THE EFFECT OF ECOLOGICAL DIFFERENTIATION ON GENETIC RECOMBINATION IN THE ENTEROBACTERIA

by

Adam Christopher Retchless

B.S. in Biological Sciences, Carnegie Mellon University, 2000

Submitted to the Graduate Faculty of

Arts and Sciences in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy in Biological Sciences

University of Pittsburgh

2010

UNIVERSITY OF PITTSBURGH

ARTS AND SCIENCES

This dissertation was presented

by

Adam Retchless

It was defended on

July 30, 2010

and approved by

Roger W. Hendrix, PhD; Distinguished Professor, Department of Biological Sciences

Valerie Oke, PhD; Lecturer, Department of Biological Sciences

Stephen J. Tonsor, PhD; Associate Professor, Department of Biological Sciences

Dannie Durand, PhD; Associate Professor, Department of Biological Sciences, Carnegie Mellon University

Dissertation Advisor: Jeffrey G. Lawrence, PhD; Professor, Department of Biological Sciences

THE EFFECT OF ECOLOGICAL DIFFERENTIATION ON GENETIC RECOMBINATION IN THE ENTEROBACTERIA

Adam C. Retchless, PhD

University of Pittsburgh, 2010

The existence of distinct species of life is generally explained by the genetic process of reproduction without recombination between populations and/or the ecological process of adaptation to different environments. Both processes affect prokaryotes, and have shaped existing genomes. Here, we use comparative genomic techniques to evaluate the dynamics of divergence among species of the Enterobacteriaceae. Bacteria such as *Escherichia coli* preferentially acquire allelic variants from closely related organisms (*i.e.* other *E. coli*) rather than from more diverged bacteria. Ecological differences between donor and recipient affect the probability of allelic variants becoming fixed across the recombining population. We examine the history of recombination among groups of genomes that no longer recombine with each other, but retain sufficient conservation of ancestral nucleotide sequences to allow recombination to be inferred. From these analyses, we conclude that substantial levels of recombination occurred between *E. coli* and diverging lineages even after some regions of the genomes had acquired many nucleotide differences. We identify two evolutionary radiations leading to *E. coli* where the disparity among loci confounds the phylogenetic relationships among species, as evidenced by topological incongruence among gene trees. The forces affecting recombination, reflected in both pairwise divergence and topologically informative sites, vary across regions of the genome measuring tens of kilobases. To examine the relationship between ecological differentiation and genetic recombination, we characterize differences that could be responsible for ecological differentiation among these species. Some of the loci with the most apparent functional differences (*i.e.* the gain and loss of genes) are associated with the greatest levels of sequence divergence between species, consistent with the hypothesis that ecological divergence interferes with homologous recombination, and therefore drives sequence divergence and genetic isolation. To investigate the role of more subtle ecological differentiation, we develop a statistical framework to evaluate codon usage bias of each protein-coding gene, taking into account the stochastic balance between codon selection, which is driven by the need for high expression, and mutational biases. This tool will be useful in future studies examining codon selection as contribution to diversification among the ecologically diverse species of Enterobacteriaceae.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

PREFACE

Terminology:

ACE: Adaptive Codon Enrichment

BSC : Biological Species Concept

CAI : Codon Adaptation Index

dS, dN : Divergence (synonymous, non-synonymous)

dsDNA : double-stranded DNA

HGT : Horizontal Gene Transfer

K_a, K_s : Estimated substitution count (amino acid, synonymous)

MLST : Multi-Locus Sequence Typing

MYr : Millions of years

ORF : Open Reading Frame

Prokaryote: Bacteria and Achaea

Previously published material has been incorporated into this document:

Chapter 1:
Lawrence, J. G. and A. C. Retchless (2009). "The interplay of homologous recombination and horizontal gene transfer in bacterial speciation." Methods Mol Biol **532**: 29-53. Available Online.

Chapter 2:
Retchless, A. C. and J. G. Lawrence (2007). "Temporal fragmentation of speciation in bacteria." Science **317**(5841): 1093-1096. Available Online.

Chapter 3:
Retchless, A. C. and J. G. Lawrence (2010). "Phylogenetic incongruence arising from fragmented speciation in enteric bacteria." Proc Natl Acad Sci U S A **107**(25): 11453-11458. Available Online.

Acknowledgements:

The time that I have spent at the University of Pittsburgh has been the most intense and rewarding period of my intellectual development, and many people have contributed to my success here. Foremost, my advisor Jeff Lawrence has never failed to provide a stimulating environment, including a steady stream of ideas for new experimental approaches to examine important issues. The other members of the lab group have also been essential to my development, bringing interesting publications to my attention, participating in critical examination of these publications, prompting me to clearly explain and justify my own research, and inviting me to discuss the implications of their own research. In particular, Heather Hendrickson and Kristen Butela provided extensive feedback during experiment design and manuscript preparation. Hans Wildschutte, Rajeev Azad, Bryan Goddard, Tom Seiflein, Sarah Hainer, and Robin Monteverde contributed to this atmosphere by providing their own perspectives on both biology and life outside of the lab

I am grateful to Carol LaFave and Melanie Popa for their guidance while acting as a teaching assistant. I have also benefited from the activities of the Teaching Club within the department, and wish that I had made more time to participate in their programs. Fortunately, Kristen was able to give me an overview of many of the issues being discussed in their meetings.

The research faculty of this department has, of course, been instrumental in my scientific development. In particular, the members of my dissertation committee have provided the most feedback. I thank Dannie Durand (of CMU) and Steve Tonsor for their time both in the classroom and in individual meetings. When I met them, I was probably no more knowledgeable about their fields of expertise than are the sophomore undergraduates in their departments, but they were patient enough to walk me through some pretty basic issues that I didn’t initially grasp. Roger Hendrix has pushed me to think more deeply about the relationship between genes and organisms. Roger has also assembled a great team of researchers, whom I have been fortunate to have daily interactions with. Valerie Oke has always encouraged rigor in my

analyses. Also, during first-year rotations, she provided my initial exposure to issues of microbial diversity, development, and ecology. While Susan Kalisz was not a member of my dissertation committee, she has provided valuable insight on issues ranging from population biology to what it means to do science. Tony Bledsoe provided help interpreting the results of phylogenetic analyses. Eric Polinko provided interesting and useful thoughts regarding the teaching of bioinformatics and troubleshooting of computer problems.

My fellow graduate students have been excellent classmates and colleagues. John Paul and April Randle helped me to place my own work in the context of eukaryotic species concepts. The departmental staff has made it possible for me to get my work done. I have received the most direct help from Cathy Barr, Pat Dean, Crystal Petrone, and Hernan “Bud” Brizuela, but I also appreciate that the other administrators and technicians keep the department running smoothly.

A large number of people were instrumental in preparing me for graduate research. Among my many mentors prior to matriculation at the University of Pittsburgh were Mike Freeling, David Braun, Jenn Fletcher, Cristel Carles, Sarah Hake, Katherine Krolikowski, Jon Jarvik, Cheryl Telmer, Dave Malehorn, Susan Henry, Vincent Bruno, Jun Yang, Manuel Villa-Garcia, Susan Dowd, and Janna Patton-Vogt. I would have never sought the opportunity to work with the above scientists were it not for Dave Phoebus at Dulaney High School, who first kindled my fascination with biology.

Of course, none of this would have been possible without my family. My grandparents, mother, father, and older brother guided me, encouraged me, and acted as role models throughout everything. My wife, Beth, has been by my side since the days when we were both applying to graduate school. There is no way that I could list all of the ways that she has contributed to my scientific development, let alone personal development. She is my partner in everything.

1.0 INTRODUCTION

Humans naturally categorize life-forms, placing similar organisms into named groups and recognizing differences between groups. Children as young as four months of age are able to categorize animals such as cats and dogs in a way that distinguishes between the groups even as they recognize different individuals within each group (Quinn 2002), demonstrating an intuitive species concept. The identification of organisms is a fundamental step in deciphering that organism's biology; the power of classification is the implicit understanding of what that organism is likely to do, or is capable of doing, based on past experiences with similar organisms.

However, biological species are not simply collections of similar objects. As the theory of spontaneous generation lost influence (Strick 2000; Wilkins 2004), the idea of species acquired a genetic component – each generation of a species was the continuation of previous generations. Evolutionary theories proposed that changes could occur over generations, and species could split, creating new species. Darwin proposed that divergence would be a frequent result of natural selection, as competition with similar organisms (*i.e.* members of the same species) would only be relieved by the evolution of distinct traits that enabled exploitation of new resources (Darwin 1859).

Darwin's theory required that the reproductive incompatibilities observed between species be able to develop gradually as one species split into two. To argue for the plausibility of this process, he attacked the idea that species necessarily had strict reproductive barriers between them. To this end, he cited several examples where reproductive isolation failed to clearly distinguish between species and the varieties within species (Darwin 1859). Instead, Darwin proposed that reproductive incompatibilities between species were fundamentally the same as those found within species, being different only in degree (Darwin 1859).

1.1 SPECIES DELIMITATION

Despite the necessary ambiguity that accompanies any attempt to discern when one species has split into two, several attempts have been made to conceptualize the essence of species identity (de Queiroz 2005). An influential line of investigation has focused on the genetic independence of species, most famously described in Mayr's "biological species concept" (BSC) (Mayr 1942; Mayr 1963). Mayr proposed that species' members share a common gene pool, and that frequent genetic exchange among groups of con-specific individuals provided genotypic (and thus phenotypic) cohesion within species. Therefore, the inability to exchange genetic material is the definition of separate species.

A contrasting perspective has focused on ecological differentiation, emphasizing the role of selection on the organism as a whole over the reassortment of genetic diversity (van Valen 1976). Here, each species experiences constant selection for a particular phenotype, purging any mal-adaptive variants that arise, whether from mutational processes or gene exchange. The distinguishing trait of species is their ability to sustainably coexist, since neither can outcompete the other across its entire range.

Attempts to reconcile these contrasting species concepts (among many others) have recast genetic isolation and ecological differentiation as contingent properties of the species, which is itself a fundamental unit of biological organization (de Queiroz 2005). An alternative reaction to the vast diversity of species concepts and the difficulty in applying any of them is to deny the pre-eminence of the species category, instead viewing it as part of a continuum with other levels of population structure created by a variety of mechanisms (Mallet 2008).

The above debates have focused primarily on multicellular eukaryotes, many of which can only reproduce by mating with another individual. Microbes were exempted from these debates in part due to the difficulty of characterizing their physiology, ecology, and relatedness, but more fundamentally because their asexual mode of reproduction appeared to make the logic of Mayr's species concept inapplicable (Mayr 1963), particularly for Bacteria and Archaea, (henceforth collectively referred to as "prokaryotes" in reference to similarities in genetic transmission and ecology).

Early prokaryotic classification schemes incorporated little information, as prokaryotes had few morphological traits and broad geographical distribution, making it difficult to place them within the evolutionary perspective of a species concept. For those prokaryotes that could be reliably identified (*e.g.* grown in pure culture), species were defined by a phenotypic approach, as used in the first edition of *Bergey's Manual of Systematic Bacteriology*. The development of molecular evolutionary theory produced phylogenetic techniques that allowed prokaryotes to be clustered according to DNA sequence similarities, thereby placing them into an evolutionary context (Fox, Stackebrandt et al. 1980; Woese, Kandler et al. 1990). Within this framework, molecular divergence can be used to split prokaryotes into different species if they pass a threshold, such as < 70% reassociation by DNA-DNA hybridization, or < 97% identity at

16S rRNA genes (Gevers, Cohan et al. 2005). Whole-genome sequence data has introduced the possibility of resolving prokaryotic relationships with even greater detail based on measures such as the average nucleotide identity of aligned sequences, or the number of genes shared between two genomes (Konstantinidis, Ramette et al. 2006).

The use of these thresholds allows prokaryotes to be classified into manageable units of biodiversity. However, this definition of species does not equate the species with any special biological properties—the species is just one step along a continuum of molecular divergence among organisms. To apply species concepts, this molecular sequence data must be used to infer the population dynamics that are the basis for genetic and ecological species concepts (reviewed in Gevers, Cohan et al. 2005; Doolittle and Zhaxybayeva 2009).

1.1.1 Prokaryotic species as gene pools

The possibility that Mayr's BSC could be applied to prokaryotes arose with the recognition that some bacteria acquire DNA from closely related strains at a meaningful frequency. The history of allele transfer between strains was apparent in the conflicting phylogenies of genes scattered around the *Escherichia coli* chromosome (Dykhuizen and Green 1991). While this challenged the established paradigm of prokaryotic evolution based on binary fission and vertical inheritance, it was consistent with genetic behaviors observed in the laboratory (described below), where a segment of DNA can be introduced into the chromosome and replace a similar native sequence. This process is often called "recombination," alluding both to the role of the homologous recombination machinery in catalyzing allele replacement, and to the population genetic process of reducing linkage disequilibrium (Dykhuizen and Green 1991). Recombination in *E. coli* permits advantageous alleles (even entire pathogenicity islands, see Schubert, Darlu et al. 2009) to spread beyond their original genomic context and reassort with alleles at other loci, defining the limits of the species according to the BSC (Guttman and Dykhuizen 1994).

Recombination in prokaryotes does not involve the fusion of haploid genomes, but rather the unidirectional transfer of small fragments of DNA between donor and recipient. Here, DNA may be moved between cells by one of three mechanisms (Ochman, Lawrence et al. 2000). Transduction occurs when bacteriophages mistakenly package host DNA into their capsids instead of virus DNA. When this particle finds a target cell, the DNA – limited in size to a fragment which will fit in the capsid – is injected. Transformation occurs when a prokaryotic cell imports fragment of naked DNA directly from the environment; this is common among bacteria which consume DNA as a source of food. Conjugation can transfer chromosomal genes when a plasmid integrates into its host chromosome and then begins its process of replication and transfer into another host. Plasmid DNA is transferred directly between the cytoplasm of the donor cell into the cytoplasm of the recipient, thus requiring prolonged cell-cell contact. Conjugation can move large portions of chromosomal DNA between cells.

After the DNA has been injected into the cytoplasm of the recipient cell, it is subjected to four important processes. First, restriction endonucleases will cleave almost all incoming DNA fragments, with the exception of DNA arriving from a cell expressing the same *hsd*-encoded restriction/modification system, whose cleavage sites have thus been protected. Given the variability in *hsd* genes within and among bacterial species (Barcus, Titheradge et al. 1995; O'Neill, Chen et al. 1997; Murray 2000), this exception is rare, even within named species. Second, exonucleases will degrade the dsDNA ends of the resulting fragments. These two processes act in concert to reduce the size of incoming DNA fragments and, most often, prevent the DNA from integrating into the recipient chromosome. Third, RecA-mediated homologous recombination may occur, whereby the incoming DNA fragment – reduced in size through the action of nucleases (Milkman, Raleigh et al. 1999) – is integrated in to the chromosome, replacing the resident allele at its cognate position. This requires nucleotide sequence identity between regions of incoming and resident DNA, so the presence of mismatches reduces the probability of successful recombination (Shen and Huang 1986). Fourth, if no region of similarity exists between the incoming and resident DNA, illegitimate recombination may occur, placing the arriving DNA anywhere in the chromosome or, alternatively, site-specific recombinases (*e.g.*, phage integrases) may catalyze recombination into specific locations.

The Dykhuizen and Green operational interpretation of Mayr's biological species concept was retrospective, using the patterns of genetic diversity among individuals to delineate species boundaries (Dykhuizen and Green 1991). They proposed that within species, the relationships among individuals as inferred from different genes would not be congruent, but between species the phylogenies would be congruent. This model works well when applied to some groups of bacteria. For example, different genes among different strains of the enteric bacteria *Escherichia coli* or *Salmonella enterica* show different relationships, reflecting homologous recombination within these groups (Dykhuizen and Green 1991; Milkman 1997). Yet phylogenies are congruent among more-diverged bacteria, implying that homologous recombination does not readily exchange genes across the boundaries of these named species (Daubin, Moran et al. 2003; Wertz, Goldstone et al. 2003). The population structures of many prokaryotes have been examined by Multi-Locus Sequence Typing (MLST) – wherein alleles at a handful of shared loci are sequenced (Maiden, Bygraves et al. 1998) – clearly indicating that many named prokaryotic species have appreciable rates of homologous recombination among constituent strains (Feil, Maiden et al. 1999; Feil, Smith et al. 2000; Feil, Holmes et al. 2001; Feil and Spratt 2001; Whitaker, Grogan et al. 2005; Hanage, Fraser et al. 2006; Papke, Zhaxybayeva et al. 2007).

However, the applicability of this recombination-based species concept depends on whether individuals from recombining populations can generally be classified into one recombining group or another. Among the highly recombinogenic *Neisseria*, MLST analyses have identified "fuzzy species", where the named species (classified phenotypically) broadly correspond to the identified genotypic clusters, but some individual genomes occupy intermediate genotypic space and some loci seem to have recombined across the species boundaries (Hanage, Fraser et al. 2005). Other reports have suggested that diverging lineages have experienced a sudden increase in recombination following an ecological shift, and may be losing their distinct character (Didelot, Achtman et al. 2007; Sheppard, McCarthy et al. 2008).

These ambiguities in delimiting species at the population level are not surprising given that the mechanisms of gene transfer do not suggest any clear criteria for distinguishing species, such as reproductive incompatibilities among eukaryotes. For example, bacteriophage do not seem to respect bacterial species boundaries – showing specificity for some strains within species even as they infect across species also (Sullivan, Waterbury et al. 2003). More importantly, bacterial species may be infected by numerous bacteriophages, each with a different host range. For example, *E. coli* is infected both by bacteriophage lambda, which has difficulty infecting other enteric bacteria due to differences in the LamB receptor protein, and bacteriophage P1, which infects many enteric bacteria. Indeed, genes encoding the P1 tail-fiber proteins have been used to create vectors for mutagenesis across numerous enteric bacterial species (Roncero, Sanderson et al. 1991). Even geographic barriers – while clearly slowing down migration and/or recombination in prokaryotes (Whitaker, Grogan et al. 2003; Whitaker, Grogan et al. 2005; Papke, Zhaxybayeva et al. 2007) – is not an absolute isolating mechanism for most bacteria. Clear-cut isolation only seems to occur when a single clone is isolated from all conspecifics (Dykhuizen 2000), such as obligate intracellular symbionts like *Buchnera* (Sullivan, Waterbury et al. 2003).

One thoroughly studied proposal for an isolating mechanism is the reliance of homologous recombination on high levels of sequence identity between the donor and recipient molecules. Studies of this mechanism have revealed that there is no clear threshold of sequence divergence beyond which homologous recombination ceases to act (Zawadzki, Roberts et al. 1995; Vulic, Dionisio et al. 1997; Majewski and Cohan 1999; Vulic, Lenski et al. 1999; Springer, Sander et al. 2004). Instead, the efficacy of recombination decreases as an exponential function of sequence divergence. Simulations have examined the dynamics that would arise as a population diversifies and sequence divergence inhibits recombination (Falush, Torpdahl et al. 2006; Hanage, Spratt et al. 2006; Fraser, Hanage et al. 2007). These models have suggested that sequence divergence could undermine the cohesive effects of recombination under the right conditions, allowing populations to diverge indefinitely, and providing a tipping point that is equated to “speciation” (Fraser, Hanage et al. 2007), even though low levels of recombination would occur even past this point.

However, such models are easily undermined by the complicated processes that determine sequence divergence in actual bacterial populations. Most notable is that sequence divergence varies greatly around the genome, as the effect of mutation and selection can vary by locus. Much of this variation is the result of gene expression levels, which is inversely associated with substitution rates (both synonymous and non-synonymous), and is probably effected through a combination of mutational effects and different levels of purifying selection (Eyre-

Walker and Bulmer 1995; Stoletzki and Eyre-Walker 2007). In a less systematic manner, selective sweeps purge diversity at associated loci (Guttman and Dykhuizen 1994), while divergent selection maintains excess diversity (Wildschutte and Lawrence 2007). What's more, a single recombination event between divergent strains can create large differences in divergence among neighboring loci, resulting in substantial differences in recombination efficiency among loci (Demerec and Ohta 1964). A consequence of this diversity is that mismatch induced speciation would not apply to the entire genome, but only to those regions of the genome that have passed the divergence tipping point.

A further limit to the idea of a speciation tipping point caused by sequence divergence is that it could always be negated by a few recombination events that decreased sequence divergence. Natural selection could be sufficient to cause the occasional "cross-species" allele to reach fixation in the recipient population (Didelot, Achtman et al. 2007; Sheppard, McCarthy et al. 2008), or the mismatch repair system responsible for recombination interference could be inactivated, allowing high levels of recombination even among diverged sequences (Demerec and Ohta 1964; Vulic, Dionisio et al. 1997).

Strains with defective mismatch-repair machinery are called "mutators" due to their tendency to accumulate mutations, along with their increased tendency to integrate foreign DNA into their chromosome. These mutator strains may arise frequently and persist in a population because they carry no immediate fitness detriment but can enable rapid adaptation to changing environments. They are observed both in laboratory evolution experiments (Negri, Morosini et al. 2002; Lenski, Winkworth et al. 2003) and in the wild (del Campo, Morosini et al. 2005). Evidence of recombination in natural populations indicates that bacterial lineages may pass through periods of elevated recombination due to the disruption of these genes, during which they rapidly diversify due to the import of novel gene sequences (Hanage, Fraser et al. 2009), ultimately returning to infrequent recombination when the anti-mutator genes are repaired by recombination (Denamur, Lecointre et al. 2000; Brown, LeClerc et al. 2001).

1.1.2 Prokaryotic species as ecotypes

As described above, natural selection can shape the patterns of gene exchange among prokaryotic genomes, leading to both convergence and divergence. This fact, combined with the fact that prokaryotic populations can expand greatly without gene exchange (Fraser, Hanage et al. 2005), has contributed to arguments that ecological factors are the essential forces behind bacterial evolution, and gene exchange is incidental to the definition and identification of species (Cohan 2001).

Levin (1981), extending ideas of Atwood *et al.* (1951), proposed periodic selection as a mechanism through which microbial population could retain similarity and avoid mutation-driven diversification. Here, when selectively-beneficial mutations sweep a population, the entire chromosome "hitchhikes" with it, thus purging variability at all loci. The resulting clonal expansion of the strain bearing the beneficial allele is bounded only by its ability to out-compete similar strains lacking this allele. Cohan and colleagues have termed such lineages as ecotypes (Cohan 2001; Gevers, Cohan et al. 2005; Cohan and Perry 2007; Koeppel, Perry et al. 2008), arguing that their sweeps purge genetic variability only within ecologically-identical strains. Hence, periodic selection events result in genotypic cohesion in a bacterial population.

In many ways, ecotypes have properties that are associated with species. Similarity is maintained among individuals in a population by an active process, groups are clearly differentiated from one another by ecological distinctiveness, and there is a mechanism (fixation of beneficial mutations) that can lead to lineage separation. Thus ecotypes could be considered one of the most fundamental units of organization of bacterial strains. But ecotypes may not be sufficiently stable to warrant identification as a truly distinct form of life.

From a genetic perspective, the scope of an ecotype – that is, the boundaries of the population encompassed by a periodic selection event – is a function of the nature of the beneficial mutation driving periodic selection. Mutations of small benefit would define a narrow ecotype, whereas those with greater benefits would purge variability among a group of more diverse strains. Distinguishing among ecotypes has proven difficult even in laboratory model systems, where closely related ecotypes have demonstrated the ability to invade each other's niche by adaptive processes (Dykhuizen and Dean 2004). The stability of ecotypes may also be undermined by source-sink processes, illustrated by opportunistic infections, where variants of a parental ecotype regularly invade a new environment, and adapt to it, before going extinct and opening the environment to recolonization from the parental species (Chattopadhyay, Feldgarden et al. 2007).

Another complication in the ecotype view of species, both conceptually and practically, is that the evolutionary potential of an ecotype may be strongly influenced by the limited collection of genomes that it can acquire genetic material from by recombination. Beneficial alleles that arise within an ecotype may also spread to a much larger, and more diverse, set of strains via homologous recombination (Guttman and Dykhuizen 1994; Guttman and Dykhuizen 1994; Cohan 2001). By preventing the gradual divergence of ecotypes, recombination could obscure the phylogenetic patterns produced by periodic selection (Levin 1981).

Indeed, the maintenance of diversity among several ecotypes within a recombining population may be necessary to generate the diversity that permits homologous recombination to act as a source of evolutionary novelty for recipients. Considering the interaction between ecological differentiation and genetic exchange in clonally reproducing populations, we can imagine two interacting scales of population structure. At the larger scale is the recombining population, and within it are several ecotypes. The ecotype populations may expand and contract as they encounter good growth conditions or acquire adaptive traits, while exchanging alleles with other ecotypes within the recombining population. New ecotypes may arise and go extinct frequently, while the recombining population is maintained for a longer period due to its greater inclusivity.

1.2 ADAPTATION IN PROKARYOTES

Bacteria exhibit an extraordinary ability to adapt to novel environments, and even to specialize within a seemingly homogenous environment. Upon being introduced into novel laboratory environments, *E. coli* fitness regularly increases by 20% over the course of several months, following multiple adaptive paths (Maharjan, Seeto et al. 2006; Cooper and Lenski 2010). In spatially structured environments, clones rapidly diversify into specialists adapted to the different regions (Rainey and Travisano 1998). Even in the absence of spatial structure, specialists develop from a shared ancestor, possibly due to metabolic specialization (Rozen and Lenski 2000). This metabolic specialization may or may not be stable, as strains previously adapted to feeding on different sugars do not necessarily manage to coexist when co-cultured with both sugars present (Dykhuizen and Dean 2004).

Adaptation in the wild is more difficult to trace, but closely related Bacillus isolates from an "evolution canyon" show subtle adaptations to growth on north-facing and south-facing slopes, which vary drastically in solar exposure (Connor, Sikorski et al. 2010). More drastic adaptations are apparent in response to exposure to man-made antibacterial compounds (Holt, Parkhill et al. 2008).

Much of the physiological variation seen between species can be traced to the patterns of gene presence/absence among bacteria. Within species, there is extensive variation in both the number and identity of genes, sometimes producing noticeable phenotypic differences. For instance, the first three *E. coli* genomes sequenced ranged in size from 4288 to 5063 protein coding genes, but only shared 2996 of these genes (Welch, Burland et al. 2002). The genomes of *Shigella* species are distinguished from *E. coli* primarily by the acquisition of particular pathogenicity genes (Lan and Reeves 2001). Even within a single serotype, *E. coli* O157:H7, variable presence was detected for over 200 of the 4753 genes from the Saki strain that were examined in four other strains of the same serotype (Wick, Qi et al. 2005). Gene presence/absence polymorphisms have been observed to spread among *E. coli* by recombination (Schubert, Darlu et al. 2009).

Many laboratory adaptations of bacteria have been observed to result from the gain or loss of genes, though gene gain necessarily occurs under extremely unnatural conditions, such as selection for phenotypes of interest following transformation with libraries constructed from metagenomic DNA (Handelsman 2004). Another commonly identified avenue of adaptation is through amino acid substitutions that produce modified biochemical activities in gene products (Lunzer, Miller et al. 2005). Widespread positive selection for amino acid substitutions has been inferred from several phylogenetic studies of bacterial genomes (Orsi, Sun et al. 2008; Lefebure and Stanhope 2009; Soyer, Orsi et al. 2009). Changes in gene expression patterns are also commonly cited as the target for natural selection (often a consequence of amino acid changes in or inactivation of regulatory proteins). A long-term study of *E. coli* adaptation to laboratory conditions resulted in many fitness-increasing changes in gene expression, which emerged independently in parallel replicates of evolving populations (Cooper, Remold et al. 2008). Adaptation of the regulatory system is also apparent in comparisons between different pathogens within the Enterobacteriaceae; for instance, the PhoPQ system activates different (yet overlapping) sets of genes in *Yersinia pestis* and *Salmonella enterica* (Perez, Shin et al. 2009).

One form of adaptation that is difficult to observe in the laboratory, despite being widespread in nature, is codon adaptation. Selection on synonymous codons produces systematic biases among the open reading frames found in a genome, where the frequency of certain codons increase relative to their synonyms as a result of selection for that codon (Ikemura 1981). While codon selection is not the only selective force that affects the nucleotide identity of synonymous sites, it is the primary selective force in many bacteria, with the less-preferred codons existing as a result of mutation and genetic drift (Bulmer 1991; Smith and Eyre-Walker 2001). This bias increases in tandem with the expression level of the ORF, indicating stronger selection in these genes (Sharp, Bailes et al. 2005). Once codon usage is optimized in a gene, purifying selection reduces population polymorphism and divergence between species (Sharp, Emery et al. 2010).

The genes encoding core physiological processes often exhibit high frequencies of preferred codon usage. Aside from widely conserved, highly expressed genes (*e.g.,* ribosomal proteins (Sharp, Bailes et al. 2005)), enrichment for preferred codon usage is also seen in genes that are distinctive to particular groups of bacteria (*e.g.,* photosynthesis genes in cyanobacteria (Karlin and Mrazek 2000)), suggesting that codon selection acts beyond those genes that are essential for all organisms.

While differences in preferred codon usage have been noted among orthologous genes (Karlin and Mrazek 2000), these differences have not been examined quantitatively and therefore the extent of such changes is unknown. However, since codon bias tracks the gene expression level, changes in codon selection may be common, regardless of whether the change in gene expression is a consequence of regulatory changes or of simple environmental changes.

All of the above modes of adaptation have the potential to produce population structures that may be identified as species – groups of bacteria sharing similarity due to their shared ancestry and adaptation to the same niche. However, such structure could be limited to particular loci within the genome, if recombinational processes are able to spread alleles into genomes that are adapted to alternative (but possibly overlapping) niches. The interplay of homologous recombination and adaptation in these organisms is as yet unresolved, and likely to play out in myriad ways, due to the vast ecological and genetic diversity of prokaryotes (Doolittle and Zhaxybayeva 2009).

2.0 TEMPORAL FRAGMENTATION OF SPECIATION IN BACTERIA

The proper identification and delineation of bacterial species plays a critical role in medical diagnosis, food safety, epidemiology and bioterrorism. Human responses are guided by perceptions of the biological properties and capabilities of a named species, as well as an understanding of their natural variability and potential to change. The Biological Species Concept (BSC) considers a species to be a group of organisms that readily exchange genetic information only with each other (Mayr 1942). In eukaryotes, recombination – here defined as allelic exchange – is often tied to reproduction, whereby meiosis is followed by the karyogamy of two entire haploid genomes. Consequently, as new species arise, genetic isolation would occur simultaneously for all loci, meaning that all pairs of orthologous genes would be diverging for approximately the same amount of time. While bacterial speciation is a complex process (Cohan 2001; Lawrence 2002; Gevers, Cohan et al. 2005), the BSC has also been applied to bacteria such as *Escherichia coli* (Dykhuizen and Green 1991). Here, recombination involves the occasional, unidirectional transfer of small DNA fragments from one strain into the homologous locus of another. Because only a small portion of the genome is transferred, orthologs would have diverged for differing amounts of time (Figure 2.1AB). Inter-species transfer is limited by mismatch repair systems, which reject recombination when donor and recipient sequences are not nearly identical (Vulic, Dionisio et al. 1997). Yet this process does not speak to how recombination ceases within a group of recombining strains – wherein allelic differences are few – thereby allowing two genetically distinct groups to form.

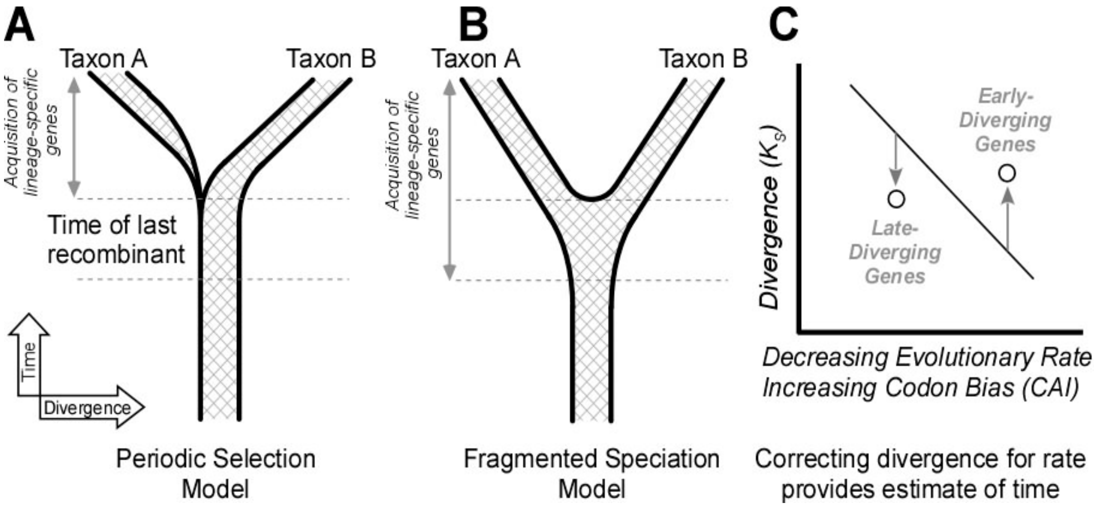

A
Taxon A
Taxon B
Acquisition of lineage-specific genes
Time of last recombinant
Time
Divergence
Periodic Selection Model
B
Taxon A
Taxon B
Acquisition of lineage-specific genes
Fragmented Speciation Model
C
Divergence (K_s)
Early-Diverging Genes
Late-Diverging Genes
Decreasing Evolutionary Rate
Increasing Codon Bias (CAI)
Correcting divergence for rate provides estimate of time

Figure 2.1 Models of bacterial lineage diversification

(**A**) Rapid isolation model for bacterial speciation. A population divides into two, each of which adapts to a particular niche without further genetic input from the other. While recombination (cross-hatching) may produce different times of divergence between genes in taxa A & B (region between horizontal, dashed lines), no cross-population recombination occurred after lineage-specific genes were acquired (vertical arrow). (**B**) Temporal fragmentation model for bacterial speciation. Ecological diversification, involving lineage-specific gene acquisition, occurs in the context of genetic exchange (vertical arrow). (**C**) The number of synonymous substitutions (K_s) between homologous sequences is a function of both the rate of substitutions (estimated by CAI) and the amount of time that the sequences have been diverging.

Given the vast range of recombination rates seen for bacterial populations (Feil and Spratt 2001; Hanage, Fraser et al. 2006), we propose two models for lineage separation following the emergence of and selection for a differentially-adapted genotype. First, nucleotide substitutions and lineage-specific loci could be acquired quickly relative to the rate of recombination (Falush, Torpdahl et al. 2006). Under this model, genetic isolation would be established at approximately the same time for all orthologs (Figure 2.1A). Alternatively, niche-specific changes may be acquired more slowly relative to the rate of recombination, and gene conversion events would continue at loci unlinked to niche-defining genes (Figure 2.1B) (Fraser, Hanage et al. 2007). Here, variability-purging selective sweeps (Guttman and Dykhuizen 1994) would occur only at loci that are unlinked to genes imparting ecological distinctiveness, because recombinants would likely be poorly-adapted to either environment and be counter-selected (Lawrence 2002). Thus, alleles may undergo selective sweeps across ‘species’ boundaries when not proximate to niche-specific loci; over time, all loci would become genetically isolated as mismatches accumulate and the number of niche-specific loci increases. This fragmented speciation model further predicts that early-diverging genes will be linked to loci that interfered with effective inter-lineage recombination, such as those encoding niche-specific traits or those subject to diversifying or frequency-dependent selection (Milkman 1997; Wildschutte, Wolfe et al. 2004).

2.1 SEQUENCE DIVERGENCE VARIATION AMONG SITES RESULTING FROM HOMOLOGOUS RECOMBINATION

To detect temporal fragmentation of speciation, we must first distinguish between early- and late-diverging orthologues. Since divergence is a function of both time and evolutionary rate, time may be estimated from divergence once evolutionary rate is determined (Figure 2.1C). At synonymous sites, evolutionary rate can be estimated from the codon adaptation index (CAI), an intragenomic, time-independent measure of selection (Sharp and Li 1987). Divergence at synonymous sites is measured as K_s; because K_s decreases as CAI increases (Figure 2.1C), CAI can be used to generate the expected value of K_s if divergence times are uniform among genes (see Methods). Early-diverging orthologs will have larger K_s values than expected because more time has elapsed since their divergence, and late-diverging orthologs will have lower values than expected (Figure 2.1C).

We applied this method to the genomes of *Escherichia coli* and *Salmonella enterica*; recombination is common within either taxon (Guttman and Dykhuizen 1994; Feil, Holmes et al. 2001; Falush, Torpdahl et al. 2006), while inter-species recombination is inhibited (Daubin, Moran et al. 2003). To assemble a robust data set, we analyzed genes with orthologues present in each of three different *E. coli* and *S. enterica* genomes representing the most diverse available genome sequences for these species (see Methods). These strains share a chromosomal backbone of 2677 sets of orthologs. CAI and between-species K_s were computed for protein-coding genes and the relationship was fit by polynomial regression (Figure 2.2). As expected, increasing selection for preferred codons (high CAI) is generally reflected by lower divergence (K_s). We ignored 527 pairs of genes with <50 synonymous sites, whose K_s values were in saturation, or where the relationship between CAI and K_s was unclear (Figure 2.2). The effect of map position on K_s (Sharp, Shields et al. 1989) was estimated by treating CAI-corrected-K_s as a linear function of the gene's distance from the *E. coli* K12 replication origin (Figure 2.2 inset). Ultimately, relative divergences of 2150 genes along the chromosomal backbone (Figure 2.3) were calculated as the ratio of observed K_s to that expected from CAI and map position (see Methods).

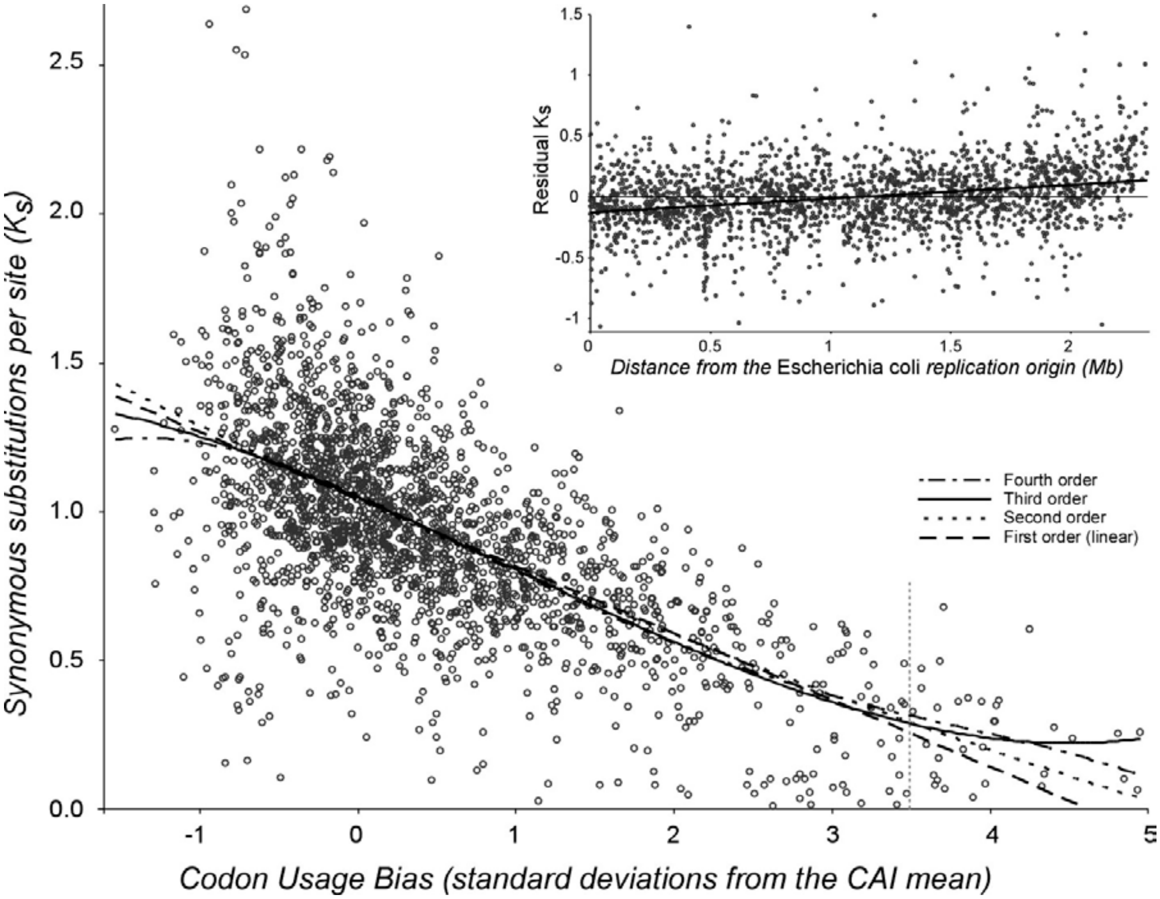


Figure 2.2 Influences on synonymous substitution rate.

Synonymous substitutions as a function of mean codon bias of the ORFs, with polynomial least-squares regression lines. The vertical line indicates the value of CAI above which the relationship between CAI on Ks was unclear. Inset: Scatter plot of third-order regression residuals as a function of distance from the E. coli K12-MG1655origin, with a linear least-squares regression line.

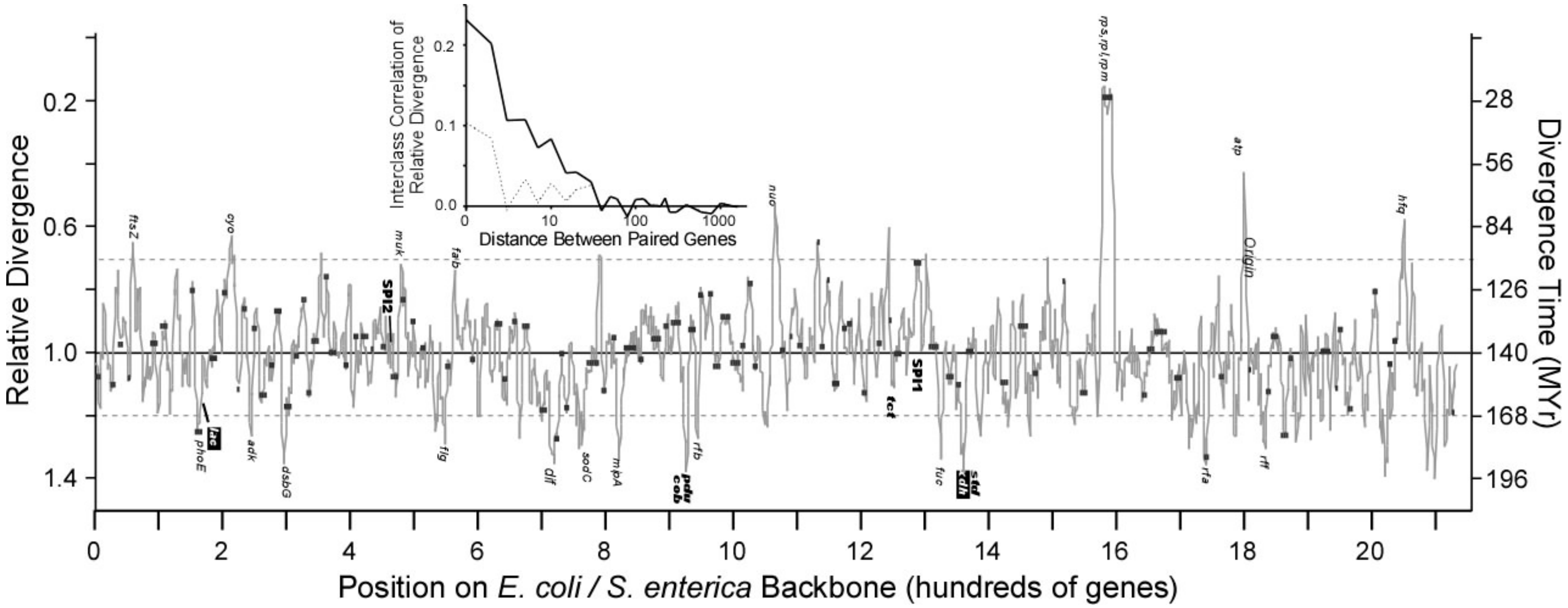


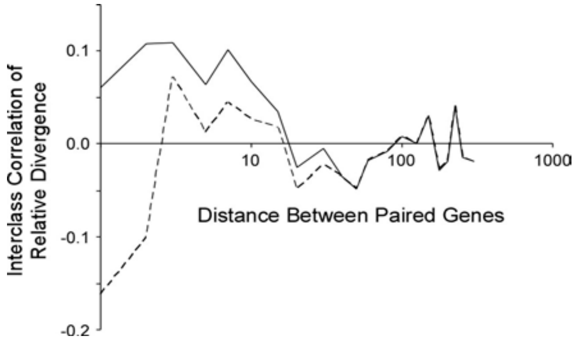
Figure 2.3 Time of divergence of chromosomal regions

Relative divergence for orthologs is plotted against E. coli K12-MG1655 chromosomal position, averaged across a 7-gene window; dark lines indicate divergence times of regions longer than 6 genes. Dashed lines delineate 95% of the range of divergence values. Shared loci are noted in italics; Escherichia- and Salmonella-specific are noted at their corresponding location on the backbone in inverse and bold-faced type. Inset: Intraclass correlations of relative divergence for gene pairs as a function of distance; solid line, all gene pairs; dotted line, gene pairs not within runs of consecutive genes transcribed in the same direction.

While stochastic variation in the accumulation of substitutions will account for much of the variability in relative divergence, genes that have recombined more recently will tend to have lower values. To detect this footprint of recombination, we rely on a mechanistic constraint of bacterial gene exchange: physically-proximate genes will be transferred in the same recombination events. As a result, early- and late-diverging genes will not be randomly distributed throughout the genome, but will cluster in regions defined by the most recent exchange between the lineages. Therefore, physical association among genes with K_s values higher or lower than expected can be taken as evidence for recombination. The scale of recombination regions was estimated from the correlation of relative divergence values for pairs of orthologs (Figure 2.3 inset, solid line). Adjacent genes showed a strong intraclass correlation (ICC=24%, $P < 10^{-24}$, F-test) which decreased as pairs of orthologs became more distantly situated on the chromosomal backbone, becoming undetectable when position differed by more than 20 genes (~32 kb). These results are consistent with the boundary of recombination interference observed for the *rfb* locus (Milkman and Bridges 1993; Milkman, Jaeger et al. 2003). To remove any correlation in K_s among cotranscribed genes resulting from transcription-associated repair or selection for mRNA stability, ICCs were recalculated having excluded comparisons between genes consecutively transcribed in the same direction. Despite the decreased sample size, a significant correlation (11%, $P < 10^{-2}$, F-test) extended to the same distance (Figure 2.3 inset, dashed line). These data show that genes diverged at significantly different times at different locations within the *E. coli/S. enterica* chromosomes.

Regions of potential recombination events were delineated using an agglomerative clustering algorithm (see Methods), identifying regions wherein variability in relative divergence was lowest; the most robust segments – maximum SE=0.013; each longer than 6 genes, covering 49% of genes – appear as black bars in Figure 2.3. If *E. coli* and *S. enterica* genes have been diverging for ~140 million years (MYr) on average (Ochman and Wilson 1988), the distribution of divergence times shows that genetic isolation developed over a period of ~70 MYr (Figure 2.3, region between dashed lines). As expected, among the first regions to diverge were those containing shared genes that produce surface structures, such as the *rfa*, *rfb, rff, flg*, *mipA* and *phoE* genes, which are often subject to frequency-dependent or diversifying selection. Other early-diverging regions are associated with differences in gene content, such as those adjacent to the *S. enterica cbi*, *pdu*, *std,* and *tct* operons, and the *E. coli lac* and *xdh* operons (Figure 2.3), most of which encode physiological functions that distinguish the two species (Rambach 1990). In contrast, the regions around *Salmonella* Pathogenicity Islands SPI1 and SPI2 diverged more recently, suggesting that they do not represent the initial differences separating *E. coli* and *S. enterica*. Even though relative divergence has been corrected for evolutionary rate, the major peaks in Figure 2.3 consistently represent clusters of genes with high CAI values. We posit that these slowly evolving regions offer longer stretches of DNA free of substitutions, thereby postponing the establishment of recombination barriers.

As a control, we compared the genomes of *Buchnera aphidicola* strains APS and Sg whose 489 conserved protein-coding genes show divergence similar to *E. coli* / *S. enterica* comparisons (mean K_s of 0.89, and 0.97, respectively). *Buchnera* are *recA*-deficient intracellular endosymbionts believed to rarely recombine (Tamas, Klasson et al. 2002). We would expect lineage diversification to affect all loci simultaneously and analysis of these genomes showed no significant correlation in relative divergence



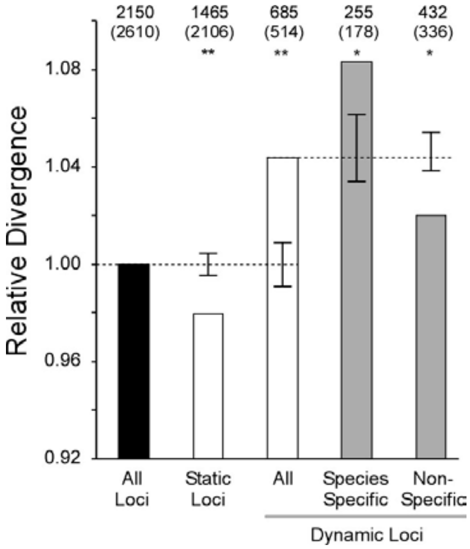
The solid line denotes ICCs calculated from all gene pairs of a given distance. The dotted line denotes ICC calculated having excluded comparisons between genes that are within a single run of consecutive genes transcribed in the same direction.

Figure 2.4 Correlation of relative coalescence times for pairs of genes based on their relative positions in the *Buchnera* backbone

for adjacent genes (Figure 2.4). To control for the lower sample size, we examined all regions of the *E. coli / S. enterica* genomes with equal numbers of genes; these regions showed significant ICCs that were invariably stronger than the *Buchnera* value, suggesting that the lack of correlation for *Buchnera* reflects a lack of recombination.

2.2 CHANGES IN GENE CONTENT AND HOMOLOGOUS RECOMBINATION

The Fragmented Speciation Model (Figure 2.1B) predicts that genetic and ecological differentiation developed even as high levels of recombination continued at loci not conferring ecological distinctiveness. Niche-specific traits often arise by gene gain or loss, where altered physiology allows cells to thrive in conditions that are hostile to parental strains (Ochman, Lawrence et al. 2000). If recombination between incipient *E. coli* and *S. enterica* continued at some loci even after lineage-specific loci had arisen, then the regions around the lineage-specific genes should be among the first to be genetically isolated, because recombination in those regions would have eliminated the gene-content differences at those loci. Conversely, if the differences in gene content developed only after inter-lineage recombination had effectively ceased, then these genes should be distributed without regard to the divergence time of the surrounding region.



Bars show the mean relative divergence of sets of orthologues (numbers above bars) classified according to adjacency to loci that distinguish genomes (number of loci in parentheses). Error bars show 1 SE for the distribution of randomized samples. *, P < 0.01; **, P < 0.000001.

Figure 2.5 Relative divergence based on region character

We defined a locus as a pair of genes in the *E. coli/S. enterica* comparative backbone. There were 514 dynamic loci (685 genes; some genes contributed to 2 loci), at which some *E. coli* strain contained a gene between the conserved pair that was absent from all *S. enterica* genomes, or vice versa (see Methods); the remaining 2106 static loci showed no insertion/deletion events (Figure 2.5, white bars). Genes at static loci have an average divergence time 4.4% younger than the average for the entire genome ($P < 10^{-5}$ by randomization), likely because the longer stretches of uninterruptible, slowly-evolving genes allow for continued recombination. A fraction of dynamic loci (178 loci, Table S2) show species-specific differences, whereat the conserved gene pair was interrupted in the three strains of one species by genes absent from the other species; these loci would include those whereat differences arose while the *E. coli* and *S. enterica* lineages were diverging. Other dynamic loci – *e.g.*, those where only a single strain shows a difference – would have arisen only after recombination had effectively ceased between the two lineages. Genes adjacent to species-specific loci are 6.2% older than genes adjacent to other dynamic loci ($P < 10^{-2}$ by randomization; Figure 2.5, gray bars); thus, species-specific genes are not distributed randomly in the chromosomal backbone but are found preferentially in the older regions, indicating that the incipient *E. coli* and *S. enterica* lineages continued to participate in recombination at loci unlinked to lineage-specific genes.

2.3 CONCLUSIONS

In contrast to the rapid formation of eukaryotic species boundaries, which are generally established within a couple million years (Coyne and Orr, 2004), the ~70 MYr time frame over which genetic isolation evolved between *E. coli* and *S. enterica* represents a temporal fragmentation of speciation. Because separate lineages arise within populations that continue to recombine at some loci for tens of millions of years, relationships among species inferred from few loci may underestimate their underlying complexity. Taxa may show different relationships depending on which genes are compared. Long periods of partial genetic isolation allows extant, named species – such as *Escherichia coli* itself – to contain multiple nascent species. Although one can observe recombination at some genes within *E. coli* as a whole, strains also have niche-specific loci that may act as genetic progenitors for the creation of new species. That is, a clean distinction between intra- and inter-specific variability may not be possible (Hanage, Fraser et al. 2005), and clearly-defined species cannot represent newly-formed lineages. Therefore the Dykhuizen and Green species concept (Dykhuizen and Green 1991) – gene phylogenies are congruent among representatives of different species, but are incongruent among members of the same species – works to delineate long-established species, but fails to recognize incipient species.

3.0 PHYLOGENETIC INCONGRUENCE ARISING FROM FRAGMENTED SPECIATION IN ENTERIC BACTERIA

At first glance, prokaryotes appear to have simple, well-ordered relationships resulting from asexual reproduction and divergence by mutation. However, homologous recombination between closely related strains can lead to complex, non-clonal relationships (Dykhuizen and Green 1991). Recombination has implications that are so profound that its potential within populations is often taken to be the definitive feature of species. Mayr's Biological Species Concept (BSC) frames species in the context of reproductive barriers, whereby only conspecific individuals exchange genes; individuals that fail to recombine represent different species (Mayr 1942). Despite its formulation for sexual eukaryotes, Dykhuizen and Green (Dykhuizen and Green 1991) proposed that the BSC could apply to bacteria; operationally, phylogenies of orthologous genes would be identical for strains of different species but demonstrably different for strains within species due to recombination. Several studies have applied these criteria to multi-locus phylogenetic analysis of prokaryotes, supporting the notion that there are distinct groups of organisms experiencing recombination within each group but not between them (Wertz, Goldstone et al. 2003).

Despite these results, the BSC may not be generally applicable to prokaryotes, even for taxa that undergo high rates of recombination. While eukaryotic recombination affects all genes during meiosis, recombination in prokaryotes involves only a small fragment of DNA being introduced into a cell via transformation, phage-mediated transduction or plasmid conjugation. In both prokaryotes and eukaryotes, recombinants may be counter-selected when genomic incompatibilities reduce hybrid fitness. In eukaryotes, this inhibits gene exchange genome-wide (Rieseberg, Wood et al. 2006), while in prokaryotes recombination interference affects only that small portion of the genome that carries the incompatible DNA. Similarly, antirecombination driven by sequence divergence and the mismatch repair system causes hybrid sterility in eukaryotes such as *Saccharomyces* (Greig 2009), while in prokaryotes it simply prevents the integration of the particular sequence that has diverged from the recipient (Shen and Huang 1986).

As a result, barriers to recombination in bacteria are more modular than in eukaryotes. Consider a bacterial population freely recombining at all loci. Subpopulations can develop through genetic isolation of only a few loci, driven by ecology or sequence divergence; such subpopulations – recombining at many loci but genetically isolated at other loci – could be numerous within a larger population. Consistent with this fragmented speciation model (Lawrence 2002), several Multilocus Sequence Analysis (MLSA) studies identified closely related populations that appear to be recombining at some loci but remain genetically isolated at others (Spratt, Bowler et al. 1992; Hanage, Fraser et al. 2005). In addition, a comparison of *Escherichia* and *Salmonella* genomes revealed extensive variation in the level of sequence divergence across regions of the chromosome, suggesting that many regions experienced homogenizing recombination as much as 70 million years after other regions had become isolated (Chapter 2). Critically, excessively diverged regions were clustered around the loci where gene gains or losses distinguish *Escherichia* from *Salmonella;* such adaptive changes in gene inventory could have contributed to ecological differentiation within the recombining ancestral population and been the focus of selection against recombination.

If recombination barriers are imparted gradually as populations split, they may not be complete before each descendent population splits again. The stepwise acquisition of genetic isolation at different locations around the chromosome would lead to differing phylogenies of orthologous genes, resulting in the lack of clear organismal relationships. Alternatively, recombination could cease for all loci instantly when each population splits, as suggested for *Yersinia pestis* (Dykhuizen 2000). In this instant speciation model, all recombination events occur prior to the acquisition of genome-wide genetic isolation. Any phylogenetic incongruence would result from the partitioning of ancestral variation among descendent lineages, which would confound our ability to discern otherwise robust organismal relationships. Here, we test these models directly.

3.1 PHYLOGENETIC DISCORDANCE IN THE ENTEROBACTERIACEAE DOES NOT REFLECT ONGOING RECOMBINATION

To identify taxa with potentially confounded relationships, we looked within the well characterized species-rich clade of enteric bacteria. To establish a reference phylogeny, we aligned a core genome containing 1174 orthologous ORFs in each of 17 genomes (Table S4), with <15% of aligned sites having gaps in any genome. A NeighborNet (Bryant and Moulton 2004) analysis of the concatenated codon alignment of these genes shows conflicting phylogenetic signals among these genomes (Figure 3.1). Regions with conflicting signal may reflect the incongruent histories among genes due to recombination (Lawrence and Retchless 2010). Examining each gene independently by maximum likelihood (ML), there is near universal support for the separation of *Erwinia*, *Dickeya, Pectobacterium, Serratia,* and *Yersinia* from *Cronobacter* and the other genomes (>99% of those alignments with a single topology in the 90% confidence limit). These taxa are used as outgroup genomes in subsequent tests.

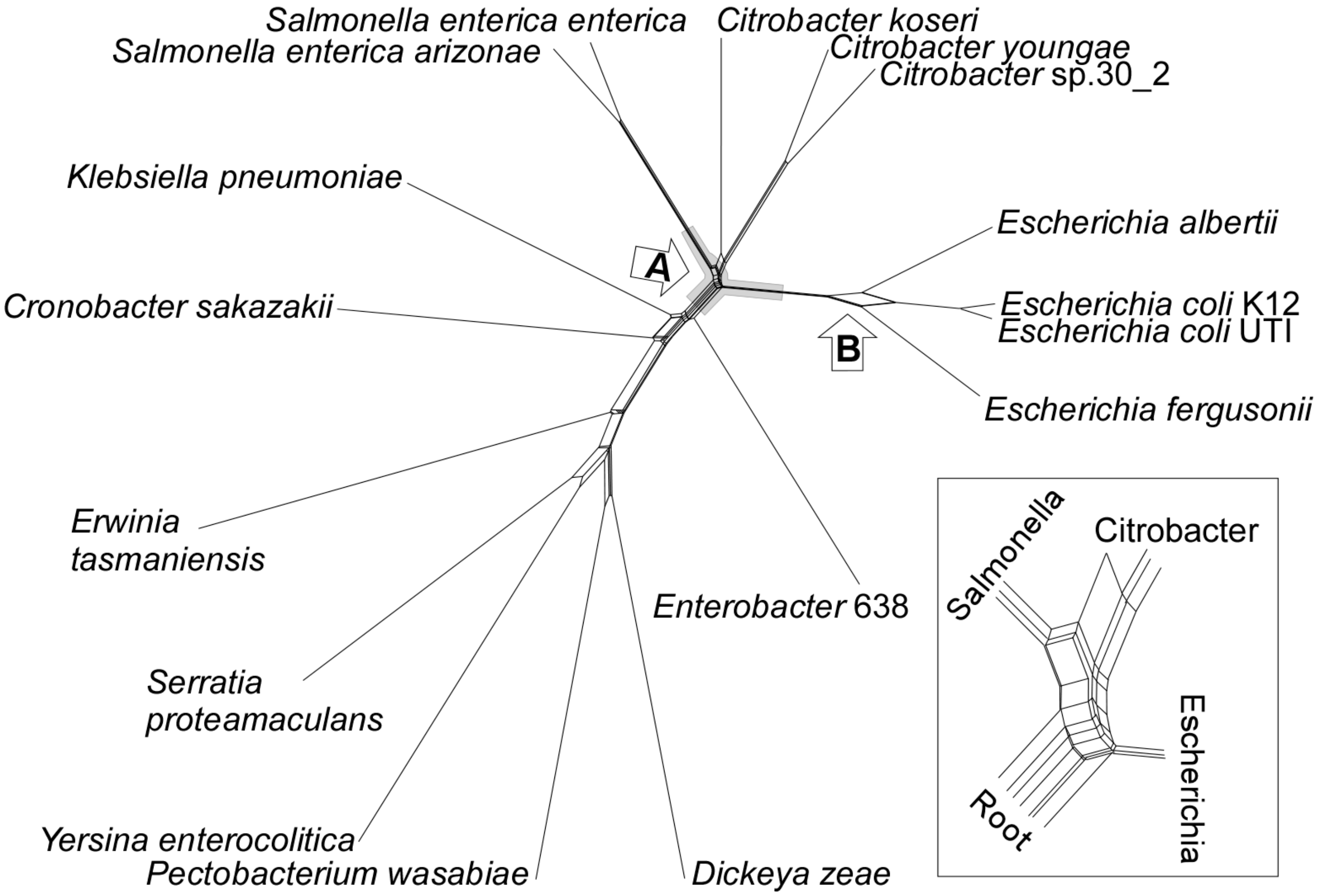


Figure 3.1 Phylogenetic network of enteric bacteria

The NeighborNet (Bryant and Moulton 2004) dendrogram was calculated by SplitsTree. Shaded region indicates the range for placement of the node separating *Escherichia coli* and *Salmonella enterica* according to relative divergence analysis (Chapter 2). Inset focuses on the divergence of *Escherichia*, *Salmonella* and *Citrobacter* (arrow A).

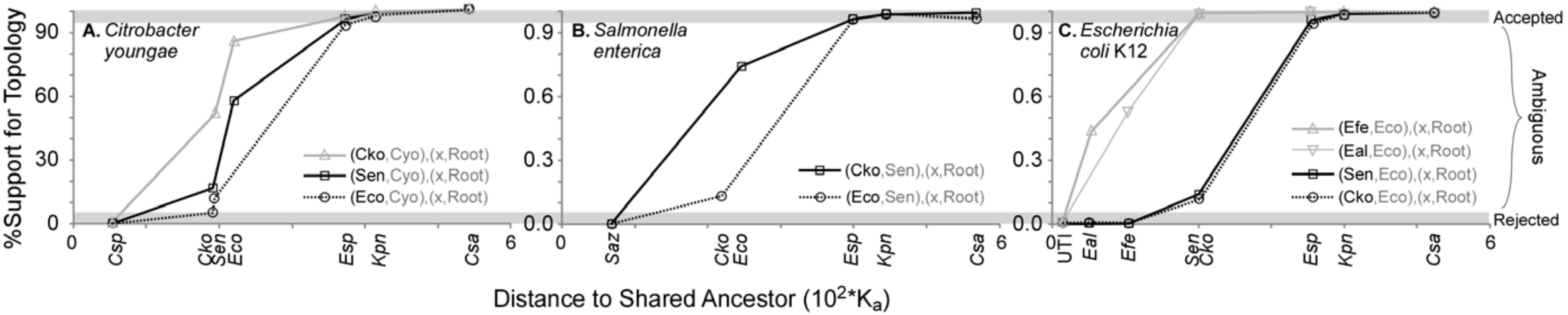


Figure 3.2 Percent of genes supporting clades

Each panel shows how often a reference genome was found to be the sister taxon to a second genome (defining each trendline, see legend) in a 4-taxa ML phylogeny. Each phylogeny included the reference pair, a constant outgroup and each of a series of test genomes. Results are plotted according to the distance from the reference genome to the node leading to the test genome on a neighbor-joining tree based on estimates of amino acid substitution counts (K_a). Gene counts were limited to those ORF alignments that generated substantial likelihood support for a single topology (90% CI, SH test). Quartet composition is listed to the right of each chart; "x" represents the test genome, which is identified on the distance axis. **Cko**; *Citrobacter koseri*; **Csp**, *C. sp*. 30_2; **Cyo**, *C. youngae*; **Csa**, *Cronobacter sakazakii*; **Eal**, *E. albertii*; **Eco**, *E. coli* MG1655; **Efe**, *E. fergusonii;* **Esp**, *Enterobacter sp*. 683; **Kpn**, *K. pneumonia*; **Saz**, *S. enterica arizonae*; **Sen**, *S. enterica* LT2; **UTI**, *E. coli* UTI89.

Using this outgroup, we examined reference pairs of taxa in a quartet analysis to test the robustness of their relationship with respect to an additional taxon (Figure 3.2). For high confidence phylogenies, the additional taxon should be either a clear outgroup (supporting the reference pair as sister taxa) or a clear ingroup (rejecting the pair as sister taxa). As expected from the NeighborNet results, virtually all genes supported the *Escherichia/Salmonella* pair when it was evaluated with either *Klebsiella* or *Cronobacter*, and virtually no gene supported this pair when evaluated with *E. albertii*, *E. fergusonii* (Figure 3.2C) or *S. arizonae* (Figure 3.2B). However, two regions of the NeighborNet phylogeny show substantial conflicting signal (Figure 3.1 regions A & B). The divergence of *Escherichia* and *Salmonella* has supported a fragmented speciation model (Chapter 2), whereby chromosomal regions became genetically isolated during an extended timeframe (shaded area in Figure 3.1). This range includes the nodes representing the divergence of the *Citrobacter* lineages, suggesting that the relationship between these three genera will be ambiguous (Figure 3.1, inset). This theme was reinforced by the individual gene quartet analyses, where relationships between *Escherichia* and *Salmonella* were ambiguous – with the pair being neither widely accepted nor widely rejected – when evaluated with *Citrobacter koseri* (14% accepted) or *C. youngae* (30% accepted). A similar pattern was observed for the *C. youngae/Salmonella* clade (Figure 3.2A).

Escherichia species represent one of the most recent radiations of bacteria recognized phylogenetically as separate species. In contrast to the divergence of *Escherichia*, *Salmonella* and *Citrobacter*, which have been separated for tens of millions of years, the three species of *Escherichia* are likely in the final throes of genetic isolation. MSLA data (Walk, Alm et al. 2009) is consistent with very low levels of recombination between otherwise distinct groups of *Escherichia* with divergence comparable to *E. coli* and *E. fergusonii*. The vast majority of genes in our analysis also supports the monophyly of *E. coli* K12 with *E. coli* UTI89 (Figure3.2C) as well as the monophyly of the *Escherichia* relative to other genera; however, the relationships between the three *Escherichia* species remain unclear. The *E. coli/E. albertii* clade was supported by 53% of genes and the alternative *E. coli/E. fergusonii* clade by 44% of genes. Thus these taxa represent the genesis of the phylogenetic ambiguity that plagues the relationships of *Escherichia*, *Salmonella* and *Citrobacter.*

These gene-based quartets provided no evidence for recent recombination between species of different genera, indicating that (for the genomes analyzed here) any substantial gene flow was limited to the time periods before the genera diversified into extant species. Yet with both radiations, the remaining phylogenetic incongruence may be interpreted several ways. The conflicting signal may simply represent noise, especially when inferring the relationships between the more distantly related *Escherichia*, *Salmonella* and *Citrobacter.* Alternatively, conflicting phylogenetic signal may reflect maintenance of ancestral polymorphism, whereby incomplete lineage sorting leads to ambiguous phylogenetic relationships even when genetic isolation occurs instantly for all genes (Pamilo and Nei 1988). Lastly, the fragmented speciation model predicts conflicting phylogenetic signal due to stepwise acquisition of barriers to recombination.

3.2 ROBUST ALTERNATIVE RELATIONSHIPS AMONG BACTERIAL GENERA

One expects the ratio of phylogenetic signal to noise to be weakest for very short branches, so one may posit that the inference of conflicting topologies described above simply reflects the lack of support for the ‘true’ organismal phylogeny. To test this hypothesis we determined if the support for alternative topologies is stronger than expected. Likelihood analyses were performed on alignments of sequences from 14 genomes representing the maximal available diversity among *Escherichia*, *Salmonella*, *Citrobacter* and *Klebsiella*, while maintaining the monophyly of each group (6, 4, 2, and 2 genomes, respectively; Table S4). The use of multiple taxa for each clade increased the signal-to-noise ratio. While 2028 potentially orthologous ORFs were present in each of the 14 genomes, we removed 705 unreliably aligned ORFs (>5% of their multiple sequence alignment contained gaps), 14 potentially paralogous ORFs for which syntenic neighboring ORFs could not be reliably identified, and 165 ORFs for which the monophyly of each of the four genera was not confidently supported by Bayesian analysis. For the 1144 remaining ORFs, the three possible relationships of the four genera were evaluated by codon-based maximum likelihood, holding the relationships within each genus fixed as defined by the Bayesian analysis.

A substantial number of alignments supported each of the three topologies (Figure 3.3AB) and several lines of evidence ruled out stochastic and systematic errors as the basis for these incongruent results. Among those alignments generating strong bootstrap support for a topology (Figure 3.3A), where bootstrap support thresholds provide conservative estimates of accuracy (Hillis and Bull 1993; Taylor and Piel 2004), no topology had the level of support expected if it were the true topology for all genes (dashed line). Furthermore, an excess of alignments rejected each topology with high confidence, indicating strong phylogenetic information at the gene level despite the widespread incongruence between genes (Figure 3.3B; $p < 0.01$ for all categories where individual genes are rejected at $p <= 0.25$; binomial test using threshold gene p-values as the expected probabilities). Unlike what would be expected for an unambiguous organismal phylogeny, large fractions of alignments reject the *Escherichia/Citrobacter* clade, the *Escherichia/Salmonella* clade and, for high confidence alignments (P < 0.25), the *Citrobacter/Salmonella* clade.

The proportions of alignments supporting each topology were robust to subsampling guided by statistical confidence and to a variety of other subsampling techniques that would purge different varieties of stochastic and systematic errors. Support for a single topology did not arise when we removed potentially mismatched orthologs, alignments with gaps or few informative sites, or any alignment generating inconsistent phylogenetic results using a codon-position model, other outgroups or fewer taxa (Table S1).

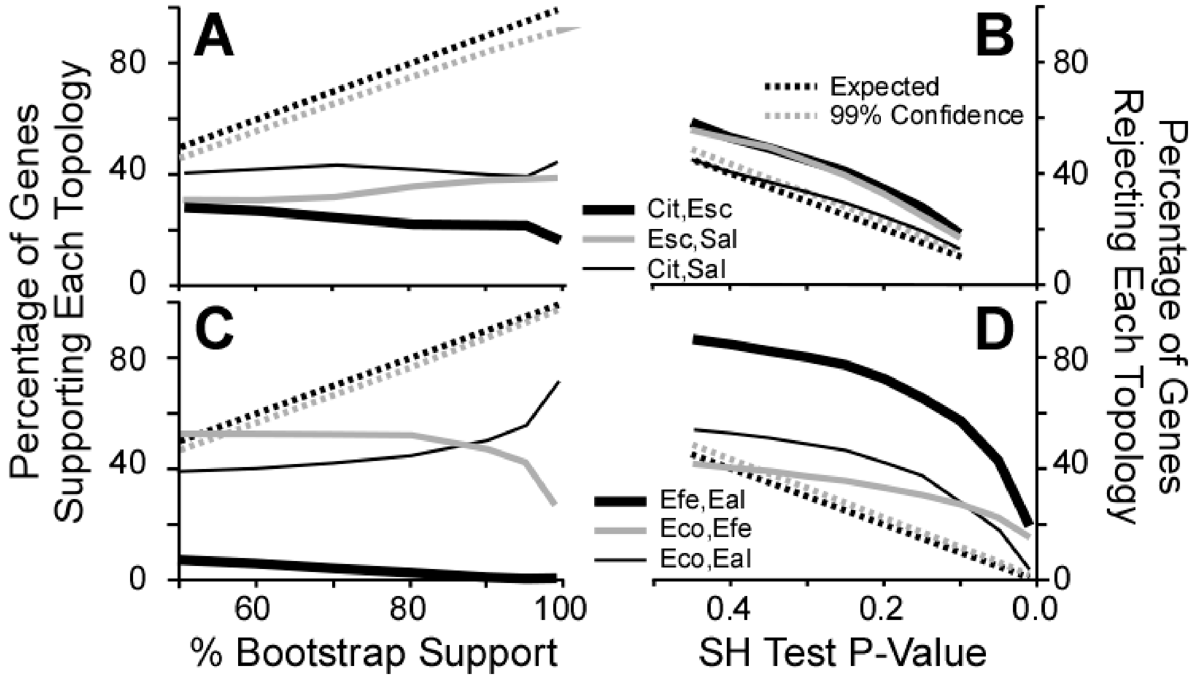


Figure 3.3 Phylogenetic discordance at all confidence levels

Measures of confidence on ML tests for individual genes do not match expectations for a genome-wide topology, either for the relationship among *Citrobacter*, *Escherichia*, and *Salmonella* (AB) or *E. albertii*, *E. coli*, and *E. fergusonii* (CD). Bootstrap support is expected to correspond to accuracy (AC), and SH test p-values provide expected frequencies of topology rejection (BD). Support for (AC) or rejection of (BD) alternative clades is indicated by trendlines. Within species quartets are in Fig. S2.

3.3 ROBUST ALTERNATIVE RELATIONSHIPS AMONG THE *ESCHERICHIA*

The same set of 1144 alignments were tested according to a codon-position maximum likelihood model applied to each of the three topologies involving *E. albertii*, *E. coli*, and *E. fergusonii*, with an outgroup comprising two genomes each of *Salmonella* and *Klebsiella*. As with the original tests (Figure 3.2C), roughly equal portions of alignments supported the clustering of *E. coli* with either *E. albertii* or *E. fergusonii*, while *E. albertii* and *E. fergusonii* rarely clustered together (Figure 3.3CD). None of the topologies had support from a sufficient number of genes to justify the hypothesis that it is the single true topology and the others are artifactual (Figure 3.3C), while each topology was rejected more often than expected by chance (Figure 3.3D). Support for a single topology did not arise when we removed potentially mismatched orthologs, or alignments with gaps, few informative sites, or that did not produce identical results when *Klebsiella*, *Citrobacter*, or *Salmonella* were used as single outgroups (Table S2).

Interestingly, alignments supporting the *E. albertii/E. fergusonii* clade are rare (Figure 3.3CD), illustrating the complexity of the isolation process. *E. fergusonii* and *E. albertii* may have arisen from small populations that rarely encountered each other, but continued to recombine with *E. coli.* Alternatively, ecological differentiation may be greater between *E. fergusonii* and *E. albertii* than between either of them and *E. coli,* suppressing recombination between them more. The former explanation is consistent with elevated substitution rates in the lineages leading to *E. fergusonii* and *E. albertii* relative to *E. coli* observed previously (Walk, Alm et al. 2009), but is not exclusive of the latter explanation. Also of note, while the *E. albertii/E. coli* clade is favored by those genes with strongest bootstrap support, a greater number of genes support the *E. fergusonii/E. coli* clade (Figure 3.3CD). This could reflect the idiosyncratic nature of the fragmented speciation process, suggesting either that some loci in the *E. fergusonii* genome became isolated from the *E. albertii/E. coli* gene pool exceptionally early, or the *E. albertii/E. coli* gene pool maintained coherence at these loci until relatively recently.

As a final test of the robustness of the incongruence in both quartet analyses, we attempted to identify hidden likelihood support (Gatesy and Baker 2005) for a congruent topology by concatenating those alignments supporting each of the three topologies tested, then repeating the maximum likelihood analysis. In each case, we found unambiguous support for the topology that the genes had individually supported (no alternate topology within 99% confidence interval). To guard against the analysis being dominated by a few genes with the strongest support for the given topology, we repeated the analysis using only those genes that had 50-60% bootstrap support for the given topology. Again, we recovered unambiguous support for each topology except the *E. albertii/E. fergusonii* clade, which produced 64% bootstrap support for itself, but could not reject the *E. fergusonii/E. coli* clade.

3.4 CLUSTERING OF PHYLOGENETIC SIGNAL WITHIN THE CHROMOSOME

The above results suggest that no single phylogeny is appropriate to describe the relationship between *Escherichia, Salmonella* and *Citrobacter*, or between the three species of *Escherichia*. If recombination between nascent species were responsible for the phylogenetic incongruence, then the phylogenetically informative sites should be clustered in their respective genomes according to the topology that they support. To evaluate clustering, we concatenated single-gene, high-quality alignments of genes with reliable identification of neighboring ORFs. Supporting sites in the alignment were defined as those for which one topology is more parsimonious than the other two. Parsimony criteria were applied to nucleotide, amino acid and synonymous codon alignments, and two analyses were performed to measure the clustering of sites supporting each topology within the 1309 ORFs.

A runs test for randomness in the order of supporting sites indicated highly significant clustering of supporting sites by topology ($p << 10^{-10}$, Table 3.1). That is, for both the *Escherichia/Salmonella/Citrobacter* and *E. albertii/E. coli/E. fergusonii* clades, sites supporting each of the conflicting topologies were clustered in these genomes. To test if sites supporting each topology were clustered, we repeated the runs test by omitting each topology in turn; significant clustering was still observed (Table S3, $p < 10^{-5}$). To investigate the scale of clustering as a function of distance between supporting sites, pairs of sites were binned according to distance, calculating the frequency that both sites within the binned distance range supported the same topology (Figure 3.4). For all analyses, there was a clear enrichment of sites supporting the same topology over the frequency expected from genome averages. This is most noticeable for the analysis of *Escherichia* species, where phylogenetic signal is expected to be the strongest; clustering is apparent both within and between genes (Figure 3.4ABC). Clustering is detectable in the *Escherichia/Citrobacter/Salmonella* analysis (Figure 3.4D) though less apparent due to the accumulation of noise.

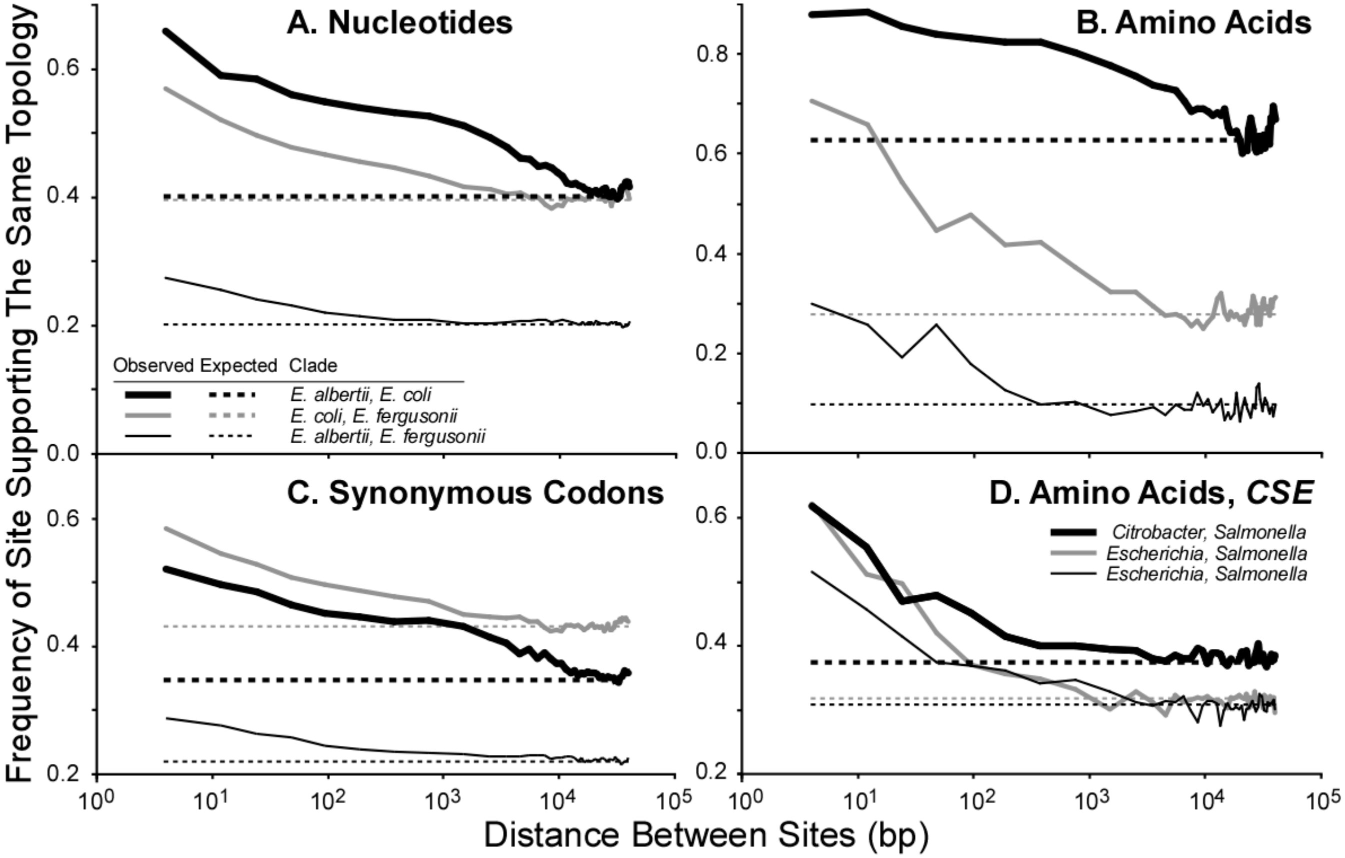


Figure 3.4 Chromosomal clustering of parsimony informative sites supporting each topology

Each site supporting a distinct topology was compared against each other site and the observation binned according to the distance between sites. Trendlines report proportions of observations where a site supporting a given topology was paired with a site supporting the same topology. Expected values are derived from the genome-wide proportion of sites supporting that topology. Data is plotted at the midpoint of the bin range. Within species quartets are in Fig. S3.

Table 1 Occurrence of sites supporting the same topology

	Topology[1]			Runs of Sites			
	1	**2**	**3**	**Exp**	**SD**	**Obs**	**Z**
ESC Nuc	19539	19886	18829	38827	114	35684	27.6
ESC Pro	2382	2879	2442	5117	41	4600	12.5
ESC Syn	12229	11666	11621	23672	89	22844	9.3
EEE Nuc	9687	19298	18950	30718	102	24327	62.6
EEE Pro	471	3074	1361	2558	29	1660	31.0
EEE Syn	6950	11009	13658	20357	83	17159	38.7

[1] **ESC1**=Cit,Esc; **ESC2**=Cit,Sal; **ESC3**=Esc,Sal; **EEE1**=Eal,Efe, **EEE2**=Eal,Eco; **EEE3**=Eco,Efe

Columns: 1,2,3: count of sites supporting the specified topology; Exp, Expected number of runs if sites are arranged randomly; SD: standard deviation of expectation; Obs: Observed number of runs; z: z value (Obs-Exp)/SD. All values are highly significant (p << 0.001)

3.5 PHYLOGENETIC INCONGRUENCE DOES NOT REFLECT INCOMPLETE LINEAGE SORTING

The clustering of informative sites is evidence that recombination produced different phylogenies for different regions of these genomes, thus eliminating the possibility of recovering an unambiguous organismal phylogeny from the sequence data. Yet this evidence is not sufficient to determine whether recombination occurred at some loci subsequent to the genetic isolation at other loci (fragmented speciation), or an ancestral population split into two descendent populations, one of which split again before its ancestral polymorphisms had been resolved (incomplete lineage sorting). By the latter model, one topology would reflect the history of population splitting (the species topology), and the other two would represent the diversity of the original population (Pamilo and Nei 1988). To evaluate this "instant speciation" model empirically, we compared the observed diversity of the extant populations to the inferred diversity of the ancestral population that could have generated the incongruent phylogenies (Fig S1). Putative ancestral diversity was measured as the length of the innermost branch connecting all four genera on a maximum likelihood tree (internal branch). The diversity within an extant population was measured as half of the distance separating any two genomes in that population (terminal branch). Four *E. coli* genomes were selected to include maximal diversity (Touchon, Hoede et al. 2009), as were 3 *S. enterica enterica* genomes. As reported above, widespread support for the monophyly of *E. coli* indicates that *E. albertii* and *E. fergusonii* do not recombine freely with *E. coli*; therefore they were not included in measures of extant diversity. Higher levels of congruence supported the distinctness of *S. enterica enterica* (subspecies I) relative to *S. enterica arizonae* (>97.5% of alignments where no alternate topology is within the 90% confidence limit).

If incomplete lineage sorting were to produce the observed phylogenetic ambiguity between *Escherichia*, *Salmonella* and *Citrobacter*, then the terminal and internal branch-lengths on maximum likelihood trees with non-species topologies would be comparable because both would represent within-species diversity. Only trees with the 'true' species topology could have accumulated extra divergence along the internal branch. For all three topologies we counted the number of trees for which the internal branch was longer than the terminal branch. To provide a conservative estimate, we focused on the pair of *E. coli* genomes that most often provided a terminal branch measurement that exceeded the length of the internal branch. Still, the average terminal branch of UMN026/UTI89 was generally longer than the internal branch. This was true for each topology, in 93.0% (346/372), 94.8% (292/308), and 95.7% (443/463) of the trees that clustered *Escherichia* with *Salmonella, Escherichia* with *Citrobacter* or *Citrobacter* with *Salmonella*, respectively (Fig. S1A). Diversity within *S. enterica enterica* was always smaller than the ancestral variation in at least than 98% of trees.

To determine how often the internal branch would be longer than the terminal branch if genetic isolation were imposed simultaneously for all genes, we simulated the evolution of these taxa according to the best ML tree. This guide tree was modified so that the internal branch length was equal to the average terminal branch length between strains UMN026 and UTI89 (Fig. S1B). We repeated the quartet analysis using 100 simulations of the set of 1144 genes (Fig. S1C). The internal branch on a ML tree of the simulated data was longer than the average terminal branch for 60 ± 1 % of the genes (maximum value 63.9%). Comparable values were found when the genes supporting each topology were analyzed separately (Table S5; maximum value 66.8% for the 308 genes supporting the *Escherichia/Citrobacter* clade). Therefore, our data indicate that measured ancestral diversity for the *Escherichia/Salmonella/Citrobacter* split far exceeds diversity found in extant species of *E. coli* and *S. enterica*. Because the data suggest that all three topologies represent the 'true' species topology, we reject incomplete lineage sorting as the mechanism leading to phylogenetic incongruence. Similar results were found using the internal branch of the *E. albertii/E. coli/E. fergusonii* split when the genes supporting each of the two dominant topologies were analyzed (Table S5); the rare gene alignments that supported an *E. albertii /E. fergusonii* clade produced branch lengths within the expected distribution, which is consistent with the lack of prolonged recombination between these genes in these lineages, as was suggested above.

3.6 QUESTIONING THE TREE OF LIFE

Significant phylogenetic incongruence was observed between bacterial taxa (Figure 3.1). This incongruence could have reflected noise, recent recombination between otherwise genetically isolated populations, or the random assortment of ancestral diversity following instant acquisition of genetic isolation. Above, we provided evidence rejecting these alternatives; therefore, another model, such as the stepwise acquisition of genetic isolation (fragmented speciation), must be invoked to explain the data. In accordance with this model, previous data for

E. coli and *S. enterica* show that genetic isolation occurred at different times for different genes, driven by adaptive change (Chapter 2). The fragmented speciation model suggests that organismal phylogenies cannot be deduced from gene phylogenies since genes have different evolutionary histories. Given the vast diversity of prokaryotes (Dykhuizen 1998), groups of ambiguously related taxa produced by rapid evolutionary radiations may be common.

The number and density of such problematic relationships can only increase as more microbial diversity is characterized. In the most extreme interpretation, this would invalidate the Tree of Life hypothesis, which is founded on the idea that extant taxa have unambiguous relationships (Doolittle and Bapteste 2007). Phylogenetic trees of organisms serve as frameworks for interpreting evolutionary change; characteristics of ancestral taxa are inferred by coalescence and serve as platforms for interpreting changes in descendent taxa. Yet our data suggest that such ancestral taxa may not have existed, and inferences that require them, *e.g.*, any utilization of parsimony, would fail. For example, if one accepts an organismal phylogeny that places *Escherichia* as an outgroup to the *Citrobacter/Salmonella* clade, then any feature in common between *E. coli* and *Salmonella* would be interpreted as a parallel gain or a loss from *Citrobacter* (Fig. S4). Alternatively, this feature may have been shared by the two taxa throughout the fragmented speciation process, since a distinct taxon ancestral to the *Citrobacter*/*Salmonella* clade need not exist (Lawrence and Retchless 2010). Given these complications, the Tree of Life, in demanding a strict bifurcating relationship among descendent taxa, cannot form the basis for rigorous examination of bacterial diversity.

3.7 QUESTIONING BACTERIAL SPECIES CONCEPTS

The patterns of incongruence that we identified suggest that populations of potentially recombinogenic bacteria are neither freely recombining nor genetically isolated at all loci as required by the BSC. The implication is that any apparently freely-recombining population actually comprises many partially genetically-isolated subpopulations. As a result, Mayrian species boundaries cannot be defined rigorously by gene flow because extant species will include numerous ecological protospecies that are in partial genetic isolation, leading to ambiguity in the relationships among derived taxa as shown above. Moreover, ecotypes are not a good basis for species identification, as closely related ecotypes can still experience substantial gene flow, causing the evolution of one ecotype to influence the trajectory of another as in *Neisseria* (Spratt, Bowler et al. 1992) or *Campylobacter* (Sheppard, McCarthy et al. 2008). Given the complexity of bacterial gene exchange, we are unlikely to identify any rules for identifying the threshold beyond which two populations are destined to follow separate paths. Historical evidence for recombination does not necessitate ongoing potential for recombination.

Thus species concepts may not apply to bacteria (Doolittle and Zhaxybayeva 2009), even if phenotypically distinct groups of related bacteria are readily identifiable. Such concepts connect patterns of phenotypic diversity in groups of organisms to the evolutionary forces acting upon those organisms' constituent genes. Forces that lead to cohesion within sexual eukaryotic populations act upon all genes in concert; as a result, the history of such organisms is reflected in the collective history of their genes. Their sexual systems simplify species conceptualization by producing mating barriers that affect entire genomes at once (Rieseberg, Wood et al. 2006). In contrast, evolutionary forces do not act on all bacterial genes in unison; recombination may be successful at some loci, but be counterselected at others. The evolutionary independence of bacterial genes afforded by position-specific gene exchange generates incongruence among gene trees. Therefore, a species concept attributing the unambiguous species delineation to the action of a particular evolutionary process may be unattainable in bacteria.

One response would be for taxonomy to embrace the pluralistic nature of bacterial taxa, placing strains into more than one species (Bapteste and Boucher 2009), or abandoning species names altogether for a less hierarchical approach (Lawrence, Hatfull et al. 2002). Yet one could argue that bacterial species names carry the greatest practical impact, placing organisms into defined groups that are utilized for agriculture, biotechnology, epidemiology, public health, disease diagnosis, and bioterrorism. Indeed, the public policy impact of such ambiguity and fluidity in the characterization of bacteria may preclude the widespread adoption of such a classification system. Barring this approach, then, what is left is the necessary use of practical definitions in the absence of a feasible species concept. Such definitions would encompass collections of bacteria that are phenotypically similar by criteria that are subjectively important to the classifiers, leading to both narrowly-defined (*e.g., Bacillus anthracis*) and broadly-defined (*e.g., E. coli*) groups. The ease by which many medically relevant taxa can be classified suggests that it can be an effective approach. While this lacks the elegance and satisfaction of groupings driven by biological processes, the absence of strong theoretical underpinning to their delineation does not detract from their utility.

4.0 QUANTIFICATION OF CODON SELECTION FOR COMPARATIVE BACTERIAL GENOMICS

It has long been recognized that protein coding sequences show nonrandom, organism-specific patterns of codon usage (Grantham, Gautier et al. 1980). Codon usage bias is most pronounced in highly expressed genes (Ikemura 1981), where codon preferences are associated with the tRNA abundance within the cytoplasm (Ikemura 1981). Measurement of codon selection is of interest because the extent to which different genes use the preferred codons is predictive of their expression levels. Comparative studies of codon selection have provided insight into the population structure and lifestyle of organisms (Sharp and Li 1987; Karlin and Mrazek 2000; Rocha 2004; Sharp, Bailes et al. 2005; Vieira-Silva and Rocha 2010).

Numerous statistics have been devised to measure variation in codon selection among Open Reading Frames (ORFs) within a genome, yet none fully account for the evolutionary dynamics that shape codon usage bias. The simplest metrics evaluate how much the codon usage frequencies of a gene deviate from expected frequencies. These methods, such as the Effective Numbers of Codons (ENC) and the ENC′ (Wright 1990; Novembre 2002), incorporate no information about the fitness differences among synonymous codons. The logic of these metrics has been expanded by Karlin (Karlin and Mrazek 2000) and Supek (Supek and Vlahovicek 2005), comparing each gene's codon usage to both genome-wide codon frequencies (representing mutational tendencies) and to codon frequencies in a defined set of genes believed to experience strong codon selection. However, these “reference point” approaches have been criticized for being designed such that genes with the most extreme biases in terms of preferred or non-preferred codons would not be assigned the most extreme values (Henry and Sharp 2007).

An alternative approach is to assign a score to each codon based upon inferences regarding the typical fitness advantage of the codon relative to its synonyms. The simplest such statistics summarize the optimal codon frequency for each amino acid (F_{op} (Ikemura 1981) and CBI (Bennetzen and Hall 1982)) while more complicated “scoring table” methods incorporate additional information about the relative importance of non-optimal codons (*e.g.* CAI (Sharp and Li 1987), tAI (dos Reis, Savva et al. 2004), GCB (Merkl 2003)). Use of a scoring table weights the statistic so that it is influenced more by those amino acids for which the synonymous codons have a greater perceived fitness difference. Such statistics may still be normalized to assure that the amino acid composition of a protein does not influence them, allowing them to reflect variation in synonymous codon usage only. One method for normalizing across amino acids is to compare the score of the observed codons against the maximum possible score for an ORF with the same amino acid composition (CAI, tAI). This normalization produces a uniform maximum score for all ORFs regardless of amino acid composition, but does not normalize non-optimal codons across amino acids, allowing the final statistic to be influenced by amino acid composition for the majority of ORFs containing many non-optimal codons (Sharp, Emery et al. 2010). Alternatively, the statistic can be calculated as the unweighted average of the statistic for each amino acid (Eyre-Walker 1996), but this ignores the differences in information content arising from different abundances of amino acids in a protein.

Despite the power of these statistical methods, none of them quantify the patterns of variation that are expected to arise from mutation-selection balance, which is the primary explanation for the occurrence of non-optimal codons (Bulmer 1991; Smith and Eyre-Walker 2001). The selection-mutation-drift theory of synonymous codon usage describes an equilibrium condition where preferred and non-preferred codons occur in proportions determined by mutational biases, selection, and the effective population size. Recent studies have calculated the parameters of this model explicitly (Sharp, Bailes et al. 2005; dos Reis and Wernisch 2009; Sharp, Emery et al. 2010), but only codons for two-fold degenerate amino acids were analyzed, limiting the information available to make inferences about individual genes. To date, no analytical method accounts for the variation in the codon usage statistic that arises from the stochastic nature of the selection-mutation-drift model.

Here, we expand upon the scoring table class of methods by introducing a new statistic that incorporates a stochastic model allowing ORFs to be evaluated in terms of their deviation from a null expectation of codon composition in genes lacking strong codon selection. This allows us not only to measure the impact of selection against the background of mutational bias, but to normalize the values assigned to non-preferred codons of different amino acids so that each amino acid is expected to contribute an equal score under the null model. By deriving the expected distributions of the statistic under a null hypothesis about codon frequencies, our statistical framework provides a means to compare the strength of codon selection within and between genomes.

Below, we describe a statistic for summarizing the codon usage of a protein coding sequence. The raw statistic is the sum of values assigned to each of the codons in the sequence and may be normalized according to its expected distribution. Normalized scores for individual genes can be combined to summarize the magnitude of codon selection operating on the entire genome. We compare our measure to previously described codon usage statistics, both conceptually and empirically.

4.1 RELATIVE ADAPTIVENESS OF SYNONYMOUS CODONS

To quantify enrichment of a codon among genes experiencing codon selection, we define a score (δ) for each codon *cdn* in a manner similar to Merkl's CB (Merkl 2003) as,

$$\delta_{ij} = \log \frac{f_o(cdn_{ij})}{f_n(cdn_{ij})} ,$$

where cdn_{ij} is the *j*th codon of the *i*th amino acid and $f(cdn_{ij})$ is the expected frequency of that codon among its synonyms in genes that have (f_o) or have not (f_n) been optimized by codon selection. Use of the logarithm enables us to summarize the codon optimization of a gene or set of genes as the sum of the individual scores of the codons comprising the gene, generating the Summed Codon Bias (*SCB*). To facilitate examination of the stochastic properties of the *SCB*, it is calculated as the sum of the composite scores for each amino acid (α), which are determined from the scores of their constituent codons as,

$$\alpha_i = \sum_{j=1}^{N_i} C_{ij}\delta_{ij} \text{ and}$$

$$SCB_{gene} = \sum_{i=1}^{20} \alpha_i = \sum_{i=1}^{20} \sum_{j=1}^{N_i} C_{ij}\delta_{ij} ,$$

where N_i is the number of synonyms and C_{ij} is the count of that codon within the gene being analyzed. Merkl (Merkl 2003) argues that statistics of this form are the optimal test statistics for distinguishing between two populations. Here, we use the sum because it has convenient properties, described below, which we will use to normalize this continuous statistic. The *SCB* is related to other codon optimization statistics by different normalization routines. Merkl's GCB (Merkl 2003) is the length-normalized form of the *SCB*. The logarithm of the CAI (Sharp and Li 1987) can be derived from the *SCB* by calculating δ_{ij} with a non-optimized table (f_n) showing no bias among synonymous codons, then subtracting *SCB* from the maximum possible value available given its amino acid composition, and dividing by the number of codons in the ORF, ignoring methionine and tryptophan.

Crucially, scoring tables created from δ_{ij} reveal which codons increase in frequency among the most optimized proteins, and to what degree. This is different from the RSCU values that are used to calculate the CAI (Sharp and Li 1987), which reflect simply the abundance of codons in optimized genes without reference to their abundance in non-optimized genes. Codons with greatest abundance in optimized genes may not have experienced the strongest selection for enrichment and, in the worst cases, may actually be disfavored. This adjustment to the estimate of codon adaptiveness should have the greatest effect in genomes where nucleotide composition shows the greatest deviation from equal usage. To examine the effect of this difference between *SCB* and CAI, we evaluated multiple genomes by constructing f_o from a set of 40 protein-coding genes whose products comprise the ribosome and other parts of the translation apparatus (see Methods, (Sharp, Bailes et al. 2005)) and constructing f_n from all protein coding sequences in the genome. Accounting for the biases in f_n creates substantial changes in δ relative to the values obtained otherwise (Table 4.1), even changing estimates of which codon is most preferred. In

Pseudomonas putida (67% GC), for four amino acids, the synonymous codons that are enriched among ribosomal proteins and translation elongation factors are not the same as the synonymous codons that are most abundant among those proteins. These effects are also observed in genomes with less bias in nucleotide composition, such as *Bacillus subtilis* (44% GC) and *Escherichia coli* (51% GC), each of which had one amino acid where the enriched codon is not the most abundant codon.

Table 2 The effect of mutational biases on codon scores

								Pseudomonas putida		
		Escherichia coli MG1865			*Bacillus subtilis* 168			KT2440		
Residue	Codon	All	40	Dif	All	40	Dif	All	40	Dif
Lys	AAG	0.303	0.380	1.000	0.427	0.189	0.441	1.000	1.000	0.634
Lys	AAA	1.000	1.000[a]	0.800	1.000	1.000	1.000	0.385	0.607	1.000
Pro	CCG	1.000	1.000	1.000	1.000	0.279	0.127	1.000	1.000	0.409
Pro	CCA	0.358	0.183[b]	0.511	0.439	0.962	1	0.2817	0.689	1.000
Pro	CCT	0.295	0.206	0.697	0.659	1.000	0.693	0.2374	0.557	0.959
Pro	CCC	0.231	0.017	0.074	0.206	0.039	0.086	0.4627	0.151	0.134
Thr	ACG	0.613	0.082	0.050	0.652	0.233	0.140	0.264	0.046	0.078
Thr	ACA	0.290	0.094	0.121	1.000	0.606	0.238	0.104	0.054	0.232
Thr	ACT	0.374	1.000	1.000	0.392	1.000	1.000	0.137	0.307	1.000
Thr	ACC	1.000	0.924[c]	0.346	0.386	0.026	0.026	1.000	1.000	0.448
Val	GTG	1.000	0.229	0.160	0.906	0.168	0.185	1.000	0.646	0.117
Val	GTA	0.415	0.545	0.916	0.695	0.629	0.904	0.201	0.399	0.361
Val	GTT	0.698	1.000	1.000	1.000	1.000	1.000	0.181	1.000	1.000
Val	GTC	0.587	0.139	0.166	0.904	0.157	0.174	0.572	0.798	0.253

a. Red cells indicate that this table overestimates selection against this codon and incorrectly denotes it as the preferred codon.
b. Yellow cells indicate that this table overestimates selection against this codon.
c. Green cells indicate that this table underestimates selection against this codon.

Table 4.1. Normalized Synonymous Codon Usage values (NSCU, (Sharp and Li 1987)) for select amino acids, illustrating the effect of calculating these values relative to the genome-wide codon composition. NSCU is the frequency of each codon divided by the frequency of the most common of its synonyms. For each genome (listed at top), three columns represent NSCU values calculated for the frequency of codons in the entire genome (All, f_n), in 40 highly expressed proteins (40, f_o), and the ratio of the two (Dif, f_o/f_n)

4.2 NORMALIZATION OF CODON BIAS STATISTICS TO A THEORETICAL DISTRIBUTION

Rigorous interpretation of any codon bias statistic depends upon knowledge of its distribution given expected synonymous codon usage frequencies. Issues as simple as discerning if one protein coding sequence is more enriched for optimal codons than another cannot be resolved unless we know what values of the summary statistic are expected to occur for protein coding sequences that vary in amino acid composition but not synonymous codon frequencies. Likewise, unless the variance of the summary statistic is known, variation between genes cannot be inferred to result from differences in the strength of selection rather than being due to the stochastic nature of mutation and drift.

If the null hypothesis is that a protein coding sequence has not been shaped by selection for optimal codons, then the table of expected codon frequencies for each amino acid is equivalent to f_n, above. We will use genome-wide codon composition as estimates of f_n. To estimate the distribution of the *SCB* expected for a given protein coding sequence, we first estimate the sampling distribution of the composite score for each amino acid (α). The expected contribution of each amino acid is the count (*C*) of that amino acid, multiplied by the weighted average of the scores of each of its codons (δ), so that

$$E(\alpha_i) = C_i \sum_{j=1}^{N_i} P_{ij}\delta_{ij}$$

and

$$E(SCB) = \sum_{i=1}^{20} E(\alpha_i) = \sum_{i=1}^{20} C_i \sum_{j=1}^{N_i} P_{ij}\delta_{ij} \ ,$$

where P_{ij}, the probability of observing that codon at random, is the frequency of that codon among genes lacking selection, equivalent to $f_n(cdn_{ij})$. In our null model, the identity of the codon at each site is independent of that at other sites, meaning that the variance of the *SCB* is the sum of the variance for each site, so that

$$V(\alpha_i) = C_i \left[\sum_{j=1}^{N_i} P_{ij}\,{\delta_{ij}}^2 - \left(\sum_{j=1}^{N_i} P_{ij}\delta_{ij} \right)^2 \right]$$

and

$$V(SCB) = \sum_{i=1}^{20} V(\alpha_i).$$

Being the sum of several independent random variables, the *SCB* has an approximately normal distribution according to the Central Limit Theorem (Sheskin 2007). Many statistical tests assume a normal distribution, so we will describe a statistic derived from that distribution. The Adaptive Codon Enrichment (ACE) is the difference between the observed *SCB* and the expected *SCB* for a protein coding sequence:

$$\text{ACE} = SCB - E(SCB).$$

This may be normalized in two ways. First, it may be presented as a standard deviation score or z-value as,

$$\text{ACE}_z = \frac{\text{ACE}}{\sqrt{V(SCB)}} = \frac{SCB - E(SCB)}{\sqrt{V(SCB)}} \ .$$

Alternatively, the ACE may be unit normalized so that it reflects the deviation averaged per codon in the coding sequence as,

$$\mathrm{ACE}_u = \mathrm{ACE} \Big/ \sum_{i=1}^{20} \sqrt{V(\alpha_i) C_i} \ .$$

Because each amino acid provides different amounts of variance to the final score, normalization takes into account the variance contributed by each amino acid rather than simply dividing by the length of the protein coding sequence. This is equivalent to calculating ACE$_u$ as the average of the z-value for each individual codon. Notably, the ACE is indifferent to the inclusion or exclusion of methionine and tryptophan codons because, having only single codons, they always make an equal contribution to the observed and expected value and do not contribute to the variance. This is in contrast to statistics that are sensitive to the frequency with which the most preferred codon occurs, such as the CAI, where methionine and tryptophan are explicitly ignored (Sharp and Li 1987).

To validate that ACE statistics can be treated as random normal variables, we used Monte Carlo simulations to examine the properties of genes for which the *SCB* fit this assumption. Distributions were constructed from 2000 Monte Carlo samples for each ORF of *E. coli* and *P. putida*, using the expected codon distribution of the respective genome. The predicted mean and variance were universally accurate, while skewness and kurtosis produced deviations from normality that were only detectable within the GC-biased *P. putida* genome. D'Agostino's K-squared test (Sheskin 2007) identified an excess of genes having non-normal *SCB* null distributions ($p < 0.05$ for 340 of 5350 ORFs; 6.3%), although the skewness and kurtosis values were universally small (-1×10^{-3} to 8×10^{-4} and -6×10^{-4} to 6×10^{-4}, respectively) and the worst approximations were concentrated among genes with less than 100 degenerate codons (67 of 503 small ORFs being non-normal at $p < 0.05$).

4.3 PREDICTION OF GENE EXPRESSION DATA

Using existing gene expression data, we examined the predictive power of several codon selection statistics and their robustness in the face of uncertainty regarding optimal parameterization. Here we examined those methods that use information about the frequency with which each codon is used within a set of ORFs optimized for translation (f_o). A robust method will generate a consistently high level of performance when parameterized with any set of ORFs for which the codon usage bias has been shaped by codon selection. We selected three datasets of transcript abundance data for evaluation: *Escherichia coli* (Bernstein, Khodursky et al. 2002)(Figure 4.1 AB), *Pseudomonas aeruginosa* (Waite, Paccanaro et al. 2006) (Figure 4.1 CD), and *Saccharomyces cerevisiae* (Dudley, Aach et al. 2002) (Figure 4.1 EF). These include both bacteria and eukaryotes, with genomic nucleotide compositions ranging from strongly AT biased to strongly GC biased.

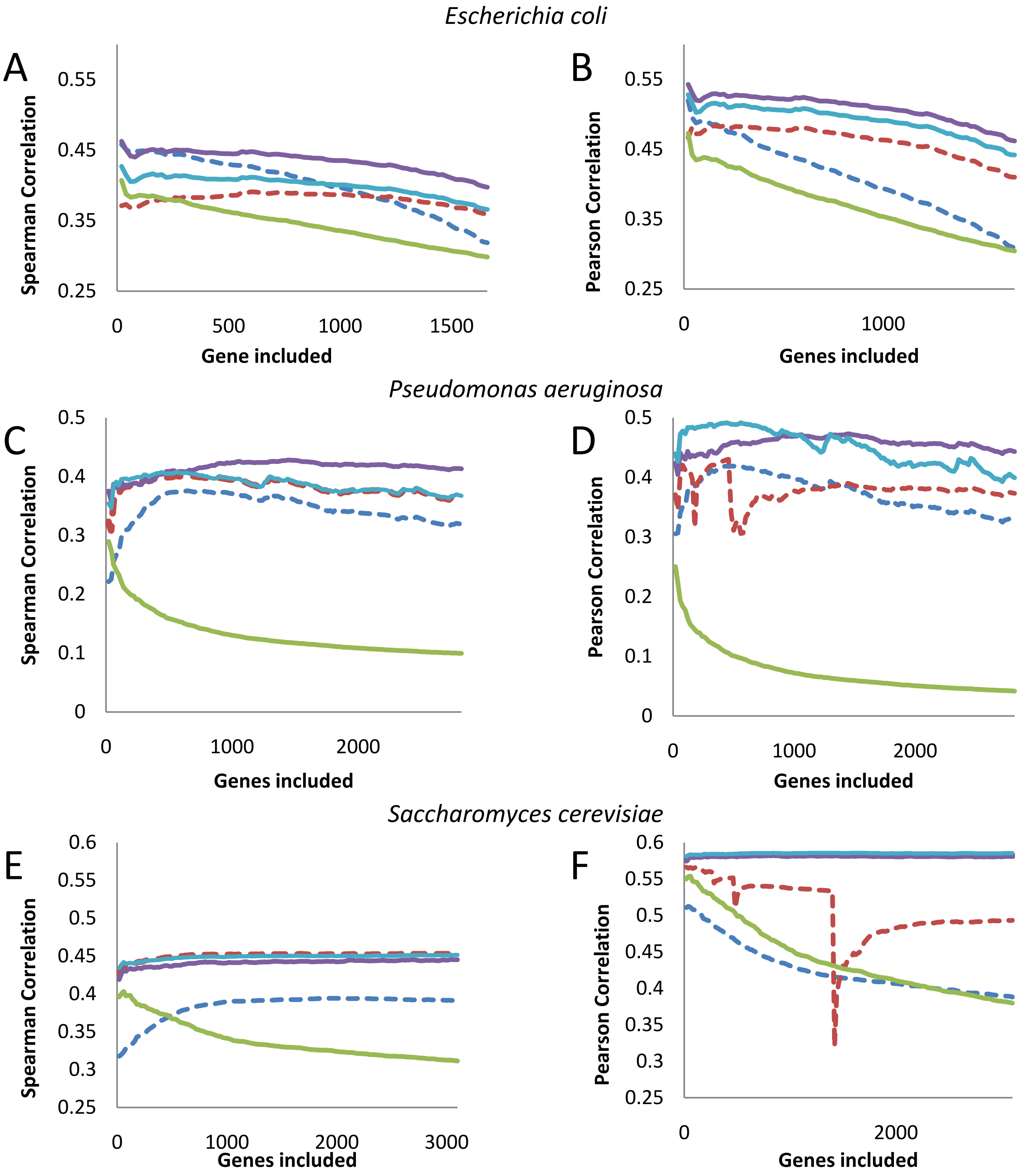


Figure 4.1 Correlation coefficients of codon selection statistics and transcript abundance data.

CAI, green line; E, blue dashes; MELP, red dashes; GCB, purple line; ACE, blue line. Abscissa: the number of genes included in the highly expressed gene set (f_o). Ordinate: Correlation coefficient between respective codon statistic and transcript abundance level for each gene (see main text). From top to bottom: *E. coli*, *P. aeruginosa*, *S. cerevisiae*

For each dataset, we examined the correlation of the transcript abundance data relative to each codon optimization statistic (CAI (Sharp and Li 1987), GCB (Merkl 2003), length-normalized ACE (this study), Karlin's E (Karlin and Mrazek 2000), and MELP (Supek and Vlahovicek 2005)) when the codon statistic was calibrated against the most abundant transcripts from the same dataset. Here, our intention is not to actually predict the transcript abundance data, but to evaluate the behavior of each method under optimal conditions. By using the dataset that the statistics are tested against, we avoid any semi-arbitrary decisions in parameterization that may inadvertently favor one method over another. To examine how each statistic responds to decreased precision in identifying the optimal genes, the number of genes contributing codons to f_o was gradually increased, 20 at a time until it included half of all genes. For the statistics that require an estimate of codon usage in the absence of codon selection (f_n), we used the codon composition of the entire genome.

The correlation between CAI and expression was generally weak in *S. cerevisiae* and *P. aeruginosa* (Figure 4.1 C-F), which is expected given that these genomes exhibit strong biases in their nucleotide composition and CAI does not incorporate any information about this bias. For this reason, the author of the CAI suggested that the it may not be applicable to highly biased genomes such as *P. aeruginosa* (Grocock and Sharp 2002). The other four methods, taking the nucleotide composition into account, perform much better on biased genomes (Figure 4.1 C-F).

These four methods perform comparably when evaluated with Spearman's (rank) correlation (Figure 4.1 A,C,E), but show differences when evaluated with Pearson's correlation (Figure 4.1 B,D,F). For the calculation of Pearson's correlation, a logarithmic function was applied to the transcript abundance, E, MELP, and CAI, because they generally performed better after this transformation, and the GCB and ACE are intrinsically calculated with logarithms.

Pearson correlations are generally higher than Spearman correlations, indicating that there is some proportionality between these statistics and gene expression levels. The highest correlations were typically produced by the GCB and ACE (Figure 1, purple and blue lines), and these correlations are most robust to the decreased resolution of the set of “highly expressed genes”. The length-normalized ACE performed similarly to the ACE_u, so the later was not displayed in the graphs. The MELP also performed rather well, but exhibited erratic behavior when examined with the Pearson correlation; this effect arises from the fact that the component metric (the MILC (Supek and Vlahovicek 2005)) can approach zero, producing extreme values for MELP (genes with negative scores were excluded from the correlation calculation).

The ability of the ACE to predict gene expression levels in *P. aeruginosa* with such high accuracy (ρ = 0.65, 5543 genes, using the 100 most highly expressed genes to construct f_o) is surprising in light of previous studies suggesting that there is little codon selection acting in this genome (Sharp, Bailes et al. 2005). Grocock and Sharp found that codon variation in *P. aeruginosa* was primarily due to the presence of genes with atypical nucleotide composition (presumably recently acquired), with a secondary trend due to codon selection (Grocock and Sharp 2002). Recently acquired genes tend to be expressed weakly so that, even in the absence of codon selection, a statistic that simply discriminated between native and foreign genes would be expected to correlate with expression levels. We tested whether this factor contributed to the high correlation by limiting the analyses to the 1678 genes that are likely to be native to *P. aeruginosa* because orthologs were detected in each of four other diverse *Pseudomonas* species: *P. mendocina, P. stutzeri, P. entomophila,* and *P. putida* (mean dS > 1.25 for each of the 10 pairs, where dS is synonymous divergence estimated by the method of (Yang and Nielsen 2000)). For the 1677 genes in this set that also had transcript abundance values, the correlation coefficient actually increased to 0.75 when the 100 most highly expressed genes were used to construct f_o, indicating that most of this correlation is indeed due to codon selection.

4.4 SUMMARIZING GENOMIC CODON SELECTION

The level of codon selection may vary between genomes and several approaches have been implemented to measure these differences (Lawrence 2001; dos Reis, Savva et al. 2004; Rocha 2004; Dethlefsen and Schmidt 2005; Sharp, Bailes et al. 2005). These studies have found that codon selection increases in genomes with greater numbers of tRNA- and rRNA-encoding genes, suggesting that codon adaptation is associated with genomic structures that minimize generation time under optimal growth conditions (Sharp, Emery et al. 2010; Vieira-Silva and Rocha 2010).

Unlike other measure of gene-level codon usage bias, the ACE lends itself naturally to estimates of genome-wide codon selection. A χ^2 distribution is defined as the sum of the squares of samples from a standard normal distribution. Therefore, we can calculate a normalized χ^2 statistic for each genome – measuring the degree to which selection has moved codon usage away from that expected by mutation alone – by calculating the average of the squared z-scores for each gene g, as

$$ACE\chi^2 = \frac{1}{N}\sum_{g=1}^{N} ACE_{z_g}^{2}$$

In the absence of codon selection, values should approach 1.0, where genes, on average, share the same codon usage (Sheskin 2007). The Monte Carlo simulations described above confirmed that when all ORFs share the same codon composition, the $\mathrm{ACE_z}$ distribution for the genome has a mean of zero, and a variance of one, resulting in a normalized χ^2 of one. The χ^2 statistic is suitable for summarizing across several genes, since when it is generated from a set of genes experiencing uniformly strong codon selection, it will be proportional to the total number of codons evaluated, because the contribution of each amino acid to the z score is proportional to the square root of the number of codons encoding that amino acid.

To examine the behavior of this statistic, we quantified selection in three diverse bacterial genomes under a variety of analytical assumptions. First, we examined *Buchnera aphidicola*, which is generally believed to experience very little codon selection. The ACEχ^2 for the entire genome (563 ORFs) ranged from 1.64 to 1.94, depending on whether f_o was calculated using the orthologs of the 27 *E. coli* genes used to calibrate the original CAI, or the orthologs of 40 *E. coli* translational genes (see Methods, see Fig. S5 for distribution of ACE$_z$). Limiting the analysis to the 498 ORFs shared with *E. coli* K12 (for both f_n and ACEχ^2) resulted in an altered ACEχ^2 of 1.42 and 1.96 for the two f_o tables.

The choice of genes for f_n and ACEχ^2 has a greater impact in larger genomes with a greater number of horizontally acquired genes. For example, in *E. coli* K12 the ACEχ^2 for the entire genome (4149 ORFs) ranged from 9.8 to 10.3 for the different f_o tables. Limiting the analysis to the 2628 ORFs shared with *E. fergusonii* and *E. albertii* (for both f_n and ACEχ^2) increased ACEχ^2 scores to 11.79 and 11.3 for the two different f_o tables. A more extreme restriction to the 498 genes shared with *Buchnera* increased the scores to 18.7 and 17.7.

The increase in scores generated by the more restricted set of widely conserved genes is due to the tendency of widely conserved genes (those shared with *Buchnera*) to be highly expressed and therefore experience strong codon selection. In fact, this increase in ACEχ^2 is moderated by the simultaneous adjustment of f_n to represent a population that is more strongly influenced by codon selection, thereby decreasing the ACE of those genes with the greatest optimization. The impact of f_n is apparent from comparing ACEχ^2 for the 486 *E. coli* ORFs that have matches in each of the other three genomes (*Buchnera*, *E. fergusonii*, and *E. albertii*). When using the 40 translational proteins for f_o, ACEχ^2 increases from 17.9 for the *Buchnera/Escherichia* f_n, to 24.9 for the *Escherichia* f_n, to 29.2 for the *E. coli* f_n. The ACEχ^2 calculated from the broader set of 2628 Escherichia orthologs is not as responsive to the composition of f_n, being 11.3 for the *Escherichia* f_n, and 11.7 for the *E. coli* f_n. The variability of the ACEχ^2 as a result of the genes selected for f_n and ACEχ^2 illustrates the need to consider carefully the composition of these sets before interpreting the ACEχ^2. One approach for cross-genome comparisons is to select orthologous genes for all steps of the analysis (f_o, f_n, ACEχ^2), which is the approach that will be used below.

As a final examination of the behavior of $\mathrm{ACE}\chi^2$, we considered *P. aeruginosa*. Grocock and Sharp (Grocock and Sharp 2002) demonstrated that highly expressed genes exhibit distinctive codon usage in this genome, which was verified above with the high correlation between ACE and transcript abundance. But Sharp's attempt (Sharp, Bailes et al. 2005) to estimate the strength of codon selection on 40 translational proteins revealed no selection (S = -0.019). This was attributed to the fact that S was calculated based on the codons for only four amino acids, which were not the amino acids for which the synonymous codons were enriched in the highly expressed genes of *P. aeruginosa* (Sharp, Bailes et al. 2005). Because ACE incorporates information from all synonymous codons, this limitation should be avoided.

To examine this, we repeated the above analysis where the f_o table was constructed with incrementally increasing sets of genes having the most abundant transcripts. Using all genes, $\mathrm{ACE}\chi^2$ ranged from 3.7 to 5.5, which is noticeably greater than the value expected in the absence of selection (1.0) or the value obtained for *Buchnera aphidicola* (~2.0). A more accurate measure should be obtained by limiting the analysis (for both f_n and $\mathrm{ACE}\chi^2$) to the 1678 genes with orthologs in the four other *Pseudomonas* species. Here, $\mathrm{ACE}\chi^2$ ranged from 6.1 to 7.3, providing additional evidence of codon selection since the codon usage variation in this set of native genes is revealed by calibration on the most highly expressed genes.

To compare the strength of selection between different genomes, we examined a broad set of 15 genomes from the Enterobacteriaceae (Table 4.2), limited to those genomes where the average synonymous divergence (Yang and Nielsen 2000) was greater than 1.0 for all pairwise comparisons. An f_n table was constructed for each of the genomes based on its contribution to the 634 sets of putative orthologs, where each gene was the pairwise reciprocal best match to the gene in each genome, so that each gene had a one-to-one relationship to a gene in each other genome. The f_o table of each genome was constructed from the genes found in the same set of putative orthologs as the 40 *E. coli* translational genes (see Methods). The $\mathrm{ACE}\chi^2$ for the 634 genes shared among these 15 genomes ranged from 3.0 for *Hamiltonella defensa*, a secondary endosymbiont of aphids with a reduced genome, to 15.7 for *E. coli* (Table 4.2).

Previous studies have identified a correlation between genome-scale codon selection and the number of tRNA genes in a genome, suggesting that the two may be causally linked. The $\mathrm{ACE}\chi^2$ is likewise correlated to tRNA gene copy number (Figure 4.2, Table 4.2). This effect is most pronounced when comparing *H. defensa* to the other genomes, but is still substantial even when this outlier is excluded ($r^2 = 0.31$). Reanalysis of these genomes without *H. defensa* permitted the inclusion of 989 sets of putative orthologs (used for both f_n and $\mathrm{ACE}\chi^2$), but only changed $\mathrm{ACE}\chi^2$ values slightly (Figure 4.2, Table 4.2). The values from this larger set of genes are slightly lower than the values from the previous analysis, but strongly correlated ($r^2 = 0.98$), indicating that cross-genome comparisons are robust to the exact set of orthologs examined.

Table 3 Properties of 15 genomes from Enterobacteriaceae.

Organism	ORF count	% coding	tRNA count	$\mathrm{ACE}\chi^2$(14)	$\mathrm{ACE}\chi^2$ (15)
Dickeya dadantii	3970	85.21	74	8.3	9.7
Edwardsiella tarda	3535	85.44	95	12.1	14.7
Enterobacter sp. 638	4115	87.8	83	12.8	14.2
Cronobacter sakazakii	4255	89.1	80	13.0	15.1
Pectobacterium atrosepticum	4472	85.94	76	9.2	10.6
Erwinia tasmaniensis	3427	85.08	81	8.2	10.0
Escherichia coli	4149	85.2	86	14.6	15.9

Photorhabdus luminescens	4683	81.09	85	7.3	8.3
Proteus mirabilis	3607	84.36	83	10.3	12.0
Serratia proteamaculans	4891	87.26	85	9.6	10.9
Sodalis glossinidius	2432	50.91	69	5.3	6.5
Yersinia enterocolitica	3978	83.6	81	10.0	11.0
Hamiltonella defensa	2094	80.41	42	*N/A*	3.1
Xenorhabdus bovienii	4260	85.64	83	8.7	10.4
Pantoea ananatis	4237	87.95	67	8.9	10.5

Columns: ORF, number of annotated open reading frames on main chromosome; % coding, total nucleotide length of annotated ORFs as a percentage of the nucleotide length of the main chromosome; tRNA count, number of annotated tRNA genes on main chromosome; ACEχ^2(14), the normalized chi-square ACE for the 989 genes found in the 14 larger genomes; ACEχ^2 (15), the normalized chi-s;uare ACE for the634 genes found in all 15 genomes

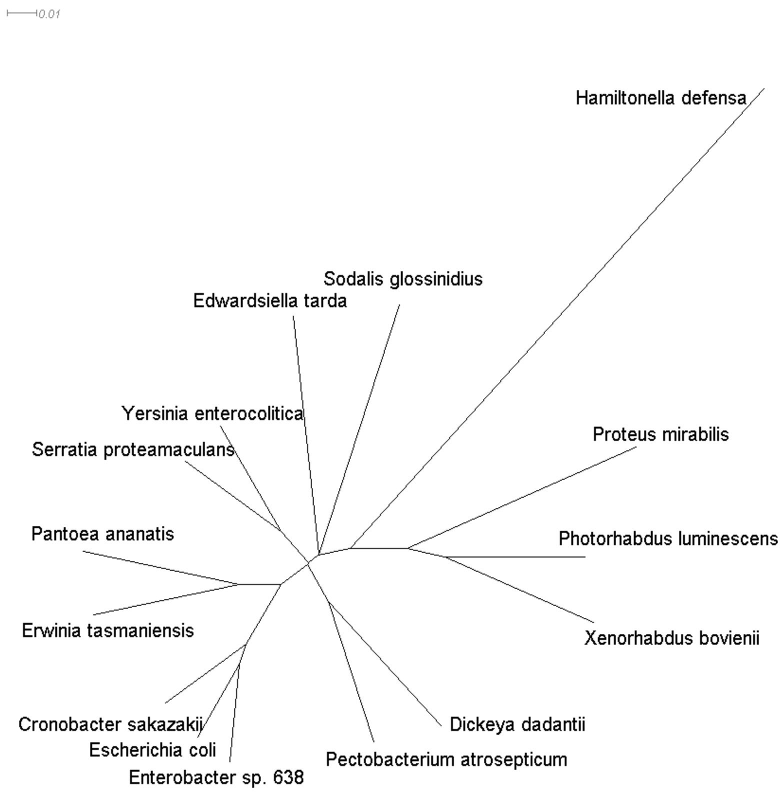


Figure 4.2 Phylogenetic relationship of 15 Enterobacteria

Neighbor Joining tree constructed from concatenated amino-acid alignment of 93 core genes from 15 Enterobacteria, where each multiple sequence alignment had fewer than 1% of positions with gaps. Constructed by ClustalW(Larkin, Blackshields et al. 2007)

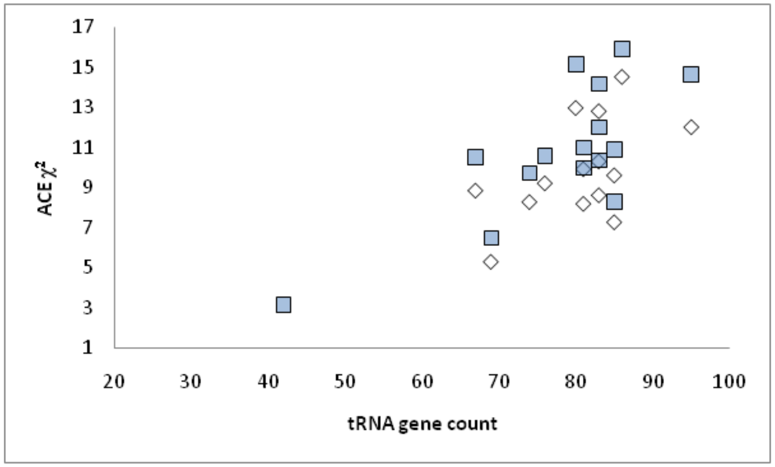


Figure 4.3 Association of ACEχ^2 with tRNA gene copy number

Scatter plot of tRNA count vs. ACEχ^2 scores (see Table 4.2). The filled squares represent ACEχ^2(15), while the diamonds represent ACEχ^2(14).

4.5 INTERPRETATION OF ACE

The ACE does not measure the magnitude of selection (*s*) on codon choice but rather the magnitude of the effect of codon selection on codon choice. While being strongly correlated to *s*, it actually addresses a slightly different issue. We have taken care to remove the influence of amino-acid composition from the ACE to provide a better prediction of physiological parameters such as gene expression levels. In contrast, an estimate of *s* should be sensitive to the amino acid composition, and a direct estimate of codon selection will likely provide better estimates of population diversity parameters such as the patterns of polymorphism (Sharp, Emery et al. 2010). Moreover, the ACE is a linear function of codon frequency; for an amino acid encoded by two codons, the contribution to ACE is directly proportional to the frequency of the preferred codon (*P*). In contrast, selection is a non-linear function of *P* ($N_e s = \log[(kP)/(1-P)]$) where k represents the mutational balance).

The ACE uses an estimate of the codon composition specified as arising from mutational processes alone. We constructed a single table to reflect these codon frequencies, implicitly assuming that a uniform process is acting upon all genes in the genome. This assumption is reasonable for bacteria once recently introduced genes are excluded, aside from subtle strand variation and origin-to-terminus gradients (Ochman 2003), and the model could be refined to accommodate such variance by creating separate f_o and f_n tables for leading and lagging strands, or for origin-proximal and terminus-proximal genes and interpolating the values to estimate f_o and f_n according to chromosomal position. This assumption of mutational uniformity is more severely undermined in some eukaryotes genomes that harbor isochores, wherein separate tables would need to be created.

4.6 VARIANCE IN THE ACE

We modeled the stochastic distribution of the ACE as though each gene had a constant amino acid composition and each amino acid could be encoded by any of its encoding codons with a probability given by genome-wide substitution parameters. Of course, amino acids will vary stochastically in a constant regime of mutation and selection, and modeling such variation may increase the expected variance of the ACE, though the normalization across amino acids should minimize any variance introduced by amino acid substitutions. Regardless of that correction, amino acid composition should not be modeled as a simple random variable because selective pressures acting on amino acid substitutions clearly are not uniform across the length of the protein. Substitution tables may provide some guidance for simulating the variation in amino acid composition that may be expected among ORFs experiencing identical mutational and selective pressures.

Selection acting on synonymous substitutions varies among sites within ORFs (Eyre-Walker and Bulmer 1995; Cannarozzi, Schraudolph et al. 2010; Tuller, Carmi et al. 2010). The ACE is robust to this complication layered on top of the mutation-selection-drift model, and can be interpreted as being proportional to the number of sites under strong selection for use of the globally preferred codon. Such variation in the strength of selection among sites would reduce the variance in the ACE and other codon bias statistics.

4.7 COMPARISONS OF CODON SELECTION ACROSS GENOMES

The ACEχ^2 is fundamentally different from previous attempts to quantify variation in the strength of codon selection between genomes. Three recently proposed methods have focused on a small fraction of the ORFs in each genome (*e.g.* ribosomal proteins) and used the deviation of their codon usage from the genome-wide average as an estimate of the efficacy of selection in each genome (Rocha 2004; Dethlefsen and Schmidt 2005; Sharp, Bailes et al. 2005). These methods have two assumptions that can be examined with the methods proposed here. First, they use the whole genome codon composition to estimate the equilibrium arising from mutational processes; second, they interpret the strength of selection on a particular subset of ORFs as being representative of, or proportional to, the strength of selection acting on all ORFs in the genome. In contrast, our ACE χ^2 statistic can be calculated from all genes believed to be long-term residents of the genome. While ACE χ^2 is correlated with tRNA copy number, as are these other statistics, differences between the statistics could indicate that the pre-selected subset of genes contributes more or less to the total codon selection experienced by a genome.

4.8 EXTENDING THE ACE FRAMEWORK TO OTHER ANALYSES.

The scoring table class of statistics includes several where the scoring table values (*i.e.*, δ_{ij}) are generated by calculations that do not consider the codon composition of actual genes [*e.g.*, the tAI (dos Reis, Savva et al. 2004) and eAI (Najafabadi, Lehmann et al. 2007)], instead relying on estimates arising from knowledge of molecular mechanisms. All of these approaches are amenable to the statistical analysis described for the ACE. While these reductionist approaches aim to reveal the same general phenomenon as the CAI or ACE, they are limited by uncertainty in how physiological properties such as tRNA abundance affect selection on codon usage. Even if knowledge of tRNA modifications and their effect on codon selection were perfect, these reductionist approaches would only illuminate the nature of contemporary selection, since they rely on descriptions of the current tRNA repertoire and suite of modifications. Yet tRNA genes are gained and lost from the genome (Withers, Wernisch et al. 2006) and the nature of their modifications also changes. Thus, mechanistic predictions of translational efficiency cannot capture a gene's history of codon selection as effectively as statistics that examine the codons themselves.

5.0 CONCLUDING REMARKS ON PROKARYOTIC BIODIVERSITY

Species concepts provide researchers with an intellectual framework within which to think about the relationships among organisms. Using these concepts to identify the definitive units of biodiversity remains problematic. This may be especially problematic for prokaryotes, with their immense diversity of genetic systems and ecological relationships (Doolittle and Zhaxybayeva 2009). Many prokaryotes are highly (though facultatively) recombinogenic, while others are essentially non-recombinogenic. Prokaryotes also follow different rules for different kinds of recombination, being able to acquire new genes from essentially any other organism, even as the transfer of allelic variants is limited to closely related organisms – usually. Prokaryote ecology also ranges from free-living organisms living in complex ecosystems, to obligate endosymbionts which are in many ways more like organelles than organisms.

The ecology of prokaryotes is just beginning to come to light. The vast diversity of many microbial communities creates opportunities for myriad ecological interactions, including predatory, competitive, and mutualistic. Each of these interactions creates opportunities for ongoing adaptation, which can happen quite quickly in prokaryotes. The spatial and temporal scale of these interactions is largely unexplored (Hunt, David et al. 2008; Vos, Birkett et al. 2009), leaving open the question of how barriers to migration contribute to biodiversity (Papke and Ward 2004; Green, Bohannan et al. 2008). Human transportation technology may even be changing the biogeography of microbes in a fundamental way (Gevers, Cohan et al. 2005).

The nature of prokaryotic recombination is also just beginning to be appreciated. The potential for genetic exchange is well described, but the factors that influence its actualization are poorly understood. Genetic and environmental conditions can greatly impact the frequency with which DNA is incorporated into a genome, and once there, selection can act on both the gene's compatibility with the rest of the genome and with the ecological niche occupied by the organism. For instance, many genomic fragments have been found to be toxic to the *E. coli* strain used during shotgun sequencing, and this toxicity depends on interactions with the gene expression machinery of the host genome (Sorek, Zhu et al. 2007).

Other components of the genome (*e.g.* chaperones, nucleoid-structuring proteins) are known to modify the behavior of specific genes in the genome; their diversity could play a large role in determining which recombination events are evolutionarily consequential. Beyond that, the fate of novel genetic material will depend on whether it modifies the existing genome in a manner that permits it to exploit new opportunities. Ultimately, genetic exchange and natural selection seem bound to interact in myriad complicated manners that prevent biodiversity and evolutionary processes from falling into a single, clear cut paradigm such as those provided by species concepts.

6.0 MATERIALS AND METHODS

6.1 GENERAL METHODS

6.1.1 Orthologous and distinctive genes:

For each pair of genomes, orthologous genes and distinctive genes were identified by performing BLASTP (proteins) or BLASTN (structural RNAs) searches with each predicted gene product against the products of all genes in the other genome, using default settings, ignoring the maximum expected value (Altschul, Madden et al. 1997). Genes that were pairwise reciprocal best matches (RBM) in all genomes being compared were used to construct the shared chromosomal backbone (table S1), in which gene adjacency was set to mirror the cross-species consensus. Where no cross-species consensus was available, adjacency was set to mirror that of a particular reference genome, and all analyses were repeated using a reference gene order from each taxon. If BLAST identified no possible orthologs in another genome, then the gene was listed as being distinctive between the two genomes (table S2), unless the query sequence did not contain enough information to confidently identify orthologs, as evidenced by an inability to identify itself with a BLAST e-value better than 10^{-5} in a search of its own genome. Matches were considered to be possible orthologs if they met any of the following conditions: were a RBM; had the same neighbor among RBMs in the compared genomes; or had amino acid similarity higher than a threshold chosen to distinguish between orthologs and paralogues (table S3), based on the distribution of similarity scores for RBMs in conserved locations and RBMs in disparate locations in the two genomes, respectively.

6.1.2 CAI and Divergence measurements:

Divergence of bacterial genes is inversely proportional to the strength of codon usage bias (Sharp and Li 1987). The Codon Adaptation Index (CAI) of each protein coding gene was measured after the method of Sharp and Li (1987); NSCU tables were constructed from the most highly-biased genes that contributed 13000 codons; highly-biased genes were initially identified using the P2 measure of codon bias(Gouy and Gautier 1982) and iteratively refined using the CAI. CAI values for *E. coli*, *S. enterica* and *Buchnera* genes were calculated using the NSCU tables from the K12, LT2 and APS genomes, respectively. CAI values were normalized to the mean CAI of all genes in that genome, reported in terms of standard deviations from the mean, and averaged across all members within a set of orthologous genes (table S1). Because the CAI is a surrogate measure of evolutionary rate, we would eliminate any gene from the analysis whose CAI values differed significantly among strains, possibly showing differences in evolutionary rates among lineage. We did not detect any such genes among those shared among the 6 strains of *E. coli* and *Salmonella* tested.

Synonymous substitutions were estimated for orthologous protein-coding sequences by the method of Li (1993), using the BLOSUM90 substitution matrix (Henikoff and Henikoff 1992) both to perform a local sequence alignment (Smith TF 1981) and to assign weights to alternative substitution pathways when aligned codons differed by more than one base. The variance in K_s is a decreasing function of the number of synonymous sites involved, so estimations of Ks were limited to genes containing at least 50 synonymous sites in the MG1655/LT2 alignment. To reduce the number of paralogous comparisons, estimations of K_s were limited to genes in orthologous sets where all RBMs had amino acid similarity higher than a threshold chosen to distinguish between orthologs and paralogs (table S3), based on the distribution of similarity scores for RBMs in conserved locations and RBMs in disparate locations in the two genomes, respectively. K_s for a set of orthologous genes is the mean of all K_s measurements between genes in that set from different species, except when K_s could not be calculated for a pair due to saturation of synonymous substitutions.

For genomes with biased nucleotide composition, divergence was estimated as dN and dS as implemented in the PAML program YN00 (Yang and Nielsen 2000).

6.2 METHODS OF CHAPTER 2

6.2.1 Genomes and software:

Genome sequences were retrieved from NCBI (ftp://ftp.ncbi.nih.gov/genomes/Bacteria/). *Salmonella enterica* was represented by strains LT2 (NC_003197.1), Ty2 (NC_004631.1), and ATCC 9150 (NC_006511.1); *Escherichia coli* was represented by strains K12-MG1655 (NC_000913.2), EDL933 (NC_002655.2), and CFT073 (NC_004431.1); *Buchnera aphidicola* was represented by strains APS (NC_002528.1) and Sg (NC_004061.1). Genes were retrieved using the documentation provided by NCBI RefSeq (Pruitt, Tatusova et al. 2005). All analyses were performed with DNA Master and DeltaKs; available from http://cobamide2.bio.pitt.edu/.

6.2.2 Calculation of relative divergence values for orthologous gene pairs:

Expected K_s values were generated as a function of CAI and distance from the origin of replication. Polynomial least-squares regressions were performed on the K_s/CAI data set (Figure 2.2A). A preliminary expected K_s (EKS) was calculated for each orthologous gene set using the CAI and the best regression curve chosen based on the following criteria: having a strong correlation between EKS and K_s; being a monotonic function over the relevant range of CAI values; and providing only positive EKS values. Calculations of EKS were limited to genes with CAI values within the range where the various polynomial regression functions generated similar EKS values (Figure 2.2A; vertical dashed line). Because K_s is also influenced by distance from the replication origin (Sharp, Shields et al. 1989),the residual from the K_s/CAI regression (*i.e.* K_s-EKS) was regressed as a linear function of the distance from the replication origin in the *E. coli* MG1655 genome (table S1). The corrected EKS (CEKS) is the sum of the EKS of a gene and its expected EKS based on its distance from the origin of replication, which represents an estimate of the rate of synonymous substitutions over the average divergence time of genes in the genome. Therefore, relative divergence is calculated as K_s /CEKS (table S1).

6.2.3 Identifying recombination regions:

To determine if relative divergence values were correlated at a given distance along the shared chromosomal backbone, every orthologous set was paired with the set at that distance; an analysis of variance was performed to determine if these pairs had significantly different means from each other. Between-pair and within-pair variability were measured and used to calculate intraclass correlation (ICC) and an F-statistic.

Regions with similar relative divergence values were identified using an agglomerative clustering algorithm that began with each orthologous set as a separate cluster, then iteratively merged a pair of adjacent clusters so as to minimize the total variability within clusters. This was repeated while monitoring the percentage of gene pairs at each distance that were included in the same cluster; clustering was terminated when the proportions at different distances most closely reflected the proportions of the ICCs for orthologous sets those distances.

6.2.4 Chromosomal regions associated with changes in gene content:

A locus is defined as pairs of orthologous genes which are adjacent (*ab*). For static loci, genes *ab* are adjacent in all 6 *E. coli* and *S. enterica* genomes. For dynamic loci, *ab* are adjacent in one genome while interrupted (*axb*) in another genome, where *x* represents gene(s) not present in the first genome (table S2). Species-specific dynamic loci show genotype *ab* for the 3 strains of one species, and genotype *axb* for the 3 strains of the other species (table S2); non-specific dynamic loci show other patterns. Because no outgroup was defined, locus *axb* may arise from locus *ab* by insertion, or locus *ab* may arise from locus *axb* by deletion. The average relative divergence is calculated for all genes *ab* in that class. Statistical significance of differences between average divergence values was determined by resampling from the parent set of loci and counting the portion of resampled sets with average divergence values equal to greater than the value from original sample.

6.3 METHODS OF CHAPTER 3

6.3.1 Ortholog identification.

Annotated open reading frames were translated and used as BLASTP queries to search databases composed of ORFs from each of the other genomes (e < 1) followed by semiglobal alignment. Sets of putative orthologs were assembled from those ORFs where each was a reciprocal best match with the others. Analyses of 17 enteric genomes (Figs. 1, 2) used alignments with >65% similarity; 14 genome analyses (Figures 3.3, 3.4; Table 3.1) used alignments with >70% sequence similarity. Multiple sequence alignments (MSA) were produced with ClustalW and back-translated to codon alignments. At least five syntenic genes must have been identified to establish orthology.

6.3.2 Genomes

The sequences of *Citrobacter* sp. 30_2, *C. koseri*, *C. youngae; Cronobacter sakazakii, Dickeya zeae, Klebsiella pneumoniae* strains 342 and MGH 78578, *Enterobacter* sp. 638, *Erwinia* *tasmaniensis, Escherichia coli* MG1655, UMN026, UTI89, and IAI39, *E. fergusonii, E. albertii, Pectobacterium wasabiae, Salmonella enterica enterica* LT2, CT18, and CVM19633, *S. enterica arizonae*, *Serratia proteamaculans* and *Yersinia enterocolitica* were downloaded from NCBI. Accession numbers appear in Table S4.

6.3.3 Quartet analyses

For each analysis, we evaluated the relationships among four groups of genomes for each orthologs. Alignments for each gene were subject to maximum likelihood analysis for each of the three possible topologies, while the relationships within each group were specified if the group comprised more than two genomes. The root in Figure 3.2 was specified according to the dominant topology in the NeighborNet tree (Figure 3.1), and the relationships among *Salmonella* and *Escherichia* strains in Figure 3.3AB were specified using MrBayes (Ronquist and Huelsenbeck 2003).

6.3.4 Maximum likelihood on ORFs

The topologies were evaluated by the PAML package (Yang 2007), using each MSA in turn, generating Resampling Estimated Log Likelihood (RELL) bootstrap support and Shimodaira-Hasegawa (SH) test p-values (Shimodaira and Hasegawa 1999). Our simulations (below) support bootstrap thresholds as a conservative estimate of accuracy. Codon-based ML used a single omega parameter and the Miyata geometric amino acid substitution probabilities. Codon-position ML constructed an HKY85 nucleotide substitution model for each codon position. This model is computationally efficient relative to the codon model, and has been shown to have similar accuracy for both closely and distantly related sequences (Ren, Tanaka et al. 2005).

6.3.5 Simulation of instant speciation

Simulated sequences were generated by the Evolver program in PAML. Test trees were generated by a codon-position nucleotide model. Using actual sequences, this produced results comparable to the full codon model, but was much less computationally intensive. All parameters were based upon the actual MSA being tested. The input tree was identical to the ML tree generated from the actual MSA, except that the innermost branch was set to be the same length as other branches in the tree that were being compared to the innermost branch. Sequence length was set to be identical to the number of sites aligned across all sequences of the MSA. Codon proportions were identical to the frequencies observed across all sequences in the MSA, and the kappa and omega parameters were set according to the parameters estimated by the YN00 program, with pairwise values averaged together by first averaging all pairs of genomes between any two groups within the quartets, then averaging the six pairwise values for the four groups within the quartet analysis.

6.4 METHODS OF CHAPTER 4

6.4.1 Sets of highly expressed genes for f_o.

Pre-selected sets of highly expressed genes were taken from previous literature. The set of 40 ribosomal proteins and translation elongation factors (Sharp, Bailes et al. 2005) included the genes *tufA, tsf, fusA, rplA-rplF, rplI-rplT* and *rpsB-rpsT*. The set of 27 highly expressed proteins (Sharp and Li 1987) included the genes *tufA, tufB, tsf, fusA, rplA, rplC, rplJ, rplK, rplL, rplQ, rpsA, rpsB, rpsG, rpsJ, rpsL, rpsO, rpsT, rpsU, rpmB, rpmG, rpmH, lpp, ompA, ompC, ompF, recA*, and *dnaK*.

6.4.2 Genomes used.

Acquired from the NCBI RefSeq website: *Pseudomonas aeruginosa PA01, P. mendocina ymp, P. stutzeri A1501, P. entomophila L48, and P. putida KT2440; Escherichia coli* K12 MG1665, *E. fergusonii* and *E. albertii; Buchnera aphidicola* APS. *Saccharomyces cerevisiae* S288c, *Dickeya dadantii, Edwardsiella tarda, Enterobacter sp. 638, Cronobacter sakazakii, Pectobacterium atrosepticum, Erwinia tasmaniensis, Photorhabdus luminescens, Proteus mirabilis, Serratia proteamaculans, Sodalis glossinidius, Yersinia enterocolitica, Hamiltonella defensa, Xenorhabdus bovienii, Pantoea ananatis*

SUPPLEMENTAL MATERIAL

Table S1 Robustness of incongruence for *Escherichia/Salmonella/Citrobacter* radiation.

This table reports the number of alignments (and percentage of all alignments) that support each topology after various filters have been applied. Filters are as follows: “RELL bootstrap support” = alignments than support the maximum likelihood topology with RELL bootstrap support above the specified value (Figure 3.3A). “Informative nucleotide sites” = alignments containing at least the specified number of informative sites among the eight genomes (two genomes each of *Escherichia, Salmonella, Citrobacter,* and *Klebsiella*). “Complete alignment”= alignments where the entire sequence of all 14 genomes was aligned in the multiple sequence alignment (other analyses require 95% alignment). “Consensus among analyses with taxon resampling (18 tests)” =Alignments supporting the same topology for the four genera regardless of the genomes used (one test using all 14 genomes, and one test using two representatives from each genus ,16 tests using a single representative of each genus). “No unmatched paralogs” = exclusion of all alignments where any gene had > 55% amino acid sequence similarity to a gene in any other genome that did not have a reciprocal best match in the first gene’s genome. “Consensus including *Enterobacter* outgroup (26 tests)”= Addition of another 8 analyses that used *Enterobacter* sp. 638 as a single outgroup along with a single genome from each of the three test genera. All other categories involve the application of two filters at the same time.

Filter	**Topology[1] count**			**Total**	**Portion**		
	1	**2**	**3**		**1**	**2**	**3**
None	457	327	360	**1144**	40%	29%	31%

RELL bootstrap support > 0.5	414	290	317	**1021**	41%	28%	31%
> 0.6	334	214	245	**793**	42%	27%	31%
> 0.7	244	138	180	**562**	43%	25%	32%
> 0.8	153	82	130	**365**	42%	22%	36%
> 0.9	73	40	69	**182**	40%	22%	38%
> 0.95	41	23	40	**104**	39%	22%	38%
> 0.99	16	6	14	**36**	44%	17%	39%
Informative nucleotide sites > 50	446	320	351	**1117**	40%	29%	31%
> 100	424	293	312	**1029**	41%	28%	30%
> 150	368	255	269	**892**	41%	29%	30%
> 200	290	196	209	**695**	42%	28%	30%
> 250	216	130	161	**507**	43%	26%	32%
> 300	143	93	108	**344**	42%	27%	31%
> 350	93	62	69	**224**	42%	28%	31%
> 400	66	45	44	**155**	43%	29%	28%
Complete alignment (all 14 sequences)	169	115	136	**420**	40%	27%	32%
Consensus among analyses with taxon resampling (18 tests)	142	78	106	**326**	44%	24%	33%
Consensus & alignment	50	32	37	**119**	42%	27%	31%

No unmatched paralogs (55% similarity)	207	161	202	**570**	36%	28%	35%
No unmatched paralogs (55% similarity) & consensus	64	42	51	**157**	41%	27%	32%
Consensus including *Enterobacter* outgroup (26 tests)	64	25	67	**156**	41%	16%	43%
***Enterobacter* consensus and alignment**	16	11	22	**49**	33%	22%	45%

1. Topologies: 1=(Cit,Esc),Sal; 2=(Cit,Sal),Esc; 3=(Esc,Sal),Cit

Table S2 Robustness of incongruence for Escherichia radiation.

This table reports the number of alignments (and percentage of all alignments) that support each topology after various filters have been applied. Filters are as follows: "RELL bootstrap support" = alignments than support the maximum likelihood topology with RELL bootstrap support above the specified value (Figure 3.3C). "Informative nucleotide sites" = alignments containing at least the specified number of informative sites among the seven genomes (three *Escherichia*, and two each of *Salmonella* and *Klebsiella*). "Complete alignment (14)"= alignments where the entire sequence of all 14 genomes was aligned in the multiple sequence alignment (other analyses require 95% alignment). "Complete alignment (7)"= as above, but limited to the 7 genomes in the phylogenetic analysis. "Consensus among analyses with different outgroups (3 tests)" =Alignments supporting the same topology for the three species regardless of the single outgroup used (*Salmonella enterica arizonae*, *Citrobacter youngae*, or *Klebsiella pneumoniae* str. 342). "No unmatched paralogs" = exclusion of all alignments where any gene had > 55% amino acid sequence similarity to a gene in any other genome that did not have a reciprocal best match in the first gene's genome. All other categories involve the application of two filters at the same time.

Filter	**Topology[1] count**			**Total**	**Portion**		
	1	**2**	**3**		**1**	**2**	**3**
None	602	449	93	**1144**	53%	39%	8%
RELL bootstrap support > 0.5	579	439	83	**1101**	53%	40%	8%

> 0.6	533	407	65	**1005**	53%	40%	6%
> 0.7	469	370	43	**882**	53%	42%	5%
> 0.8	397	337	21	**755**	53%	45%	3%
> 0.9	276	292	7	**575**	48%	51%	1%
> 0.95	195	253	4	**452**	43%	56%	1%
> 0.99	70	186	3	**259**	27%	72%	1%
Informative nucleotide sites > 50	583	437	85	**1105**	53%	40%	8%
> 100	510	388	73	**971**	53%	40%	8%
> 150	399	319	51	**769**	52%	41%	7%
> 200	273	226	32	**531**	51%	43%	6%
> 250	177	143	18	**338**	52%	42%	5%
> 300	104	95	5	**204**	51%	47%	2%
> 350	69	68	2	**139**	50%	49%	1%
> 400	41	48	1	**90**	46%	53%	1%
Complete alignment (for all 14 sequences)	229	144	47	**420**	55%	34%	11%
Complete alignment (for 7 included	236	157	48	**441**	54%	36%	11%

sequences)							
Consensus among analyses with different outgroups (3 tests)	435	342	27	**804**	54%	43%	3%
Consensus and alignment (14)	158	110	12	**280**	56%	39%	4%
No unmatched paralogs (55% similarity)	319	204	47	**570**	56%	36%	8%

1. Topologies: 1=(Eal,Efe),Eco; 2=(Eal,Eco),Efe; 3=(Eco,Efe)Eal

Table S3 All classes of topology-supporting sites are clustered

Two-topology runs tests for randomness indicate significant clustering of topology-supporting sites, demonstrating that clustering in the three-category runs test cannot be attributed to the clustering of a single topology. Runs tests were performed on the same concatenated alignments while ignoring each category of topology-supporting sites in turn. All z-scores occur with $p < 10^{-5}$ in random sequences.

		Topology[1]			Run of Sites			
		1	2	3	Expected	Variance	Observed	z score
ESC Nuc	All 3	19539	19886	18829	38827.0	12941.2	35684	-27.6
	1&2	N/A	19886	18829	19712.0	9854.0	18016	-17.1
	1&3	19539	N/A	18829	19178.4	9584.7	17626	-15.9
	1&4	19539	19886	N/A	19344.1	9663.6	17722	-16.5
ESC Prot	All 3	2382	2879	2442	5117.2	1703.8	4600	-12.5
	1&2	N/A	2879	2442	2608.0	1291.1	2375	-6.5
	1&3	2382	N/A	2442	2412.6	1204.9	2150	-7.6
	1&4	2382	2879	N/A	2643.6	1311.6	2338	-8.4
ESC Syn	All 3	12229	11666	11621	23671.9	7889.3	22844	-9.3
	1&2	N/A	11666	11621	11941.9	5966.4	11388	-7.2
	1&3	12229	N/A	11621	11918.3	5954.0	11586	-4.3
	1&4	12229	11666	N/A	11644.5	5821.0	11287	-4.7
EEE Nuc	All 3	9687	19298	18950	30717.8	10432.2	24327	-62.6
	1&2	N/A	19298	18950	12900.1	5739.8	11389	-19.9
	1&3	9687	N/A	18950	12821.4	5738.9	10856	-25.9
	1&4	9687	19298	N/A	19123.4	9559.7	12861	-64.1
EEE Prot	All 3	471	3074	1361	2558.1	840.9	1660	-31.0
	1&2	N/A	3074	1361	817.8	187.9	700	-8.6
	1&3	471	N/A	1361	700.8	266.8	601	-6.1
	1&4	471	3074	N/A	1887.7	802.0	992	-31.6
EEE Syn	All 3	6950	11009	13658	20356.9	6832.5	17159	-38.7
	1&2	N/A	11009	13658	8521.8	4042.1	7864	-10.3
	1&3	6950	N/A	13658	9213.3	4117.4	8060	-18.0
	1&4	6950	11009	N/A	12192.3	6024.6	9071	-40.2

1. ESC1=Cit,Esc; ESC2=Cit,Sal; ESC3=Esc,Sal; EEE1=Eal,Efe, EEE2=Eal,Eco; EEE3=Eco,Efe

Table S4 Chromosomes used in this study.
Seventeen were used for the analyses in Figures 1&2 (Analysis A). Fourteen were used to examine the Escherichia/Salmonella/Citrobacter split in Figures 3&4 (Analysis B). Seven were used to analyze the E. albertii/E. coli/E. fergusonii split (Analysis C).

Organism	Strain	Accession	RefSeq Annotation Date	Analyses
Escherichia coli	K12 MG1655	NC_000913	7/30/2009	ABC
E. coli	UMN026	NC_011751	11/10/2009	_B_
E. coli	UTI89	NC_007946	4/28/2009	AB_
E. coli	IAI39	NC_011750	11/10/2009	_B_
E. fergusonii	ATCC 35469	NC_011740	5/7/2009	ABC
E. albertii	TW07627	NZ_ABKX00000000	03/03/2008*	ABC
Salmonella enterica subsp. enterica	LT2 (Typhimurium)	NC_003197	4/30/2009	ABC
S. enterica subsp. enterica	CT18 (Typhi)	NC_003198	11/10/2009	_B_
S. enterica subsp. enterica	CVM19633 (Schwarzengrund)	NC_011094	4/29/2009	_B_
S. enterica subsp. arizonae	serovar 62:z4,z23	NC_010067	4/30/2009	ABC
Citrobacter koseri	ATCC BAA-895	NC_009792	5/7/2009	A__
C. sp.	30_2	NZ_ACDJ00000000	2/18/2009*	AB_
C. youngae	ATCC 29220	NZ_ABWL00000000	10/01/2008*	AB_
Klebsiella pneumoniae	342	NC_011283	4/28/2009	ABC
K. pneumoniae	MGH 78578	NC_009648	5/1/2009	_BC
Cronobacter sakazakii	ATCC BAA-894	NC_009778	4/26/2009	A__
Enterobacter sp.	638	NC_009436	5/7/2009	A__
Erwinia tasmaniensis	Et1/99	NC_010694	5/7/2009	A__
Serratia proteamaculans	568	NC_009832	5/7/2009	A__
Yersinia enterocolitica	8081	NC_008800	11/12/2009	A__
Pectobacterium wasabiae	WPP163	NC_013421	11/11/2009	A__
Dickeya zeae	Ech1591	NC_012912	7/7/2009	A__

*Release date of non-RefSeq list of coding regions

Table S5 Test of incomplete lineage sorting model.

Maximum Likelihood (ML) trees generated from actual sequence data were compared to ML trees generated from sequences simulated by Monte Carlo codon evolution. The comparison is based on the frequency with which the internal branch is longer than terminal branch for each of the 1144 multiple sequence alignments (Figure S1). Multiple sequence alignments are analyzed by topology because according to the incomplete lineage sorting model, one topology represents the species topology and is expected to have internal (between species) branches that are longer than the terminal (within species) branches. A total of 100 simulations were performed.

Topology		**Frequency that ML tree had longer internal branch than terminal branch**			**Significance**
ID	Count	Actual sequences (count)	Simulation mean ± s.d.	Simulation maximum	
Cit,Esc	308	94.8% (292)	61.5% ± 2.5%	66%	p < 0.01
Cit,Sal	372	93.0% (346)	60.0% ± 2.5%	64%	p < 0.01
Esc,Sal	463	95.7% (443)	60.6% ± 2.5%	65%	p < 0.01
Eal, Efe	66	34.8% (23)	40.9% ± 5.1%	56%	Not significant
Eal,Eco	434	88.2% (383)	70.1% ± 1.8%	75%	p < 0.01
Eco,Efe	644	78.7% (507)	61.2% ± 1.7%	64%	p < 0.01

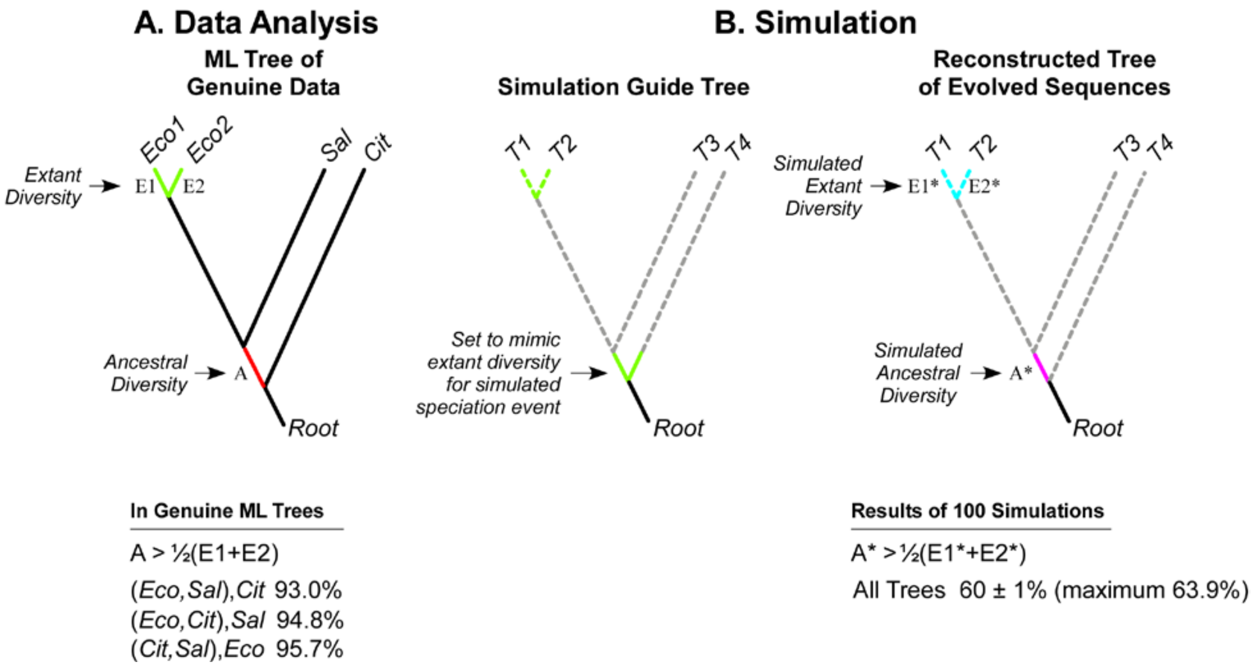


Figure S1 Schematic of test for incomplete lineage sorting.

(**Panel A, top)** The incomplete lineage sorting model proposes that incongruence among gene trees is the result of recombination that occurred in a freely recombining ancestral population prior two non-overlapping but closely timed processes producing genetic isolation at all loci (*i.e.* speciation). To test the plausibility of this model, we assumed that the ancestral recombining population would be comparable to *E. coli* (Fig. S2), then we examined the branch lengths on the best Maximum Likelihood tree and asking if extant diversity of *E. coli* (green terminal branches E1 and E2, separating Eco1 from Eco2) is large enough to account ancestral diversity inferred from the internal branches (red branch A, separating Sal, Cit and Eco) of the non-species topologies. (**Panel A, bottom**) Only for topologies representing the species relationships would the internal branch (A) regularly be longer than the terminal branch (one half of E1 + E2), yet all three topologies were represented by trees where the internal branch was regularly longer than the terminal branch. (**Panel B, top**) A molecular evolution simulation (100 replicates using the

Evolver program in the PAML package; of a codon model and guide tree derived for each multiple sequence alignment) generated expected values for the frequency with which terminal branch length would exceed internal branch length under this model. The guide tree was identical to the ML tree for the actual multiple sequence alignment, except that the internal branch was set to be the same length as the terminal branch (green branches). Evolver created a root sequence with properties similar to the observed sequence then evolved that sequence by a Monte Carlo substitution process for the lengths of time indicated by the tree according to substitution parameters inferred from the sequences. Recombination is not explicitly modeled, but is implicit in the inferred phylogenies of the *E. coli* strains that were used as guide trees (Fig. S2). The resulting simulated sequences were then subjected to the same quartet analysis as the actual sequence data, generating a distribution of values for how many genes generated trees where the internal branch length exceeded the terminal branch lengths. (**Panel B, bottom**; Table S7) This test is conservative to the extent that the estimates of terminal branch lengths are larger than implied by the population model that likely gave rise to the internal branch length. Gene tree incongruence is most pronounced -- and the internal branch length the shortest -- if the two speciation events are essentially simultaneous. This would result in a single allele from the ancestral population becoming fixed in each of the three descendant species, as long as each speciation event produced one low diversity (essentially clonal) population. Alternatively, the internal branch length would be longest -- and gene tree incongruence minimized – if the second speciation event occurred late enough after the first event that all alleles had become fixed in the intermediate species. Similarly, the internal branch length would be increased if, following the second speciation event, substantial ancestral diversity were to persist in each of the descendant populations and intragenic recombination were to occur. Conversely, recombination among alleles shared across the internal branch would shorten the internal branch (Penny, White et al. 2008). To account for the potential for these processes to inflate the internal branch length, we sampled the two most diverged strains among the *E. coli*, as though all of the alleles in one strain had become fixed (ignoring the incompatibility of this assumption with the observed levels of incongruence).

1. Penny D, White WT, Hendy MD, & Phillips MJ (2008) A bias in ML estimates of branch lengths in the presence of multiple signals. *Mol Biol Evol* 25(2):239-242.

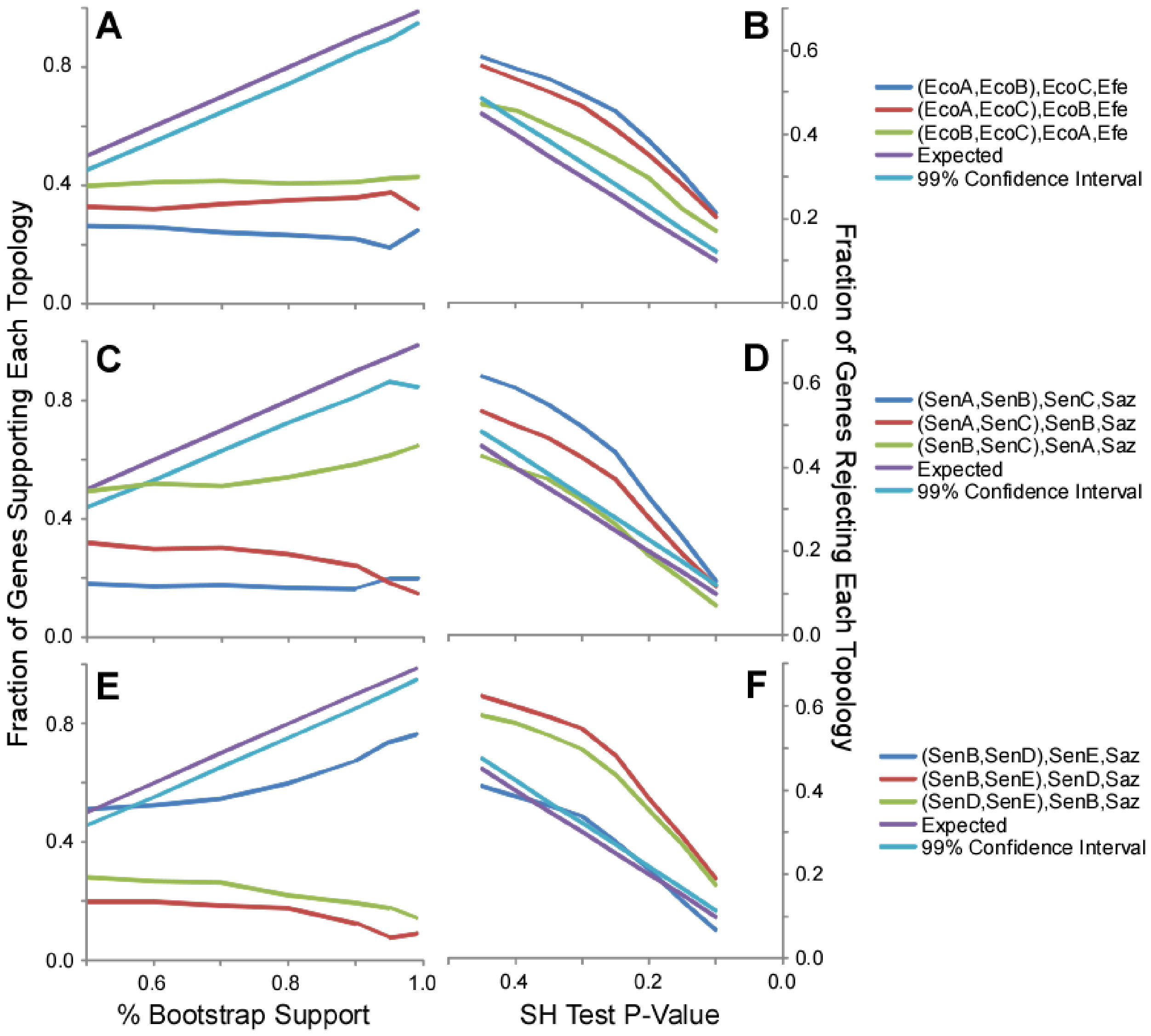


Figure S2 Phylogenetic discordance within extant recombining populations.
As in Figure 3.3, bootstrap support is expected to correspond to accuracy (**ACE**), and SH test p-values provide expected frequencies of topology rejection (**BDF**). Support for (**ACE**) or rejection of (**BDF**) alternative clades is indicated by trendlines. Quartets of strains were analyzed using *E. fergusonii* as an outgroup for *E. coli* strains (**AB**) and *S. enterica arizonae* as an outgroup for analysis of *S. enterica enterica* strains (**C-F**). Genomes are designated as follows: EcoA, *Escherichia coli* K12 MG1655; EcoB, *E. coli* IAI39; EcoC, *E. coli* UMN026; Efe, *E. fergusonii* ATCC 35469; SenA, *Salmonella enterica enterica* Schwarzengrund CVM19633 ; SenB, *S. enterica enterica* Typhi CT18; SenC, *S. enterica enterica* Typhimurium LT2; SenD, *S. enterica enterica* Paratyphi A ATCC 9150; SenE, *S. enterica enterica* Choleraesuis SC-B67; Saz, *S. enterica arizonae* 62:z4.

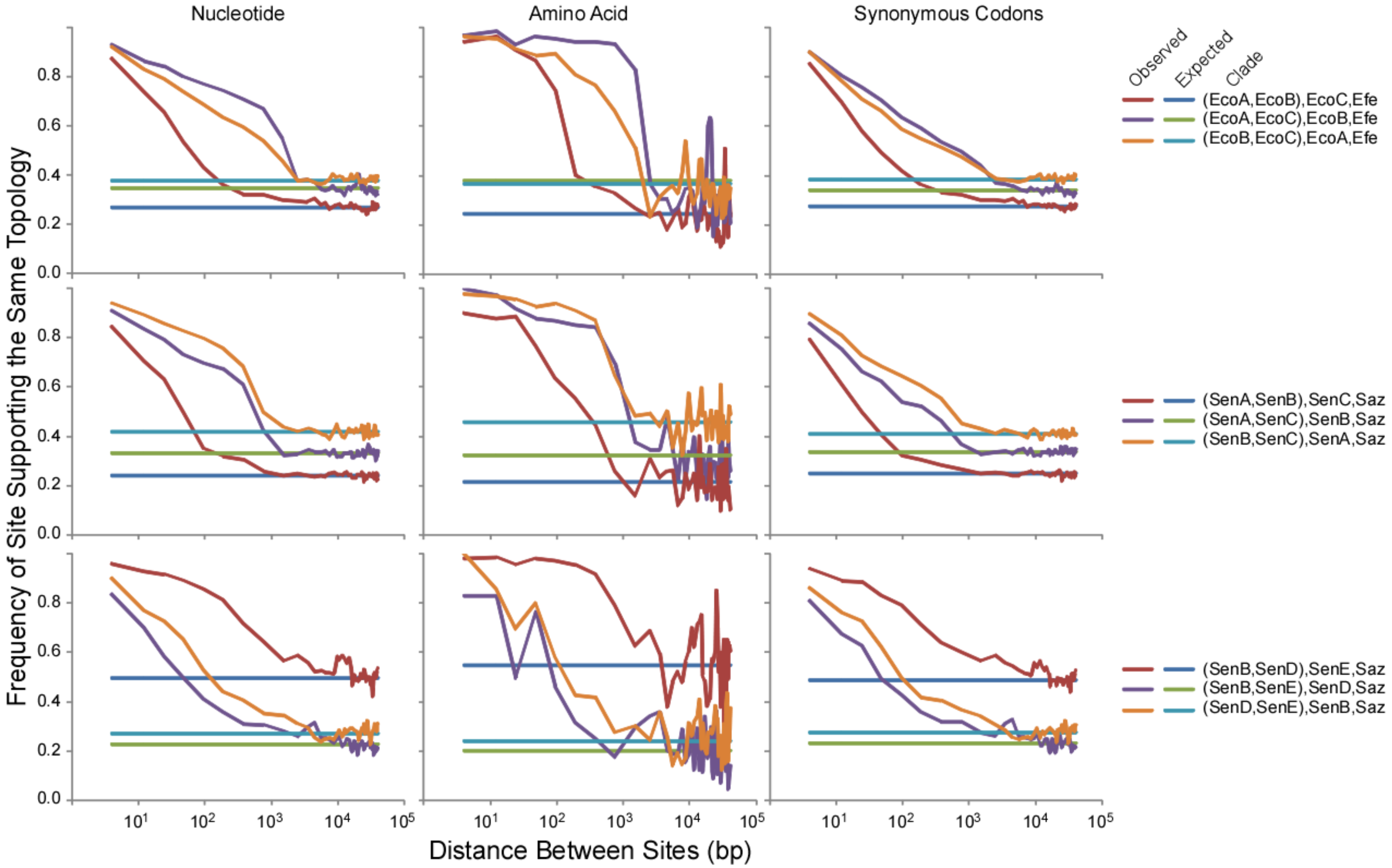


Figure S3 Chromosomal clustering of parsimony informative sites supporting each topology. As in Figure 3.4, each site supporting a distinct topology was compared against each other site and the observation binned according to the distance between sites. Trendlines report proportions of observations where a site supporting a given topology was paired with a site supporting the same topology. Expected values are derived from the genome-wide proportion of sites supporting that topology. Data is plotted at the midpoint of the bin range. Quartets are the same as in Fig. S2.

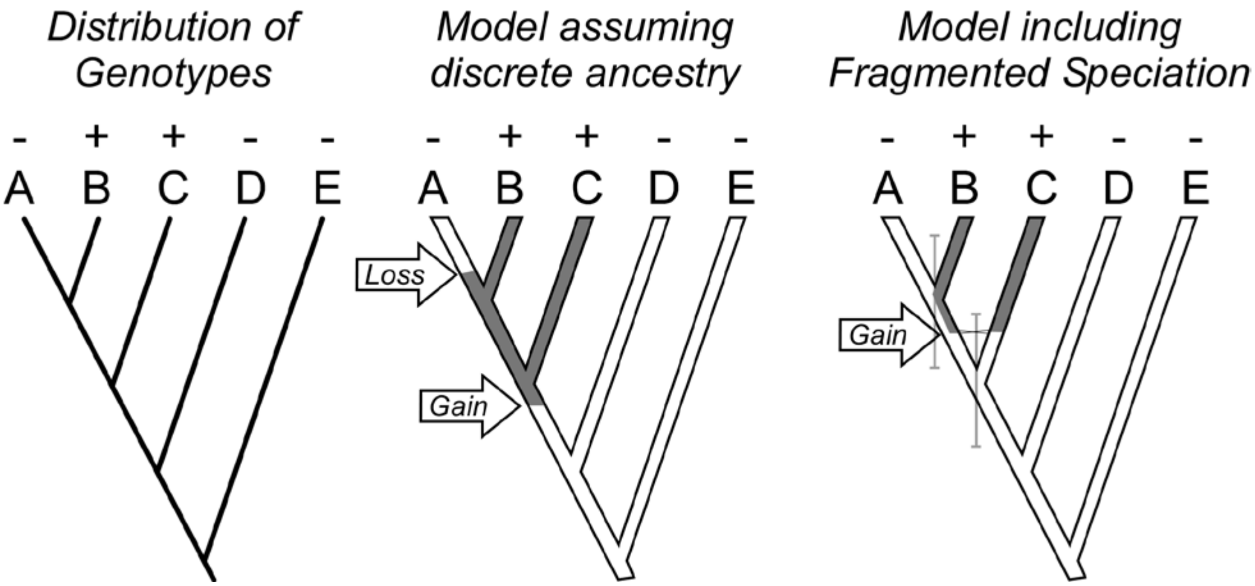

Distribution of Genotypes
Model assuming discrete ancestry
Model including Fragmented Speciation
- + + - -
A B C D E
- + + - -
A B C D E
Loss
Gain
- + + - -
A B C D E
Gain

Figure S4 Misinterpretation of character state data.

At left, the distribution of genotypes depicted on an inferred species tree. If one assumes independence of lineage separation events, multiple events are required to explain the existing variation (center). Yet the fragmented speciation process can lead to the retention of ancestral diversity in the emergent taxa, leading to the apparent discontinuous distribution of genotypes (right).

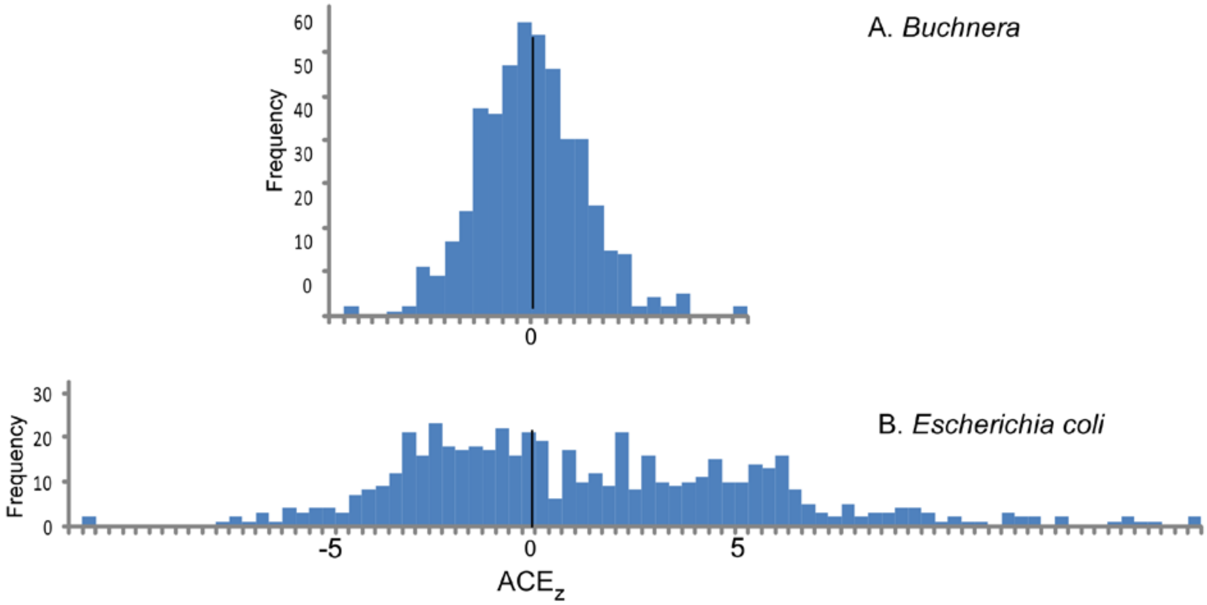


Figure S5 Distribution of ACE$_z$ scores

Distribution of ACE$_z$ scores for 498 orthologs identified between *Buchnera aphidicola* (A) and *Escherichia coli* (B).

BIBLIOGRAPHY

Altschul, S. F., T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller and D. J. Lipman (1997). "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs." Nucl. Acids Res. **25**(17): 3389-3402.

Atwood, K. C., L. K. Schneider and F. J. Ryan (1951). "Periodic selection in Escherichia coli." Proc Natl Acad Sci U S A **37**(3): 146-155.

Bapteste, E. and Y. Boucher (2009). "Epistemological impacts of horizontal gene transfer on classification in microbiology." Methods Mol Biol **532**: 55-72.

Barcus, V. A., A. J. Titheradge and N. E. Murray (1995). "The diversity of alleles at the *hsd* locus in natural populations of *Escherichia coli*." Genetics **140**(4): 1187-1197.

Bennetzen, J. L. and B. D. Hall (1982). "Codon selection in yeast." J Biol Chem **257**(6): 3026-3031.

Bernstein, J. A., A. B. Khodursky, P. H. Lin, S. Lin-Chao and S. N. Cohen (2002). "Global analysis of mRNA decay and abundance in Escherichia coli at single-gene resolution using two-color fluorescent DNA microarrays." Proc Natl Acad Sci U S A **99**(15): 9697-9702.

Brown, E. W., J. E. LeClerc, B. Li, W. L. Payne and T. A. Cebula (2001). "Phylogenetic evidence for horizontal transfer of mutS alleles among naturally occurring Escherichia coli strains." J Bacteriol **183**(5): 1631-1644.

Bryant, D. and V. Moulton (2004). "Neighbor-net: an agglomerative method for the construction of phylogenetic networks." Mol Biol Evol **21**(2): 255-265.

Bulmer, M. (1991). "The selection-mutation-drift theory of synonymous codon usage." Genetics **129**(3): 897-907.

Cannarozzi, G., N. N. Schraudolph, M. Faty, P. von Rohr, M. T. Friberg, A. C. Roth, P. Gonnet, G. Gonnet and Y. Barral (2010). "A role for codon order in translation dynamics." Cell **141**(2): 355-367.

Chattopadhyay, S., M. Feldgarden, S. J. Weissman, D. E. Dykhuizen, G. van Belle and E. V. Sokurenko (2007). "Haplotype diversity in "source-sink" dynamics of Escherichia coli urovirulence." J Mol Evol **64**(2): 204-214.

Cohan, F. M. (2001). "Bacterial species and speciation." Syst Biol **50**(4): 513-524.

Cohan, F. M. and E. B. Perry (2007). "A systematics for discovering the fundamental units of bacterial diversity." Curr Biol **17**(10): R373-386.

Connor, N., J. Sikorski, A. P. Rooney, S. Kopac, A. F. Koeppel, A. Burger, S. G. Cole, E. B. Perry, D. Krizanc, N. C. Field, M. Slaton and F. M. Cohan (2010). "Ecology of speciation in the genus Bacillus." Appl Environ Microbiol **76**(5): 1349-1358.

Cooper, T. F. and R. E. Lenski (2010). "Experimental evolution with E. coli in diverse resource environments. I. Fluctuating environments promote divergence of replicate populations." BMC Evol Biol **10**: 11.

Cooper, T. F., S. K. Remold, R. E. Lenski and D. Schneider (2008). "Expression profiles reveal parallel evolution of epistatic interactions involving the CRP regulon in Escherichia coli." PLoS Genet **4**(2): e35.

Darwin, C. (1859). On the Origin of Species.

Daubin, V., N. A. Moran and H. Ochman (2003). "Phylogenetics and the cohesion of bacterial genomes." Science **301**(5634): 829-832.

de Queiroz, K. (2005). "Ernst Mayr and the modern concept of species." Proc Natl Acad Sci U S A **102 Suppl 1**: 6600-6607.

del Campo, R., M. I. Morosini, E. G. de la Pedrosa, A. Fenoll, C. Munoz-Almagro, L. Maiz, F. Baquero and R. Canton (2005). "Population structure, antimicrobial resistance, and mutation frequencies of Streptococcus pneumoniae isolates from cystic fibrosis patients." J Clin Microbiol **43**(5): 2207-2214.

Demerec, M. and N. Ohta (1964). "Genetic Analyses of Salmonella Typhimurium X Escherichia Coli Hybrids." Proc Natl Acad Sci U S A **52**: 317-323.

Denamur, E., G. Lecointre, P. Darlu, O. Tenaillon, C. Acquaviva, C. Sayada, I. Sunjevaric, R. Rothstein, J. Elion, F. Taddei, M. Radman and I. Matic (2000). "Evolutionary implications of the frequent horizontal transfer of mismatch repair genes." Cell **103**(5): 711-721.

Dethlefsen, L. and T. M. Schmidt (2005). "Differences in codon bias cannot explain differences in translational power among microbes." BMC Bioinformatics **6**: 3.

Didelot, X., M. Achtman, J. Parkhill, N. R. Thomson and D. Falush (2007). "A bimodal pattern of relatedness between the Salmonella Paratyphi A and Typhi genomes: convergence or divergence by homologous recombination?" Genome Res **17**(1): 61-68.

Doolittle, W. F. and E. Bapteste (2007). "Pattern pluralism and the Tree of Life hypothesis." Proc Natl Acad Sci U S A **104**(7): 2043-2049.

Doolittle, W. F. and O. Zhaxybayeva (2009). "On the origin of prokaryotic species." Genome Res **19**(5): 744-756.

dos Reis, M., R. Savva and L. Wernisch (2004). "Solving the riddle of codon usage preferences: a test for translational selection." Nucleic Acids Res **32**(17): 5036-5044.

dos Reis, M. and L. Wernisch (2009). "Estimating translational selection in eukaryotic genomes." Molecular Biology and Evolution **26**(2): 451-461.

Dudley, A. M., J. Aach, M. A. Steffen and G. M. Church (2002). "Measuring absolute expression with microarrays with a calibrated reference sample and an extended signal intensity range." Proc Natl Acad Sci U S A **99**(11): 7554-7559.

Dykhuizen, D. E. (1998). "Santa Rosalia revisited: why are there so many species of bacteria?" Antonie Van Leeuwenhoek **73**(1): 25-33.

Dykhuizen, D. E. (2000). "Yersinia pestis: an instant species?" Trends Microbiol **8**(7): 296-298.

Dykhuizen, D. E. and A. M. Dean (2004). "Evolution of specialists in an experimental microcosm." Genetics **167**(4): 2015-2026.

Dykhuizen, D. E. and L. Green (1991). "Recombination in *Escherichia coli* and the definition of biological species." J. Bacteriol. **173**(22): 7257-7268.

Eyre-Walker, A. (1996). "Synonymous codon bias is related to gene length in Escherichia coli: selection for translational accuracy?" Mol Biol Evol **13**(6): 864-872.

Eyre-Walker, A. and M. Bulmer (1995). "Synonymous substitution rates in enterobacteria." Genetics **140**(4): 1407-1412.

Falush, D., M. Torpdahl, X. Didelot, D. F. Conrad, D. J. Wilson and M. Achtman (2006). "Mismatch induced speciation in Salmonella: model and data." Philos Trans R Soc Lond B Biol Sci **361**(1475): 2045-2053.

Feil, E. J., E. C. Holmes, D. E. Bessen, M. S. Chan, N. P. Day, M. C. Enright, R. Goldstein, D. W. Hood, A. Kalia, C. E. Moore, J. Zhou and B. G. Spratt (2001). "Recombination within natural populations of pathogenic bacteria: short-term empirical estimates and long-term phylogenetic consequences." Proc. Natl. Acad. Sci., USA **98**(1): 182-187.

Feil, E. J., M. C. Maiden, M. Achtman and B. G. Spratt (1999). "The relative contributions of recombination and mutation to the divergence of clones of Neisseria meningitidis." Mol Biol Evol **16**(11): 1496-1502.

Feil, E. J., J. M. Smith, M. C. Enright and B. G. Spratt (2000). "Estimating recombinational parameters in Streptococcus pneumoniae from multilocus sequence typing data." Genetics **154**(4): 1439-1450.

Feil, E. J. and B. G. Spratt (2001). "Recombination and the population structures of bacterial pathogens." Annu Rev Microbiol **55**: 561-590.

Fox, G. E., E. Stackebrandt, R. B. Hespell, J. Gibson, J. Maniloff, T. A. Dyer, R. S. Wolfe, W. E. Balch, R. S. Tanner, L. J. Magrum, L. B. Zablen, R. Blakemore, R. Gupta, L. Bonen, B. J. Lewis, D. A. Stahl, K. R. Luehrsen, K. N. Chen and C. R. Woese (1980). "The phylogeny of prokaryotes." Science **209**(4455): 457-463.

Fraser, C., W. P. Hanage and B. G. Spratt (2005). "Neutral microepidemic evolution of bacterial pathogens." Proc Natl Acad Sci U S A **102**(6): 1968-1973.

Fraser, C., W. P. Hanage and B. G. Spratt (2007). "Recombination and the nature of bacterial speciation." Science **315**(5811): 476-480.

Gatesy, J. and R. H. Baker (2005). "Hidden likelihood support in genomic data: can forty-five wrongs make a right?" Syst Biol **54**(3): 483-492.

Gevers, D., F. M. Cohan, J. G. Lawrence, B. G. Spratt, T. Coenye, E. J. Feil, E. Stackebrandt, Y. Van de Peer, P. Vandamme, F. L. Thompson and J. Swings (2005). "Re-evaluating prokaryotic species." Nat. Rev. Microbiol. **3**: 733-739.

Gouy, M. and C. Gautier (1982). "Codon usage in bacteria: correlation with gene expressivity." Nucleic Acids Res **10**(22): 7055-7074.

Grantham, R., C. Gautier, M. Gouy, R. Mercier and A. Pave (1980). "Codon catalog usage and the genome hypothesis." Nucleic Acids Res **8**(1): r49-r62.

Green, J. L., B. J. Bohannan and R. J. Whitaker (2008). "Microbial biogeography: from taxonomy to traits." Science **320**(5879): 1039-1043.

Greig, D. (2009). "Reproductive isolation in Saccharomyces." Heredity **102**(1): 39-44.

Grocock, R. J. and P. M. Sharp (2002). "Synonymous codon usage in Pseudomonas aeruginosa PA01." Gene **289**(1-2): 131-139.

Guttman, D. S. and D. E. Dykhuizen (1994). "Clonal divergence in Escherichia coli as a result of recombination, not mutation." Science **266**(5189): 1380-1383.

Guttman, D. S. and D. E. Dykhuizen (1994). "Detecting selective sweeps in naturally occurring *Escherichia coli*." Genetics **138**(4): 993-1003.

Hanage, W. P., C. Fraser and B. G. Spratt (2005). "Fuzzy species among recombinogenic bacteria." BMC Biol **3**: 6.

Hanage, W. P., C. Fraser and B. G. Spratt (2006). "The impact of homologous recombination on the generation of diversity in bacteria." J Theor Biol **239**(2): 210-219.

Hanage, W. P., C. Fraser, J. Tang, T. R. Connor and J. Corander (2009). "Hyper-recombination, diversity, and antibiotic resistance in pneumococcus." Science **324**(5933): 1454-1457.

Hanage, W. P., B. G. Spratt, K. M. Turner and C. Fraser (2006). "Modelling bacterial speciation." Philos Trans R Soc Lond B Biol Sci **361**(1475): 2039-2044.

Handelsman, J. (2004). "Metagenomics: application of genomics to uncultured microorganisms." Microbiol Mol Biol Rev **68**(4): 669-685.

Henikoff, S. and J. G. Henikoff (1992). "Amino Acid Substitution Matrices from Protein Blocks." Proc Natl Acad Sci U S A **89**(22): 10915-10919.

Henry, I. and P. M. Sharp (2007). "Predicting gene expression level from codon usage bias." Mol Biol Evol **24**(1): 10-12.

Hillis, D. M. and J. J. Bull (1993). "An Empirical Test of Bootstrapping as a Method for Assessing Confidence in Phylogenetic Analysis." Syst Biol **42**(2): 182-192.

Holt, K. E., J. Parkhill, C. J. Mazzoni, P. Roumagnac, F. X. Weill, I. Goodhead, R. Rance, S. Baker, D. J. Maskell, J. Wain, C. Dolecek, M. Achtman and G. Dougan (2008). "High-throughput sequencing provides insights into genome variation and evolution in Salmonella Typhi." Nat Genet **40**(8): 987-993.

Hunt, D. E., L. A. David, D. Gevers, S. P. Preheim, E. J. Alm and M. F. Polz (2008). "Resource partitioning and sympatric differentiation among closely related bacterioplankton." Science **320**(5879): 1081-1085.

Ikemura, T. (1981). "Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes." J Mol Biol **146**(1): 1-21.

Ikemura, T. (1981). "Correlation between the abundance of Escherichia coli transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the E. coli translational system." J Mol Biol **151**(3): 389-409.

Karlin, S. and J. Mrazek (2000). "Predicted highly expressed genes of diverse prokaryotic genomes." J Bacteriol **182**(18): 5238-5250.

Koeppel, A., E. B. Perry, J. Sikorski, D. Krizanc, A. Warner, D. M. Ward, A. P. Rooney, E. Brambilla, N. Connor, R. M. Ratcliff, E. Nevo and F. M. Cohan (2008). "Identifying the fundamental units of bacterial diversity: a paradigm shift to incorporate ecology into bacterial systematics." Proc Natl Acad Sci U S A **105**(7): 2504-2509.

Konstantinidis, K. T., A. Ramette and J. M. Tiedje (2006). "The bacterial species definition in the genomic era." Philosophical Transactions of the Royal Society B: Biological Sciences **361**(1475): 1929-1940.

Lan, R. and P. R. Reeves (2001). "When does a clone deserve a name? A perspective on bacterial species based on population genetics." Trends Microbiol **9**(9): 419-424.

Larkin, M. A., G. Blackshields, N. P. Brown, R. Chenna, P. A. McGettigan, H. McWilliam, F. Valentin, I. M. Wallace, A. Wilm, R. Lopez, J. D. Thompson, T. J. Gibson and D. G. Higgins (2007). "Clustal W and Clustal X version 2.0." Bioinformatics **23**(21): 2947-2948.

Lawrence, J. G. (2001). "Catalyzing bacterial speciation: correlating lateral transfer with genetic headroom." Syst Biol **50**(4): 479-496.

Lawrence, J. G. (2002). "Gene transfer in bacteria: speciation without species?" Theor. Popul. Biol. **61**(4): 449-460.

Lawrence, J. G., G. F. Hatfull and R. W. Hendrix (2002). "Imbroglios of viral taxonomy: genetic exchange and failings of phenetic approaches." J Bacteriol **184**(17): 4891-4905.

Lawrence, J. G. and A. C. Retchless (2010). "The myth of bacterial species and speciation." Biology and Philosophy **In Press**.

Lefebure, T. and M. J. Stanhope (2009). "Pervasive, genome-wide positive selection leading to functional divergence in the bacterial genus Campylobacter." Genome Res **19**(7): 1224-1232.

Lenski, R. E., C. L. Winkworth and M. A. Riley (2003). "Rates of DNA sequence evolution in experimental populations of Escherichia coli during 20,000 generations." J Mol Evol **56**(4): 498-508.

Levin, B. R. (1981). "Periodic selection, infectious gene exchange and the genetic structure of *E. coli* populations." Genetics **99**(1): 1-23.

Li, W. H. (1993). "Unbiased estimation of the rates of synonymous and nonsynonymous substitution." J. Mol. Evol. **36**(1): 96-99.

Lunzer, M., S. P. Miller, R. Felsheim and A. M. Dean (2005). "The biochemical architecture of an ancient adaptive landscape." Science **310**(5747): 499-501.

Maharjan, R., S. Seeto, L. Notley-McRobb and T. Ferenci (2006). "Clonal adaptive radiation in a constant environment." Science **313**(5786): 514-517.

Maiden, M. C., J. A. Bygraves, E. Feil, G. Morelli, J. E. Russell, R. Urwin, Q. Zhang, J. Zhou, K. Zurth, D. A. Caugant, I. M. Feavers, M. Achtman and B. G. Spratt (1998). "Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms." Proc Natl Acad Sci U S A **95**(6): 3140-3145.

Majewski, J. and F. M. Cohan (1999). "DNA sequence similarity requirements for interspecific recombination in Bacillus." Genetics **153**(4): 1525-1533.

Mallet, J. (2008). "Hybridization, ecological races and the nature of species: empirical evidence for the ease of speciation." Philos Trans R Soc Lond B Biol Sci **363**(1506): 2971-2986.

Mayr, E. (1942). Systematics and the Origin of Species. New York, Columbia University Press.

Mayr, E. (1963). Animal species and evolution. Cambridge, Mass, Belknap Press.

Merkl, R. (2003). "A survey of codon and amino acid frequency bias in microbial genomes focusing on translational efficiency." Journal of Molecular Evolution **57**(4): 453-466.

Milkman, R. (1997). "Recombination and population structure in *Escherichia coli*." Genetics **146**: 745-750.

Milkman, R. and M. M. Bridges (1993). "Molecular evolution of the Escherichia coli chromosome. IV. Sequence comparisons." Genetics **133**(3): 455-468.

Milkman, R., E. Jaeger and R. D. McBride (2003). "Molecular evolution of the *Escherichia coli* chromosome. VI. Two regions of high effective recombination." Genetics **163**(2): 475-483.

Milkman, R., E. A. Raleigh, M. McKane, D. Cryderman, P. Bilodeau and K. McWeeny (1999). "Molecular evolution of the Escherichia coli chromosome. V. Recombination patterns among strains of diverse origin." Genetics **153**(2): 539-554.

Murray, N. E. (2000). "Type I restriction systems: sophisticated molecular machines (a legacy of Bertani and Weigle)." Microbiol. Mol. Biol. Rev. **64**(2): 412-434.

Najafabadi, H. S., J. Lehmann and M. Omidi (2007). "Error minimization explains the codon usage of highly expressed genes in Escherichia coli." Gene **387**(1-2): 150-155.

Negri, M. C., M. I. Morosini, M. R. Baquero, R. del Campo, J. Blazquez and F. Baquero (2002). "Very low cefotaxime concentrations select for hypermutable Streptococcus pneumoniae populations." Antimicrob Agents Chemother **46**(2): 528-530.

Novembre, J. A. (2002). "Accounting for background nucleotide composition when measuring codon usage bias." Mol Biol Evol **19**(8): 1390-1394.

O'Neill, M., A. Chen and N. E. Murray (1997). "The restriction-modification genes of *Escherichia coli* K-12 may not be selfish: they do not resist loss and are readily replaced by alleles conferring different specificities." Proc. Natl. Acad. Sci., USA **94**(26): 14596-14601.

Ochman, H. (2003). "Neutral mutations and neutral substitutions in bacterial genomes." Mol Biol Evol **20**(12): 2091-2096.

Ochman, H., J. G. Lawrence and E. Groisman (2000). "Lateral gene transfer and the nature of bacterial innovation." Nature **405**: 299-304.

Ochman, H., J. G. Lawrence and E. A. Groisman (2000). "Lateral gene transfer and the nature of bacterial innovation." Nature **405**(6784): 299-304.

Ochman, H. and A. C. Wilson (1988). "Evolution in bacteria: evidence for a universal substitution rate in cellular genomes." J. Mol. Evol. **26**: 74-86.

Orsi, R. H., Q. Sun and M. Wiedmann (2008). "Genome-wide analyses reveal lineage specific contributions of positive selection and recombination to the evolution of Listeria monocytogenes." BMC Evol Biol **8**: 233.

Pamilo, P. and M. Nei (1988). "Relationships between gene trees and species trees." Mol Biol Evol **5**(5): 568-583.

Papke, R. T. and D. M. Ward (2004). "The importance of physical isolation to microbial diversification." FEMS Microbiol Ecol **48**(3): 293-303.

Papke, R. T., O. Zhaxybayeva, E. J. Feil, K. Sommerfeld, D. Muise and W. F. Doolittle (2007). "Searching for species in haloarchaea." Proc Natl Acad Sci U S A **104**(35): 14092-14097.

Penny, D., W. T. White, M. D. Hendy and M. J. Phillips (2008). "A bias in ML estimates of branch lengths in the presence of multiple signals." Mol Biol Evol **25**(2): 239-242.

Perez, J. C., D. Shin, I. Zwir, T. Latifi, T. J. Hadley and E. A. Groisman (2009). "Evolution of a bacterial regulon controlling virulence and Mg(2+) homeostasis." PLoS Genet **5**(3): e1000428.

Pruitt, K. D., T. Tatusova and D. R. Maglott (2005). "NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins." Nucl. Acids Res. **33**(suppl_1): D501-504.

Quinn, P. C. (2002). Young Infants' Categorization of Humans Versus Nonhuman Animals: Roles for Knowledge Access and Perceptual Process. Building Object Categories in Developmental Time. D. H. R. Lisa Gershkoff-Stowe, Lawrence Erlbaum Associates.

Rainey, P. B. and M. Travisano (1998). "Adaptive radiation in a heterogeneous environment." Nature **394**(6688): 69-72.

Rambach, A. (1990). "New plate medium for facilitated differentiation of *Salmonella* spp. from *Proteus* spp. and other enteric bacteria." Appl. Environ. Microbiol. **56**(1): 301-303.

Ren, F., H. Tanaka and Z. Yang (2005). "An empirical examination of the utility of codon-substitution models in phylogeny reconstruction." Syst Biol **54**(5): 808-818.

Rieseberg, L. H., T. E. Wood and E. J. Baack (2006). "The nature of plant species." Nature **440**(7083): 524-527.

Rocha, E. P. (2004). "Codon usage bias from tRNA's point of view: redundancy, specialization, and efficient decoding for translation optimization." Genome Res **14**(11): 2279-2286.

Roncero, C., K. E. Sanderson and M. J. Casadaban (1991). "Analysis of the host ranges of transposon bacteriophages Mu, MuhP1, and D108 by use of lipopolysaccharide mutants of Salmonella typhimurium LT2." J. Bacteriol. **173**(16): 5230-5233.

Ronquist, F. and J. P. Huelsenbeck (2003). "MrBayes 3: Bayesian phylogenetic inference under mixed models." Bioinformatics **19**(12): 1572-1574.

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.

Schubert, S., P. Darlu, O. Clermont, A. Wieser, G. Magistro, C. Hoffmann, K. Weinert, O. Tenaillon, I. Matic and E. Denamur (2009). "Role of intraspecies recombination in the spread of pathogenicity islands within the Escherichia coli species." PLoS Pathog **5**(1): e1000257.

Sharp, P. M., E. Bailes, R. J. Grocock, J. F. Peden and R. E. Sockett (2005). "Variation in the strength of selected codon usage bias among bacteria." Nucleic Acids Res **33**(4): 1141-1153.

Sharp, P. M., L. R. Emery and K. Zeng (2010). "Forces that influence the evolution of codon bias." Philos Trans R Soc Lond B Biol Sci **365**(1544): 1203-1212.

Sharp, P. M. and W.-H. Li (1987). "The codon adaptation index - a measure of directional synonymous codon usage bias, and its potential applications." Nucleic Acids Res. **15**: 1281-1295.

Sharp, P. M. and W. H. Li (1987). "The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications." Nucleic Acids Res **15**(3): 1281-1295.

Sharp, P. M. and W. H. Li (1987). "The rate of synonymous substitution in enterobacterial genes is inversely related to codon usage bias." Mol Biol Evol **4**(3): 222-230.

Sharp, P. M., D. C. Shields, K. H. Wolfe and W. H. Li (1989). "Chromosomal location and evolutionary rate variation in enterobacterial genes." Science **246**(4931): 808-810.

Shen, P. and H. V. Huang (1986). "Homologous recombination in Escherichia coli: dependence on substrate length and homology." Genetics **112**(3): 441-457.

Sheppard, S. K., N. D. McCarthy, D. Falush and M. C. Maiden (2008). "Convergence of Campylobacter species: implications for bacterial evolution." Science **320**(5873): 237-239.

Sheskin, D. J. (2007). Handbook of Parametric and Nonparametric Statistical Procedures, Chapman \& Hall/CRC.

Shimodaira, H. and M. Hasegawa (1999). "Multiple Comparisons of Log-Likelihoods with Applications to Phylogenetic Inference." Mol Biol Evol **16**(8): 1114-1116.

Smith, N. G. and A. Eyre-Walker (2001). "Why are translationally sub-optimal synonymous codons used in Escherichia coli?" J Mol Evol **53**(3): 225-236.

Smith TF, W. M. (1981). "Identification of common molecular subsequences." J Mol Biol. **147**(1): 195-197.

Sorek, R., Y. Zhu, C. J. Creevey, M. P. Francino, P. Bork and E. M. Rubin (2007). "Genome-wide experimental determination of barriers to horizontal gene transfer." Science **318**(5855): 1449-1452.

Soyer, Y., R. H. Orsi, L. D. Rodriguez-Rivera, Q. Sun and M. Wiedmann (2009). "Genome wide evolutionary analyses reveal serotype specific patterns of positive selection in selected Salmonella serotypes." BMC Evol Biol **9**: 264.

Spratt, B. G., L. D. Bowler, Q. Y. Zhang, J. Zhou and J. M. Smith (1992). "Role of interspecies transfer of chromosomal genes in the evolution of penicillin resistance in pathogenic and commensal Neisseria species." J Mol Evol **34**(2): 115-125.

Springer, B., P. Sander, L. Sedlacek, W. D. Hardt, V. Mizrahi, P. Schar and E. C. Bottger (2004). "Lack of mismatch correction facilitates genome evolution in mycobacteria." Mol Microbiol **53**(6): 1601-1609.

Stoletzki, N. and A. Eyre-Walker (2007). "Synonymous codon usage in Escherichia coli: selection for translational accuracy." Mol Biol Evol **24**(2): 374-381.

Strick, J. E. (2000). Sparks of Life. Cambridge, Massachusetts, Harvard University Press.

Sullivan, M. B., J. B. Waterbury and S. W. Chisholm (2003). "Cyanophages infecting the oceanic cyanobacterium Prochlorococcus." Nature **424**(6952): 1047-1051.

Supek, F. and K. Vlahovicek (2005). "Comparison of codon usage measures and their applicability in prediction of microbial gene expressivity." BMC Bioinformatics **6**: 182.

Tamas, I., L. Klasson, B. Canback, A. K. Naslund, A.-S. Eriksson, J. J. Wernegreen, J. P. Sandstrom, N. A. Moran and S. G. E. Andersson (2002). "50 Million Years of Genomic Stasis in Endosymbiotic Bacteria." Science **296**(5577): 2376-2379.

Taylor, D. J. and W. H. Piel (2004). "An assessment of accuracy, error, and conflict with support values from genome-scale phylogenetic data." Mol Biol Evol **21**(8): 1534-1537.

Touchon, M., C. Hoede, O. Tenaillon, V. Barbe, S. Baeriswyl, P. Bidet, E. Bingen, S. Bonacorsi, C. Bouchier, O. Bouvet, A. Calteau, H. Chiapello, O. Clermont, S. Cruveiller, A. Danchin, M. Diard, C. Dossat, M. E. Karoui, E. Frapy, L. Garry, J. M. Ghigo, A. M. Gilles, J. Johnson, C. Le Bouguenec, M. Lescat, S. Mangenot, V. Martinez-Jehanne, I. Matic, X. Nassif, S. Oztas, M. A. Petit, C. Pichon, Z. Rouy, C. S. Ruf, D. Schneider, J. Tourret, B. Vacherie, D. Vallenet, C. Medigue, E. P. Rocha and E. Denamur (2009). "Organised genome dynamics in the Escherichia coli species results in highly diverse adaptive paths." PLoS Genet **5**(1): e1000344.

Tuller, T., A. Carmi, K. Vestsigian, S. Navon, Y. Dorfan, J. Zaborske, T. Pan, O. Dahan, I. Furman and Y. Pilpel (2010). "An evolutionarily conserved mechanism for controlling the efficiency of protein translation." Cell **141**(2): 344-354.

van Valen, L. (1976). "Ecological Species, Multispecies, and Oaks." Taxon **25**(2/3): 233-239.

Vieira-Silva, S. and E. P. Rocha (2010). "The systemic imprint of growth and its uses in ecological (meta)genomics." PLoS Genet **6**(1): e1000808.

Vos, M., P. J. Birkett, E. Birch, R. I. Griffiths and A. Buckling (2009). "Local adaptation of bacteriophages to their bacterial hosts in soil." Science **325**(5942): 833.

Vulic, M., F. Dionisio, F. Taddei and M. Radman (1997). "Molecular keys to speciation: DNA polymorphism and the control of genetic exchange in Enterobacteria." Proc. Natl. Acad. Sci., USA **94**(18): 9763-9767.

Vulic, M., R. E. Lenski and M. Radman (1999). "Mutation, recombination, and incipient speciation of bacteria in the laboratory." Proc. Natl. Acad. Sci., USA **96**(13): 7348-7351.

Waite, R. D., A. Paccanaro, A. Papakonstantinopoulou, J. M. Hurst, M. Saqi, E. Littler and M. A. Curtis (2006). "Clustering of Pseudomonas aeruginosa transcriptomes from planktonic cultures, developing and mature biofilms reveals distinct expression profiles." BMC Genomics **7**: 162.

Walk, S. T., E. W. Alm, D. M. Gordon, J. L. Ram, G. A. Toranzos, J. M. Tiedje and T. S. Whittam (2009). "Cryptic lineages of the genus Escherichia." Appl Environ Microbiol **75**(20): 6534-6544.

Welch, R. A., V. Burland, G. Plunkett, 3rd, P. Redford, P. Roesch, D. Rasko, E. L. Buckles, S. R. Liou, A. Boutin, J. Hackett, D. Stroud, G. F. Mayhew, D. J. Rose, S. Zhou, D. C. Schwartz, N. T. Perna, H. L. Mobley, M. S. Donnenberg and F. R. Blattner (2002). "Extensive mosaic structure revealed by the complete genome sequence of uropathogenic Escherichia coli." Proc Natl Acad Sci U S A **99**(26): 17020-17024.

Wertz, J. E., C. Goldstone, D. M. Gordon and M. A. Riley (2003). "A molecular phylogeny of enteric bacteria and implications for a bacterial species concept." J Evol Biol **16**(6): 1236-1248.

Whitaker, R. J., D. W. Grogan and J. W. Taylor (2003). "Geographic barriers isolate endemic populations of hyperthermophilic archaea." Science **301**(5635): 976-978.

Whitaker, R. J., D. W. Grogan and J. W. Taylor (2005). "Recombination shapes the natural population structure of the hyperthermophilic archaeon *Sulfolobus islandicus*." Mol. Biol. Evol. **22**(12): 2354-2361.

Wick, L. M., W. Qi, D. W. Lacher and T. S. Whittam (2005). "Evolution of genomic content in the stepwise emergence of Escherichia coli O157:H7." J Bacteriol **187**(5): 1783-1791.

Wildschutte, H. and J. G. Lawrence (2007). "Differential Salmonella survival against communities of intestinal amoebae." Microbiology **153**(Pt 6): 1781-1789.

Wildschutte, H. K., D. M. Wolfe, A. Tamewitz and J. G. Lawrence (2004). "Protozoan predation, diversifying selection and the evolution of antigenic diversity in *Salmonella*." Proc. Natl. Acad. Sci., USA **101**(29): 10644-10649.

Wilkins, J. S. (2004, April 26, 2004). "Spontaneous Generation and the Origin of Life." The TalkOrigins Archive Retrieved 7/10, 2010, from http://www.talkorigins.org/faqs/abioprob/spontaneous-generation.html.

Withers, M., L. Wernisch and M. dos Reis (2006). "Archaeology and evolution of transfer RNA genes in the Escherichia coli genome." RNA **12**(6): 933-942.

Woese, C. R., O. Kandler and M. L. Wheelis (1990). "Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya." Proc Natl Acad Sci U S A **87**(12): 4576-4579.

Wright, F. (1990). "The 'effective number of codons' used in a gene." Gene **87**(1): 23-29.

Yang, Z. (2007). "PAML 4: phylogenetic analysis by maximum likelihood." Mol Biol Evol **24**(8): 1586-1591.

Yang, Z. and R. Nielsen (2000). "Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models." Mol Biol Evol **17**(1): 32-43.

Zawadzki, P., M. S. Roberts and F. M. Cohan (1995). "The log-linear relationship between sexual isolation and sequence divergence in *Bacillus* transformation is robust." Genetics **140**(3): 917-932.