4171	<i>miaA</i>	-0.3	1.0	0.1	-0.4	δ -(2)-iso-pentenylpyrophosphate tRNA-adenosine transferase	<i>yjeFE-amiB-mutL-miaA-hfq-hflXKC</i>	Information transfer; RNA related; RNA modification
b2597	<i>raiA</i>	-0.3	1.8	0.3	0.0	cold shock protein associated with 30S ribosomal subunit	<i>raiA</i>	Information transfer; protein related; translation, cell structure; ribosome; cold-shock protein
b1463	<i>nhoA</i>	-0.2	-2.0	-0.3	-0.3	N-hydroxyarylamine O-acetyltransferase	<i>nhoA</i>	Metabolism
b1466	<i>narW</i>	-0.5	-2.8	-1.1	-1.5	nitrate reductase 2 (NRZ), δ -subunit (assembly subunit)	<i>narZYWV</i>	Metabolism; Anaerobic respiration C & energy metabolism; information transfer; protein related; chaperoning, folding
b1467	<i>narY</i>	-0.9	-3.0	-1.3	-1.5	nitrate reductase 2 (NRZ), β -subunit	<i>narZYWV</i>	Metabolism; Anaerobic respiration energy production/transport; e^- acceptor; carbon; cell structure; membrane
b0430	<i>cyoC</i>	0.4	-0.8	1.2	0.7	cytochrome o ubiquinol oxidase subunit III	<i>cyoABCDE</i>	Metabolism; Aerobic respiration; C & energy metabolism; energy production/transport; e^- acceptor; Primary Active Transporters; Oxidoreduction-driven Active Transporters; H^+ -translocating Cytochrome Oxidase (COX)

b0431	<i>cyoB</i>	0.4	-1.1	0.8	0.6	cytochrome o ubiquinol oxidase subunit I	cyoABCDE	Metabolism; Aerobic metabolism; e^- acceptor; C & energy metabolism; Primary Active Transporters; Oxido-reduction-driven Active Transporters; H ⁺ -translocating Cytochrome Oxidase (COX)
b0432	<i>cyoA</i>	0.3	-1.2	0.7	0.9	cytochrome o ubiquinol oxidase subunit II	cyoABCDE	Metabolism; Aerobic respiration; e^- transport; C & energy metabolism; Primary Active Transporters; Oxido-reduction-driven Active Transporters; H ⁺ -translocating Cyt oxidase (COX)
b0754	<i>aroG</i>	0.1	-1.5	0.2	0.2	3-deoxy-D-arabino-heptulosonate-7-phosphate synthase, phenylalanine repressible	aroG	Metabolism; Building block biosynthesis; amino acids; phenylalanine
b1073	<i>flgB</i>	1.3	-1.1	2.0	0.7	flagellar component of cell-proximal portion of basal-body rod	flgBCDEFGHIJ	Metabolism; Biosynthesis; <u>flagellum</u> macromolecules (cellular constituent) cell processes; motility (incl. chemotaxis, energytaxis, aerotaxis, redoxaxis), cell structure;
b1074	<i>flgC</i>	1.3	-1.0	1.9	0.6	flagellar component of cell-proximal portion of basal-body rod	flgBCDEFGHIJ	Metabolism; Biosynthesis; <u>flagellum</u> macromolecules (cellular constituent), cell processes; motility (incl. chemotaxis, energytaxis, aerotaxis, redoxaxis), cell structure
b1075	<i>flgD</i>	1.5	-1.4	2.0	0.7	flagellar hook assembly protein	flgBCDEFGHIJ	Metabolism; Biosynthesis; <u>flagellum</u> macromolecules (cellular constituent); cell processes; motility (incl. chemotaxis, energytaxis, aerotaxis, redoxaxis), cell structure;
b1076	<i>flgE</i>	1.3	-1.5	2.0	0.7	flagellar hook protein	flgBCDEFGHIJ	Metabolism; Biosynthesis; <u>flagellum</u> macromolecules (cellular constituent); cell processes; motility (incl. chemotaxis, energytaxis, aerotaxis, redoxaxis), cell structure;
b1077	<i>flgF</i>	1.1	-1.1	1.4	0.5	flagellar component of cell-proximal portion of	flgBCDEFGHIJ	Metabolism; Biosynthesis;

						basal-body rod		<u>flagellum</u> macromolecules (cellular constituent); cell processes; motility (incl. chemotaxis, energytaxis, aerotaxis, redoxaxis), cell structure;
b1078	<i>flgG</i>	1.3	-1.0	1.5	0.7	flagellar component of cell-distal portion of basal-body rod	flgBCDEFGHIJ	Metabolism; Biosynthesis; <u>flagellum</u> macromolecules (cellular constituent), cell processes; motility (incl. chemotaxis, energytaxis, aerotaxis, redoxaxis), cell structure;
b1079	<i>flgH</i>	1.1	-0.7	1.3	0.6	flagellar protein of basal- body outer-membrane L ring	flgBCDEFGHIJ	Metabolism; Biosynthesis; <u>flagellum</u> macromolecules (cellular constituent), cell processes; motility (incl. chemotaxis, energytaxis, aerotaxis, redoxaxis) , cell structure; membrane, cell structure;
b1941	<i>fliI</i>	1.7	-0.6	2.0	0.7	flagellum-specific ATP synthase	fliFGHIJK (fliFp)	Metabolism; Biosynthesis; <u>flagellum</u> ; C & energy metabolism; ATP-H ⁺ motive force interconversion, cell structure
b1944	<i>fliL</i>	0.9	-0.3	1.2	0.3	flagellar biosynthesis protein	fliLMNOPQR (fliLp1)	Metabolism; Biosynthesis; <u>flagellum</u> macromolecules (cellular constituent); cell processes; motility (incl. chemotaxis, energytaxis, aerotaxis, redoxaxis) , cell structure;
b1945	<i>fliM</i>	1.6	-0.6	1.6	1.4	flagellar motor switching and energizing component	fliLMNOPQR (fliLp1)	Metabolism; Biosynthesis; <u>flagellum</u> macromolecules (cellular constituent), cell processes; motility (incl. chemotaxis, energytaxis, aerotaxis, redoxaxis), cell structure;
b2553	<i>glnB</i>	0.2	-0.6	0.5	0.2	regulatory protein P-II for glutamine synthetase	glnB	Metabolism; Building block biosynthesis; <u>amino acids</u> ; glutamine, information transfer; RNA related; transcriptional, regulation; <u>post- transcriptional regulation</u> ;

									inhibition/activation of enzymes
b2378	<i>lpxP</i>	0.6	2.2	0.8	0.4	palmitoleoyl-acyl carrier protein (ACP)-dependent acyltransferase	<i>lpxP</i>	Metabolism; Biosynthesis; <u>lipopolysaccharide</u> ; macro-molecules (cellular constituent) lipid A, cell processes; adaptation to thermal stress; cell structure; membrane, cell structure; surface antigens (ECA, O antigen of LPS)	
b0750	<i>nadA</i>	-0.5	-1.9	-0.9	-0.8	quinolinate synthase, subunit A	<i>nadA-pnuC</i>	Metabolism; Building block biosynthesis; NAD+ carrier	
b2476	<i>purC</i>	0.0	-1.5	-0.5	0.1	phosphoribosylaminoimidazole-succinocarboxamide synthetase	<i>purC</i>	Metabolism; Building block biosynthesis; <u>purine biosynthesis</u>	
b0494	<i>tesA</i>	0.0	1.1	0.1	-0.1	multifunctional acyl-CoA thioesterase I and protease I and lysophospholipase L1	<i>tesA</i>	Metabolism; Building block biosynthesis; <u>fatty acid & phosphatidic acid</u>	
b4069	<i>acs</i>	3.9	0.0	-1.5	-0.3	acetyl-CoA synthetase	<i>acs-yjcHG</i>	Metabolism; Building Block Biosynthesis; fatty acid and phosphatidic acid metabolism; acetate utilization; central intermediary metabolism;	
b1896	<i>otsA</i>	-2.4	-0.7	-2.6	-2.4	trehalose-6-phosphate synthase	<i>otsBA</i>	Metabolism; central intermediary metabolism; glucose metabolism, cell processes; osmotic stress adaptation;	
b1702	<i>pps</i>	0.0	-1.8	0.3	0.6	phosphoenolpyruvate synthase	<i>pps</i>	Metabolism; gluconeogenesis central intermediary metabolism	
b2091	<i>gatD</i>	1.8	-0.5	1.3	0.8	galactitol-1-phosphate dehydrogenase, Zn-dependent and NAD(P)-binding	<i>gatYZABCD</i>	Metabolism; C utilization; <u>galactitol</u>	
b2095	<i>gatZ</i>	2.0	-1.2	1.6	1.3	D-tagatose 1,6-bisphosphate aldolase 2, subunit	<i>gatYZABCD</i>	Metabolism; C utilization; <u>tagatose</u>	
b2096	<i>gatY</i>	1.8	-1.4	1.2	1.0	D-tagatose 1,6-bisphosphate aldolase 2, catalytic subunit	<i>gatYZABCD</i>	Metabolism; C utilization; <u>tagatose</u>	
b0161	<i>degP</i>	-0.4	1.9	-1.1	-0.9	serine endoprotease (protease Do), membrane-associated	<i>degP</i>	Metabolism; MacroMolecule Degradation; proteins/peptides/glycopeptides, cell processes; adaptation to thermal stress;	
b1844	<i>exoX</i>	0.0	1.1	0.3	0.2	DNA exonuclease X	<i>exoX-yobB</i>	Metabolism; macromolecule degradation; information transfer; <u>DNA repair, DNA degradation</u>	
b2193	<i>narP</i>	-0.1	1.1	0.0	-0.1	DNA-binding response regulator in 2-component regulatory system with NarQ or NarX	<i>narP</i>	Transcriptional Regulation; Information Transfer; <u>Transcriptional activator/repressor</u> ; C & energy metabolism; anaerobic respiration; RNA related;	

b3461	<i>rpoH</i>	-0.2	0.8	-0.5	-0.7	RNA polymerase, σ^{32} (σ^H) factor	rpoH	Information transfer; Transcriptional Regulation; RNA related; \square factors, anti- \square -factors; stimulon, cell processes; adaptation to stress; temperature extremes
b1663	<i>mdtK</i>	-0.5	0.4	-0.4	-0.4	multidrug efflux system transporter	mdtK	Transport; Multi Antimicrobial Extrusion (MATE) Family; <u>Electrochemical potential driven transporters</u> ; Porters (Uni-, Sym- and Antiporters); cell processes; cell structure; protection; drug resistance/sensitivity, membrane
b0929	<i>ompF</i>	1.2	-1.7	1.3	0.0	outer membrane porin 1a (la;b:F)	ompF	Transport; \square -barrel porins (Outer Membrane Porin (OMP) Functional Superfamily); <u>Solute:Sodium Symporter (SSS) Family</u> ; Channel-type Transporters; General Bacterial Porin (GBP) Family, cell structure; hydrophilic molecule
b3540	<i>dppF</i>	0.1	1.6	1.1	1.0	dipeptide transporter; ATP-binding component of ABC superfamily	dppABCDF	Primary Active Transporters; ATP-binding Cassette (ABC) Superfamily + ABC-type Uptake Permeases; C utilization; <u>amino acid transport</u> ; PP _i Bond (ATP, GTP, P ₂) Hydrolysis-driven Active Transporters; Metabolism
b3454	<i>livF</i>	0.1	-0.1	0.6	-1.5	leucine/isoleucine/valine transporter subunit; ATP-binding component of ABC superfamily	livKHMGF	Primary Active Transporters; Building block biosynthesis: (isoleucine/valine/leucine); amino acid transport/metabolism); PP _i Bond (ATP, GTP, P); ATP-binding Cassette (ABC) Superfamily + ABC-type
b3455	<i>livG</i>	0.1	-0.3	0.6	-1.6	leucine/isoleucine/valine transporter subunit; ATP-binding component of ABC superfamily	livKHMGF	Primary Active Transporters; building block biosynthesis: (isoleucine/valine/leucine); amino acid metabolism; PP _i Bond (ATP, GTP, P), ATP-binding Cassette (ABC) Superfamily + ABC-type
b1017	<i>efeU₂</i>	0.4	-0.9	0.6	1.1	C-terminal fragment of ferrous iron permease	efeBO	Transporter pseudogene

						(pseudogene)		
b1018	<i>efeO</i>	0.2	-0.6	0.7	0.9	component of a tripartite ferrous iron transporter	efeBO	Transport
b3915	<i>fieF</i>	-0.3	0.6	-0.1	-0.2	zinc transporter	fieF	Transport; Cation Diffusion Facilitator (CDF) Family; <u>Electro-chemical potential driven transporters</u> ; Porters (Uni-, Sym- and Antiporters); cell structure; membrane
b0763	<i>modA</i>	0.2	1.7	-0.4	-0.6	molybdate transporter subunit; periplasmic-binding component of ABC superfamily	modABC	Primary Active Transporters; Metabolism; Building Block Biosynthesis; cofactor, small molecule carrier (Mo); PPi Bond (ATP, GTP, P2) Hydrolysis-driven Active Transporters; ATP-binding Cassette (ABC) Superfamily + ABC-type
b0764	<i>modB</i>	0.5	1.2	0.0	-0.1	molybdate transporter subunit; membrane component of ABC superfamily	modABC (modAp1)	Transport; ATP-binding Cassette (ABC) Superfamily + ABC-type Uptake Permeases; Primary Active Transporters; PPi Bond (ATP, GTP, P2) Hydrolysis-driven Active Transporters; membrane component, cell structure;
b1469	<i>narU</i>	-1.3	-4.7	-1.1	-1.9	nitrate/nitrite transporter	narU	Transport; <u>Major Facilitator Superfamily (MFS)</u> ; N metabolism; Electrochemical potential driven; Porters (Uni-, Sym-, Anti-porters); cell structure; membrane;
b4067	<i>actP</i>	5.1	-0.2	-0.1	-0.3	acetate permease	acs-yjcHG (acsp1)	Transport; <u>Electrochemical potential driven transporters</u> ; Porters (Uni-, Sym-, Antiporters); cell structure; membrane
b2092	<i>gatC</i>	1.6	-1.7	1.1	0.4	galactitol-specific enzyme IIC component of PTS	gatYZABCD (gatYp)	Transport; (PEP-dependent PTS) C utilization; Group Translocators; Phosphotransferase Systems: PTS Galactitol (Gat) Family, cell structure; membrane; Metabolism;
b2093	<i>gatB</i>	1.7	-1.4	1.3	0.7	galactitol-specific enzyme IIB component of PTS	gatYZABCD (gatYp)	Transport; (PEP-dependent PTS) C utilization; Group Translocators; Phosphotransferase Systems: PTS Galactitol (Gat) Family, cell structure; membrane;

								Metabolism
b2094	<i>gatA</i>	2.0	-1.5	1.5	1.0	galactitol-specific enzyme IIA component of PTS	gatYZABCD (gatYp)	Transport; (PEP-dependent PTS) C utilization; Group Translocators; Phospho-transferase Systems: PTS Galactitol (Gat) Family, cell structure; membrane; Metabolism
b3453	<i>ugpB</i>	-0.9	1.6	-0.4	-0.4	periplasmic-binding component of an ABC superfamily glycerol-3-phosphate transporter	ugpBAECQ	Transport; Building block biosynthesis; fatty acid/ phosphatidic acid metabolism; C & energy metabolism; aerobic and anaerobic respiration; central intermediary metabolism: glycerol.
b4482	<i>yigE</i>	-0.1	2.4	-0.5	-0.5	predicted protein	yigE	
b1829	<i>htpX</i>	0.1	1.8	0.1	0.0	predicted endopeptidase	htpX	Cell processes; adaptation to thermal stress; cell structure
b3055	<i>htrG</i>	-0.4	0.9	-0.5	-0.6	predicted signal transduction protein (SH3 domain)	htrG-cca (htrGp2)	Cell Structure; membrane
b1113	<i>ycfS</i>	0.2	1.2	-0.3	-0.6	conserved protein	ycfS	Cell wall remodeling
b1255	<i>yciC</i>	0.3	0.8	-0.5	0.0	predicted inner membrane protein	yciC	Cell Structure; membrane
b1473	<i>yddG</i>	-0.2	-1.2	-0.2	-0.3	predicted methyl viologen efflux pump	yddG	Cell Structure; membrane
b1806	<i>yeaY</i>	0.2	1.5	-0.3	0.0	predicted lipoprotein	yeaZY (yeaZp)	Cell Structure; membrane
b3471	<i>yhhQ</i>	0.0	2.3	0.2	0.0	conserved inner membrane protein	yhhQ	Cell Structure; membrane
b3955	<i>yijP</i>	0.4	-0.9	0.6	0.4	conserved inner membrane protein	yijP	Cell Structure; membrane
b3095	<i>yqjA</i>	0.0	2.0	-0.3	-0.4	conserved inner membrane protein	yqjAB (yqjAp1)	Cell Structure; membrane
b1080	<i>flgI</i>	1.4	-0.5	1.7	0.9	predicted flagellar basal body protein	flgBCDEFGHIJ	Metabolism; Biosynthesis; flagellum; macromolecules (cellular constituent) cell processes; motility (incl. chemotaxis, energytaxis, aerotaxis, redox taxis), cell structure;
b1057	<i>yceJ</i>	-0.2	0.7	-0.3	-0.4	predicted cytochrome b561	yceJ	Metabolism; macromolecules (cellular constituent) biosynthesis; lg. molecule carriers; cytochromes
b2458	<i>eutD</i>	1.3	-0.4	1.0	0.9	predicted phosphotransacetylase subunit	eutDMPQST	Metabolism; C utilization; amines
b0925	<i>ycbB</i>	-0.5	1.4	-0.1	-0.2	predicted carboxypeptidase	ycbB	Metabolism; Macromolecule Degradation; proteins/peptides/glycopeptides
b1477	<i>yddM</i>	-0.7	-1.9	-0.5	-0.5	predicted DNA-binding transcriptional regulator	yddM	Regulation; Transcriptional Regulation, cell

								processes: defense/survival
b3082	<i>ygiM</i>	0.2	1.4	-0.1	-0.5	predicted DNA-binding transcriptional regulator	ygiMN	Regulation; Transcriptional Regulation;
b0377	<i>sbmA</i>	0.3	1.5	0.2	0.0	predicted transporter	sbmA-yaiW (sbmAp)	Transport; Primary Active Transporters; PP _i bond (ATP, GTP, P ₂) Hydrolysis-driven Active Transporters;
b0495	<i>ybbA</i>	-0.1	1.0	0.1	-0.1	predicted transporter subunit: ATP-binding component of ABC superfamily	ybbA	Transport; ATP-binding Cassette (ABC) Superfamily + ABC-type Uptake Permeases; Primary Active Transporters; PP _i Bond (ATP, GTP, P ₂) Hydrolysis-driven Active Transporters; ATP binding <u>cytoplasmic component</u>
b0805	<i>fiu</i>	-0.1	-1.4	-0.7	0.8	predicted iron outer membrane transporter	fiu	Transport Fe acquisition, cell processes; adaptation to stress; cell structure; membrane
b1902	<i>yecI</i>	0.2	2.8	0.6	0.7	predicted ferritin-like protein	ftnB (ftnBp2)	Transport; Fe acquisition; cell processes; adaptation to stress;
b4068	<i>yjcH</i>	5.4	-0.2	-0.1	-0.3	conserved inner membrane protein; acetate transport	acs-yjcHG (acsp2)	
b1843	<i>yobB</i>	0.0	1.9	-0.1	-0.2	conserved protein	exoX-yobB	
b0378	<i>yaiW</i>	0.0	1.2	0.2	0.1	predicted DNA-binding transcriptional regulator	sbmA-yaiW (sbmAp)	
b1056	<i>yceI</i>	0.0	1.4	-0.1	-0.5	secreted protein	yceI	
b1254	<i>yciB</i>	0.2	0.8	-0.4	-0.1	predicted inner membrane protein	yciB	
b1464	<i>yddE</i>	-0.2	-2.5	-0.1	-0.1	conserved protein	yddE	
b1535	<i>ydeH</i>	-0.2	2.2	-0.3	-0.6	conserved protein	ydeH	
b1846	<i>yebE</i>	0.0	3.0	-0.5	0.0	conserved protein	yebE	
b2602	<i>yfiL</i>	-0.8	1.4	-0.9	-0.7	predicted protein	yfiL	
b2761	<i>ygcB</i>	0.9	-0.5	1.5	0.4	conserved protein, member of DEAD box family	ygcB	
b1452	<i>yncE</i>	0.2	-6.0	-0.2	0.2	conserved protein	yncE	
b1436	<i>yncJ</i>	0.0	1.6	0.0	0.0	predicted protein	yncJ	
b3096	<i>yqjB</i>	-0.3	1.7	-0.5	-0.8	conserved protein	yqjAB (yqjAp1)	
b3097	<i>yqjC</i>	-1.8	0.3	-2.0	-2.0	conserved protein	yqjCDEK	
b3098	<i>yqjD</i>	-2.0	-0.2	-2.2	-2.0	conserved protein	yqjCDEK	
b3099	<i>yqjE</i>	-1.9	-0.3	-2.2	-2.0	conserved inner membrane protein	yqjCDEK	
b3207	<i>yrbL</i>	-1.2	0.3	-1.5	-1.0	predicted protein	yrbL	
b4217	<i>ytfK</i>	-0.5	0.7	-0.4	-0.4	conserved protein	ytfK	

Supplementary Table 4- Primers used for qRT-PCR

Name	Sequence (5' →3')
acsF	TCGTCGCTGCGCATTCT
acsR	CTCGTTGCCGATTTTTTTCC
acs probe	FAM 5' TTCCGTGGGCGAGCCAATT 3' BHQ
flgBf	GACGCCTCCTACCGCAGAA
flgBr	CGTTCGCGATCCATATCGA
flgB probe	FAM 5' ATTCCGGACCAGCCTTCGC 3' BHQ
lamB probe	FAM 5' CACAACAGAATGACTGGGAAGCTACCGATC 3' BHQ
lamBF	CGACACTAACGTGGCCTATTCC
lamBR	GCCATTCGATCAGGTTTTTACC
mdaB probe	FAM 5' CATGATGTCCGCATCGTTCGCG 3' BHQ
mdaBF	GCACACTGCGCGACCTT
mdaBR	ACTTCCGCTTTGACATCGTAGTC

Supplementary Table 5 – Sequencing primers

Name	Sequence (5' → 3')
acsseqF	ACCGTTACCGACTCGCATC
acsseqI1	TCGATACCTGGTGGCAGAC
acsseqI2	TGATGTGGTGGCGATTTATATG
acsseqR	GGAGCAGCCGTTTGTTCAG
cyaF1	TCGCCATCAACTTGTCTTTG
cyaI1	GCACTATCACCATCCGCTAA
cyaI2	TGGCAGCTCTACAAGAGTATCG
cyaI3	GTATAACCGCGGCCAAA
cyaI4	GAAACCGGGCGTTTCAAG
cyaR	CAGGCGGGTGAAACAGTC
glpKf	CGCACGTTTCGGGACTAC
glpKIf	CGGAACCACATACACACCAT
glpKr	CGCTGTAATATGACTACGGGACA
glpRf	AATGACGCGGATCGGCTA
glpRr	GGGTTAGCCGTGGGTTTAG
lamBseqf	TAAGCACCCACAAAACACA
lamBseqr	CTGCTGATAAACAGAGGACGAT
malTf1	AGGTTTCTGGCCGACCTTAT
malTf2	GAGCTGCCGAAAATCCAC
malTprom	ACAACGTTATCGCTAGTTTGC
malTr1	CGACAGTTCGCTATGGTTGA
malTr2	CGGTGCGGTTTAGTTTGATA
mglDf1	TGATTGCCAGTGCCTTCAC
mglDf2	ATCACATTGTTAAGATACTGTGAAA
mglDI	CCCCAGCAGTTCAACCATC
mglDr	GCTCTGGCGTCAGTTAACTTTG
mleF1 ¹	CTGAATGCTCTCAGGTGAGG
mleR1 ¹	CTCCACCGTTATGCTTCAC
ptaLf	CGGCGGTAACGAAAAGAGG
ptaLr	GGCAGTCAGAGATTCGATCC
ptaRf	CCTGCAGAGCTTCAACCTG

ptaRr	AAGGATTAATGCAAATTAAGAGAAT
ptsGLf	CCCGTCTGTTTCACATCGAC
ptsGLr	CAAACGGTACCAGGCAAC
ptsGRf	TCTTTACTGGCGTTGTGCTG
ptsGRr	ACCGGCACGTATCAATTC
rpoSF1 ²	CGGACCTTTTATTGTGCACA
rpoSF ²	CTGTAAACGGCCGAAGAAGA
rpoSR1 ²	TGATTACCTGAGTGCCTACG
spoTLf	GCCAGGAACAGCAAGAGC
spoTLr	TGCTCTTTATAAGCCAGTGC
spoTLr2	CCTTCCGGTGTGAAAACGTA
spoTRf	CCAGTACTACCGCACAAATCC
spoTRr	CGCAGATGCGTGCATAAC

Microarray comparative genomic hybridization of *Escherichia coli* from human and animal hosts

Kinnersley, M. , Rosenzweig, F. and W. Holben*

Abstract

Escherichia coli is a genetically diverse model prokaryote that is capable of both pathogenic and commensal associations with mammalian hosts. Because it is widespread and easily cultivated, it has also been commonly used as an indicator organism for tracking the origin of fecal water contamination. The successful application of *E. coli* for this purpose is predicated on the assumption that isolates recovered from contaminated water will harbor a genetic signature indicative of the host from which they originated. In this study, we compared two fingerprinting methods used for *E. coli* based microbial source tracking (repetitive element PCR and pulsed-field gel electrophoresis, or PFGE) with whole genome profiles obtained via microarray comparative genome hybridization for natural isolates of *E. coli* from humans, bear, cattle and deer. Our results show that patterns of gene presence or absence were more useful for distinguishing *E. coli* isolates from different sources than traditional fingerprinting methods, particularly in the case of human strains. In addition, a number of differences in genome composition that demarcated one host from another involved virulence-associated genes and occurred in regions of the *E. coli* chromosome previously shown to

be “hot spots” for the integration of horizontally acquired DNA. The data presented here suggest that despite the high level of diversity between isolates as measured by PFGE fingerprints, the human-derived strains that were examined comprise a distinct ecotype and as a result a number of potential library-independent source tracking markers for *E. coli* could be identified.

Introduction

The study of prokaryotic genome composition can provide useful insight into many aspects of microbial ecology and evolution, and help elucidate the general mechanisms by which relatively simple genomes evolve. The widespread application of whole-genome sequencing has rapidly expanded our understanding of the principles that govern microbial genome evolution, particularly at the subspecies level. It is now commonly accepted that prokaryotic species are comprised of a number of genetically distinct ecotypes that share a core genome content but vary in their complement of accessory genes (Snel, Bork et al. 2002; Lawrence and Hendrickson 2005). While overall genome size remains relatively constant within a species, the accessory genome content can be shaped over time by gene loss, gene formation and horizontal gene transfer (Snel, Bork et al. 2002). Comparative studies of sequenced genomes across many taxa have revealed that, in general, gene loss is the predominant mechanism by which prokaryotic genome content is modified over long evolutionary time periods (Snel, Bork et al. 2002).

Escherichia coli, one of the most comprehensively studied prokaryotes both in regard to population structure and genetics, is an ideal model organism with which to study how environment differentially influences genome composition at the sub-species level.

Because it is a familiar inhabitant of the lower intestinal tract of mammals and causative agent of diarrhegenic and urogenital illness, genetic variation in both commensal and pathogenic strains of *E. coli* has been studied extensively. As a result, it is widely accepted that the *E. coli* species is highly diverse, has a largely clonal population structure and consists of 5 major phylogenetic groups (A, B1, B2, D and E), with most pathogenic isolates residing in group B2 and groups A and B1 showing a rough association with carnivores and herbivores, respectively (Whittam 1996; Baldy-Chudzik, Mackiewicz et al. 2008; Jauregui, Landraud et al. 2008). It has been further suggested that much of the extant genetic diversity that exists between populations of *E. coli* may be the result of niche adaptation (i.e. to different host species) (Maynard-Smith 1991; Reeves 1992; Ihssen, Grasselli et al. 2007). On the other hand, an increasing number of studies have shown that only a small proportion of genetic variation in *E. coli* can reliably be attributed to host species effects (Souza, Rocha et al. 1999; Gordon 2001; Gordon and Cowling 2003).

The widespread use of *E. coli* as an indicator organism for tracing the animal origin of fecal water contamination is predicated on the assumption that isolates recovered from contaminated water will harbor a genetic signature indicative of the host from which they originated. Further, successful implementation of large-scale library-based source tracking programs necessitate that this signature is detectable using coarse measures of genome content such as rep-PCR and enzyme-based fingerprinting. While these methods have met with some success at grouping isolates from known host sources, the choice of *E. coli* as an indicator has been questioned on the basis of two observations: First, the total amount of genetic variation in *E. coli* that can be reliably attributed to host species is

rather low (~6%), and second, transmission into the external environment (as must occur for all isolates recovered from contaminated water) may result in the disproportionate survival or even reproduction of certain genotypes (Whittam 1989; Gordon and Lee 1999; Souza, Rocha et al. 1999; Gordon 2001; Gordon, Bauer et al. 2002; Barnes and Gordon 2004; Anderson, Whitlock et al. 2005; Ishii and Sadowsky 2008). Nevertheless, because monitoring of fecal coliforms is already an established part of many water quality assessment programs, there remains widespread interest in continuing the use of *E. coli* for source tracking.

If adaptation to a novel environment or host is the result of the gain, loss or change of one or a few genes, then it is plausible that this level of genetic variation, while certainly of adaptive significance, would not be captured using traditional fingerprinting methods. The recent identification of single locus host-specific genetic markers for *E. coli* from the feces of geese and ducks supports this hypothesis and suggests that it is possible to find similar markers for other host species, including humans (Hamilton, Yan et al. 2006; Yan, Hamilton et al. 2007). Thus, the extent to which routine fingerprinting methods accurately reflect genomic content at the gene level and the degree to which this variation is influenced by host species affiliation are largely unanswered questions. To address these issues, we used array comparative genome hybridization (a-CGH) to assess the gene content of *E. coli* from different host species and compared the results to strain relationships as determined by rep-PCR, PFGE and PCR-based ECOR phylogenetic group assignment. We found that patterns of gene presence or absence were more useful for distinguishing *E. coli* isolates from different sources than traditional fingerprinting methods, particularly in the case of those strains from human sewage.

In addition, a substantial number of genes that distinguish *E. coli* from different host species were identified that may be useful for the development of microbial source tracking markers in the future.

Materials and Methods

Strains, media and culture conditions

The strains used in this study were isolated as part of a larger microbial source tracking study in the Many Glacier region of Glacier National Park, USA. Fecal samples were collected from deposited material estimated to be less than 24 hours old using sanitized metal spatulas. Raw human sewage was collected directly from the outlet pipe leading from the Many Glacier Hotel to a sewage settling pond surrounded by an electric fence to exclude wildlife. Both sample types were sealed in sterile polypropylene containers, promptly placed on ice for transport back to the laboratory and processed within 6-8 hours of collection.

E. coli were isolated according to USEPA method 1603 (<http://www.epa.gov/nerlcwww/1603sp02.pdf>). Fecal samples were homogenized in 1X PBS (pH 7.4), diluted and filtered onto sterile 0.2 μ M membranes. Sewage samples were filtered directly without dilution. Membranes were placed onto modified mTEC agar (BBL) filtrate side up and incubated at 35°C for 2 hours followed by a 22-24 hour incubation at 44.5°C. Isolated red or magenta colonies were picked from the filters, struck onto nutrient agar (Difco), incubated overnight at 35°C and subjected to standard biochemical tests to confirm species identity (indole +, EC gas +, citrate - and oxidase -).

Confirmed *E. coli* were grown overnight in nutrient broth liquid culture and stored at -80°C in 20% glycerol.

ECOR group PCR

ECOR group PCR was performed according to the method of Clermont et al. with Qiagen HotStar taq using whole cell suspensions (prepared as described for rep-PCR fingerprinting, below) and a 3-step PCR program instead of the published 2-step rapid cycling protocol: 94°C for 15 min, 30 cycles of 94°C for 1 min, 55°C for 1 min. 30 sec, 72°C 1 min., followed by final extension at 72°C for 2 min (Clermont, Bonacorsi et al. 2000). PCR products were run on a 2% agarose gel using standard practices and stained with ethidium bromide.

Rep-PCR fingerprinting

Rep-PCR fingerprints were generated using the BoxA1R primer (5' CTACGGCAAGGCGACGCTGAC 3', (Versalovic 1998)) following the procedure of Dombek et al. available at (<http://www.ecolirep.umn.edu/boxfingerprint.shtml>) (Dombek, Johnson et al. 2000). Approximately 5 µl of *E. coli* colony material was resuspended in 100 µl of sterile PCR grade water and 2 µl of this suspension was added to a 50 µl PCR reaction containing 1X Gitscher buffer, 2.5 µl DMSO, 0.08 mg/mL BSA, 2.5 mM dNTPs, 0.3 µg BoxA1R primer, and 1.25 U Qiagen HotStar taq polymerase (Valencia, CA). PCR Reactions were cycled in a MJ Research PTC 100 thermocycler using the following program: 94°C for 15 minutes, 95°C for 2 minutes, 30 cycles of 94°C for 3 seconds, 92°C for 30 seconds, 50°C for 1 minute, 65°C for 8 minutes, followed by 1 cycle of 65°C for 8 minutes. Completed reactions were subjected to electrophoresis on 1.5%

agarose gels at 70V for 15 hours in 0.5X TAE buffer at 4°C and stained with 0.125 µg/mL ethidium bromide before photographing. Gel images were captured using a CCD digital camera system and the QuantityOne software v 4.5.2 (BioRad Inc, Hercules CA).

Pulsed-Field Gel Electrophoresis

The preparation of PFGE plugs was essentially as described in the Bio-Rad CHEF DR-II applications guide (Hercules, CA). A single colony of *E. coli* was grown in 25 mL of LB to an A_{600} of 0.7-0.8 and treated with 180 µg/mL chloramphenicol for 20 minutes. 1 mL of this suspension was pelleted and resuspended in 250 µl of buffer consisting of 10mM Tris, pH 7.2, 20 mM NaCl and 50 mM EDTA, mixed with an equal volume of 2% low-melt agarose and pipetted into plug molds. Plugs were treated with lysozyme (10 mM Tris, pH 7.2, 50 mM NaCl, 0.2% sodium deoxycholate, 0.5% sodium lauryl sarcosine, 1 mg/mL lysozyme) for 1 hour, washed and treated with proteinase K (100 mM EDTA, pH 8.0, 0.2% sodium deoxycholate, 1% sodium lauryl sarcosine, 1 mg/mL proteinase K) overnight, and washed with a buffer containing PMSF prior to storage at 4°C. Half-plugs were digested with 25 U of NotI, XbaI or ICeul (NEB) overnight prior to loading on an agarose gel. NotI and XbaI treated samples were run on a 1.2% gel at 12°C and 6V/cm for 16.6 hours with a 0.5-25 second switch time and 12 hours with a 30-60 second switch time. ICeul plugs were run on a 1% gel at 12°C at 6V/cm for 23 hours with a switch time of 5-200seconds.

Fingerprint analysis

Fingerprints were analyzed using the BioNumerics software package from Applied Maths (Sint-Martens-Latem, Belgium). Invitrogen's 1Kb ladder was used to standardize BoxA1R fingerprints across gels and the yeast chromosome, lambda and low range markers from NEB as well as the *H. wingei* chromosome standard from Bio-Rad were used for PFGE gels. In both sets of analyses, only bands inside of the useable range of the markers were considered for analysis. Pearson (curve-based) and Jaccard (band-based) similarity coefficients were used to construct similarity matrices using the lowest optimization settings that gave the highest scores between duplicate samples run on different gels on different days. These corresponded to 0.3% for the BoxA1R fingerprints, 1.0% for NotI PFGE, 0.9% for XbaI PFGE and 1.0% for ICeul PFGE. All dendrograms were constructed with the UPGMA clustering method. Fragment sizes for the ICeul fingerprints calculated using the aforementioned markers as a reference and added together to determine genome size estimates.

Nucleic acid extraction

Genomic DNA for a-CGH was extracted from overnight cultures grown in Luria Broth as described by Syn and Swarup (Syn and Swarup 2000) with slight modifications as follows: subsequent to DNA precipitation, spun pellets were treated with 50µg/mL DNase-free RNase A and incubated at 37°C for 30 minutes. Samples were then re-extracted once with phenol:chloroform (3:1), once with phenol:chloroform (1:1), twice with chloroform, reprecipitated and then resuspended in TE, pH 8.0.

Array-based Comparative Genome Hybridization (a-CGH)

The design and construction of the microarrays used in this study are as described elsewhere. Arrays were printed using a Virtek (formerly ESI) Vision Arrayer on Corning Gaps II aminosilane coated slides in 3X SSC (Waterloo, ON, Canada)(Kinnersley 2009).

Comparative Genome Hybridization was performed using the protocol developed by the J. Craig Venter Institute (<http://pfgre.tigr.org/protocols/protocols.shtml>) with two minor modifications as follows. First, 5 μ g of genomic DNA was sonicated to an average fragment length of 2-5 kb using a Branson Digital Sonifier at 11% amplitude for 1.1 seconds. Second, a final concentration of 0.5 mM, 1:1 aa-dUTP:dTTP labeling mixture was used in the Klenow reaction. Finally, prior to hybridization, slides were blocked in 5X SSC, 0.1% SDS, 1% Roche Blocking Reagent rather than BSA (Roche Applied Science, Mannheim, Germany).

Array image processing and statistical methods

Hybridized arrays were scanned using an Axon 4000B scanner (Molecular Devices, Sunnyvale, CA) and the resulting images were analyzed using a combination of GenePix Pro 6.0, the freely available TIGR TM4 software suite (www.tm4.org) and Microsoft Excel. Spots with an intensity: background ratio > 1.5 and overall intensity > 350 in the reference channel and an intensity:background ratio of > 1.0 in the experimental channel were considered acceptable for downstream processing. Local background was subtracted for each spot, the corresponding \log_2 ratios were normalized using total intensity normalization, and replicate spots were averaged using TIGR MIDAS. Genes that had missing data (i.e. unacceptable spots) for more than half of the samples were

excluded from the downstream analysis leaving 3993 ORFs in the final data set. Each sample was hybridized twice and the results averaged in Microsoft Excel after processing with MIDAS. A strict cutoff of \log_2 ratio > 0.9 or < 0.9 was applied for the determination of gene amplifications and gene absences, respectively. A full genome character matrix was created in Microsoft Excel in which each \log_2 value was replaced with a -1, 0 or 1 to indicate gene absence, presence or amplification. Hierarchical clustering and bootstrapping on this character matrix was conducted in TIGR MeV using the Euclidian distance measure, the average linking method and 100 bootstrap replicates.

Data archiving

Data will be available through the NIH GEO database.

Results

Bacterial Strains

E. coli isolates from the feces of humans, bear, white-tail deer and domestic cattle were collected in Western Montana as part of a larger rep-PCR based microbial source tracking study in Glacier National Park, USA. As the number of isolates that could effectively be processed using array-CGH was limited, we chose to maximize the diversity sampled within each host group and avoid clones by selecting 3 strains from each host type that had unique BOXA1R rep-PCR fingerprints. Source information, ECOR group assignments, plasmid content and genome size estimates are shown in Table 1.

Genome sizes estimates ranged from ~ 4.56-5.08 Mb , which is well within the range of typically reported for *E. coli* (Bergthorsson and Ochman 1995). No discernable relationship was found between genome size, plasmid content and host source.

Rep-PCR fingerprinting

In order to assess the ability of each fingerprinting method to accurately classify samples according to their host groups, we used two clustering approaches routinely used for microbial source tracking fingerprint assessment based on Pearson product-moment and Jaccard similarity scores. The former takes into account both the position and intensity of bands while the latter considers only band presence or absence. For each method, optimization settings (the degree to which fingerprints are allowed to shift in order to find the best match) were determined independently for each type of fingerprint by choosing the lowest setting that gave the highest similarity between duplicate samples run on different gels on different days.

In general, both the Pearson and Jaccard correlations were only moderately successful at grouping BoxA1R rep-PCR fingerprints by host source (Figure 1A). This result was not entirely unexpected as isolates were specifically chosen to maximize the rep-PCR fingerprint diversity within each host group. Pearson similarity scores ranged from 55% at the base of the dendrogram to 90.5% for the cluster uniting two of the three cow isolates, CI and CIII. Same-species isolates were successfully paired in three other instances: HI with HII, HIII with the human-derived laboratory reference strain K12-MG1655, and BI with BIII. In no case did all strains from the same host species form a three-member monophyletic cluster. Fingerprints from cow and deer *E. coli* had the

highest overall similarity (united at node "C,D"), followed by human isolates at node "H" and bear strains at the internal root. However, the Pearson similarity measure did reasonably well at grouping fingerprints from herbivores together and separating them from those of omnivores (bold face type, Figure 1A).

Jaccard similarity scores for the same BoxA1R fingerprints were smaller than those obtained with the Pearson correlation (48.2% to 70.8%). The corresponding dendrograms bore little similarity to one another in regard to topology with the exception of the consistent pairing of BI with BIII and K12-MG1655 strain with another human isolate.

To address whether the presence or absence of any single band could be considered diagnostic for a particular host group, bands were assigned to forty-two different classes according to size. Eight bands were shared by all isolates. No bands that were unique to the human, bear or deer strains were identified.

PFGE fingerprinting

To determine if the rep-PCR fingerprinting results were consistent with other coarse measures of genome composition, we performed PFGE on all twelve isolates. The advantage of using PFGE over other typing techniques is that it consistently produces highly discriminatory fingerprints that correlate well with established methods of determining evolutionary relationships such as MLST (Harbottle, White et al. 2006; Johnson, Arduino et al. 2007). While the application of PFGE to large microbial source tracking studies has been limited, it has consistently performed well at classifying

unknown samples by host source compared to other library-based techniques (Griffith, Weisberg et al. 2003; Myoda, Carson et al. 2003).

We implemented the same analyses used for the rep-PCR fingerprints on PFGE fingerprints generated with three different restriction enzymes: XbaI, NotI and ICeul. As a whole, none of the enzymes were able to reliably discriminate isolates from different host sources, regardless of the similarity measure used (Figure 1B, C and D). Furthermore, combining banding patterns for all three enzymes into a single measure using the composite data set function in BioNumerics did not improve resolution (data not shown). NotI and XbaI PFGE similarity scores were substantially lower than those found using rep-PCR (ranging from 17-70%) due to the high discriminatory power of this technique (Casarez, Pillai et al. 2007). Scores with the rare-cutter ICeul were somewhat higher as the corresponding fingerprints contained fewer bands that were highly conserved. The most cohesive grouping from a single source was that of the XbaI cow fingerprints clustered by Jaccard similarity in which (Figure 1 B). Conversely, the BI and BIII isolates showed the least overall similarity to the other isolates, and one or the other was frequently placed at the base of the dendrogram.

Band-matching analysis identified a total of 32 unique bands for XbaI, 21 for NotI and 14 for ICeul. However, none of these were diagnostic for host source.

Array-CGH

To better understand the relationship between diversity as measured by standard fingerprinting techniques and genome content at a finer scale, we performed microarray Comparative Genome Hybridization on all twelve isolates using the laboratory K12 strain

MG1655 as a reference. The primary advantage to using comparative genome hybridization over other measures of diversity is its ability to simultaneously measure the presence or absence of all *E. coli* K12 reference genes in the genome of interest.

However, genes unique to any of the wild isolates were not detected.

Whole-genome fingerprinting -Out of 4098 genes represented on the array, 3993 were reliably detected in at least three-quarters of the samples. To assess whether whole-genome "fingerprints" could be used to group isolates by host source, the Euclidian distance metric was applied to a 3993-member character matrix consisting of gene presence/absence data for all twelve wild strains. The resulting similarity values were used to construct a dendrogram that could then be compared to those generated by the other fingerprinting methods (as shown in Figure 2, versus Figure 1). Surprisingly, this simple approach reliably clustered all three of the human isolates together into a single, well-supported group. In addition, two out of three strains from both cow and deer clustered together (DI with DIII and CII with CIII). The remaining ruminant *E. coli* samples, DII and CII, showed the highest similarity to one another and were consistently placed in the same clade as the other cow and deer samples. The bear isolates were the least resolved: BI and BIII bore little resemblance to either the human or ruminant strains while BII repeatedly clustered with cow and deer. Thus, with this limited yet genetically diverse sample set, genome-wide gene presence/absence data appears to be better suited to distinguishing human, cow and deer *E. coli* from one another than traditional fingerprinting techniques, but is of limited utility for isolates derived from bear.

Genome characteristics shared by all wild isolates- To better understand what large-scale processes or environmental factors have shaped the genome content in our

isolates, we looked for global patterns of gene loss or amplification. A total of sixty-nine genes were scored as absent from all twelve natural isolates compared to the laboratory reference K12- MG1655 (see Figure 2, Table 2 and Supplementary Table 1). Over two-thirds (71%) of these are bacteriophage or bacteriophage-related. This observation is consistent with previously published reports of a-CGH on a variety of both commensal and pathogenic *E. coli* and likely represents acquisition of phage DNA by the K12 reference subsequent to its domestication in laboratory culture (Ochman and Jones 2000; Dobrindt, Agerer et al. 2003; Ihssen, Grasselli et al. 2007). The remaining 29% belong to a variety of functional categories including central intermediary metabolism, cell structure and transport (Table 2).

By contrast, no genes were uniformly amplified across all of the strains. Only 67 putative amplifications were detected in 43 different open-reading frames distributed across 10 of the isolates (Table 3, Supplementary Table 2). As was noted for missing genes, the majority of amplifications (70%) were also bacteriophage or transposon related and were primarily found only in the human isolates. Two non-repetitive DNA amplifications that we observed were the possible copy number increase of *holE* (the DNA polymerase III θ subunit) in isolate HIII and the apparent duplication of the putative oxidoreductase system *ydhTUXV* in strain DII.

Genes that differentiate isolates by host - To identify genes that contributed to the successful clustering of strains by host, we used K-means clustering implemented in TIGR MeV to search for those ORFs whose presence/absence pattern was diagnostic for at least two out of three strains from each animal. The results are displayed as hierarchical clustering diagrams in Figures 3 and 4. Out of 110 genes captured by this

analysis, 84 were “diagnostic” for the human strains, 21 for bear and 5 for cow. An additional four were identified that showed a unique pattern for two-thirds of deer isolates but are displayed separately as their inclusion in the dendrogram reduced the overall bootstrap support values for the deer cluster (Figure 4). In general, the smaller data set of putative diagnostic genes performed better at clustering isolates by source than the whole-genome profile. Specifically, all three deer isolates and two of the bear samples were successfully united with high bootstrap values using the smaller data set. This clustering did not change when the input order of the samples was randomized, confirming that the grouping was not a software-generated artifact. Interestingly, the inclusion of the four genes whose patterns were unique to the deer isolates negatively impacted the cohesiveness of the deer cluster. Thus, in regard to the reduced data set, the deer isolates may be united more by their dissimilarity to other strains than their similarity to one another.

The 84 genes with presence/absence patterns unique to the human isolates can be further divided into three main categories: (1) those that are amplified relative to or at the same level as the reference strain in all three human isolates but are absent in non-human samples, (2) those that are present in **two** out of three human isolates and absent in all others and (3) those that are absent in two out of the three human strains but are present in all others (Figure 3).

Twenty-three genes fall into the first category: increased or reference-level copy number for HI, HII and HIII. Eleven of these are detected at levels above those present in the reference. The majority of these, 10, are copies of the mobile insertion element IS1 (Figure 3). This result is not surprising in and of itself, given that transposition of IS1

elements in *E. coli* can be induced by a number of environmental cues including extended starvation and carbon limitation (Kharat, Coursange et al. 2006). In this case, however, the relative amplification of IS1 in all three human isolates is in marked contrast to the decrease of the element in the non-human strains. To determine whether IS1 elements in the non-human strains were present in lower copy numbers or were absent altogether, we performed PCR on all of the isolates using IS1-specific primers. Four strains (HI, HII, HIII and CI) in addition to the reference gave a positive result while the remaining isolates were negative (Supplementary Figure 1, panel A). These results are entirely consistent with the array data as isolate CI does possess a single copy of *insA* (*insA-3*) as shown in Figure 3. Aside from insertion elements, only one other gene (*yeaJ*, a predicted di-guanylate cyclase that is involved in down-regulating motility and initiating biofilm formation) shows the same pattern of amplification. Moreover, no copy number changes are detected for the genes immediately up- or downstream of *yeaJ*, suggesting that the duplication affects only *yeaJ*.

The remaining 12 genes in the first category are present in all three human isolates at K12 reference levels but absent or reduced in other strains. Six of these encode the CRISPR associated cascade genes (*cas1,2BCDE*) that constitute a primitive RNA-mediated immune system that may protect prokaryotes against bacteriophage and repetitive DNA infection (Figure 3) (Haft, Selengut et al. 2005). The absence of these cascade genes in all of the non-human samples was unexpected, as the core genes (*cas 1,2,3 and 4*) are widely distributed across prokaryotic genera and appear to be propagated via lateral gene transfer (Jansen, Embden et al. 2002; Haft, Selengut et al. 2005). While certainly not essential, strains that possess the CRISPR/*cas* loci may have

an adaptive advantage in certain environments over those that do not which further suggests that the human isolates are adapted to an environment that is distinct from the one experienced by the animal isolates.

Twenty ORFs fell into the second category of human-specific genes- those found in two out of the three human strains but absent in all other isolates. This subset contains a number of insertion element and prophage genes but also contains the functionally-relevant iron dicitrate transporter subunits Fec B, C and E. The presence of this locus in HI and HII and its absence in the other isolates was confirmed by PCR (Supplementary Figure 1). The *fec* locus is non-essential, but enables those strains that possess it to scavenge extracellular iron bound to citrate. Interestingly, the *fec* locus was also identified in four out of five human *E. coli* isolates during a screen of human-specific bands from a modified rep-PCR BoxA1R fingerprint digested with the restriction enzyme BamHI (unpublished results). Examination of the genome sequence of the K12 MG1655 reference confirmed that this operon is indeed flanked by BoxA1R repeats.

Finally, the third category of genes with unique distribution patterns in the human isolates consists of 39 ORFs from a variety of functional groups that are absent in two of the human strains but present in all other samples. The most remarkable feature of this subset is that it contains three large blocks of genes involved in the production of cellular structures for surface attachment: the exopolysaccharide colanic acid (M-antigen), the *E. coli* common pilus (ECP), and type 1 fimbriae. Each set of genes is located proximal to prophage loci present in MG1655 which suggests that these areas of the genome are recombinatorial "hot spots" (Figure 3) (Dobrindt, Agerer et al. 2003; Le Gall, Darlu et al. 2005).

In contrast to the large number of genes that were diagnostic for the human isolates, a relatively small number (30) were useful for distinguishing bear, cow and deer *E. coli* (Figure 4). In regard to the bear strains, twenty-one ORFs were absent in two out of three bear isolates whereas no genes that were diagnostic for all three could be found. BI and BIII both lacked the *yra IJK* operon (encoding putative fimbriae synthesis genes) and a series of unrelated genes in the vicinity of the transcriptional regulator Hsc62. BII and BI were only associated by the absence of a single gene of unknown function. An even smaller number of genes (5) had similar presence/absence patterns for the cow isolates. One, *ybaY*, a predicted outer membrane lipoprotein, was absent in all three cow strains. An additional three (*ybcV*, *nirD* and *yagK*) were absent only in BI and BIII but present in all of the other samples. Finally, four genes, three of which were prophage related, were uniquely found in DI and DIII while a fifth (*lacA*) was missing from both DII and DIII but was present in the other isolates. Taken as a whole, these results suggest that the presence or absence of certain genes can distinguish bear, deer and cow *E. coli* from one another, but that an even greater number appear to discriminate human isolates from non-human ones.

Discussion

E. coli is widely recognized as both a physiologically and genetically diverse species (Selander and Levin 1980; Hartl and Dykhuizen 1984; Whittam 1996; Ochman and Jones 2000). A portion of this variation is thought to be the result of niche adaptation (Selander and Levin 1980; Reeves 1992; Turner, Lewis et al. 1997; Gordon and Cowling 2003; Weissman, Chattopadhyay et al. 2006; Gauger, Leatham et al. 2007),

although it is not clear if adaptation to the particular intestinal environment of different host species plays a significant role (Souza, Rocha et al. 1999; Gordon 2001).

To better understand the relationship between host source and genome composition, we applied microarray comparative genome hybridization to *E. coli* isolated from human sewage, bear, cow and deer feces, and compared the results to those obtained with traditional fingerprinting methods.

Fingerprinting vs. a-CGH

Rep-PCR fingerprinting is a technique that is widely utilized in the construction of microbial source tracking databases. It generally produces fingerprint libraries that can be used to correctly classify isolates from known host sources about 60-90% of the time. On the other hand, the application of PFGE to microbial source tracking has been limited. PFGE is more commonly used in epidemiological tracking studies due to its high degree of precision and accuracy (Williams, Isaacs et al. 2000; Stoeckel, Mathes et al. 2004; Casarez, Pillai et al. 2007; Denny, Bhat et al. 2008). In practice, the number of unique fingerprints identifiable by PFGE is higher than that of other source tracking methods such as rep-PCR (Casarez, Pillai et al. 2007). Furthermore, PFGE patterns appear to correlate well with established methods of determining evolutionary relationships such as MLST (Harbottle, White et al. 2006; Johnson, Arduino et al. 2007). Thus, we expect that the relationship between PFGE fingerprint patterns more closely approximates the true genetic relationships between strains than the PCR-based fingerprinting techniques.

The twelve isolates used in this study exhibited a high level of genetic diversity as measured by both methods. Isolates from the same host source showed little similarity

to one another and no band classes that distinguished one host source from another could be identified using either technique. Given the high discriminatory power of PFGE, it seems clear that all of the strains collected from each host source represent distinct lineages and are not direct descendants of one another.

On the other hand, genome-wide gene presence/absence data was much more successful at grouping isolates by host source. Reducing the data set to include only those genes common to at least two out of three strains from each animal further improved clustering. Taken together, our results suggest that while much of the genome composition across isolates is conserved, there are a number of genetic differences that distinguish *E. coli* from different animal sources.

Genomic differences that distinguish isolates by source

The most striking feature uniting all three strains from a single source was the shared amplification of the insertion element IS1 in the human/sewage isolates. As the PFGE fingerprints for all three human strains were distinct, we cannot assume that this similarity is due to common ancestry. The movement of IS1 in the *E. coli* genome is not unusual, and while we cannot conclusively rule out the possibility that transposition events occurred after the initial isolation in the laboratory (although all isolates were handled in a similar fashion and were passaged in rich media as little as possible prior to storage at -80°C), IS insertion or deletion events frequently cause mutations of adaptive significance (Badia, Ibanez et al. 1998; Stentebjerg-Olesen, Chakraborty et al. 2000; Barker, Pruss et al. 2004; Schneider and Lenski 2004; Zhong and Dean 2004; Leatham, Stevenson et al. 2005; Fernandez, Gil et al. 2007; Gauger, Leatham et al. 2007). Thus,

to one another and no band classes that distinguished one host source from another could be identified using either technique. Given the high discriminatory power of PFGE, it seems clear that all of the strains collected from each host source represent distinct lineages and are not direct descendants of one another.

On the other hand, genome-wide gene presence/absence data was much more successful at grouping isolates by host source. Reducing the data set to include only those genes common to at least two out of three strains from each animal further improved clustering. Taken together, our results suggest that while much of the genome composition across isolates is conserved, there are a number of genetic differences that distinguish *E. coli* from different animal sources.

Genomic differences that distinguish isolates by source

The most striking feature uniting all three strains from a single source was the shared amplification of the insertion element IS1 in the human/sewage isolates. As the PFGE fingerprints for all three human strains were distinct, we cannot assume that this similarity is due to common ancestry. The movement of IS1 in the *E. coli* genome is not unusual, and while we cannot conclusively rule out the possibility that transposition events occurred after the initial isolation in the laboratory (although all isolates were handled in a similar fashion and were passaged in rich media as little as possible prior to storage at -80°C), IS insertion or deletion events frequently cause mutations of adaptive significance (Badia, Ibanez et al. 1998; Stentebjerg-Olesen, Chakraborty et al. 2000; Barker, Pruss et al. 2004; Schneider and Lenski 2004; Zhong and Dean 2004; Leatham, Stevenson et al. 2005; Fernandez, Gil et al. 2007; Gauger, Leatham et al. 2007). Thus,

the amplification IS1 in HI, HII and HIII may be indicative of ongoing adaptation to the human and/or the secondary sewage environment; IS1-mediated mutation has previously been shown to dramatically affect the ability of *E. coli* MG1655 to colonize the mouse digestive system by improving growth rate in cecal mucus and increasing its ability to catabolize certain sugars (Leatham, Stevenson et al. 2005; Gauger, Leatham et al. 2007).

The amplification of IS1 in the human strains is in marked contrast to their apparent absence in all but one of the non-human isolates. This is somewhat surprising as IS elements are thought to move horizontally between strains and IS1 in particular is present in over 90% of the isolates in the ECOR collection (a collection of 72 strains from a variety of animals and humans worldwide) (Sawyer, Dykhuizen et al. 1987; Hartl and Sawyer 1988). Thus, the lack of IS1 in the nonhuman isolates suggests that the distribution and/or opportunity for transmission of the element in their natural habitats is low. Additional information is necessary to determine whether the presence or absence of any particular IS1 element would be a useful marker for microbial source tracking.

Our results also show that all three human strains share a set of contiguous, functionally related genes, *cas1,2BCDE*, that are absent in all other strains tested. These genes are part of a recently described RNA-mediated defense system that protects the cell against viral infection (Barrangou, Fremaux et al. 2007; Brouns, Jore et al. 2008). While certainly not essential, the *cas* system is likely to be advantageous in certain niches and is often acquired horizontally from members of other prokaryotic genera (Haft, Selengut et al. 2005). Some *cas* genes appear to have indispensable roles in fundamental cellular processes such as replication and may be important for adaptation to novel environments (Haft, Selengut et al. 2005). The distribution of these genes in *E. coli* is largely

unstudied, but their associated CRISPR repeats occur in at least 15 out of 18 sequenced strains, nearly all of which are human derived (<http://crispr.u-psud.fr/crispr/CRISPRdatabase.php>). The presence of the *cas* genes only in HI, HII and HIII suggests that these strains have experienced an environmental niche conducive to acquisition and/or persistence of the *cas* loci that the non-human isolates have not. In this case, the human intestinal tract and the secondary sewage environment are equally likely candidates. Thus, the absence of the Cas proteins in the nonhuman isolates is intriguing and merits further investigation as either a human or sewage-specific marker.

Several multi-gene clusters also showed unique presence/absence patterns in the human vs. non-human strains: the ferric dicitrate iron transporter subunits *fecB, C and E* were present in 2/3 of the human strains but absent in the other isolates, while the colonic acid biosynthesis genes (*wca*), the type 1 fimbriae (*fim*) loci and the *E. coli* common pilus encoded by *matAB* were absent in 2/3 of the human isolates and present otherwise. The deletion of these genes as discrete units argues strongly for the involvement of repetitive or mobile DNA. In fact, all four of these transcription units are proximal to MG1655 prophage and tRNA loci which are known to be preferred integration sites for foreign DNA (Figure 1)(Cheetham and Katz 1995; Ochman and Jones 2000; Ochman, Lawrence et al. 2000). From this perspective, it is perhaps even more surprising that the *fec*, *fim*, *wca* and *mat* loci appear to be largely intact in all of the nonhuman isolates due to the high rate of recombination expected at these sites.

In regard to the adaptive significance of the gain or loss of these genes, it is interesting to note that the iron acquisition systems and fimbriae/pili production are both considered virulence associated traits in *E. coli* (Pouttu, Westerlund-Wikstrom et al.

2001; Wright, Seed et al. 2007; Dobrindt and Hacker 2008; Lloyd, Henderson et al. 2009). However, an increasingly large body of evidence suggests that they are also important as colonization factors in commensal and attenuated strains (Levin and Svanborg Eden 1990; Wold, Caugant et al. 1992; Grozdanov, Raasch et al. 2004; Hejnova, Dobrindt et al. 2005). Whether a gene enhances virulence or commensalism for an individual isolate is largely dependent on the environmental context (Hejnova, Dobrindt et al. 2005; Zdziarski, Svanborg et al. 2008). At least in the case of the *fim* genes, there is evidence that the fitness effects of fimbriation differ between host species (Bergsten, Wullt et al. 2005). Thus, the relationship between the environment and presence/absence of genes associated with both virulence and commensalism appears to be complex. However, as a group these genes seem to possess at least two desirable characteristics for *E. coli* microbial source-tracking markers: their distribution is variable and it may be determined largely by the environment rather than phylogenetic affiliation.

A comparatively small number of genetic differences distinguished nonhuman isolates from one another and most of these have unknown function or predicted function based solely on homology. BI and BII had the highest number of similarities and were united by the absence of two small but contiguous blocks of genes: *yraIJK* (a putative fimbrial protein and its associated chaperone) and *ybeR-hscC* (a set of six loci that contains a heat shock protein, a predicted tRNA ligase and a 6-phosphogluconate phosphatase). The *ybeR* gene region is located within ~4.5 Kb of an insertion element which suggests that, like many of the genes missing in the human isolates, its absence may be due to a transposition or recombination event. No repetitive or mobile DNA was found in the neighborhood of *yraIJK*, but its loss is consistent with the loss of fimbrial

genes in the human strains and reinforces the idea that, as a group, genes involved in attachment may prove to be useful as host source markers.

Surprisingly, very few commonalities were found between the three cow isolates and the three deer isolates. Only two genes with functional assignments (*lysP*, a lysine APC transporter, and *nirD*, a subunit of nitrite reductase) were absent from two of the cow strains and present otherwise while one gene, *lacA* (a galactoside O-acetyltransferase), was diagnostic for two out of three deer isolates. No obvious physiological or biochemical advantage or disadvantage to the loss of these proteins could be found- both LysP and NirD are functionally redundant, and the deletion of LacA has no discernable effect on the growth of *E. coli* in laboratory culture (Wilson and Kashket 1969; Lewendon, Ellis et al. 1995). Thus, further research will be necessary to determine if these genes are useful for the reliable identification of ruminant *E. coli*.

In the present study we have conducted a preliminary investigation of the effects of host species environment on genome content in *E. coli*. As further research is conducted to assess the broader applicability of potential markers identified from whole-genome comparisons, several additional variables will have to be considered. In our study, isolates were chosen to reflect a wide range of genetic diversity as measured by rep-PCR fingerprints without regard to phylogenetic group. The phylogenetic group affiliation of all three human isolates was later determined to be A and all of the cow and deer strains belong to group B1. The influence of ECOR phylogenetic group on genome content may have contributed to the cohesive clustering of the human isolates and isolate BI to the cow and deer strains. While the fact that some of the genetic characteristics that unite the human isolates (such as the loss of *fim* genes and the presence of the insertion

element IS1) have been documented to occur across several ECOR groups in other studies argues against phylogenetic group as the only factor influencing our strain relationships, we cannot say for certain that this is the case for all of the potential human markers (Sawyer, Dykhuizen et al. 1987; Zdziarski, Svanborg et al. 2008),

The effect of transition into the secondary environment on genome composition is also a confounding factor in identifying genome content that is truly reflective of adaptation to a particular host species. Even if host-specific markers for microbial source tracking can be successfully developed from fecal isolates, their potential to identify the origin of *E. coli* recovered from secondary environmental sources such as contaminated water is unclear. In this study, we have attempted to address this variable to some extent by characterizing isolates from untreated human sewage, as input from faulty septic systems or wastewater treatment facilities is the likeliest source of human-derived water contamination. However, it is considerably more difficult to recover analogous strains from wild or domestic animals and thus a direct comparison between human and animal *E. coli* secondary environment was not feasible.

Finally, the influence of DNA not represented in the K12 MG1655 chromosome (i.e unique to an individual isolate or contained on plasmids) is invisible to the type of analysis presented here. Considering the number of potentially diagnostic genes that were identified in regions of the chromosome known to accept horizontally acquired or infectious sequences (refs.), genomic or plasmid DNA that could not be measured on our arrays may be particularly rich in genes that distinguish *E. coli* from different host sources. Thus, the application of other molecular techniques such as suppression subtractive hybridization and sequencing to assess the content of the genomic

complement outside of the core genome is warranted (Diatchenko, Lau et al. 1996; Hamilton, Yan et al. 2006; Zheng, Yampara-Iquise et al. 2009).

Conclusion

We used microarray comparative genome hybridization to analyze the whole-genome content of *E. coli* isolated from human sewage and bear, cow and deer feces. Our results indicate that genomic "fingerprints" are superior to traditional fingerprinting methods for grouping genetically distinct *E. coli* isolates by host source. Many of the genes that we found that distinguished one host from the others were related to repetitive, mobile DNA (i.e insertion sequences) or involved in attachment, colonization and virulence which suggests that these classes of genes may be more likely to yield host source diagnostic loci. Overall, our results demonstrate the utility of high-throughput genomic techniques applied to natural isolates as powerful screening tools for the future development of library-independent microbial source tracking markers.

Literature Cited

- Anderson, K. L., J. E. Whitlock, et al. (2005). "Persistence and differential survival of fecal indicator bacteria in subtropical waters and sediments." Appl Environ Microbiol **71**(6): 3041-8.
- Badia, J., E. Ibanez, et al. (1998). "A rare 920-kilobase chromosomal inversion mediated by IS1 transposition causes constitutive expression of the *viaK-S* operon for carbohydrate utilization in *Escherichia coli*." J Biol Chem **273**(14): 8376-81.
- Baldy-Chudzik, K., P. Mackiewicz, et al. (2008). "Phylogenetic background, virulence gene profiles, and genomic diversity in commensal *Escherichia coli* isolated from ten mammal species living in one zoo." Vet Microbiol **131**(1-2): 173-84.
- Barker, C. S., B. M. Pruss, et al. (2004). "Increased motility of *Escherichia coli* by insertion sequence element integration into the regulatory region of the *flhD* operon." J Bacteriol **186**(22): 7529-37.
- Barnes, B. and D. M. Gordon (2004). "Coliform dynamics and the implications for source tracking." Environ Microbiol **6**(5): 501-9.

- Barrangou, R., C. Fremaux, et al. (2007). "CRISPR provides acquired resistance against viruses in prokaryotes." *Science* **315**(5819): 1709-12.
- Bergsten, G., B. Wullt, et al. (2005). "Escherichia coli, fimbriae, bacterial persistence and host response induction in the human urinary tract." *Int J Med Microbiol* **295**(6-7): 487-502.
- Bergthorsson, U. and H. Ochman (1995). "Heterogeneity of genome sizes among natural isolates of Escherichia coli." *J Bacteriol* **177**(20): 5784-9.
- Brouns, S. J., M. M. Jore, et al. (2008). "Small CRISPR RNAs guide antiviral defense in prokaryotes." *Science* **321**(5891): 960-4.
- Casarez, E. A., S. D. Pillai, et al. (2007). "Direct comparison of four bacterial source tracking methods and use of composite data sets." *J Appl Microbiol* **103**(2): 350-64.
- Cheetham, B. F. and M. E. Katz (1995). "A role for bacteriophages in the evolution and transfer of bacterial virulence determinants." *Mol Microbiol* **18**(2): 201-8.
- Clermont, O., S. Bonacorsi, et al. (2000). "Rapid and simple determination of the Escherichia coli phylogenetic group." *Appl Environ Microbiol* **66**(10): 4555-8.
- Denny, J., M. Bhat, et al. (2008). "Outbreak of Escherichia coli O157:H7 associated with raw milk consumption in the Pacific Northwest." *Foodborne Pathog Dis* **5**(3): 321-8.
- Diatchenko, L., Y. F. Lau, et al. (1996). "Suppression subtractive hybridization: a method for generating differentially regulated or tissue-specific cDNA probes and libraries." *Proc Natl Acad Sci U S A* **93**(12): 6025-30.
- Dobrindt, U., F. Agerer, et al. (2003). "Analysis of genome plasticity in pathogenic and commensal Escherichia coli isolates by use of DNA arrays." *J Bacteriol* **185**(6): 1831-40.
- Dobrindt, U. and J. Hacker (2008). "Targeting virulence traits: potential strategies to combat extraintestinal pathogenic E. coli infections." *Curr Opin Microbiol* **11**(5): 409-13.
- Dombek, P. E., L. K. Johnson, et al. (2000). "Use of repetitive DNA sequences and the PCR To differentiate Escherichia coli isolates from human and animal sources." *Appl Environ Microbiol* **66**(6): 2572-7.
- Fernandez, A., E. Gil, et al. (2007). "Interspecies spread of CTX-M-32 extended-spectrum beta-lactamase and the role of the insertion sequence IS1 in down-regulating bla CTX-M gene expression." *J Antimicrob Chemother* **59**(5): 841-7.
- Gauger, E. J., M. P. Leatham, et al. (2007). "Role of motility and the flhDC Operon in Escherichia coli MG1655 colonization of the mouse intestine." *Infect Immun* **75**(7): 3315-24.
- Gordon, D. M. (2001). "Geographical structure and host specificity in bacteria and the implications for tracing the source of coliform contamination." *Microbiology* **147**(Pt 5): 1079-85.
- Gordon, D. M., S. Bauer, et al. (2002). "The genetic structure of Escherichia coli populations in primary and secondary habitats." *Microbiology* **148**(Pt 5): 1513-22.
- Gordon, D. M. and A. Cowling (2003). "The distribution and genetic structure of Escherichia coli in Australian vertebrates: host and geographic effects." *Microbiology* **149**(Pt 12): 3575-86.

- Gordon, D. M. and J. Lee (1999). "The genetic structure of enteric bacteria from Australian mammals." Microbiology **145** (Pt 10): 2673-82.
- Griffith, J. F., S. B. Weisberg, et al. (2003). "Evaluation of microbial source tracking methods using mixed fecal sources in aqueous test samples." J Water Health **1**(4): 141-51.
- Grozdanov, L., C. Raasch, et al. (2004). "Analysis of the genome structure of the nonpathogenic probiotic *Escherichia coli* strain Nissle 1917." J Bacteriol **186**(16): 5432-41.
- Haft, D. H., J. Selengut, et al. (2005). "A guild of 45 CRISPR-associated (Cas) protein families and multiple CRISPR/Cas subtypes exist in prokaryotic genomes." PLoS Comput Biol **1**(6): e60.
- Hamilton, M. J., T. Yan, et al. (2006). "Development of goose- and duck-specific DNA markers to determine sources of *Escherichia coli* in waterways." Appl Environ Microbiol **72**(6): 4012-9.
- Harbottle, H., D. G. White, et al. (2006). "Comparison of multilocus sequence typing, pulsed-field gel electrophoresis, and antimicrobial susceptibility typing for characterization of *Salmonella enterica* serotype Newport isolates." J Clin Microbiol **44**(7): 2449-57.
- Hartl, D. L. and D. E. Dykhuizen (1984). "The population genetics of *Escherichia coli*." Annu Rev Genet **18**: 31-68.
- Hartl, D. L. and S. A. Sawyer (1988). "Why do unrelated insertion sequences occur together in the genome of *Escherichia coli*?" Genetics **118**(3): 537-41.
- Hejnova, J., U. Dobrindt, et al. (2005). "Characterization of the flexible genome complement of the commensal *Escherichia coli* strain A0 34/86 (O83 : K24 : H31)." Microbiology **151**(Pt 2): 385-98.
- Ihssen, J., E. Grasselli, et al. (2007). "Comparative genomic hybridization and physiological characterization of environmental isolates indicate that significant (eco-)physiological properties are highly conserved in the species *Escherichia coli*." Microbiology **153**(Pt 7): 2052-66.
- Ishii, S. and M. J. Sadowsky (2008). "*Escherichia coli* in the environment: Implications for water quality and human health." Microbes and Environments **23**(2): 101-108.
- Jansen, R., J. D. Embden, et al. (2002). "Identification of genes that are associated with DNA repeats in prokaryotes." Mol Microbiol **43**(6): 1565-75.
- Jauregui, F., L. Landraud, et al. (2008). "Phylogenetic and genomic diversity of human bacteremic *Escherichia coli* strains." BMC Genomics **9**: 560.
- Johnson, J. K., S. M. Arduino, et al. (2007). "Multilocus sequence typing compared to pulsed-field gel electrophoresis for molecular typing of *Pseudomonas aeruginosa*." J Clin Microbiol **45**(11): 3707-12.
- Kharat, A. S., E. Coursange, et al. (2006). "IS1 transposition is enhanced by translation errors and by bacterial growth at extreme glucose levels." Acta Biochim Pol **53**(4): 729-38.
- Kinnersley, M., Holben, W. , Adams, J. and F. Rosenzweig (2009). "Genomic analysis of an evolved polymorphism in *Escherichia coli*." in prep.
- Lawrence, J. G. and H. Hendrickson (2005). "Genome evolution in bacteria: order beneath chaos." Curr Opin Microbiol **8**(5): 572-8.

- Le Gall, T., P. Darlu, et al. (2005). "Selection-driven transcriptome polymorphism in *Escherichia coli*/*Shigella* species." Genome Res **15**(2): 260-8.
- Leatham, M. P., S. J. Stevenson, et al. (2005). "Mouse intestine selects nonmotile flhDC mutants of *Escherichia coli* MG1655 with increased colonizing ability and better utilization of carbon sources." Infect Immun **73**(12): 8039-49.
- Levin, B. R. and C. Svanborg Eden (1990). "Selection and evolution of virulence in bacteria: an ecumenical excursion and modest suggestion." Parasitology **100** **Suppl**: S103-15.
- Lewendon, A., J. Ellis, et al. (1995). "Structural and mechanistic studies of galactoside acetyltransferase, the *Escherichia coli* LacA gene product." J Biol Chem **270**(44): 26326-31.
- Lloyd, A. L., T. A. Henderson, et al. (2009). "Genomic Islands of Uropathogenic *Escherichia coli* Contribute to Virulence." J Bacteriol.
- Maynard-Smith, J. (1991). "The population genetics of bacteria." Proceedings of the Royal Society of London **245**: 37-41.
- Myoda, S. P., C. A. Carson, et al. (2003). "Comparison of genotypic-based microbial source tracking methods requiring a host origin database." J Water Health **1**(4): 167-80.
- Ochman, H. and I. B. Jones (2000). "Evolutionary dynamics of full genome content in *Escherichia coli*." EMBO J **19**(24): 6637-43.
- Ochman, H., J. G. Lawrence, et al. (2000). "Lateral gene transfer and the nature of bacterial innovation." Nature **405**(6784): 299-304.
- Pouttu, R., B. Westerlund-Wikstrom, et al. (2001). "matB, a common fimbrillin gene of *Escherichia coli*, expressed in a genetically conserved, virulent clonal group." J Bacteriol **183**(16): 4727-36.
- Reeves, P. R. (1992). "Variation in O-antigens, niche-specific selection and bacterial populations." FEMS Microbiol Lett **79**(1-3): 509-16.
- Sawyer, S. A., D. E. Dykhuizen, et al. (1987). "Distribution and abundance of insertion sequences among natural isolates of *Escherichia coli*." Genetics **115**(1): 51-63.
- Schneider, D. and R. E. Lenski (2004). "Dynamics of insertion sequence elements during experimental evolution of bacteria." Res Microbiol **155**(5): 319-27.
- Selander, R. K. and B. R. Levin (1980). "Genetic diversity and structure in *Escherichia coli* populations." Science **210**(4469): 545-7.
- Snel, B., P. Bork, et al. (2002). "Genomes in flux: the evolution of archaeal and proteobacterial gene content." Genome Res **12**(1): 17-25.
- Souza, V., M. Rocha, et al. (1999). "Genetic structure of natural populations of *Escherichia coli* in wild hosts on different continents." Appl Environ Microbiol **65**(8): 3373-85.
- Stentebjerg-Olesen, B., T. Chakraborty, et al. (2000). "FimE-catalyzed off-to-on inversion of the type 1 fimbrial phase switch and insertion sequence recruitment in an *Escherichia coli* K-12 fimB strain." FEMS Microbiol Lett **182**(2): 319-25.
- Stoeckel, D. M., M. V. Mathes, et al. (2004). "Comparison of Seven Protocols To Identify Fecal Contamination Sources Using *Escherichia coli*." Environmental Science & Technology **38**(22): 6109-6117.

- Syn, C. K. and S. Swarup (2000). "A scalable protocol for the isolation of large-sized genomic DNA within an hour from several bacteria." *Anal Biochem* **278**(1): 86-90.
- Turner, S. J., G. D. Lewis, et al. (1997). "A genomic polymorphism located downstream of the *gcvP* gene of *Escherichia coli* that correlates with ecological niche." *Mol Ecol* **6**(11): 1019-32.
- Versalovic, J. F., de Bruijn, J., and J.R. Lupiski (1998). Repetitive sequence-based PCR (rep-PCR) DNA fingerprinting of bacterial genomes *Bacterial Genomes- physical structure and analysis*. G. M. Weinstock. London, Chapman & Hall: 38-48.
- Weissman, S. J., S. Chattopadhyay, et al. (2006). "Clonal analysis reveals high rate of structural mutations in fimbrial adhesins of extraintestinal pathogenic *Escherichia coli*." *Mol Microbiol* **59**(3): 975-88.
- Whittam, T. S. (1989). "Clonal dynamics of *Escherichia coli* in its natural habitat." *Antonie Van Leeuwenhoek* **55**(1): 23-32.
- Whittam, T. S. (1996). Genetic Variation and Evolutionary Processes in Natural Populations of *Escherichia coli*. *Escherichia coli and Salmonella*. F. C. Neidhardt. Washington, D.C., ASM Press. **2**: 2708-2720.
- Williams, R. C., S. Isaacs, et al. (2000). "Illness outbreak associated with *Escherichia coli* O157:H7 in Genoa salami. *E. coli* O157:H7 Working Group." *CMAJ* **162**(10): 1409-13.
- Wilson, T. H. and E. R. Kashket (1969). "Isolation and properties of thiogalactoside transacetylase-negative mutants of *Escherichia coli*." *Biochim Biophys Acta* **173**(3): 501-8.
- Wold, A. E., D. A. Caugant, et al. (1992). "Resident colonic *Escherichia coli* strains frequently display uropathogenic characteristics." *J Infect Dis* **165**(1): 46-52.
- Wright, K. J., P. C. Seed, et al. (2007). "Development of intracellular bacterial communities of uropathogenic *Escherichia coli* depends on type 1 pili." *Cell Microbiol* **9**(9): 2230-41.
- Yan, T., M. J. Hamilton, et al. (2007). "High-throughput and quantitative procedure for determining sources of *Escherichia coli* in waterways by using host-specific DNA marker genes." *Appl Environ Microbiol* **73**(3): 890-6.
- Zdziarski, J., C. Svanborg, et al. (2008). "Molecular basis of commensalism in the urinary tract: low virulence or virulence attenuation?" *Infect Immun* **76**(2): 695-703.
- Zheng, G., H. Yampara-Iquise, et al. (2009). "Development of *Faecalibacterium 16S* rRNA gene marker for identification of human faeces." *J Appl Microbiol* **106**(2): 634-41.
- Zhong, S. and A. M. Dean (2004). "Rapid identification and mapping of insertion sequences in *Escherichia coli* genomes using vectorette PCR." *BMC Microbiol* **4**: 26.

Figures and Tables

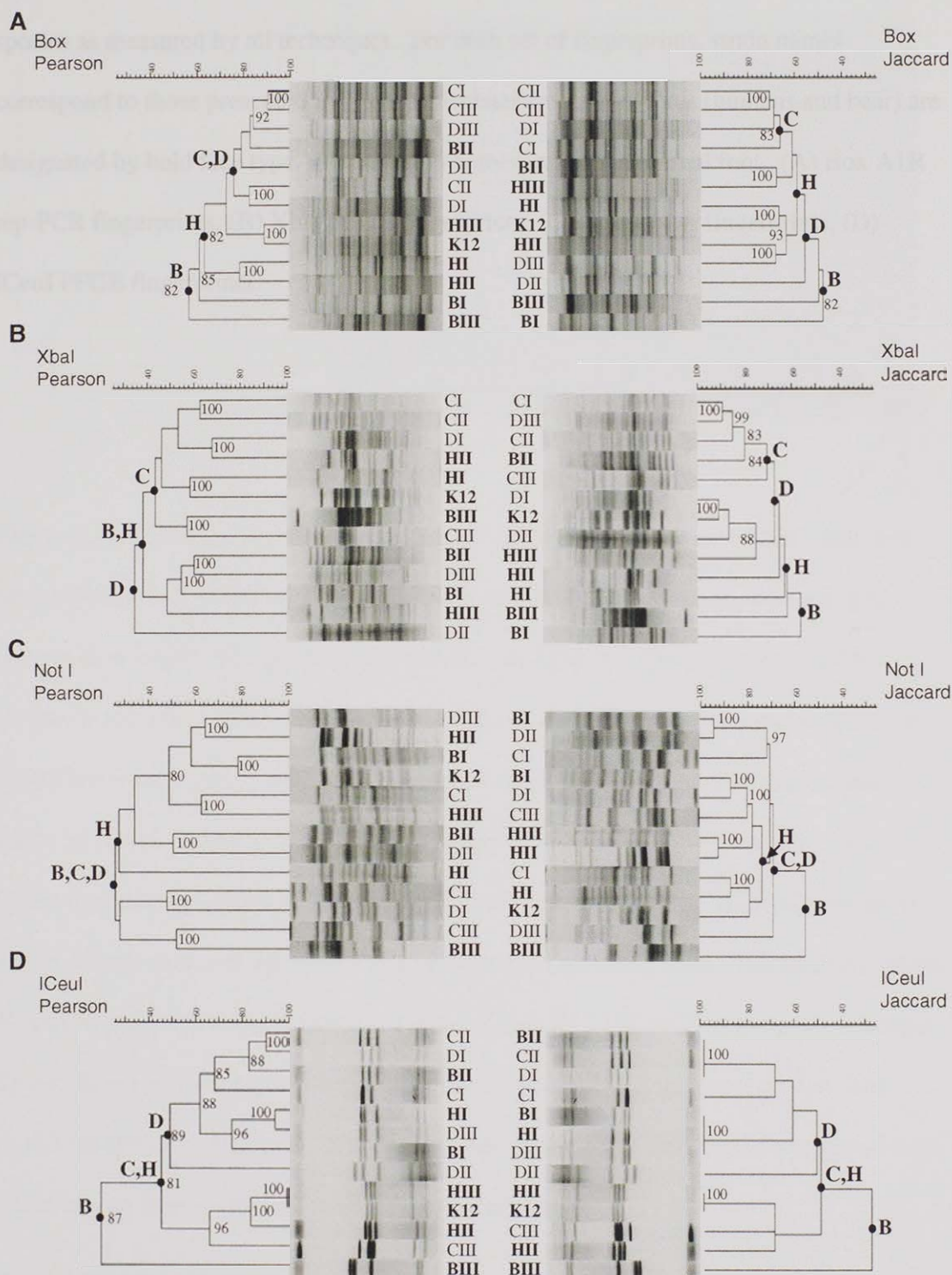


Figure 1. Clustering of rep-PCR and PFGE fingerprints using the Pearson and Jaccard similarity measures shows that there is little similarity among isolates from the same host species as measured by all techniques. For each set of fingerprints, strain names correspond to those presented in Table 1. Isolates from omnivores (humans and bear) are designated by bold face type. Strains from herbivores are in normal font. (A) Box A1R rep-PCR fingerprints, (B) XbaI PFGE fingerprints, (C) NotI PFGE fingerprints, (D) ICeul PFGE fingerprints.

Figure 2. Whole genome "fingerprints" of all 2,910 genes that were reliably detected across at least half of the isolates show better clustering of the human, cow and deer isolates than fingerprints generated by rep-PCR and PFGE. The genome for each isolate is displayed horizontally with the origin on the left. Genes that were considered shared are colored green, genes that showed an increase in copy number are red and genes for which there was no detectable change in copy number are shown in black. Genes that were not detected or considered "bad" are colored grey. Bootstrap values for 100 bootstrap replicates are shown next to their corresponding nodes. The locations of the 10 known prophages that are integrated in the *E. coli* K12 MG1655 genome are indicated by red arrows while the positions of three tRNA genes known to be integrated in "hot spots" are shown in blue type. Four gene absences that were observed in two out of three of the human isolates are indicated with black arrows.

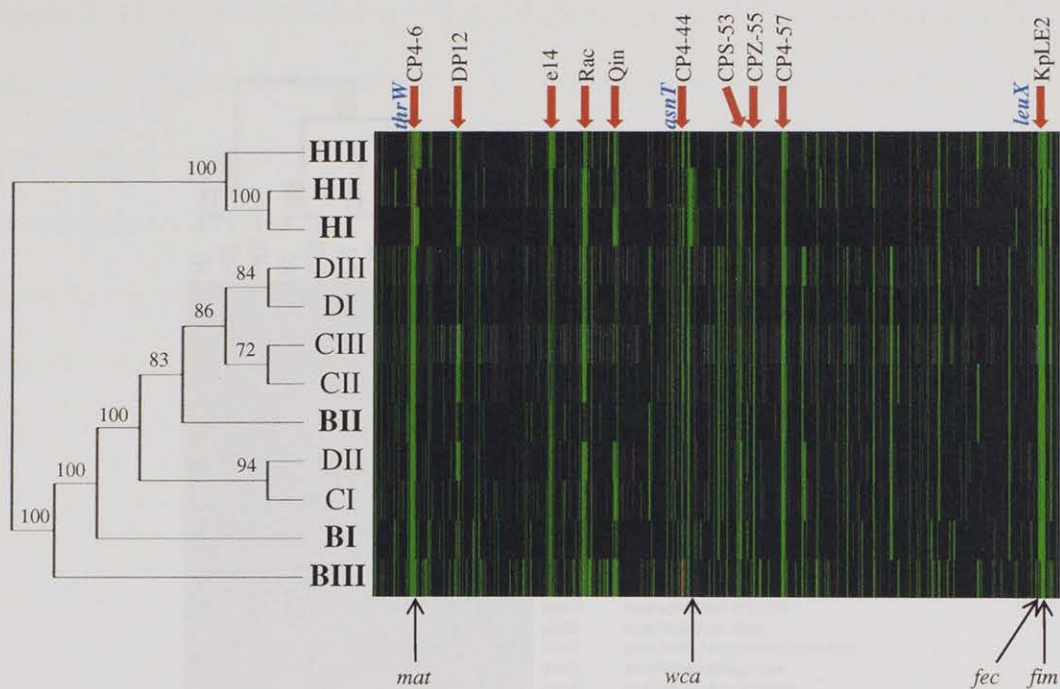


Figure 2. Whole genome “fingerprints” of all 3,993 genes that were reliably detected across at least half of the samples show better clustering of the human, cow and deer isolates than fingerprints generated by rep-PCR and PFGE. The genome for each isolates is displayed horizontally with the origin on the left. Genes that were considered absent are colored green, genes that showed an increase in copy number are red and genes for which there was no detectable change in copy number are shown in black. Genes that were not detected or considered “bad” are colored grey. Bootstrap values for 100 bootstrap replicates are shown next to their corresponding node. The locations of the 10 known prophage that are integrated in the *E. coli* K12 MG1655 genome are indicated by red arrows while the positions of three tRNA genes known to be integration “hot spots” are shown in blue type. Four gene absences that were observed in two out of three of the human isolates are indicated with black arrows.

Figure 3. Hierarchical clustering of genes that are diagnostic for at least two out of the

three housekeeping genes insA, insB, and insC. The dendrogram shows that the genes are clustered into groups based on their similarity. The bootstrap values are shown next to the branches. The genes are listed in the table below, along with their product names.

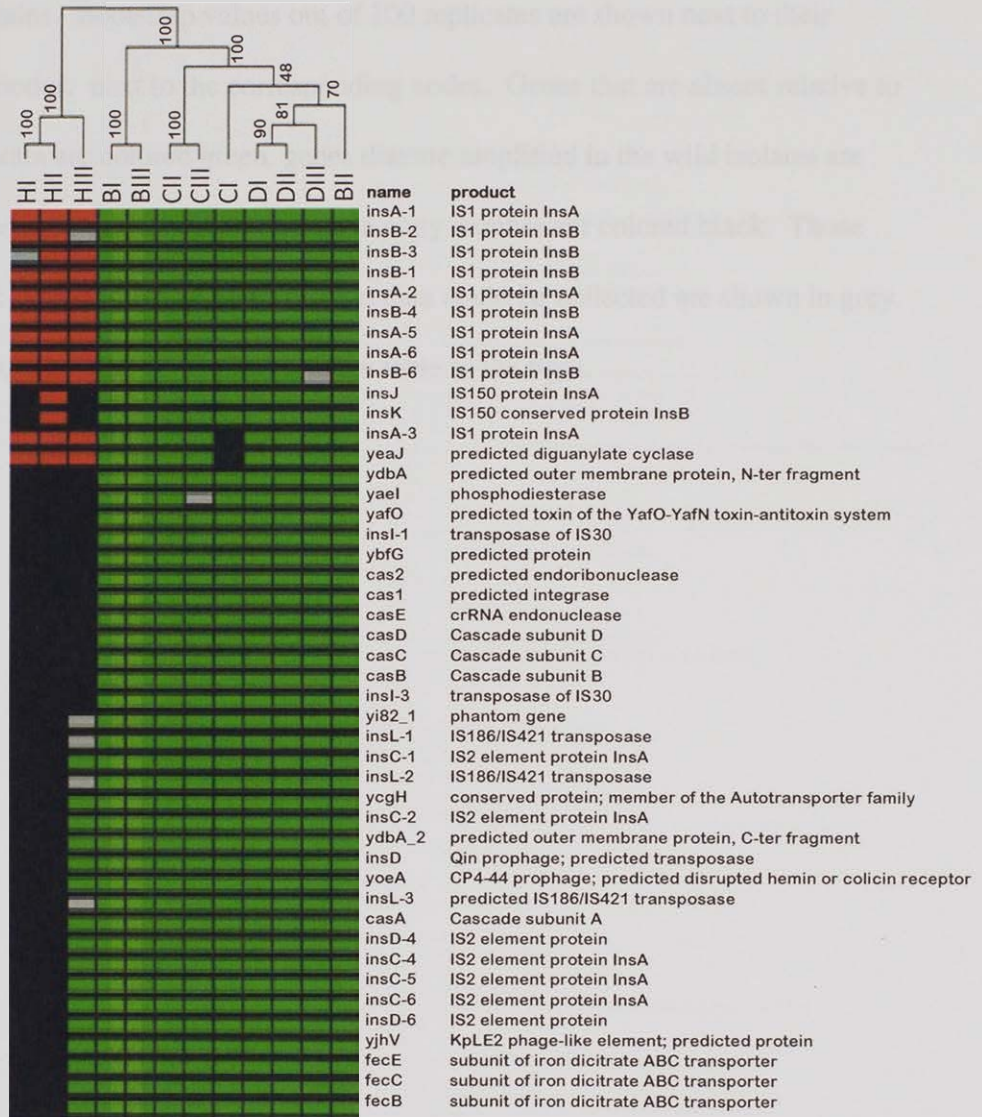


Figure 3

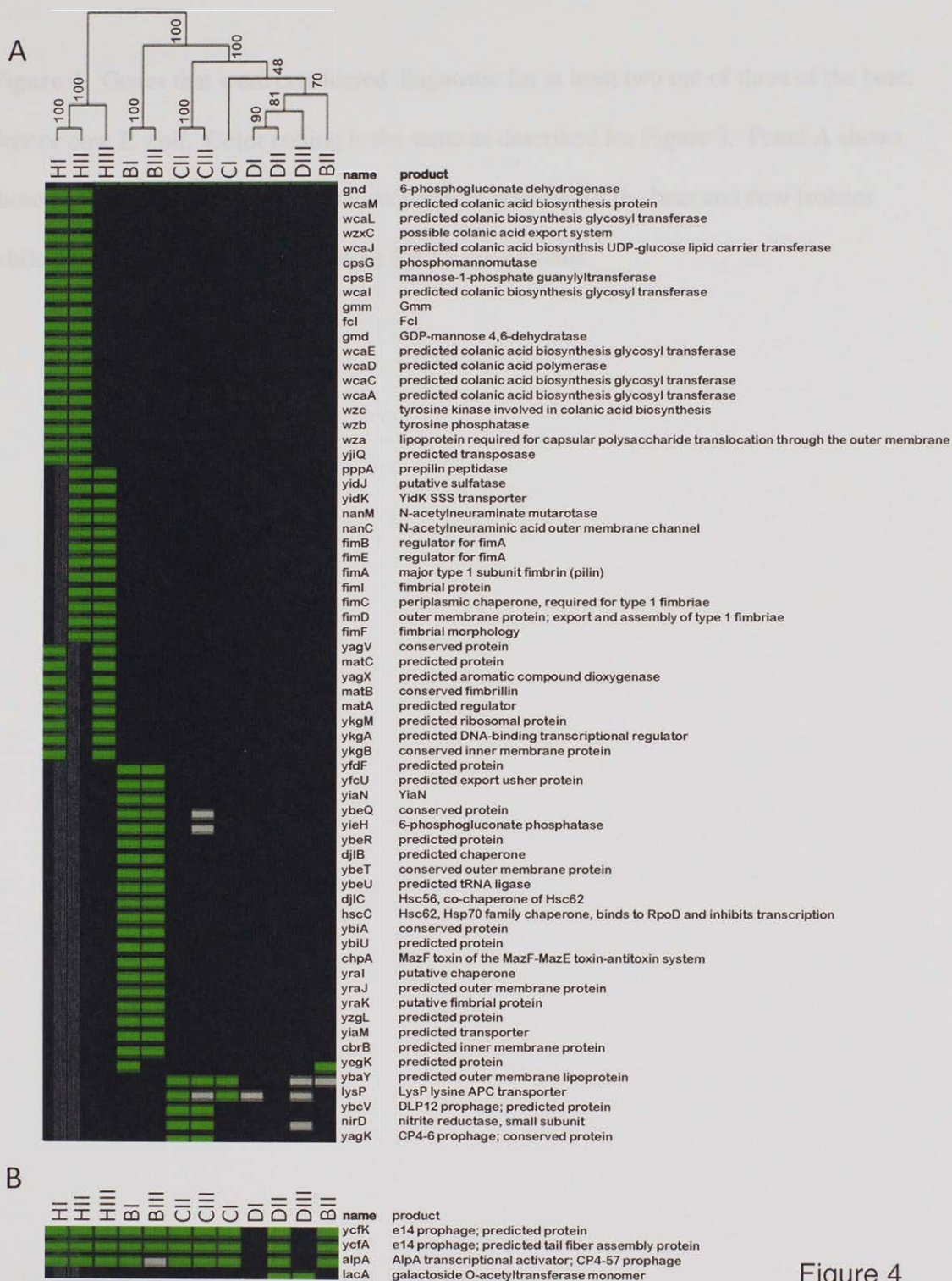


Figure 4

Figure 4. Genes that were considered diagnostic for at least two out of three of the bear, deer or cow *E. coli*. Color coding is the same as described for Figure 3. Panel A shows those genes that show unique presence/absence patterns for the bear and cow isolates while panel B shows diagnostic ORFs for the deer strains.

Gene ID	Host/Source	Presence	Diagnostic	Accession
B11	human/spawgs	A*	y	4,573
B13	human/spawgs	A*	y	4,582
B1	bear/feces	B1	y	4,692
B2	deer/feces	B2	y	4,676
B10	bear/feces	B2	n	4,679
C1	cow/feces	B1	n	4,938
C2	cow/feces	B1	y	4,664
C3	cow/feces	B1	y	4,756
D1	deer/feces	B1	y	4,588
D2	deer/feces	B1	y	4,942
D3	deer/feces	B1	y	4,670
K12	laboratory	A	y	4,615
MG1655	reference strain	A	y	44,679

Table 1. Bacterial Strains

strain	source	ECOR group	hemolysis	Estimated genome size (Kb)
HI	human/sewage	A	γ	4,603
HII	human/sewage	A*	γ	4,591
HIII	human/sewage	A*	γ	4,568
BI	bear/feces	D	γ	4,692
BII	bear/feces	B1	γ	4,616
BIII	bear/feces	B2	B	5,079
CI	cow/feces	B1	β	4,958
CII	cow/feces	B1	γ	4,661
CIII	cow/feces	B1	γ	4,756
DI	deer/feces	B1	γ	4,588
DII	deer/feces	B1	γ	4,982
DIII	deer/feces	B1	γ	4,670
K12 MG1655	laboratory reference strain	A	γ	4,615 (4,639)

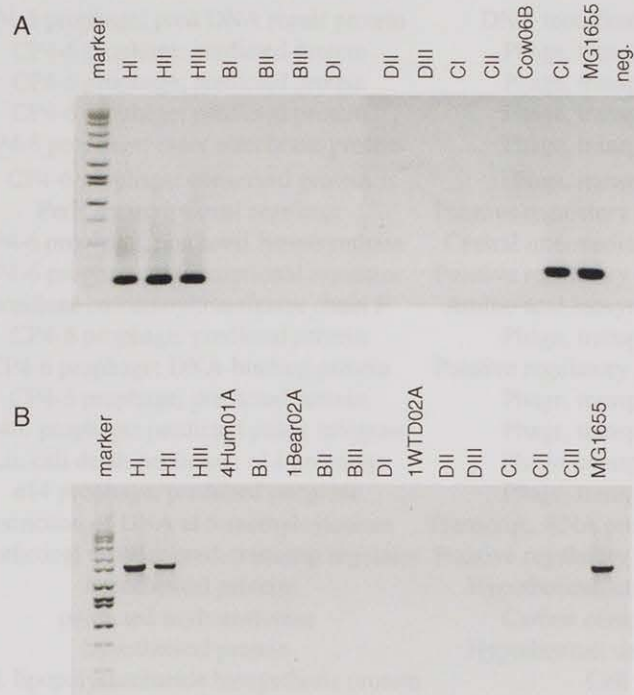
Table 2. Genes absent in all isolates

Functional Group	Number of genes (%)	Number of phage-related genes (%)
Amino acid biosynthesis and metabolism	1 (1.4%)	0
Carbon compound catabolism	2 (2.9%)	0
Cell processes (incl. adaptation, protection- phage related)	1 (1.4%)	1 (100%)
Cell structure	5 (7.2%)	1 (20%)
Central intermediary metabolism	5 (7.2%)	1 (20%)
DNA replication, recombination, modification and repair	1 (1.4%)	1 (100%)
Hypothetical, unclassified, unknown	7 (10.1%)	1 (14.2%)
Phage, transposon or plasmid	36 (52.2%)	36 (100%)
Putative chaperones	1 (1.4%)	0
Putative regulatory proteins	8 (11.6%)	7 (87.5%)
Transcription, RNA processing and degradation (phage related)	1 (1.4%)	1 (100%)
Transport and binding proteins	1 (1.4%)	0
total	69	49 (71%)

Table 3. Amplified genes

Functional Group	Number of genes (%)	Number of phage-related genes (%)
Amino acid biosynthesis and metabolism	1 (2.3%)	0
Cell processes (adaptation, protection)	3 (6.9%)	3 (100%)
Cell structure/transport	1 (2.3%)	0
DNA replication, recombination, modification and repair	3 (6.9%)	1 (33.3%)
Hypothetical, unclassified, unknown	13 (30.2%)	6 (46.1%)
Phage, transposon or plasmid	19 (44.2%)	19 (100%)
Putative regulatory proteins	1 (2.3%)	0
Structural proteins	1 (2.3%)	1 (100%)
Translation, post-translational modification	1 (2.3%)	0
total	43	30 (69.7%)

Supplementary Table 1. Genes from small isolates.



Supplementary Figure 1

Supplementary Figure 1. PCR with primers specific for the *fec* and *ins* loci confirm the comparative genome hybridization results. (A) *ins* PCR, (B) *fec* PCR.

Supplementary Table 1. Genes absent in all isolates.

locus tag	gene name	gene product	MultiFun category
b0245	ykfI	toxin of the YkfI-YafW pair	Phage, transposon, or plasmid
b0246	yafW	antitoxin of the YkfI-YafW pair	Phage, transposon, or plasmid
b0247	ykfG	CP4-6 prophage; pred DNA repair protein	DNA modification (phage related)
b0248	yafX	CP4-6 prophage; predicted protein	Phage, transposon, or plasmid
b0249	ykfF	CP4-6 prophage; predicted protein	Phage, transposon, or plasmid
b0250	ykfB	CP4-6 prophage; predicted protein	Phage, transposon, or plasmid
b0251	yafY	CP4-6 prophage; inner membrane protein	Phage, transposon, or plasmid
b0252	yafZ	CP4-6 prophage; conserved protein	Phage, transposon, or plasmid
b0254	perR	PerR transcriptional regulator	Putative regulatory proteins (phage related)
b0268	yagE	CP4-6 prophage; predicted lyase/synthase	Central intermediary met. (phage related)
b0272	yagI	CP4-6 prophage; transcriptional regulator	Putative regulatory proteins (phage related)
b0273	argF	ornithine carbamoyltransferase chain F	Amino acid biosynthesis and metabolism
b0276	yagJ	CP4-6 prophage; predicted protein	Phage, transposon, or plasmid
b0278	yagL	CP4-6 prophage; DNA-binding protein	Putative regulatory proteins (phage related)
b0279	yagM	CP4-6 prophage; predicted protein	Phage, transposon, or plasmid
b0281	intF	CP4-6 prophage; predicted phage integrase	Phage, transposon, or plasmid
b1139	lit	Lit, cell death peptidase; e14 prophage	Phage, transposon, or plasmid
b1140	intE	e14 prophage; predicted integrase	Phage, transposon, or plasmid
b1159	mcrA	restriction of DNA at 5-methylcytosines	Transcrip., RNA processing (phage related)
b1356	racR	hypothetical protein; pred. transcrip.regulator	Putative regulatory proteins (phage related)
b1358	ydaT	hypothetical protein	Hypothetical, unclassified, unknown
b1932	yedL	predicted acyltransferase	Carbon compound catabolism
b1998	yoeE	hypothetical protein	Hypothetical, unclassified, unknown
b2032	wbbK	pred. lipopolysaccharide biosynthesis protein	Cell structure
b2034	wbbI	β -1,6-galactofuranosyltransferase	Cell structure
b2037	rfbX	RfbX lipopolysaccharide PST transporter	Cell structure
b2039	rfbA	TDP-glucose pyrophosphorylase	Central intermediary metabolism
b2040	rfbD	dTDP-4-dehydrorhamnose reductase	Central intermediary metabolism
b2273	yfbN	predicted protein	Hypothetical, unclassified, unknown
b2274	yfbO	predicted protein	Hypothetical, unclassified, unknown
b2332	yfcO	predicted protein	Hypothetical, unclassified, unknown
b2333	yfcP	predicted fimbrial-like adhesin protein	Cell structure
b2336	yfcS	predicted periplasmic pilus chaperone	Putative chaperones
b2351	yfdH	CPS-53 prophage; glucosyl transferase	Cell structure (phage related)
b2352	yfdI	CPS-53 prophage; pred. membrane protein	Phage, transposon, or plasmid
b2442	intZ	CPZ-55 prophage; predicted integrase	Phage, transposon, or plasmid
b2443	yffL	CPZ-55 prophage; predicted protein	Phage, transposon, or plasmid
b2444	yffM	CPZ-55 prophage; predicted protein	Phage, transposon, or plasmid
b2445	yffN	CPZ-55 prophage; predicted protein	Phage, transposon, or plasmid
b2446	yffO	CPZ-55 prophage; predicted protein	Phage, transposon, or plasmid
b2447	yffP	CPZ-55 prophage; predicted protein	Phage, transposon, or plasmid
b2449	yffR	CPZ-55 prophage; predicted protein	Phage, transposon, or plasmid
b2450	yffS	CPZ-55 prophage; predicted protein	Phage, transposon, or plasmid
b2623	yfjH	CP4-57 prophage; predicted protein	Hypothetical, unclassified, unknown
b2625	yfjI	CP4-57 prophage; predicted protein	Phage, transposon, or plasmid
b2626	yfjJ	CP4-57 prophage; predicted protein	Phage, transposon, or plasmid
b2627	yfjK	CP4-57 prophage; conserved protein	Phage, transposon, or plasmid

b2628	yfjL	CP4-57 prophage; predicted protein	Phage, transposon, or plasmid
b2630	rnlA	CP4-57 prophage; RNase LS	Cell processes (phage-related)
b2632	yfjP	CP4-57 prophage; pred. GTP-binding protein	Phage, transposon, or plasmid
b2633	yfjQ	CP4-57 prophage; predicted protein	Phage, transposon, or plasmid
b2634	yfjR	CP4-57 prophage; pred. transcriptional regulator	Putative regulatory proteins (phage related)
b2636	yfjS	CP4-57 prophage; inner membrane protein	Phage, transposon, or plasmid
b2639	ypjL	CP4-57 prophage; pred. membrane protein	Phage, transposon, or plasmid
b2643	yfjX	CP4-57 prophage; pred. antirestriction protein	Phage, transposon, or plasmid
b2646	ypjF	CP4-57 prophage; toxin of YpjF-YfjZ system	Phage, transposon, or plasmid
b4295	yjhU	KpLE2 phage; pred. transcrip. regulator	Putative regulatory proteins (phage related)
b4296	yjhF	YjhF Gnt transporter	Carbon compound catabolism
b4297	yjhG	KpLE2 phage element; predicted dehydratase	Phage, transposon, or plasmid
b4298	yjhH	predicted lyase/synthase	Central intermediary metabolism
b4299	yjhI	KpLE2 phage; pred. transcrip. regulator	Putative regulatory proteins (phage related)
b4300	yjhJ	Putative sgc cluster transcriptional regulator	Putative regulatory proteins
b4301	sgcE	predicted epimerase	Central intermediary metabolism
b4303	sgcQ	putative nucleoside triphosphatase	Hypothetical, unclassified, unknown
b4304	sgcC	hypothetical phosphotransferase enzyme II	Transport and binding proteins
b4305	sgcX	KpLE2 element; pred. endoglucanase	Phage, transposon, or plasmid
b4306	yjhP	KpLE2 element; predicted methyltransferase	Phage, transposon, or plasmid
b4307	yjhQ	KpLE2 element; predicted acetyltransferase	Phage, transposon, or plasmid
b4308	yjhR	suppressor	Phage, transposon, or plasmid

Supplementary Table 2. Amplified genes

locus tag	gene name	amplified in	gene product	functional category
b0021	insB-1	HI,HII,HIII	IS1 protein InsB	Phage, transposon, or plasmid
b0022	insA-1	HI,HII,HIII	IS1 protein InsA	Phage, transposon, or plasmid
b0101	yacG	BII	DNA gyrase inhibitor YacG	DNA replication, modification and repair
b0258	ykfC	HIII	CP4-6 prophage; conserved protein	Phage, transposon, or plasmid
b0264	insB-2	HII,HIII	IS1 protein InsB	Phage, transposon, or plasmid
b0265	insA-2	HI,HII,HIII	IS1 protein InsA	Phage, transposon, or plasmid
b0274	insB-3	HII,HIII	IS1 protein InsB	Phage, transposon, or plasmid
b0275	insA-3	HI,HII,HIII	IS1 protein InsA	Phage, transposon, or plasmid
b0705	ybfL	HIII	predicted transposase; receptor protein	Putative regulatory prot. (transposon related)
b0988	insB-4	HI,HII,HIII	IS1 protein InsB	Phage, transposon, or plasmid
b1153	ymfQ	DIII	e14 prophage; conserved protein	Phage, transposon, or plasmid
b1272	sohB	CIII	predicted inner membrane peptidase	Translation, post-translational modification
b1372	stfR	BII	putative membrane protein	Phage, transposon, or plasmid
b1386	tynA	HIII	copper amine oxidase precursor	Amino acid biosynthesis and metabolism
b1404	insI-2	HIII	transposase of IS30	Phage, transposon, or plasmid
b1433	ydcO	DIII	predicted benzoate transporter	Cell structure, transport
b1562	hokD	CIII	Qin prophage; small toxic polypeptide	Cell processes (phage related)
b1669	ydhT	DIII	conserved protein	Hypothetical, unclassified, unknown
b1670	ydhU	DIII	predicted cytochrome	Hypothetical, unclassified, unknown
b1671	ydhX	DIII	Pred. 4Fe-4S ferredoxin-type protein	Hypothetical, unclassified, unknown
b1673	ydhV	DIII	predicted oxidoreductase	Hypothetical, unclassified, unknown
b1786	yeaJ	HI,HII,HIII	predicted diguanylate cyclase	Hypothetical, unclassified, unknown
b1842	holE	HIII	DNA polymerase III, theta subunit	DNA replication, modification and repair
b1893	insB-5	HII	IS1 protein InsB	Phage, transposon, or plasmid
b1894	insA-5	HI,HII,HIII	IS1 protein InsA	Phage, transposon, or plasmid
b1999	yeeP	BIII	CP4-44 phage; pred. GTP-binding prot.	Structural proteins (phage related)
b2000	flu	BIII	CP4-44 phage; biofilm autotransporter	Phage, transposon, or plasmid
b2001	yeeR	BIII	CP4-44 phage; pred. membrane protein	Hypothetical, unclass., (phage related)
b2002	yeeS	BII,BIII,CI,DII	CP4-44 phage; pred. DNA repair	DNA modification (phage related)
b2003	yeeT	BIII	CP4-44 phage; predicted protein	Hypothetical, unclass., (phage related)
b2004	yeeU	BIII	CP4-44 phage; antitoxin	Cell processes (phage related)
b2005	yeeV	BIII	CP4-44 phage; toxin of YeeV-YeeU	Cell processes (phage related)
b2356	yfdM	DI	CPS-53 phage; pred. methyltransferase	Hypothetical, unclass., (phage related)
b2357	yfdN	HI	CPS-53 prophage; pred. protein	Hypothetical, unclass.,(phage related)
b2359	yfdP	HI	CPS-53 prophage; pred. protein	Hypothetical, unclass.,(phage related)
b2777	ygcF	DI	conserved protein	Hypothetical, unclassified, unknown
b3444	insA-6	HI,HII,HIII	IS1 protein InsA	Phage, transposon, or plasmid
b3445	insB-6	HI,HII,HIII	IS1 protein InsB	Phage, transposon, or plasmid
b3557	insJ	HII	IS150 protein InsA	Phage, transposon, or plasmid
b3558	insK	HII	IS150 conserved protein InsB	Phage, transposon, or plasmid
b4215	ytfI	HII	predicted protein	Hypothetical, unclassified, unknown
b4275	yjgX	BIII	KpLE2 phage element; pred. protein	Hypothetical, unclass.,(phage related)
b4294	insA-7	HI	KpLE2 phage; IS1 repressor prot. InsA	Phage, transposon, or plasmid

Transcriptional profiling of *Escherichia coli* from different mammalian hostsAbstract

To determine the effect of host source on gene expression in natural populations of *E. coli*, we performed microarray-based transcriptional profiling on twelve isolates collected from the fecal material of four different mammalian species. A total of 86 significant differences in gene transcript levels were found that differentiated *E. coli* by host source. The majority of these involved loci of unknown function, transposable element genes, or genes related to cell structure and carbon utilization. The expression of insertion element IS1 was elevated in all of the human strains compared to the nonhuman animal isolates, an observation that was likely related to increased copy number of this element in the genome. By contrast, a single copy of the lactose permease LacY was present in all of the genomes, but showed higher transcript numbers only in the human-derived *E. coli*. Two virulence-associated gene clusters, *matABC* (which encodes the *E. coli* common pilus) and the *fimABCFHI* type 1 fimbriae locus, showed complex variations in transcript levels that were likely influenced by both gene deletion events and smaller-scale mutations that affected regulation. Overall, our results support the idea that *E. coli* from different environmental niches (i.e. different host species) may have characteristic gene expression patterns and demonstrate that microarray transcriptional profiling, when combined with gene copy number information, may provide useful clues about the adaptive response of *E. coli* to its natural habitat.

Introduction

The *Escherichia coli* species encompasses a diverse collection of genetically distinct ecotypes that can engage in a variety of lifestyles and is adapted to a number of different ecological niches. The determinants of niche preference in *E. coli* are poorly understood but likely involve a complex interaction between environment, genome content and gene regulation. In the mammalian intestine, it is well established that carbon source availability can have large effects on the colonization and persistence of different *E. coli* strains (Chang, Smalley et al. 2004; Fabich, Jones et al. 2008). The type and abundance of metabolic substrates that *E. coli* experiences in its natural habitat is not determined solely by host diet but can also be influenced by a number of other factors including (but not limited to) sex, age and the presence of other microbial species (Savageau 1974; Savageau 1983; Cummings and Englyst 1987).

Strain to strain variation in genome content also plays an important role in niche preference. This seems to be particularly evident for pathogenic strains of *E. coli* that colonize both inside and outside of the intestine. For example, uropathogenic isolates frequently possess specialized fimbrial adhesions that promote attachment to urinary epithelium while Shiga toxin-producing *E. coli* (STEC) are united in the possession of horizontally acquired toxin production genes (Shaikh and Tarr 2003; Ulett, Mabbett et al. 2007). However, recent genome comparisons of several commensal and pathogenic isolates has shown that many virulence associated genes are also present in the genomes of commensal isolates, which implies that regulation of gene expression is also an important determinant of niche preference (Dobrindt, Agerer et al. 2003).

To this end, a recent study done by Le Gall and colleagues showed that enteroinvasive *E. coli* (EIEC) and the obligate intracellular pathogen *Shigella* spp., which share a common niche, also exhibit a large degree of genomic and transcriptomic convergence (Le Gall, Darlu et al. 2005). Thus, variation in gene expression in these two phylogenetically distinct lineages appears to be under positive selection- a result that is not surprising considering the increasingly large number of experimental evolution studies that have demonstrated an indispensable role for regulatory mutations in the adaptive evolution of *E. coli* to novel laboratory environments (Pelosi, Kuhn et al. 2006; Cooper, Remold et al. 2008; Ferenci 2008; Kinnersley 2009).

In this study, we compared the transcriptional profiles of several genetically distinct *E. coli* isolates from 4 different animal sources to determine if adaptation to their natural host environments produced detectable patterns of convergence in gene regulation. We found that a number of gene expression differences were shared by strains from the same animal source despite the fact that they were genetically distinct from one another as measured by pulsed-field gel electrophoresis fingerprinting. Many of the transcription differences that were found were likely due to previously characterized variation in gene copy number, while others suggested the possibility of convergence at the regulatory level. Overall, these results demonstrate the potential of microarrays to detect gene expression variation in populations of *E. coli* that may be of adaptive significance in their natural habitat.

Materials and Methods

Strains, media and culture conditions

Strains used in this study are as described for Chapter 3 (Chapter 3, Table 1). Cultures for Biolog assays were grown on tryptone agar plates. Cultures for RNA extraction were grown at 37°C with shaking in M63 minimal media supplemented with 0.2% glucose to mid-late log phase ($A_{600} = 0.7-0.8$) (Silhavy 1984).

Carbon source utilization assays

The ability of each isolates to metabolize ninety-five different carbon sources was measured using the Biolog system (Hayward, CA). Cultures were struck onto tryptone agar plates from frozen glycerol stocks and grown at 37°C overnight. Colony material was harvested with a sterile cotton swab and resuspended in Biolog GN/GP inoculating fluid to a density equivalent to 59-63% transmittance as measured in a Spectronic 20 spectrophotometer. Sodium thioglycollate was added to a final concentration of 5mM and 150µl of the cell suspension was distributed into each well on the plate. Plates were incubated at 37°C for 36 hours at which time background-corrected average well color development was measured on a Molecular Devices SpectraMax M5 as the A_{590} of each test well minus the A_{590} of the negative control well. A reading of 0.15 or higher was considered a positive result. Each isolate was tested twice and discrepancies were resolved by visual inspection. Discrepancies that could not be resolved were scored as "unknown".

Nucleic acid extraction

Prior to RNA extraction, mid/late log phase cells were mixed with 1/10 volume of ethanol "stop solution" (5% phenol in 95% ethanol, <http://bugarrays.stanford.edu/protocols/rna/mRNAColi.pdf>), pelleted by centrifugation, snap-frozen in liquid nitrogen and stored at -80°C . RNA was subsequently extracted following the procedure developed by the Dunham lab found at <http://www.genomics.princeton.edu/dunham/MDyeastRNA.htm>. Cell pellets were lysed in an SDS/ hot phenol buffer (10 mM EDTA, 0.5% SDS, 10 mM Tris pH 7.4 mixed with an equal volume of acid phenol pH 4.5), vortexed and incubated at 65°C for 1 hour with mixing every 20 minutes. After separation of the phases using a phase-lock gel tube (5 Prime Inc., Gaithersburg, MD) the aqueous layer was extracted twice more with chloroform: isoamyl alcohol (24:1) and precipitated with absolute ethanol. Precipitations were centrifuged to pellet RNA, rinsed twice with RNase free 70% ethanol, dried and resuspended in 1X RQ1 RNase free DNase reaction buffer supplemented with 0.1U/ μl RQ1 RNase-free DNase (Promega, Madison WI). DNase reactions were allowed to proceed at 37°C for 1 hour after which the reactions were cleaned up using the Qiagen RNeasy Mini kit following the manufacturer's protocol. RNA quality was assessed on denaturing agarose gels and quantified using a spectrophotometer.

Array-based Transcriptional Profiling

Microarrays containing full-length open reading frames for 4,098 genes in the *E. coli* K-12 MG1655 genome were designed and fabricated as described in Chapter 3. Reverse transcription reactions and hybridizations were carried out following a protocol

developed by the J. Craig Venter Institute Pathogen Functional Genomics Center found at [ftp://ftp.jcvi.org/pub/data/PFGRC/pdf_files/protocols/M007.pdf](http://ftp.jcvi.org/pub/data/PFGRC/pdf_files/protocols/M007.pdf) with minor modifications. Briefly, 20 µg of total RNA was reverse transcribed with 9 µg of random hexamer, 0.83 mM 1:1 aa-dUTP:dTTP labeling mixture and 400U of Superscript III reverse transcriptase (Invitrogen, Carlsbad, CA). Instead of BSA, slides were blocked in 5X SSC, 0.1% SDS, 1% Roche Blocking Reagent prior to hybridization according to the protocol available at <http://www.genomics.princeton.edu/dunham/MDhomemadeDNA.pdf> (Roche Applied Science, Mannheim, Germany). Hybridized arrays were washed, and scanned using an Axon 4000B scanner (Molecular Devices, Sunnyvale, CA).

Image processing and statistical methods

Array images were analyzed using a combination of GenePix Pro 6.0, the TIGR TM4 software suite (available at <http://www.tm4.org/>) and Microsoft Excel. Image analysis and initial spot filtering was done in GenePix. Spots were considered acceptable if greater than 55% of the pixels in each channel were greater than 1 standard deviation above background, the regression ratio R^2 value was greater than 0.5, fewer than 10% of the pixels were saturated in either channel and the signal to noise ratio was greater than 2.5 for either channel. Spots that did not meet these criteria were excluded and the remaining data were converted to TIGR MEV format using Express Converter. Lowess normalization and replicate spots were averaging were done using TIGR MIDAS. Results were viewed and analyzed in TIGR MeV.

Three comparisons (including one dye-flip experiment) prepared from independent cultures grown on different days were performed for samples BII, BIII, CI, CII, DIII and HIII. Duplicate comparisons were used for samples BI, CIII, DI, HI and HII due to technical issues with the third replicate and DII was only successfully hybridized once.

Significance Analysis of Microarrays (Tusher, Tibshirani et al. 2001) (SAM) implemented in TIGR MeV was used to examine expression differences between strains using a multi-class comparison consisting of four groups. δ cutoffs were assigned by eye which gave a median false discovery rate (FDR) of 0% and a q-value of 0. The default settings for all other parameters were retained. The average (mean) \log_2 ratios for biological replicates were calculated after SAM analysis using Microsoft Excel.

Regulon Comparisons

Transcription unit, regulon and operon information was collated from the EcoCyc Database at <http://www.ecocyc.org> (Karp, Keseler et al. 2007). Predicted regulatory binding site information was obtained via TractorDB (<http://www.tractor.lncc.br>) (Gonzalez, Espinosa et al. 2005).

Results

Biolog Carbon Utilization Profiles

In order to determine whether the host intestinal environment might influence the ability of *E. coli* to metabolize different substrates, we assayed the growth of strains from bear, cow, deer and humans on various carbon sources using the Biolog GN2 microplate system (Figure 1, Supplementary Table 1). Out of the ninety-five different substrates

available in the assay, 23 were metabolized to some degree by all of the strains while 33 were consumed by none. Thirty-eight carbon sources showed some degree of variation in utilization pattern across isolates, however no strong relationships between carbon usage and host species were observed. Only two sets of isolates (BII and BIII, along with HI and HIII) showed any correspondence between a particular carbon substrate and host: BII and BIII were unique in their ability to utilize the seven substrates highlighted in green in Figure 1, and gentibiose was consumed by all of the isolates except HI, HII and BII. It is also worth noting that the two bear isolates, BII and BIII, used the largest number of carbon sources (51 and 45, respectively) out of all twelve strains tested.

Transcriptional profiling of isolates grown in glucose minimal media

Given that no strict relationship between host source and metabolic capability could be discerned from carbon utilization profiles, we also measured the transcriptional response of the same wild *E. coli* strains to cultivation in minimal media batch culture. While minimal media culture is certainly not representative of the complex growth conditions experienced by *E. coli* in its natural environment, this type of analysis has the potential to uncover strain-to-strain variation in transcriptional programs that reflect unique genetic composition or adaptation to host digestive biochemistry. To determine which genes had expression patterns that were diagnostic for host source, we performed a 4-class Significance Analysis of Microarrays (SAM): Samples were partitioned into four a priori groups (bear, cow, deer and human) and those loci that had significantly different expression in one of the groups versus the other three were identified.

4-class SAM analysis found 86 genes from 71 different transcription units whose relative expression levels distinguished at least one host source from the others (see Materials and Methods, Figure 2 and Supplementary Table 2). Ten of the major *E. coli* MultiFun functional groups were found, as depicted in Figure 3. The "unknown" category was the largest, containing 26% of the genes, followed by the transposon and phage related category (21%), cell structure (18%) and carbon utilization (14%).

All of the significant genes are shown as a clustered heat-map alongside their functional groups in Figure 2. The majority of significant expression differences depicted distinguish cow and human *E. coli* from those of bear and deer- nearly half (31 genes) were differentially expressed in the cow isolates and an additional 29 were considered diagnostic for the human strains. These diagnostic genes can be roughly divided into seven main clusters: four contain those genes with unique expression in the human strains (clusters H(1) through H(4)) and three contain genes differentially transcribed in the cow isolates (clusters C(1) through C(3))(Figure 2).

In general, the human-specific clusters are enriched for transposon and cell structure-related loci. Specifically, genes from insertion elements IS1, IS30 and IS186, along with the iron storage protein bacterioferritin (*bfr*) and the lactose permease (*lacY*) genes, show higher expression in the human strains than in the other nine isolates (Figure 2, clusters H(2) and H(4)) (Andrews, Harrison et al. 1989). Conversely, expression of structural genes necessary for the production of the *E. coli* common pilus (*matABC*), and type 1 fimbriae (*fim*) shows a trend toward down-regulation in human strains and increased relative expression in the majority of the bear, cow and deer isolates (Figure 2, clusters H(1) and H(3)). Despite the involvement of both the *mat* and *fim* genes in the

production of fimbriae, no common regulatory mechanism that could explain their unique transcription patterns in the human isolates was found (see Supplementary Table 3).

The gene clusters that distinguish the cow isolates from the other strains contain an interesting mixture of carbon utilization and information transfer loci. Three transcription units involved in central metabolism show unique expression trends for *E. coli* isolates from cow feces: enzyme IIA of the glucose-specific PTS permease (*crr*) and fructolysine-6-kinase *frlD* both show lower expression levels in all three cow strains while the TCA cycle genes, fumarase (*fumA*), succinate dehydrogenase (*sdhA,B*), succinyl CoA synthetase (*sucC*) and α -ketoglutarate dehydrogenase (*sucB*), all had somewhat higher transcript levels. Under the information transfer category, the genes for ribonuclease III (*rnb*) and the arginyl tRNA synthetase *argS* are up-regulated in the cow strains and down-regulated in the other isolates.

Expression profiling versus array-CGH

A comparison between the results presented here and the array comparative genome hybridization data for the same isolates presented in Chapter 3 revealed several similarities between genome composition and transcriptional response. In both cases insertion elements play a dominant role in distinguishing the human strains from the nonhuman isolates. CGH analysis shows that IS1 copy number is increased in HI, HII and HIII and the element is absent in all of the nonhuman strains with the exception of CI, which has a single copy of the IS1A gene, *insA-3*. Transcriptional profiling confirms that an increase in gene copy number is directly related to increased relative expression of the IS1 element in all three human strains. In addition, all five copies of the IS1 A gene

exhibit higher hybridization signal in the CI isolate- a result that is likely due to cross-hybridization of the *insA-3* transcript (Figure 2 cluster H(2) compared to Chapter 3, Figure 4). Interestingly, while HI exhibits increased transcription of IS1 genes compared to the nonhuman strains, its transcript levels are slightly lower than those found in HII and HIII. Thus, increased gene copy number is not the only determinant of IS1 expression in minimal media batch culture.

The *E. coli* common pilus encoded by the *mat* locus is another surprising example of how gene copy number influences gene transcription. In this case, the *mat* genes were absent in HI and HIII but were present in the remaining ten strains, including HII. Expression profiling shows that transcription of these genes is indeed lower in HI and HIII, but transcription is also repressed in HII despite the fact that this strain possesses the corresponding ORFs (Figure 2 cluster H(1), versus Figure 4, Chapter 3). This observation is in concordance with previously published data showing that although the *mat* genes are present in a wide variety of K12-derived and pathogenic *E. coli*, the vast majority of these isolates do not express them (Pouttu, Westerlund-Wikstrom et al. 2001).

Finally, a comparison of the a-CGH and expression data revealed a complex relationship between transcript levels and the gene presence/absence pattern for the type 1 fimbriae (*fim*) genes: they are absent in the genomes of HII and HIII, but their transcription is diminished in all three human isolates as well as BIII and CI (Figure 2 cluster H(3) versus Chapter 3 Figure 4). Thus, while absence of the *fim* genes explains their correspondingly low transcript levels for some of the strains, strain-to-strain variation in other factors that influence fimbrination clearly play a role in their expression as well.

In Chapter 3, the utility of a-CGH data for clustering of isolates by host source was investigated. We found that in comparison to traditional fingerprinting methods, patterns of gene presence and absence consistently performed better. To assess whether the same was true of transcriptional profiles, we repeated the analysis using the 86 significant genes returned in the SAM analysis. The resulting dendrogram with corresponding bootstrap values is displayed juxtaposed with the analogous a-CGH dendrogram from Chapter 3 (Figure 4). Both techniques consistently placed the three human isolates together but differed in their ability to group the bear, cow and deer strains.

Discussion

Under the premise that adaptation to a particular host environment involves a shift in the metabolic capability of *E. coli*, some researchers have attempted to deploy carbon source utilization patterns as an alternative to DNA fingerprinting for identifying potential sources of fecal water contamination (Meyer, Appletoft et al. 2005). In this study we assayed the ability of our twelve strains to grow on 95 different carbon sources using the Biolog system. We were unable to find any pattern of carbon source utilization useful for determining the animal origin *E. coli*. Thus, our data do not support the contention that carbon source utilization is likely to be a useful indicator of host source affiliation in *E. coli*.

As no discernable correlation between animal source and metabolic capability could be found, we also measured the transcriptional response of the same wild strains to cultivation in minimal media batch culture (i.e. a “common garden”) in an attempt to

uncover signatures of adaptation to the four host environments. Although it is difficult to concretely tie patterns of gene expression in aerobic glucose batch culture to the biochemically complex anaerobic environment of the rumen or intestine, several transcriptional differences were found that consistently distinguished cow and human *E. coli* from those of bear and deer which merit further discussion.

The majority of diagnostic transcripts were derived from transposon, cell structure and carbon utilization genes. One carbon utilization gene, lactose permease, was up-regulated in all of the human isolates compared to *E. coli* from other hosts. This observation appears to be inconsistent with the traditional view of how lactose is processed by the mammalian digestive system. The majority of ingested lactose is presumed to be absorbed by the host in the small intestine, where lactase levels are highest. Thus, little free lactose passes into the colon where *E. coli* normally resides (as reviewed in (Montgomery, Krasinski et al. 2007) (Savageau 1974). However, as all mammals age, they produce less lactase due to the decrease in milk consumption after weaning (Dahlqvist, Hammond et al. 1963; Buller, Kothe et al. 1990; Montgomery, Krasinski et al. 2007; Ingram, Mulcare et al. 2009). Despite this decline, humans consume dairy products throughout adult life. Given that approximately 65% of the worldwide population produces little to no lactase, it is likely that in many people, *E. coli* encounters undigested lactose on a regular basis. Thus, elevated expression of lactose permease, even under non-inducing conditions, may be a survival advantage for human-derived *E. coli*.

In the cow isolates, the glycine/serine/alanine transporter *cycA* and several genes in the TCA cycle showed higher transcript levels than were observed in the other strains,

while *crr*, a component of several sugar PTS permeases, was lower. As the regulation of genes involved in central metabolism and sugar transport is complex and dependant on a number of factors, it is difficult to speculate how this pattern of expression might reflect adaptation to the unique environment of the ruminant digestive system. Nevertheless, because these differences likely affect the uptake and utilization of metabolizable carbon and are common to three genetically distinct isolates, these data suggest that host-specific environmental differences between cattle and other animals that affect gene regulation in *E. coli* exist. A surprisingly similar pattern of gene expression (down-regulation of PTS genes and upregulation of the TCA cycle) was observed for the chemostat-adapted isolates discussed in Chapter 2 (Chapter 2, Supplementary Figure 3). In this case, the transcriptional shift was likely due to a global regulatory mutation, possibly in the stationary-phase sigma factor RpoS. Mutations in RpoS are frequently found in both natural and experimental populations of *E. coli* and can have profound effects on fitness under a variety of environmental conditions (Herbelin, Chirillo et al. 2000; Atlung, Nielsen et al. 2002; Kandror, DeLeon et al. 2002; Seeto, Notley-McRobb et al. 2004). Whether or not a similar mutation affects the cow isolates used in this study will require further investigation. Thus, in the absence of concrete knowledge as to why these genes are differentially expressed it is difficult to determine their utility as potential host-specific markers.

We also observed the parallel upregulation of several transposon genes in *E. coli* isolates from human sewage. All six copies of insertion element IS1 present in the MG1655 genome had elevated transcript levels across all three human strains. Comparative genome hybridization (as described in Chapter 3) suggests that the increase

in IS1 transcript is due, in part, to the presence of more chromosomal copies of the element in HI, HII and HIII. Gene copy number is not the only determinant of transcript level, however, as there is some variation in expression among the human strains (i.e. HI expresses less transcript of nearly all copies of *insB* than HII and HIII) and a complex assortment of both host and environmental factors are known to influence the regulation of transposition (Craig 1996; Rouquette, Serre et al. 2004). The absence of IS1 expression in the nonhuman isolates (with the exception of CI) confirms the CGH finding that the element itself is missing from the genome of these strains. This absence may be useful for development of an IS- based molecular marker for human *E. coli*, but precludes a discussion of what the “common garden” transcriptional profiles might reveal about differences in the adaptive environment of human versus nonhuman digestive systems.

Finally, transcripts for the *E. coli* common pilus (ECP, encoded by the *mat* locus) were detected in all of the nonhuman strains but were absent in the three isolates from human sewage. The ECP has been suggested to play an important role in adherence of both pathogenic and commensal *E. coli* to intestinal mucosa. The *mat* genes were found in approximately 93 - 96% of commensal isolates, 60% of which produce detectable pili under laboratory conditions (Rendon, Saldana et al. 2007). In light of these observations, the absence of the ECP locus in HI and HIII, and its low expression in HII is surprising. Furthermore, decreased transcription of *matABC* in HII may be the result of normal transcriptional control, but it may also be due to mutations affecting the *mat* locus promoter region: these genes reside in a region of the *E. coli* genome that is prone to insertion/deletion events (Chapter 3, Figure 2)(Dobrindt, Blum-Oehler et al. 2001; Parreira and Gyles 2003; Schouler, Koffmann et al. 2004). A similar phenomenon was

noted for the type 1 pilus (*fim*) gene cluster which was absent in two of human strains (HII and HIII) but down regulated in all three. Whether the *mat* and *fim* loci are disrupted by a deletion event or merely repressed, the behavior of these genes in the human isolates is unusual. Given that the ECP and type 1 pili are widely distributed in the species as a whole, one possible explanation is that the human isolates are adapted life in the extraintestinal secondary environment (i.e. sewage) in which the costly production of structures for epithelial attachment is unnecessary (Hahn, Wild et al. 2002). If this hypothesis is correct, our data suggest it may be possible to develop markers for the detection of sewage-adapted *E. coli*- a prospect that would greatly simplify the process of monitoring for environmental wastewater contamination.

Conclusion

In this study we performed transcriptional profiling on wild *E. coli* strains from bear, cow, deer and humans grown in a "common garden" condition. Our objective was to use expression differences manifest during laboratory culture to gain insight into the adaptive response of each group of *E. coli* to its natural habitat (i.e. the intestinal tract of different mammalian hosts). The types of adaptive mutation we reasonably expected to measure included, but were not limited to, gene duplications, gene deletions and regulatory mutations resulting in altered transcription initiation or constitutivity. Combined with knowledge of genome composition gained through microarray comparative genome hybridization, we were able to identify differences in the expression of over eighty genes that distinguished *E. coli* of one host from that of another based on a 4-class significance analysis of microarrays (SAM). Several of these differences could

be tied back to large scale changes in gene copy number, while others suggested the existence of small-scale mutations affecting gene regulation. Variation in the expression of one set of genes in particular, the *mat* locus, indicated that *E. coli* collected from human sewage may, in fact, be adapted to life in the extra-host environment. Overall, our results demonstrate the utility of microarray transcriptional profiling as a tool for exploring adaptive responses of *E. coli* to its natural habitat by assaying parallel changes in gene expression measured during growth in laboratory culture.

Literature Cited

- Andrews, S. C., P. M. Harrison, et al. (1989). "Cloning, sequencing, and mapping of the bacterioferritin gene (bfr) of *Escherichia coli* K-12." J Bacteriol **171**(7): 3940-7.
- Atlung, T., H. V. Nielsen, et al. (2002). "Characterisation of the allelic variation in the rpoS gene in thirteen K12 and six other non-pathogenic *Escherichia coli* strains." Mol Genet Genomics **266**(5): 873-81.
- Buller, H. A., M. J. Kothe, et al. (1990). "Coordinate expression of lactase-phlorizin hydrolase mRNA and enzyme levels in rat intestine during development." J Biol Chem **265**(12): 6978-83.
- Chang, D. E., D. J. Smalley, et al. (2004). "Carbon nutrition of *Escherichia coli* in the mouse intestine." Proc Natl Acad Sci U S A **101**(19): 7427-32.
- Cooper, T. F., S. K. Remold, et al. (2008). "Expression profiles reveal parallel evolution of epistatic interactions involving the CRP regulon in *Escherichia coli*." PLoS Genet **4**(2): e35.
- Craig, N. L. (1996). Transposition. *Escherichia coli* and *Salmonella*. F. C. Neidhardt. Washington, D.C, ASM Press. **2**: 2339-2362.
- Cummings, J. H. and H. N. Englyst (1987). "Fermentation in the human large intestine and the available substrates." Am J Clin Nutr **45**(5 Suppl): 1243-55.
- Dahlqvist, A., J. B. Hammond, et al. (1963). "Intestinal Lactase Deficiency and Lactose Intolerance in Adults. Preliminary Report." Gastroenterology **45**: 488-91.
- Dobrindt, U., F. Agerer, et al. (2003). "Analysis of genome plasticity in pathogenic and commensal *Escherichia coli* isolates by use of DNA arrays." J Bacteriol **185**(6): 1831-40.
- Dobrindt, U., G. Blum-Oehler, et al. (2001). "S-Fimbria-encoding determinant sfa(I) is located on pathogenicity island III(536) of uropathogenic *Escherichia coli* strain 536." Infect Immun **69**(7): 4248-56.
- Fabich, A. J., S. A. Jones, et al. (2008). "Comparison of carbon nutrition for pathogenic and commensal *Escherichia coli* strains in the mouse intestine." Infect Immun **76**(3): 1143-52.
- Ferenci, T. (2008). Bacterial Physiology, Regulation and Mutational Adaptation in a Chemostat Environment. Advances in Microbial Physiology. R. K. Poole. London, Elsevier. **53**: 169-229.
- Gonzalez, A. D., V. Espinosa, et al. (2005). "TRACTOR_DB: a database of regulatory networks in gamma-proteobacterial genomes." Nucleic Acids Res **33**(Database issue): D98-102.
- Hahn, E., P. Wild, et al. (2002). "Exploring the 3D molecular architecture of *Escherichia coli* type 1 pili." J Mol Biol **323**(5): 845-57.
- Herbelin, C. J., S. C. Chirillo, et al. (2000). "Gene conservation and loss in the mutS-rpoS genomic region of pathogenic *Escherichia coli*." J Bacteriol **182**(19): 5381-90.
- Ingram, C. J., C. A. Mulcare, et al. (2009). "Lactose digestion and the evolutionary genetics of lactase persistence." Hum Genet **124**(6): 579-91.
- Kandror, O., A. DeLeon, et al. (2002). "Trehalose synthesis is induced upon exposure of *Escherichia coli* to cold and is essential for viability at low temperatures." Proc Natl Acad Sci U S A **99**(15): 9727-32.

- Karp, P. D., I. M. Keseler, et al. (2007). "Multidimensional annotation of the *Escherichia coli* K-12 genome." *Nucleic Acids Res* **35**(22): 7577-90.
- Kinnersley, M., Holben, W. , Adams, J. and F. Rosenzweig (2009). "Genomic analysis of an evolved polymorphism in *Escherichia coli*." *in prep*.
- Le Gall, T., P. Darlu, et al. (2005). "Selection-driven transcriptome polymorphism in *Escherichia coli*/*Shigella* species." *Genome Res* **15**(2): 260-8.
- Meyer, K. J., C. M. Appletoft, et al. (2005). "Determining the source of fecal contamination in recreational waters." *J Environ Health* **68**(1): 25-30.
- Montgomery, R. K., S. D. Krasinski, et al. (2007). "Lactose and lactase--who is lactose intolerant and why?" *J Pediatr Gastroenterol Nutr* **45 Suppl 2**: S131-7.
- Parreira, V. R. and C. L. Gyles (2003). "A novel pathogenicity island integrated adjacent to the thrW tRNA gene of avian pathogenic *Escherichia coli* encodes a vacuolating autotransporter toxin." *Infect Immun* **71**(9): 5087-96.
- Pelosi, L., L. Kuhn, et al. (2006). "Parallel changes in global protein profiles during long-term experimental evolution in *Escherichia coli*." *Genetics* **173**(4): 1851-69.
- Pouttu, R., B. Westerlund-Wikstrom, et al. (2001). "matB, a common fimbrillin gene of *Escherichia coli*, expressed in a genetically conserved, virulent clonal group." *J Bacteriol* **183**(16): 4727-36.
- Rendon, M. A., Z. Saldana, et al. (2007). "Commensal and pathogenic *Escherichia coli* use a common pilus adherence factor for epithelial cell colonization." *Proc Natl Acad Sci U S A* **104**(25): 10637-42.
- Rouquette, C., M. C. Serre, et al. (2004). "Protective role for H-NS protein in IS1 transposition." *J Bacteriol* **186**(7): 2091-8.
- Savageau, M. A. (1974). "Genetic regulatory mechanisms and the ecological niche of *Escherichia coli*." *Proc Natl Acad Sci U S A* **71**(6): 2453-5.
- Savageau, M. A. (1983). "*Escherichia coli* Habitats, Cell Types, and Molecular Mechanisms of Gene Control." *The American Naturalist* **122**(6): 732-744.
- Schouler, C., F. Koffmann, et al. (2004). "Genomic subtraction for the identification of putative new virulence factors of an avian pathogenic *Escherichia coli* strain of O2 serogroup." *Microbiology* **150**(Pt 9): 2973-84.
- Seeto, S., L. Notley-McRobb, et al. (2004). "The multifactorial influences of RpoS, Mlc and cAMP on ptsG expression under glucose-limited and anaerobic conditions." *Res Microbiol* **155**(3): 211-5.
- Shaikh, N. and P. I. Tarr (2003). "*Escherichia coli* O157:H7 Shiga toxin-encoding bacteriophages: integrations, excisions, truncations, and evolutionary implications." *J Bacteriol* **185**(12): 3596-605.
- Silhavy, T. J., M.L. Berman and L.W. Enquist (1984). *Experiments with gene fusions*. Cold Spring Harbor, N.Y., Cold Spring Harbor Laboratory.
- Tusher, V. G., R. Tibshirani, et al. (2001). "Significance analysis of microarrays applied to the ionizing radiation response." *Proc Natl Acad Sci U S A* **98**(9): 5116-21.
- Ulett, G. C., A. N. Mabbett, et al. (2007). "The role of F9 fimbriae of uropathogenic *Escherichia coli* in biofilm formation." *Microbiology* **153**(Pt 7): 2321-31.

Figures

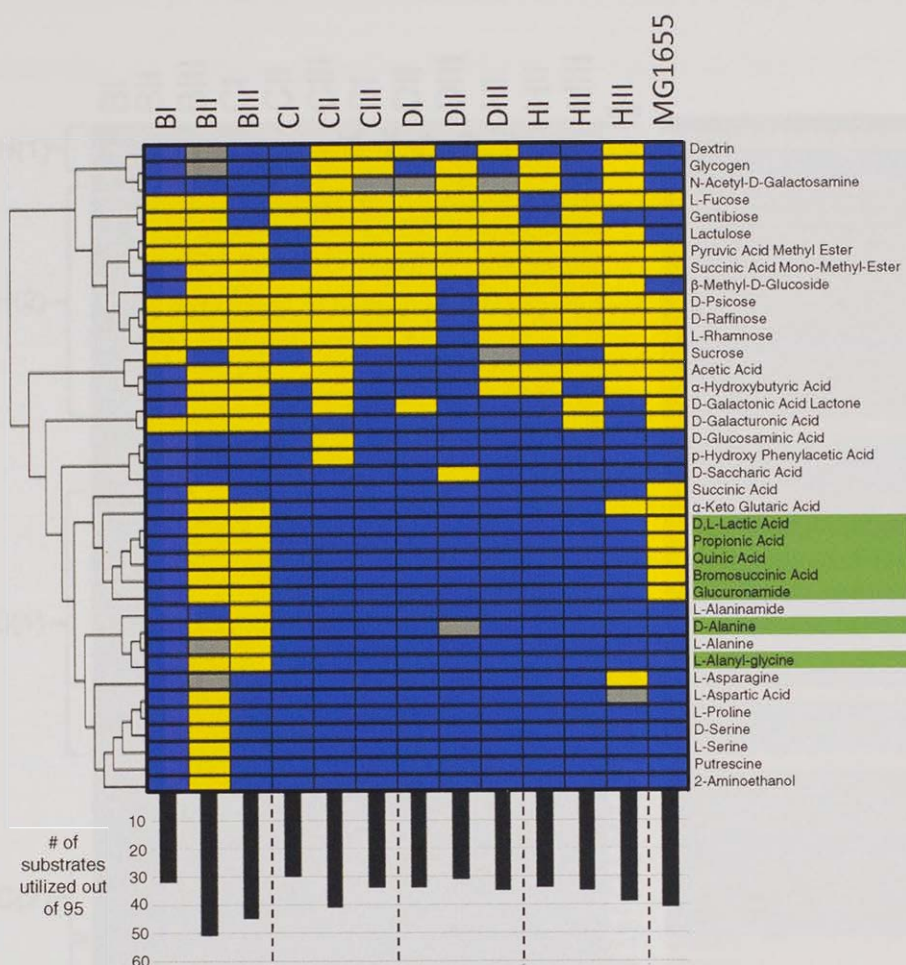


Figure 1. Carbon source utilization profiles of all twelve wild isolates for the 38 compounds that showed variable growth. Yellow boxes indicate that the substrate was utilized while blue boxes denote the absence of growth. Carbon sources that gave ambiguous results are displayed in grey. Strain names are given at the top, carbon source names are to the right and below each profile is a bar graph indicating the total number of substrates metabolized out of the 95 that were measured. Green shading denotes carbon sources that were uniquely metabolized by BII and BIII.

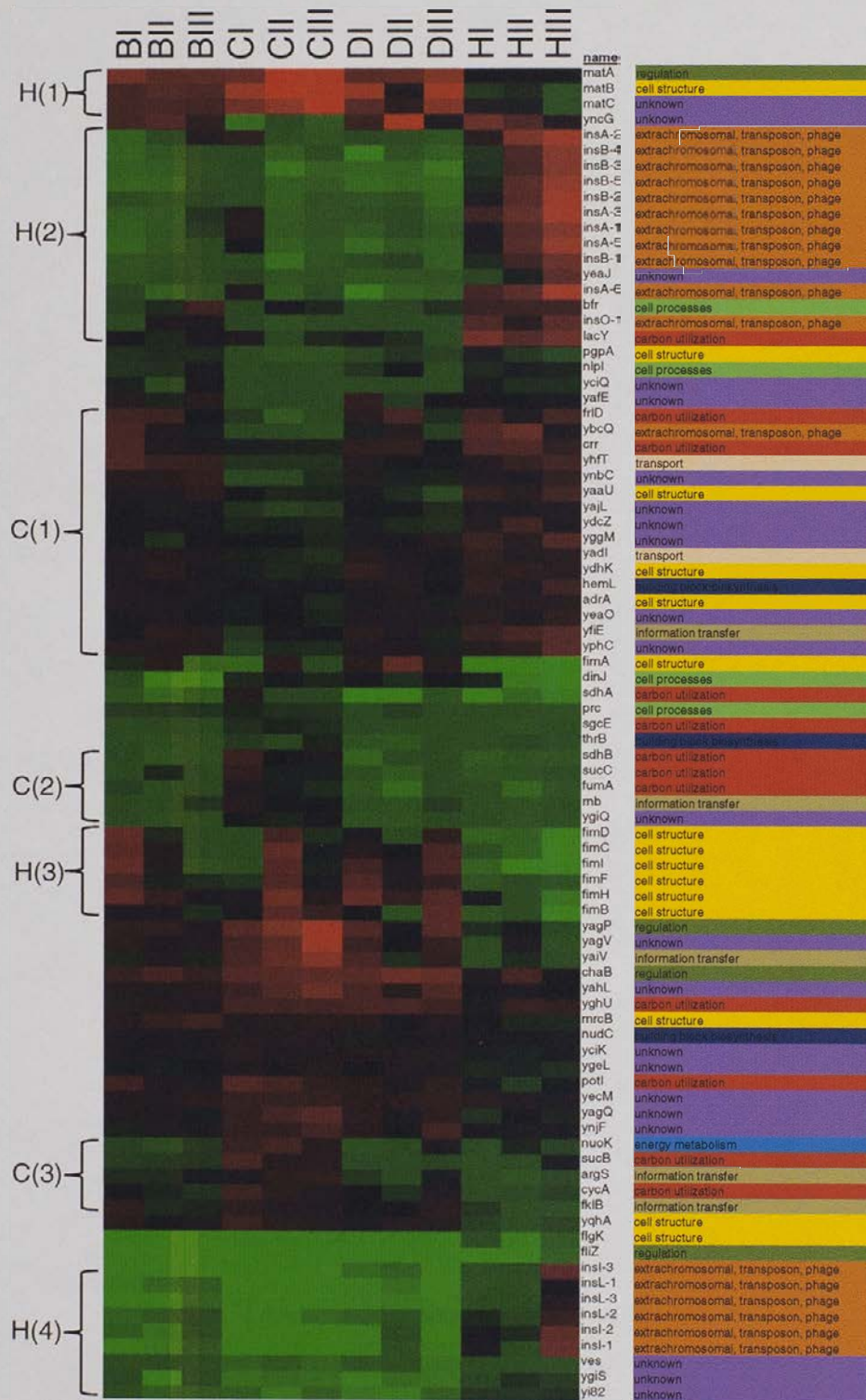
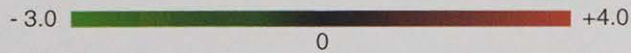


Figure 2. Heatmap showing the 86 genes that had significantly different expression patterns among the four host groups. Green shading indicates that transcription of the gene was lower than the reference ($\log_2 \text{sample/reference} < 0$) while red denotes higher expression ($\log_2 \text{sample/reference} > 0$). Strain names and functional groups are displayed on the right. Color coding of functional groups matches Figure 3. Clusters of diagnostic genes discussed in the text are numbered and indicated by brackets on the left.

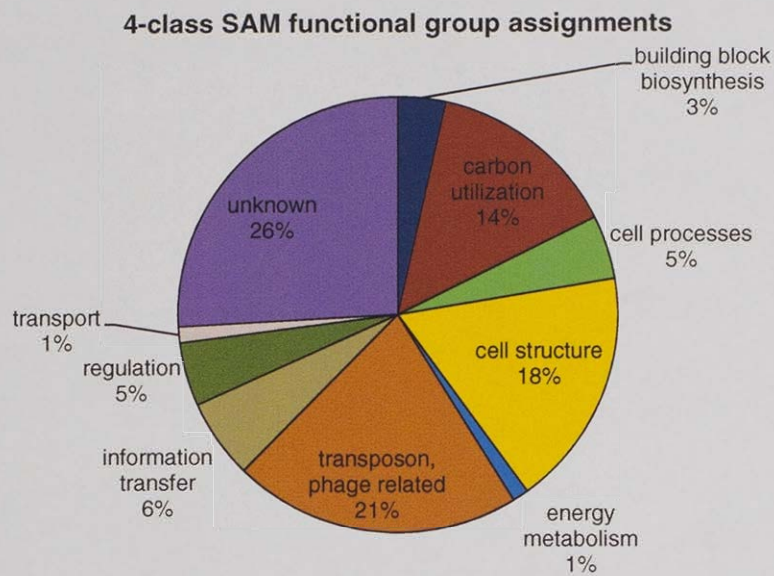


Figure 3. Pie chart depicting the distribution of significant genes from the 4-class SAM analysis by functional group.

Supplementary Table 1. Biolog results for all twelve wild isolates and *E. coli* K-12 MG1655. “+” denotes that the carbon source was metabolized, “-“ indicates no growth and “?” indicates that the result was ambiguous.

Carbon Source	Biolog well number	MG1655												
		B I	B II	B III	C I	C II	C III	D I	D II	D III	H I	H II	H III	
α -Cyclodextrin	A2													
Dextrin	A3		+	+		+		+					+	+
Glycogen	A4					+								
Tween 40	A5													
Tween 80	A6													
N-Acetyl-D-Galactosamine	A7	+	+	+		+	+	+	+	+	+	+	+	+
N-Acetyl-D-Glucosamine	A8	+	+	+	+	+	+	+	+	+	+	+	+	+
Adonitol	A9													
L-Arabinose	A10	+	+	+	+	+	+	+	+	+	+	+	+	+
D-Arabitol	A11													
D-Cellobiose	A12													
1-Erythritol	B1													
D-Fructose	B2	+	+	+	+	+	+	+	+	+	+	+	+	+
L-Fucose	B3	+	+		+	+	+	+	+	+		+	+	+
D-Galactose	B4	+	+	+	+	+	+	+	+	+	+	+	+	+
Gentibiose	B5		+										?	
α -D-Glucose	B6	+	+	+	+	+	+	+	+	+	+	+	+	+
m-inositol	B7													
α -D-Lactose	B8	+	+	+	+	+	+	+	+	+	+	+	+	+
Lactulose	B9		+	+	+	+	+	+		+	+	+	+	
Maltose	B10	+	+	+	+	+	+	+	+	+	+	+	+	+
D-Mannitol	B11	+	+	+	+	+	+	+	+	+	+	+	+	+
D-Mannose	B12	+	+	+	+	+	+	+	+	+	+	+	+	+
D-Melibiose	C1	+	+	+	+	+	+	+	+	+	+	+	+	+
β -Methyl-D-Glucoside	C2	+	+	+	+	+	+	+		+	+	+	+	+
D-Psicose	C3		+	+					?					
D-Raffinose	C4		?			+	+	+		+				+
L-Rhamnose	C5	+	+	+		+	+	+	+	+	+	+	+	+
D-Sorbitol	C6	+	+	+	+	+	+	+	+	+	+	+	+	+
Sucrose	C7		?			+	+		+		+		+	
D-Trehalose	C8	+	+	+	+	+	+	+	+	+	+	+	+	+
Turanose	C9													
Xylitol	C10													
Pyruvic Acid Methyl Ester	C11					+	?	?	+	?	+		+	
Succinic Acid Mono-Methyl-Ester	C12		?	+										
Acetic Acid	D1	+		+		+				?			+	+

Carbon Source	Biolog well number	MG1655											
		B I	B II	B III	C I	C II	C III	D I	D II	D III	H I	H II	H III
Cis-Aconitic Acid	D2												
Citric Acid	D3											+	
Formic Acid	D4	+	+	+	+	+	+	+	+	+	+	+	+
D-Galactonic Acid Lactone	D5		+	+	+	+				+	+	+	+
D-Galacturonic Acid	D6	+	+	+	+	+	+			+	+	+	+
D-Gluconic Acid	D7	+	+	+	+	+	+	+	+	+	+		+
D-Glucosaminic Acid	D8					+						+	
D-Glucuronic Acid	D9	+	+	+	+	+	+	+	+	+	+		+
α -Hydroxybutyric Acid	D10			+									
β -hydroxybutyric Acid	D11												
γ -Hydroxybutyric Acid	D12											+	
p-Hydroxy Phenylacetic Acid	E1	+	+		+	+	+	+	+	+			
Itaconic Acid	E2												
α -Keto Butyric Acid	E3												
α -Keto Glutaric Acid	E4		+	+									+
A-Keto Valeric Acid	E5												
D,L-Lactic Acid	E6		+	+									+
Malonic Acid	E7												
Propionic Acid	E8		+	+									
Quinic Acid	E9								+				
D-Saccharic Acid	E10		+										
Sebacic Acid	E11												
Succinic Acid	E12		+	+									+
Bromosuccinic Acid	F1		+										+
Succinamic Acid	F2											+	
Glucuronamide	F3		+	+		+	+	+	+	+	+		+
L-Alaninamide	F4		?										+
D-Alanine	F5		+										
L-Alanine	F6		+	+									+

Supplementary Table 2. Carbon source utilization of *Escherichia coli* strains MG1655 and MG1655Δ*ptsI* by 4-CareA334

Host groups by 4-CareA334

Carbon Source	Biolog well number	Host groups by 4-CareA334												MG1655	
		BI	BII	BIII	CI	CII	CIII	DI	DII	DIII	HI	HII	HIII		
L-Alanyl-glycine	F7	+	+	+	+	+	+	+		+	+	+	+	+	
L-Asparagine	F8		+	+		+					+	+		+	+
L-Aspartic Acid	F9		+	+											+
L-Glutamic Acid	F10														
Glycyl-L-Aspartic Acid	F11	+	+	+	+	+	+	+	+	+	+	+	+	+	+
Glycyl-L-Glutamic Acid	F12		?												
L-Histidine	G1														
Hydroxy-L-Proline	G2														
L-Leucine	G3														
L-Ornithine	G4														
L-Phenylalanine	G5														
L-Proline	G6		+												
L-Pyroglutamic Acid	G7														
D-Serine	G8	+	+	+									+		+
L-Serine	G9		+	+											+
L-Threonine	G10														
D,L-Carnitine	G11														
γ-Amino Butyric Acid	G12														
Uranic Acid	H1														
Inosine	H2	+	+	+	+	+	+	+	+	+	+	+	+	+	+
Uridine	H3	+	+	+	+	+	+	+	+	+	+	+	+	+	+
Thymidine	H4	+	+	+	+	+	+	+	+	+	+	+	+	+	+
Phenyethylamine	H5														
Putrescine	H6		+												
2-Aminoethanol	H7		+												
2,3-Butandiol	H8														
Glycerol	H9	+	+	+	+	+	+	+	+	+	+	+	+	+	+
D,L-α-Glycerol Phosphate	H10	+	+	+	+	+	+	+	+	+	+	+	+	+	+
α-D-Glucose-1-Phosphate	H11	+	+	+	+	+	+	+	+	+	+	+	+	+	+
D-Glucose-6-Phosphate	H12	+	+	+	+	+	+	+	+	+	+	+	+	+	+

Supplementary Table 2. Genes with significantly different expression between host groups by 4-class SAM.

locus tag	gene name	gene product	MultiFun category
b0003	thrB	ThrB	building block biosynthesis
b0016	insL-1	IS186/IS421 transposase	extrachromosomal, transposon, phage
b0017	yi82	phantom gene	unknown
b0021	insB-1	IS1 protein InsB	extrachromosomal, transposon, phage
b0022	insA-1	IS1 protein Ins	extrachromosomal, transposon, phage
b0045	yaaU	YaaU	cell structure
b0129	yadI	AgaX	transport
b0149	mrcB	MrcB	cell structure
b0154	hemL	glutamate-1-semialdehyde aminotransferase	building block biosynthesis
b0210	yafE	pred. SAM-dependent methyltransferase	unknown
b0226	dinJ	antitoxin of YafQ-DinJ system	cell processes
b0256	insl-1	transposase of IS30	extrachromosomal, transposon, phage
b0257	insO-1	CP4-6 prophage	extrachromosomal, transposon, phage
b0264	insB-2	IS1 protein InsB	extrachromosomal, transposon, phage
b0265	insA-2	IS1 protein InsA	extrachromosomal, transposon, phage
b0274	insB-3	IS1 protein InsB	extrachromosomal, transposon, phage
b0275	insA-3	IS1 protein InsA	extrachromosomal, transposon, phage
b0282	yagP	Pred. LYSR-type transcriptional regulator	regulation
b0283	yagQ	conserved protein	unknown
b0289	yagV	conserved protein	unknown
b0292	matC	predicted protein	unknown
b0293	matB	conserved fimbriin	cell structure
b0294	matA	predicted regulator	regulation
b0326	yahL	predicted protein	unknown
b0343	lacY	LacY lactose MFS transporter	carbon utilization
b0375	yaiV	Pred. DNA-binding transcript. regulator	information transfer
b0385	adrA	predicted diguanylate cyclase	cell structure
b0418	pgpA	phosphatidylglycerophosphatase A	cell structure
b0424	yajL	conserved protein	unknown
b0551	ybcQ	DLP12 prophage; pred. antitermination protein	extrachromosomal, transposon, phage
b0582	insL-2	IS186/IS421 transposase	extrachromosomal, transposon, phage
b0723	sdhA	succinate dehydrogenase flavoprotein	carbon utilization
b0724	sdhB	succinate dehydrogenase iron-sulfur protein	carbon utilization
b0727	sucB	SucB-S-succinylidihydroliopate	carbon utilization
b0728	sucC	succinyl-CoA synthetase, β subunit	carbon utilization
b0857	potI	PotI	carbon utilization
b0988	insB-4	IS1 protein InsB	extrachromosomal, transposon, phage
b1082	flgK	flagellar hook-filament junction protein	cell structure
b1217	chaB	predicted cation transport regulator	regulation
b1268	yciQ	predicted inner membrane protein	unknown
b1271	yciK	predicted oxidoreductase	unknown
b1286	rnb	ribonuclease II	information transfer
b1404	insl-2	transposase of IS30	extrachromosomal, transposon, phage
b1410	ynbC	predicted hydrolase	unknown

b1447	ydcZ	predicted inner membrane protein	unknown
b1454	yncG	predicted enzyme	unknown
b1612	fumA	fumarase A monomer	carbon utilization
b1645	ydhK	conserved inner membrane protein	cell structure
b1742	ves	conserved protein	unknown
b1758	ynjF	predicted phosphatidyl transferase	unknown
b1786	yeaJ	predicted diguanylate cyclase	unknown
b1792	yeaO	conserved protein	unknown
b1830	prc	tail-specific protease	cell processes
b1875	yecM	predicted metal-binding enzyme	unknown
b1876	argS	arginyl-tRNA synthetase	information transfer
b1893	insB-5	IS1 protein InsB	extrachromosomal, transposon, phage
b1894	insA-5	IS1 protein InsA	extrachromosomal, transposon, phage
b1921	fliZ	regulator of σ S activity	regulation
b2279	nuoK	NADH:ubiquinone oxidoreductase, subunit K	energy metabolism
b2394	insL-3	predicted IS186/IS421 transposase	extrachromosomal, transposon, phage
b2417	crr	Crr	carbon utilization
b2545	yphC	predicted oxidoreductase	unknown
b2577	yfiE	predicted DNA-binding transcriptional regulator	information transfer
b2856	ygeL	predicted protein	unknown
b2956	yggM	conserved protein	unknown
b2989	yghU	glutathione transferase-like protein	carbon utilization
b3002	yqhA	conserved inner membrane protein	cell structure
b3016	ygiQ	Obsolete	unknown
b3020	ygiS	predicted transporter subunit	unknown
b3163	nlpI	lipoprotein involved in cell division	cell processes
b3336	bfr	bacterioferritin monomer	cell processes
b3374	frID	fructoselysine 6-kinase	carbon utilization
b3377	yhfT	predicted inner membrane protein	transport
b3444	insA-6	IS1 protein InsA	extrachromosomal, transposon, phage
b3996	nudC	NADH pyrophosphatase	building block biosynthesis
b4207	fkIB	FKBP-type peptidyl-prolyl cis-trans isomerase	information transfer
b4208	cycA	CycA serine/alanine/glycine APC transporter	carbon utilization
b4284	insI-3	transposase of IS30	extrachromosomal, transposon, phage
b4301	sgcE	predicted epimerase	carbon utilization
b4312	fimB	regulator for fimA	cell structure
b4314	fimA	major type 1 subunit fimbrin	cell structure
b4315	fimI	fimbrial protein	cell structure
b4316	fimC	periplasmic chaperone, req. for type 1 fimbriae	cell structure
b4317	fimD	export and assembly of type 1 fimbriae	cell structure
b4318	fimF	fimbrial morphology	cell structure
b4320	fimH	minor fimbrial subunit, D-mannose adhesin	cell structure

Supplementary Table 3. List of global regulators for 4-class SAM significant genes. Yellow shading signifies that the regulator acts as a repressor while blue shading indicates activation. A brown box denotes dual activity.

gene name	sigma factor	Hns	IHF	Lrp	FNR	ArcA	CRP	Fur
matA								
matB								
matC								
yncG								
insA-2								
insB-4								
insB-3								
insB-5								
insB-2								
insA-3								
insA-1								
insA-5								
insB-1								
yeaJ								
insA-6								
bfr								
insO-1								
lacY	70	Hns					CRP	
pgpA	70							
nIpl								
yciQ								
yafE	32							
frlD								
ybcQ								
crr	70				FNR		CRP	
yhfT								
ynbC								
yaaU								
yajL								
ydcZ								
yggM								
yadI								
ydhK	24							

hemL							
adrA	70						
yeaO							
yfiE							
yphC							
fimA	70	Hns	IHF	Lrp			
dinJ							
sdhA	70		IHF	FNR	ArcA	CRP	Fur
prc							
sgcE							
thrB	70						
sdhB	70		IHF	FNR	ArcA	CRP	Fur
sucC	70		IHF	FNR	ArcA	CRP	Fur
fumA	70			FNR	ArcA	CRP	
rnb	70						
ygiQ							
fimD	70	Hns	IHF	Lrp			
fimC	70	Hns	IHF	Lrp			
fimI	70	Hns	IHF	Lrp			
fimF	70	Hns	IHF	Lrp			
fimH	70	Hns	IHF	Lrp			
fimB	70	Hns	IHF				
yagP							
yagV							
yaiV							
chaB	54						
yahL							
yghU							
mrcB							
nudC							
yciK							
ygeL							
potI	54						
yecM							
yagQ							
ynjF							
nuoK	70		IHF	FNR	ArcA		
sucB	70		IHF	FNR	ArcA	CRP	Fur
argS							
cycA							
fkIB							

yqhA							
flgK	28						
fliZ	70, 28	Hns					
insl-3							
insL-1							
insL-3							
insL-2							
insl-2							
insl-1							
ves	70						
ygiS							
yi82							

CHAPTER 5

An exploration in nitrogen cycling and plant growth

Abstract

Nutrient cycling and the interdependence of living things are often difficult concepts for younger students to grasp. In this paper we present an inquiry designed to give students hands-on experience manipulating the nitrogen cycle and measuring the effect at the chemical, microbiological and plant levels.

Introduction

Most schoolchildren are casually familiar with the process of decomposition, but few recognize the important role it plays in nutrient cycling. Part of this disconnect likely stems from the fact that decomposition typically occurs over a long period of time and is carried out by organisms that are too small to see with the naked eye. To a child, it may seem as if a rotting log simply “disappears” but the variety of life forms—from microbes, to plants, to animals—that rely on and participate in this process is enormous. Making the connection between the activity of microscopic organisms such as bacteria or fungi and the health of larger, more familiar organisms such as plants is an important step for middle school students as they transition from studying life as it pertains to individuals into a broader understanding of how these individuals function together as an ecosystem (The National Research Council, 1996).

In this investigation, students will manipulate a familiar, real-world example of nutrient cycling (composting) and examine the effects that changing a single variable (i.e. light, heat, water, etc.) has on the chemical content of the compost, the activity of microbes involved in the nitrogen cycle and, ultimately, plant growth. This experiment is intended to address the National Science Education Standards of "Science as an Inquiry" and the grade 5-8 Life Science Content Standards pertaining to populations and ecosystems. Throughout the course of the project, students will learn to collect and record data, generate simple hypotheses and relate their observations back to a simple model of the nitrogen cycle. The hands-on nature of this activity combined with the sense of ownership that caring for a microbial community and a sprouting plant will give students should reinforce the role that humans have in the health of the environment and generate a broader understanding of the interdependence of life on Earth.

Learning Goals for Students

At the end of this inquiry, students should be able to:

- Generate simple hypotheses
- Work productively in a group
- Interpret data in the context of a simple model of the nitrogen cycle
- Demonstrate an understanding of how the activity of microbes and the health of higher organisms are interconnected
- Synthesize and present scientific results to classmates

Before the Experiment

This inquiry requires that students have a basic understanding of the nitrogen cycle. Nitrogen is found in the bodies of all organisms and is essential for life. Above ground, nitrogen exists as nitrogen gas in the atmosphere. Below ground it is found in a variety of forms. Plants need nitrogen to grow, but they cannot use nitrogen gas directly from the air. As a result, most plants depend on their roots to bring it up from the soil and when they die, the nitrogen from their leaves and stems is returned to the ground. In its simplest form, the nitrogen cycle starts with the decomposition or break-down of this nitrogen-containing plant material by microbes to release stored nitrogen in the form of ammonium (NH_4^+). While new plants can take up and use ammonium, it is more often consumed by bacteria. One group of bacteria, called *Nitrosomonas*, converts the ammonium into nitrite (NO_2^-), while a second group, *Nitrobacter*, converts nitrite into nitrate (NO_3^-). It is this second form of nitrogen (nitrate) that is typically used by plants (Campbell 2002). Changing the environment in the soil (or in a compost bag) can stop or slow the growth of *Nitrosomonas* or *Nitrobacter* and cause ammonium or nitrite to build up which affects plant growth. *Nitrosomonas* and *Nitrobacter* have slightly different preferences in regard to oxygen, temperature and pH. For example, *Nitrosomonas* grows more slowly than *Nitrobacter* at low pH, while *Nitrobacter* is more affected by low levels of oxygen and low temperatures (see Figure 1) (Shammas 1986). Thus, by manipulating these parameters in a compost pile or bag, one can influence the activity of these two groups of bacteria and change the ratio of nitrogen compounds.

Student Preparation

The nitrogen cycle is best presented to students prior to the start of the experiment in a lecture-style format with a simple diagram as shown in Figure 1, followed by a question-and-answer session to reinforce the following points:

- All living things need nitrogen to grow
- Plants get their nitrogen from the soil
- When plants die, decomposers eat dead plant material and produce ammonia/ammonium (NH_3 , NH_4^+)
- *Nitrosomonas* bacteria turn ammonium into nitrite (NO_2^-)
- *Nitrobacter* bacteria turn nitrite into nitrate (NO_3^-)
- Plants take up nitrate and use it to grow- the cycle starts again!
- Changing the soil environment (i.e. changing the pH, temperature, etc.) affects the growth of *Nitrosomonas* and *Nitrobacter*, which will change the amount of ammonia, nitrite and nitrate in the soil
- Changing the amount of ammonia, nitrite and nitrate can affect plant growth

Procedure Overview

This inquiry is best done over the course of several weeks or a couple of months to complete and consists of three phases:

- (1) assembling compost bags (2-3 hours) and monitoring decomposition (30 minutes to 1 hour per week, 4-8 weeks or longer)
- (2) measuring the amount of ammonium, nitrite and nitrate produced (2-3 hours)
- (3) measuring the effect of compost treatment on plant growth (30 minutes to 1 hour for data collection 2 to 3 times per week for 2-3 weeks).

The required materials are inexpensive and can be obtained at any home and garden center and most pet stores (see Table 1). Students should also be provided with a “lab notebook” in which they can record their hypotheses and collect data (an example is given in Supplementary Figure 1).

Phase 1-Making compost

The first part of the inquiry is aimed at getting students to think about experimental design and hypothesis testing by having them create two compost bags that differ by only a single variable. After discussing the nitrogen cycle, ask students to form teams of two or three and have them discuss what variable they would like to test. This variable can be any number of things, from what materials go into the bags to where the bags are stored. Ultimately, the compost will be used to grow plants, so encourage

students to think about what factors would result in a complete nitrogen cycle versus those that would not and how this might affect plant growth.

To assemble the compost bags, send students out into the schoolyard to hunt for compostable materials. Each team will have two bags- one "control" and one "test". Both bags should contain an equal amount (a small handful) of soil to get the decomposition process started. If the students are testing the addition of a particular compostable item, then the two bags should be identical in what they contain except for this single item. If the students are testing a storage condition such as light versus dark or temperature, the two bags should contain the same materials but will be stored in different places.

After the bags are assembled, have each group measure the temperature, weight, smell and color of the compost and then store the bags in an appropriate place. The compost should remain moist and can be misted with a spray bottle if it becomes too dry (unless "dry" versus "wet" is the test variable). Each week, students should monitor the progress of their bags by taking the same measurements and recording the results in their lab notebooks. As the contents of the bags start to decompose, these parameters will change. Keeping track of these changes will help students understand how the materials they started with have transformed over time.

At the end of the composting period, it may be helpful to have students discuss how their control bag compares to their test bag. Questions for discussion include:

- Did one change more than the other?
- How can you tell (i.e. is one darker? does one smell different?)

- What does this tell you about how your treatment affected decomposition?

Phase 2- Tracking the Nitrogen Cycle

This portion of the inquiry allows students to correlate their compost treatment variable and observations with simple biochemical measurements of the intermediate compounds in the nitrogen cycle. These compounds also serve as a surrogate measure for the activity of *Nitrosomonas* and *Nitrobacter*.

Because the nitrogen cycle is also important for the establishment of healthy aquariums, there are a number of readily available test kits for monitoring ammonium, nitrite and nitrate. The easiest type for children to use in this experiment is a dipstick format, but other types of test kits can be substituted if these are not available, as long as a range of concentrations can be measured. Other tests, such as those that measure pH, may also be useful. Most kits rely on a simple color change that is easy to read using a chart provided with the strips. Before beginning, it is helpful to review the nitrogen cycle and conduct a short demonstration of the procedure.

To test the compost, have the students make a slurry for each of their bags of equal parts compost and distilled water in a paper cup. The slurry should be well mixed with a plastic spoon and allowed to settle for a minute or two before the test so that large particles fall to the bottom of the cup. Following the instructions in the kit, have each group test the liquid portion of their slurry and record the results in their lab notebooks. Depending on the experience level of the students, each measurement can be performed

three times, averaged, and the results graphed to facilitate the comparison of the treatment and control bags.

At this point, differences between the two bags should be apparent. For example, if the test condition was excess water (treatment) versus a little bit of water (control), the control bag should have a strong odor, high amounts of ammonium and very little nitrate due to the inhibition of *Nitrosomonas* and *Nitrobacter* in low oxygen environments. Using what they have learned about the nitrogen cycle, and their ammonium, nitrite and nitrate measurements, each group should be able to make a simple prediction as to whether their control or test compost will produce a healthier plant and why. Students can record their prediction in the form of a hypothesis.

Phase 3- Growing Plants

Growing plants using their experimental compost allows students to relate the measurements they recorded in phases 1 and 2 to the health of an organism they are more familiar with. It is best to start with seeds that have been pre-germinated between damp paper towels to ensure that all of the compost test pots have a viable plant. Any vegetable or flower that grows quickly is appropriate as long as it does not fix nitrogen- beets or radishes work well.

To plant the seedlings, have students mix their compost thoroughly with perlite or sand in a 1 part to 2 parts ratio in a small bucket or plastic container. Each compost bag should be mixed with perlite separately and then transferred to paper cups with holes in the bottom for drainage. Each group should have a minimum of 1 cup for the control bag

and 1 cup for the test bag. If time and materials allow, each student could be responsible for 2 plants so that within a group there are two to three replicates that can be averaged. Have the students place the seedlings gently in the cups, pat the compost mixture around the base of the plant, and put them in a warm location with moderate sunlight. Over the course of the next few weeks, each group or individual should water the plants, measure the plant height and leaf length, and make notes about the overall health of the plant (i.e. what color is it? does it appear droopy?) every few days.

Assessment

At the end of the growing period, assessment can be conducted in a variety of ways. Because one of the main objectives of this inquiry was to help students learn to synthesize and present information, a good approach is to ask each group to prepare an oral/visual presentation of their results to give to the rest of the class. This type of exercise can be combined with a question and answer session in which instructors gauge student learning by asking simple questions like “why do you think that happened?” or “what do you think would happen if...”. With this format, many students that have difficulty making the connection between the compost treatments, nitrate levels and plant growth are able to see the relationships when confronted with the information they had gathered and asked leading questions by the instructor.

Conclusion

One important aspect of understanding how ecosystems function is understanding nutrient cycling. The nitrogen cycle is often presented as a complicated diagram with a series of arrows and chemical formulas. Here we have described a simple series of hands-on experiments that allow students to manipulate the nitrogen cycle in compost and examine the effects that different treatments have on plant health. The exercises presented will give middle-school children useful practice in developing and testing simple hypotheses as well as effectively communicating scientific results while cultivating a deeper understanding of the interconnectedness of living systems.

Literature Cited

- Campbell, N. A. a. J. B. R. (2002). Biology. San Francisco, CA, Benjamin Cummings
- Shammas, N. K. (1986). "Interactions of Temperature, pH, and Biomass on the Nitrification Process." Journal (Water Pollution Control Federation) **58**(1): 52-59.
- The National Research Council. National Science Education Standards. Washington, DC, National Academy Press (1996): 155-158.

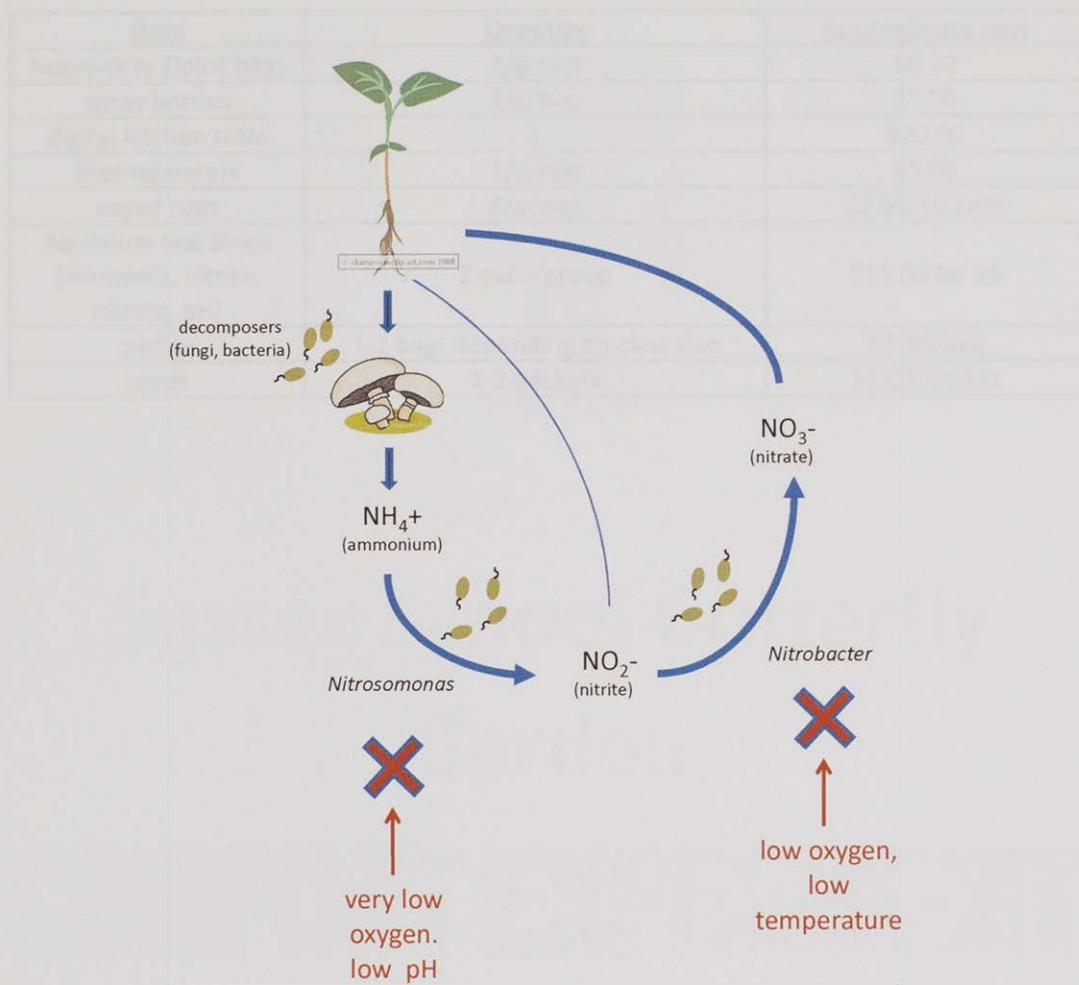


Figure 1. A simplified version of the nitrogen cycle

Table 1. Materials to be purchased

<u>Item</u>	<u>Quantity</u>	<u>Approximate cost</u>
heavy-duty Ziploc bags	2/group	\$0.20
spray bottles	1/group	\$1.50
digital kitchen scale	1	\$20.00
thermometers	1/group	\$5.00
paper cups	8/group	\$2.50/50 cups
Aquarium test strips (ammonia, nitrite, nitrate, pH)	2 each/group	\$15.00 for 25
perlite	1-2 bags depending on class size	\$5.00/bag
seeds	1-2 packets	\$1.00/packet

Name: _____

Supplementary Figure 1

Date: _____



Sussex School Butterfly Garden

COMPOST EXPERIMENT

LAB BOOK

Name: _____

Date: _____



1. Write down your variable here: _____

2. Write down what you will put in each compost bag.

Compost Experiment Hypothesis Sheet

(control)

(test)

I hypothesize that the compost bag with _____

_____ in it

will _____

3. Write down where you will store the bag and what observations you will make about the bag over the next three weeks.

because _____

Compost Experimental Design Sheet

Design an experiment to test your hypothesis.

1. Write down your **variable** here: _____

2. Write down what you will put in each compost bag.

Compost bag #1

(control)

Compost bag #2

(test)

3. Write down where you will store the bag and what observations you will make about the bag over the next three weeks.





Compost Experiment Data Sheet

Date	temperature	smell	color	weight
February 11th				

Date	temperature	smell	color	weight
		nitrate (NO ₃ ⁻) nitrite (NO ₂ ⁻)		

compost bag #1 control Date _____	total hardness total nitrate pH	compost bag #2 treatment variable _____ Date _____
	ammonia (NH₄⁺)	
	nitrate (NO₃⁻)	
	nitrite (NO₂⁻)	

Plant number	plant appearance	plant height	leaf length

CHAPTER 6

Synthesis

The study of polygenic genome composition and gene regulation can provide

Plant number	plant appearance	plant height	leaf length

CHAPTER 6

Synthesis

The study of prokaryotic genome composition and gene regulation can provide useful insight into many aspects of microbial ecology and evolution. With the mainstream application of high-throughput techniques such as microarray comparative genome hybridization, microarray transcriptional profiling and sequencing, researchers are now able to address basic questions regarding genome evolution in response to changing environments. In the work presented here, I have employed all three of these techniques in the study of both natural and experimental populations of *E. coli* with the ultimate goal of gaining a better understanding of niche adaptation and the nature of molecular variation in microbial systems.

In Chapter 2, I explored the mechanistic basis of adaptation and diversification in a polymorphic experimental population of *E. coli* that spontaneously arose after ~700 generations of glucose limitation in chemostats. This unique system afforded a tractable arena in which to investigate how intra-specific variation can be maintained by niche adaptation. My results show that mutations in both global and gene-specific regulators are primarily responsible for the stable co-existence of clones and that these mutations can have unexpected effects on gene expression when isolates are removed from the environment to which they are adapted. I also demonstrated that founder genotype can have a profound influence on evolutionary outcome, a result that is particularly pertinent to the study of diversity in natural populations. These findings serve to broaden our understanding of how microbial systems respond to environmental stressors including

nutrient limitation and competition, and may be applicable to the study of medically relevant clonal microbial populations as discussed below.

In Chapters 3 and 4, I applied microarray comparative genome hybridization and transcriptional profiling to natural strains of *E. coli* isolated from the feces of four different mammalian hosts. From an applied science perspective, the basic question of how biochemical and genetic measures of diversity are correlated with habitat differences and an assessment of how well potential biomarkers discriminate between host species is important for advancing the field of microbial source tracking. From an evolutionary standpoint, similar patterns of gene presence/absence and gene expression in phylogenetically unrelated strains can provide clues about how differences in the selective forces at work in the natural environment shape genome and transcriptome content.

My work also demonstrates that genome composition (as measured at the gene level) is a more reliable indicator of host affiliation than a number of fingerprinting methods commonly used for microbial source tracking of fecal water contamination. These data call into question the validity of using fingerprints to determine host source and suggest that the continued use of *E. coli* as an indicator organism may require the development of gene-specific molecular markers to be truly useful.

I further present evidence that all of the human derived strains show common patterns of gene presence/absence. Additional testing would be required to determine whether this phenomenon is the result of a common adaptive environment and whether such an environment was the primary habitat (i.e the human digestive system) or the secondary habitat (sewage). In either case, the result could have profound implications

for the future of *E. coli* as a wastewater indicator species. Source tracking methods that rely on genetic or physiological characteristics determined solely by phylogeny may be less likely to yield satisfactory results compared to those that exploit characteristics selected for by the environment.

Transcriptional profiling of the same wild isolates in Chapter 4 recapitulated and extended the results from Chapter 3 in that the human derived strains appear to have common patterns of gene expression as well as similar genome content. Moreover, some of these expression differences were not due solely to the presence or absence of entire open reading frames: rather, they suggested that mutations affecting regulation of certain genes may have occurred independently in all three isolates from the same host source. Future investigation into this phenomenon should include a more detailed examination of the mechanistic basis for the observed expression differences, as well as comprehensive analysis of the extent to which they are observed in larger sample sets.

Finally, in Chapter 5, I presented a portion of the work that I did as part of the ECOS program at UM. This year-long foray out of the university setting and into the public school system was illuminating on many levels. The physiological and genetic characteristics that make microorganisms a good choice for investigating many aspects of population genetics, evolution and ecology in a university laboratory also make them an ideal model system for K-8 level science education, yet I found that their application in this arena is limited. This limitation was not due to inability of students or teachers to grasp basic microbiological principles. Rather, I believe that the discrepancy stems from a perceived lack of access to materials and unfamiliarity with techniques for manipulating and cultivating microbes. In my opinion, this situation is easily remedied at the

university level with public outreach programs and expanded curriculum choices for future science teachers.

Overall, the body of work presented here expands the knowledge base in two distinct but complementary areas of research: experimental adaptive evolution and applied molecular microbial ecology. While it is often difficult or impossible to study microbes under purely natural conditions, the detailed analysis of their growth, adaptation and population dynamics in a controlled laboratory environment can be useful for building a predictive conceptual framework in which to address questions of ecological relevance. Similarly, measures of extant genetic variation in wild isolates adapted to life in the natural environment and the response of these isolates to laboratory culture conditions can be useful for refining this framework, directing future laboratory investigation and addressing real-world ecological issues such as determining the source of fecal water contamination. Thus, the integration of experimental evolution with traditional microbial ecology has the potential to lead to interesting insights that can advance both fields.

The experimental evolution study presented in Chapter 2 is one of only a few to explore the molecular basis for the *de novo* evolution of a multi-member bacterial assemblage from a single clone. The evolutionary outcome in this case appeared to be heavily influenced by mutations in global regulatory genes and the genotype of the founder strain. The repeated observation of specific regulatory mutations both within this system (such as those that affect acetyl-CoA synthetase) and between this system and analogous systems studied by other groups (such as the mutation the *rpoS*) suggest that this type of change may be a common adaptive response to novel environmental

conditions and thus perhaps not confined strictly to experimental systems. Many natural microbial populations (such as those that cause nosocomial or chronic infections) are also founded by clones (Treves, Manning et al. 1998; Notley-McRobb, King et al. 2002; Ferenci 2003; Seeto, Notley-McRobb et al. 2004; Lundin, Bjorkholm et al. 2005; King, Seeto et al. 2006; Rozen, Philippe et al. 2009). For example, chronic *Pseudomonas aeruginosa* infection of the lungs of cystic fibrosis patients frequently originates from one or a few isolates that undergo clonal expansion over the course of many years (Struelens, Schwam et al. 1993; Smith, Buckley et al. 2006). Similarly, *Helicobacter pylori* infections, the cause of most gastric ulcers, are often initiated in childhood by a single strain that persists and diversifies throughout the lifetime of an individual. Understanding why certain strains of *H. pylori* and *P. aeruginosa* are able to establish productive infections, how these pathogens adapt to novel environments and the mechanistic bases of adaptation are all avenues of research that can be informed by the results of experimental evolution studies such as the one presented here. Insights gleaned from the comparison of microbial adaptation under simplified laboratory conditions to that which takes place within an individual patient could prove to be instrumental in understanding the progression of disease as well as successful implementation of therapeutic regimens. In the case of *Pseudomonas aeruginosa*, known targets of selection during adaptation to the cystic fibrosis lung environment, much like those that were found to be responsible for adaptation to the chemostat environment, are regulatory: mutations in the aminoglycoside efflux pump regulator *mexZ* can enhance antibiotic resistance and mutations in *lasR*, a regulator of quorum sensing, may influence biofilm formation during infection.

Another key finding presented in Chapter 2 was that at least one of the strains (CV103) exhibited a different gene expression pattern when grown in the presence of the other strains versus when grown alone. This unique behavior manifest in the consortium environment suggests that this and other similar experimental microbial assemblages may be useful for studying population-level emergent properties, i.e. those properties of biological systems that are evident when individuals interact, but cannot be deduced when population constituents are studied in isolation. The advantages of using microbes for investigating emergent properties, particularly when "community" members are closely related, are numerous. Individuals can be easily manipulated and exhaustively characterized thus allowing researchers to determine the precise effects of mutation and environmental perturbation on the behavior of the system as a whole. In addition, samples can be stored long-term in a static state and quantitatively reconstituted to address the repeatability of population-level interactions. The integration of tractable experimental models such as this one into systems biology may yield valuable insights regarding the nature and evolution of emergent properties that can then be applied to questions of broader ecological importance.

In regard to the natural isolates studied in Chapters 3 and 4, the work described here identifies several genes whose presence in the genome and expression patterns appear to be associated with particular host species, especially in the case of the human derived isolates. The ecological significance of these associations remains unclear as the number of strains that were studied is relatively small and no concrete associations between host intestinal physiology and microbial genome content could be made. In this case, additional laboratory-based investigation using bioreactors that simulate the gut

environment could shed light on the adaptive advantage of some of the identified differences in genome/transcriptome content. However, regardless of their evolutionary origin, the host-specific genomic differences discovered also have the potential to be useful as library-independent molecular markers for fecal water contamination source tracking using *E. coli*. The development and application of such markers would positively impact the source tracking field as the cost of generating fingerprint libraries, currently the most commonly employed library-dependent method, is significant. However, further work is needed to establish the utility of the identified genes in this context.

In conclusion, the work I have done has answered many questions, but it has also generated many more. I would like to see future lines of investigation include a detailed analysis of the fitness effects of the deletion in CV103 and the *maltT* mutation in CV101, CV115 and CV116 as well as a comprehensive screen for genetic differences that affect glycerol metabolism in strain CV116. In regard to the wild *E. coli* populations, I believe the next step toward generating molecular markers for microbial source tracking is a thorough analysis of the distribution of potential diagnostic characters across a larger sample set. All of these experiments are imminently feasible and would further expand our understanding of how this versatile microbe adapts to both its laboratory and natural habitat.

Literature Cited

- Ferenci, T. (2003). "What is driving the acquisition of mutS and rpoS polymorphisms in Escherichia coli?" Trends Microbiol **11**(10): 457-61.
- King, T., S. Seeto, et al. (2006). "Genotype-by-environment interactions influencing the emergence of rpoS mutations in Escherichia coli populations." Genetics **172**(4): 2071-9.
- Lundin, A., B. Bjorkholm, et al. (2005). "Slow genetic divergence of Helicobacter pylori strains during long-term colonization." Infect Immun **73**(8): 4818-22.
- Notley-McRobb, L., T. King, et al. (2002). "rpoS mutations and loss of general stress resistance in Escherichia coli populations as a consequence of conflict between competing stress responses." J Bacteriol **184**(3): 806-11.
- Rozen, D. E., N. Philippe, et al. (2009). "Death and cannibalism in a seasonal environment facilitate bacterial coexistence." Ecol Lett **12**(1): 34-44.
- Seeto, S., L. Notley-McRobb, et al. (2004). "The multifactorial influences of RpoS, Mlc and cAMP on ptsG expression under glucose-limited and anaerobic conditions." Res Microbiol **155**(3): 211-5.
- Smith, E. E., D. G. Buckley, et al. (2006). "Genetic adaptation by Pseudomonas aeruginosa to the airways of cystic fibrosis patients." Proc Natl Acad Sci U S A **103**(22): 8487-92.
- Struelens, M. J., V. Schwam, et al. (1993). "Genome macrorestriction analysis of diversity and variability of Pseudomonas aeruginosa strains infecting cystic fibrosis patients." J Clin Microbiol **31**(9): 2320-6.
- Treves, D. S., S. Manning, et al. (1998). "Repeated evolution of an acetate-crossfeeding polymorphism in long-term populations of Escherichia coli." Mol Biol Evol **15**(7): 789-97.