Genomic signatures of sex, selection and speciation in the microbial world

by

B. Jesse Shapiro

B.Sc. Biology
McGill University, 2003
M.Sc. Integrative Bioscience
Oxford University, 2004

SUBMITTED TO THE PROGRAM IN COMPUTATIONAL AND SYSTEMS BIOLOGY IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY IN COMPUTATIONAL AND SYSTEMS BIOLOGY
AT THE
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

SEPTEMBER 2010

Signature of Author:...……..
Program in Computational and Systems Biology
August 6, 2010

Certified by:…….
Eric J. Alm
Assistant Professor of Civil and Environmental Engineering and Biological Engineering
Thesis Supervisor

Accepted by:……
Christopher B. Burge
Professor of Biology and Biological Engineering
Chair, Computational and Systems Biology Ph.D. Graduate Committee

Genomic signatures of sex, selection and speciation in the microbial world

by

B. Jesse Shapiro

Submitted to the Program in Computational and Systems Biology
on August 6, 2010 in Partial Fulfillment of the
Requirements for the Degree of Doctor of Philosophy in
Computational and Systems Biology

ABSTRACT

Understanding the microbial world is key to understanding global biogeochemistry, human health and disease, yet this world is largely inaccessible. Microbial genomes, an increasingly accessible data source, provide an ideal entry point. The genome sequences of different microbes may be compared using the tools of population genetics to infer important genetic changes allowing them to diversify ecologically and adapt to distinct ecological niches. Yet the toolkit of population genetics was developed largely with sexual eukaryotes in mind.

In this work, I assess and develop tools for inferring natural selection in microbial genomes. Many tools rely on population genetics theory, and thus require defining distinct populations, or species, of bacteria. Because sex (recombination) is not required for reproduction, some bacteria recombine only rarely, while others are extremely promiscuous, exchanging genes across great genetic distances. This behavior poses a challenge for defining microbial population boundaries.

This thesis begins with a discussion of how recombination and positive selection interact to promote ecological adaptation. I then describe a general pipeline for quantifying the impacts of mutation, recombination and selection on microbial genomes, and apply it to two closely related, yet ecologically distinct populations of *Vibrio splendidus*, each with its own microhabitat preference. I introduce a new tool, STARRInIGHTS, for inferring homologous recombination events. By assessing rates of recombination within and between ecological populations, I conclude that ecological differentiation is driven by small number of habitat-specific alleles, while most loci are shared freely across habitats.

The remainder of the thesis focuses on lineage-specific changes in natural selection among anciently diverged species of gammaproteobacteria. I develop two new metrics, selective signatures and slow:fast, for detecting deviations from the expected rate of evolution in 'core' proteins (present in single copy in most species). Because they rely on empirical distributions of evolutionary rates across species, these methods should become increasingly powerful as more and more microbial genomes are sampled.

Overall, the methods described here significantly expand the repertoire of tools available for microbial population genomics, both for investigating the process of ecological differentiation at the finest of time scales, and over billions of years of microbial evolution.

Thesis Supervisor: Eric J. Alm
Title: Assistant Professor of Civil and Environmental Engineering and Biological Engineering

Table of Contents **page**

List of Figures

List of Tables

Overview

This thesis has the overarching aim of harnessing whole genome sequences to gain insight into the mechanisms of ecological differentiation in the microbial world. Microbes are key players in global biogeochemical cycles, human health and disease; yet the microbial world is largely hidden from view. Even with the best microscopes and experimental techniques, it is exceedingly difficult to know the predominant selective pressures and ecological interactions at play in the wild. Microbial genome sequences provide a uniquely comprehensive and accessible record of the forces that drive microbial evolution. The products of evolution are ecologically differentiated populations or species, each under distinct regimes of natural selection. A typical approach to understanding the genetic basis of ecological adaptation is to deploy sequence-based statistical tests to identify genes under positive selection. These genes may have habitat-specific adaptive value, and may help us understand how different microbes adapt to different niches.

Gaining biological insight from microbial genome sequences and tests for selection poses several challenges. First, there are challenges arising from the enormous dynamic range of microbial evolutionary time scales: we may be interested in comparing species that diverged hundreds of millions to billions of years ago, or that diverged so recently that it is unclear if they constitute separate species or not. Second, while it was once thought that microbes do not form species in the classical sense because they reproduce clonally and do not recombine their DNA through sex, the idea is now gaining popularity that they do not form proper species because they have *too much* sex, due to their ability to exchange genes by horizontal transfer spanning great genetic distances.

Taking these difficulties into account, in the first section of this thesis I introduce the predominant forces shaping microbial evolution and assess current tests for positive selection in terms of their ability to cope with the peculiarities of microbial recombination (Introduction). A key conclusion of this section is that because sex (recombination) is in a sense optional for microbes (not required for reproduction), acquisition of DNA via homologous or illegitimate recombination, followed by its selective retention, may itself be considered an adaptive event.

In the next section (Chapter 1), I develop a pipeline for quantifying the contributions of mutation, homologous and illegitimate recombination to the process of ecological differentiation, a key step toward speciation. The core of this pipeline is STARRInIGHTS, a new method I developed to detect homologous recombination events. This chapter focuses on a population of *Vibrio splendidus* genomes, which despite their very close genetic relatedness appear to have recently split into two ecologically distinct sub-

populations, each with a different microhabitat preference in the marine water column. A key conclusion of Chapter 1 is that a small number of habitat-specific alleles, often acquired by homologous recombination, may drive ecological differentiation while most genomic loci are freely recombined across habitat boundaries. This situation is somewhat analogous to the process of sympatric speciation in animal populations, where speciation is driven by ecological, rather than physical boundaries to gene flow.

Chapters 2 and 3 turn to more ancient divergences – on the order of hundreds of millions of years – where species boundaries are more firmly established. In these chapters, I develop two new metrics for detecting changes in the strength and direction of natural selection at the protein level. Both metrics, selective signatures (Chapter 2) and the slow:fast substitution ratio (Chapter 3), exploit comparisons of multiple microbial genomes to identify proteins that deviate significantly from an empirically determined expected rate of evolution. I focus primarily on instances of lineage-specific evolutionary rate accelerations, potential indicators of positive selection. In Chapter 2, I demonstrate how proteins of similar cellular function tend to co-evolve over long evolutionary time spans; thus selective signatures is a powerful predictor of protein function. In Chapter 3, I show how different functional suites of proteins are subject to different regimes of selection in ecologically distinct lineages, and give examples from *Pseudomonas* and *Vibrio*. Selective signatures and slow:fast both have difficulty distinguishing positive selection from relaxed selective constraint, but they have the common strength of detecting *lineage-specific* evolutionary rate accelerations, which are informative of the types of proteins that have been fine-tuned specifically in a species of interest.

A unifying theme of the approaches I describe is that they use whole genome sequences, as opposed to metagenomic data. For better or for worse, this forces an organism-centered view, rather than a gene-centered view, of microbial ecology. When genes are freely recombined from genome to genome, a gene-centered view may be appropriate. Yet, for at least some period of evolutionary time, genes are coupled with one another in the same genome, and the whole genome becomes the effective substrate of natural selection. By studying whole genomes, we can achieve a better understanding of epistatic interactions among genes, and also learn about (and control for) genomewide forces such as changes in population size or mutation rate.

The thesis concludes with a brief discussion of how STARRInIGHTS, selective signatures and slow:fast are being applied for a variety of purposes in several organisms, and discusses future prospects for detecting adaptive events in microbial genomes.

Acknowledgements

I gratefully acknowledge my thesis supervisor, thesis advisory committee members, and co-authors for their contributions to this thesis. Specifically, Jonathan Friedman programmed the selection/recombination model described in the Introduction (Figure 0.1), and along with Otto Cordero, was responsible for most of the flexible genome analysis in Chapter 1 (Figures 1.2B and 1.S13 and Table 1.S2). Otto Cordero also designed and implemented the phylogenetic incongruence pre-filter for STARRInIGHTS, described in the Chapter 1 methods and Figure 1.S7. Lawrence David conducted the meta-analysis detailed in Figure 0.3. Sarah Preheim performed some of the PCR and sequencing described in Chapter 1 (Figure 1.S2 and Table 1.S3). Sonia Timberlake assembled the genomes described in Chapter 1. Dirk Gevers gathered the core gene families for the 24 *E. coli* strains used in Chapter 2. Although the rest of the work is my own, I benefited enormously from discussions with and suggestions from members of the Alm lab.

Other contributions to this thesis are harder to quantify. Thanks to my parents for making me call them once a week, periodically asking what I am working on and instilling in me the belief that the number of bicycles one should own is equal to one plus the number of bicycles one currently owns. Thanks to my sister for meeting me once a year for Pop Montreal, actually understanding what I am working on and being the gracious recipient of overflow when the bicycle count exceeded n+1. And thanks to Hillary for hanging out with me every day for at least the past 5 years and not asking what I am working on but being absolutely certain it's amazing. Also thanks for introducing me to the wonders of public television in the wee hours of the morning, being a fair critic of the aesthetic value of my figures, and doing my taxes.

Introduction:

Looking for Darwin's footprints in the microbial world

Shapiro BJ, David LA, Friedman J and Alm EJ (2009) Looking for Darwin's footprints in the microbial world. *Trends in Microbiology* **17**:196-204

Abstract

As we observe the 200th anniversary of Charles Darwin's birthday, microbiologists interested in the application of Darwin's ideas to the microscopic world have a lot to celebrate: an emerging picture of the (mostly microbial) Tree of Life at ever-increasing resolution, an understanding of horizontal gene transfer as a driving force in the evolution of microbes, and thousands of complete genome sequences to help formulate and refine our theories. At the same time, quantitative models of the microevolutionary processes shaping microbial populations remain just out of reach, a point that is perhaps most dramatically illustrated by the lack of consensus on how (or even whether) to define bacterial species. We summarize progress and prospects in bacterial population genetics, with an emphasis on detecting the footprint of positive Darwinian selection in microbial genomes.

0.1 Selection as a window into the microbial world

The microbial world is largely hidden from the naked eye, making it difficult to know the selective pressures acting on a bacterium. Nonetheless, if the genes important to fitness in a particular niche are identified, they can have a dramatic impact on our understanding of that environment. A few recent examples from marine environments help to illustrate this point. The discovery of bacteriorhodopsin in diverse marine bacteria revealed a previously unsuspected evolutionary adaptation that explained the long-standing puzzle of how such a variety of species can thrive in the nutrient poor open ocean (Beja et al, 2000, Sabehi et al, 2004). Moreover, spectral tuning of these molecules might help explain the distribution of different species across different regions and depths in the ocean (Bielawski et al, 2004). In another example, phosphorous acquisition genes in *Prochlorococcus* distribute preferentially in strains living in periodically phosphorus-limited waters, suggesting an obvious link between genetics and environmental factors (Martiny et al, 2006). Many other cases of gene-specific environmental selection, however, probably involve more subtle genetic changes than the gain or loss of an entire gene or pathway, such as amino acid substitutions at specific functional sites (Feldgarden et al, 2003).

While ongoing advances in genome sequencing technologies have made it possible to obtain complete genome sequences for entire populations of microbes, it is not clear whether genome sequences can be converted directly into evolutionary insight. One appealing and conceptually simple approach comes from the emergent field of population genomics: align the genomes of an entire population of individuals, and use the traditional tools of population genetics to pinpoint loci involved in recent Darwinian selection.

In this paper, we discuss the prospects for uncovering Darwinian selection in microbial genomes, which are becoming more readily available for a broad spectrum of medically, agriculturally and ecologically interesting organisms. We focus on genetic adaptations driven by positive selection, and the challenges involved in detecting them (Box 1). Tests for positive selection will depend on patterns of recombination, which are expected to differ between asexual microbes and sexual eukaryotes, and also among microbes depending on their lifestyles and demography. We summarize the different types of tests (Table 0.1), highlighting their relative merits under different regimes of recombination, and discuss how the interplay of positive selection and recombination affects the patterns of genetic variation in microbes.

0.2 Detecting natural selection in genomic sequence

Genome-wide scans for positively-selected loci in metazoans, especially humans, have yielded substantial insights into the functions of unknown genes, and the genetic basis of phenotypic differences between species. The general approach is to gather genome sequences of related species, compute a sequence-

based metric to quantify positive selection on each locus, and take outliers from the genome-wide distribution as candidate positively-selected loci (Black et al, 2001, Luikart et al, 2003). This approach is exemplified by a recent study of six mammalian genomes, which used the *dN/dS* metric (Table 0.1) to reveal that genes involved in immunity and sensory perception played a key role in differentiating primates and rodents (Kosiol et al, 2008). Genes acting in the same biochemical pathway were also found to undergo positive selection together, a finding we previously reported in bacteria (Shapiro & Alm, 2008) (see chapter 2). Genome-wide scans for positive Darwinian selection have also been performed on finer scales – for example, within human populations, revealing very recent positive selection in genes involved in malaria resistance (Sabeti et al, 2002), hair follicle production (Kelley et al, 2006, Sabeti et al, 2007), and lactose tolerance (Tishkoff et al, 2007).

But are these approaches, which are being applied in earnest to sexual eukaryotes, theoretically justified in bacteria where recombination might not be frequent enough to provide gene-specific resolution? To investigate this question, we first review the tools available for detecting selection in different types of populations.

0.3 Tools for near-clonal populations

In bacteria, patterns of genetic variation depend on the extent to which populations behave clonally. In a perfectly clonal population, every substitution in the genome will have arisen by mutation, never by recombination. Every adaptive allele that arises will therefore be perfectly linked to every other allele in the genome. If the goal is to distinguish adaptive loci from other mutations fixed in the clonal background, one could look for loci with an excess of functional changes (e.g, *dN/dS* and other methods discussed in the 'Detecting selection among species and higher-order groups' section), and/or that show evidence for convergent evolution.

One advantage of a perfectly clonal population, from a practical standpoint, is that genomes are related by a single phylogenetic tree, rather than a more complicated network structure that represents recombination. Alleles that arise independently multiple times in different branches, and are thus incongruous with the tree, stand out as candidate examples of convergent evolution. In a recent study of genetic variation among isolates of *Salmonella enterica* serovar Typhi, convergent evolution was observed at a few loci (Holt et al, 2008). Recombination was ruled out as the cause of the phylogenetic incongruence and convergent mutations resulted in amino acid substitutions, two of which have known adaptive value in conferring antibiotic resistance, further supporting the hypothesis of positive selection.

Sokurenko *et al.* introduced 'zonal analysis' to identify mutations of uropathogenic *Escherichia coli* associated with recent invasion of a new niche, the human urinary tract (Sokurenko et al, 2004). By definition, such mutations are recently derived, and are found near the tips of a well-resolved phylogeny. They also tended to involve repeated (convergent) amino acid changes in variable 'hotspots', and these occurred in uropathogenic but not commensal strains. The authors could thus conclude that these mutations conferred a competitive advantage in the uropathogenic niche. This type of analysis – effectively a special case of convergence testing – could be extended and generalized to identify recently selected loci across the genome, even when the population structure and/or selection pressure is obscure.

0.4 Tools for sexual populations

A suite of population-genetic tests for non-neutral patterns of evolution have been developed over the past 20 years, and often used to detect positively-selected loci in humans or other sexual eukaryotes. Although many of these tests are sensitive to various deviations from neutral evolution (Table 0.1), they primarily detect selective sweeps. These tests can be divided into two main classes: (i) Tajima's *D* and related diversity-based tests of deviations from the neutral allele-frequency spectrum (Tajima, 1989, Zeng et al, 2006) or (ii) long-range haplotype (LRH) and related tests (Sabeti et al, 2007). Diversity-based tests identify alleles that are at unusually high frequency (suggesting a selective sweep of a single beneficial allele) or intermediate frequency (diversifying selection maintaining multiple alleles in the population). This class of test can be applied to aligned homologous sequences of any length, assuming that sites within the sequence are completely linked (no recombination between them). Thus, diversity-based statistics should be computed within short windows of DNA across the genome. Meanwhile, haplotye-based tests model the decay of linkage disequilibrium (LD) with physical distance in the genome to identify haplotypes that are at unexpectedly high frequency for their age, indicating a recent or ongoing selective sweep.

Both classes of tests should in theory be applicable to populations of bacteria in which homologous recombination among strains is rampant. Diversity-based tests have been applied to a variety of bacterial populations, including cyanobacteria (Mes et al, 2006), *Buchnera* and related insect endosymbionts (Herbeck et al, 2003), *Neisseria* (Jolley et al, 2005), and *Pseudomonas* (Guttman et al, 2006). In a broad study spanning seven bacterial phyla, Hughes found that Tajima's *D* tends to be lower in nonsynonymous sites than synonymous sites, implying purifying selection on slightly deleterious mutations that lead to amino acid changes (Hughes, 2005). Hughes also observed that this difference between sites implies that recombination must be occurring at some level in order to allow sites to evolve independently. This

implies that diversity-based tests have potential to pinpoint selected loci, provided that they are separated from the clonal background by recombination.

Meanwhile, haplotype-based tests have not been applied to bacterial populations - perhaps because population sampling has not been performed at sufficient resolution to capture very recent selective sweeps. Even frequently recombining bacteria differ from sexual eukaryotes in two major ways: (i) in bacteria, recombination occurs by gene conversion rather than crossing-over and (ii) recombination is decoupled from reproduction. As a result of gene conversion, linkage between nearby loci is expected to be higher than between distant loci in bacteria, making haplotype-based tests valid, in principle, over short genomic distances. But unlike in organisms that recombine by crossing-over, linkage is expected between all loci *not* involved in a gene conversion event regardless of their physical distance on the chromosome (McVean et al, 2002). In other words, a non-recombinant locus might be linked to another distant locus, but unlinked to nearby loci that have undergone gene conversion. This type of pattern, called a clonal frame (Milkman & Bridges, 1990), is more likely to occur when gene conversion size fragments are small (e.g. on the order of ~500 bp in *Helicobacter pylori* (Falush et al, 2001)), when recombination is infrequent, or when only certain combinations of two distant alleles are tolerated, creating linkage between them, with free recombination in the intervening region (Wiehlmann et al, 2007). Such epistatic interactions among alleles could thus affect patterns of LD, and in principle represent an important determinant of recombination frequency (Kondrashov & Kondrashov, 2001).

0.5 Can recombination maintain diversity in the face of selective sweeps?

Whether diversity- and haplotype-based tests are able to distinguish adaptive mutations within a population depends on the balance between the opposing forces of positive selection (purging diversity as a new allele approaches fixation) and recombination (maintaining diversity by unlinking distant regions of the genome from a selective sweep). The ratio *r/s* can be used as a shorthand to express this balance between recombination (*r*) and selection (*s*) in a population, and *r/s* is rarely, if ever, measured. Much more commonly measured is the *r/m* ratio, which assesses the relative likelihood that a polymorphic site arises by recombination versus mutation (*m*). *r/m* varies widely among bacteria (e.g. *r/m* is ~ 5-80 in *Neisseria meningitidis*, ~ 50 in *Streptococcus pneumoniae*, ~ 10-50 in *E. coli*, and ~ 1-3 in *Bacillus* (Feil et al, 1999, Didelot & Falush, 2007)), but is generally larger than 1 (per site), suggesting that recombination is quite strong relative to mutation in many species. But the mutation rate is universally quite low in bacteria: on the order of 10^{-10} mutations per site per generation (Drake, 1991). Even if recombination generates much more diversity than mutation, is it a sufficiently strong diversity generating force to maintain diversity in the face of selection? Selective coefficients for adaptive mutations might be

quite high: on the order of 0.01 or higher (Barrett et al, 2006). Thus, adaptive mutations might become fixed before recombination has time to act, leading to genome-wide purges of diversity (Atwood et al, 1951, Cohan, 2001) sometimes referred to as periodic selection. This would render diversity- and haplotype-based tests powerless to discern the selected locus against the background of uniformly low diversity.

So exactly how much recombination is needed to overcome the purging of diversity that can result from periodic selection? Figure 0.1A illustrates the level of diversity at a single locus in a population after an adaptive mutation at a distant gene locus becomes fixed, as a function of r and s. When recombination rates are low, the resulting diversity at the distant (non-selected) locus is effectively purged resulting in a clonality of 1, which we define as Σp_i^2, where p_i is the frequency of the i^{th} allele at the non-selected locus (also called Simpson's diversity index in ecology). When recombination rates are high, diversity at the non-selected locus is maintained either through the generation of new mutations or retention of initial diversity. Simplifying matters, the trends are only weakly dependent on the population size (Figure 0.1B) when expressed in the natural variables ρ (=Nr, where N is the population size and r is the per gene or locus per generation recombination rate) and σ (=Ns, where s is the relative fitness advantage conferred by the adaptive mutation).

As illustrated graphically in Figure 0.1, and as previously modeled by others (*e.g.*, (Majewski & Cohan, 1999, Cohan & Koeppel, 2008)), there exist regimes of recombination and selection that allow an adaptive mutation to purge diversity locally in the genome (at loci near the adaptive site), without substantially reducing diversity elsewhere in the genome. For this panmictic scenario to be viable, r must be quite large and/or s must be small. High r/s or $r > s$ can be used to roughly describe this scenario, but this shorthand does not do justice to the complex relationship between diversity, r and s (Figure 0.1A). Panmixis could result from a relatively high population recombination rate, similar to those observed in promiscuous groups such as *Neisseria* or *Helicobacter* (Falush et al, 2001, Feil et al, 1999). Conversely, if $r << s$, a population will be effectively clonal. This type of clonal population has been observed repeatedly in the context of long-term experimental evolution studies of *E. coli*. In these studies, adaptive point mutations are successively fixed in a clonal background, with little or no contribution of recombination (Rozen et al, 2005, Hegreness et al, 2006, Blount et al, 2008).

0.6 Patterns of recombination and their interplay with selection

Bacterial lineages show substantial variation in population structure - ranging from essentially clonal (*e.g. Salmonella*) to panmictic (*e.g. Neisseria gonorrhoeae*) (Smith et al, 1993). In many recombining lineages

including *Campylobacter* (Fearnhead et al, 2005) *Neisseria* (Jolley et al, 2005), *Helicobacter* (Falush et al, 2001), and *E. coli* (Mau et al, 2006), LD decays with distance in the genome. To illustrate this distance-dependent decay of LD, we show LD between pairs of genes located throughout the *E. coli* genome (Figure 0.2). Genes up to ~20 kilobases apart on the chromosome show LD, but this LD drops off around 100 kb. However, LD never decays to zero in bacteria because some fraction of very distant loci will remain together in the clonal frame (McVean et al, 2002). Patterns of LD have a great impact on tests for selection (Table 0.1, Figure 0.4), and it is thus important to quantify these patterns in the population of interest.

How can distance-dependent LD be explained? First, some recombination must be occurring although the clonal frame is still evident since a fraction of even very distant loci may be linked (Didelot & Falush, 2007). Second, the majority of recombinant fragments introduced by gene conversion are probably small (Fearnhead et al, 2005, Mau et al, 2006, Guttman & Dykhuizen, 1994, McKane & Milkman, 1995). Third, it follows from the simulation results that r must be large and/or s must be small. This leads to three different scenarios:

1. the strains in question form an effectively sexual population (r is large),
2. selection coefficients are low or selection is infrequent (s is small), or
3. there are ecological barriers to selective sweeps but not recombination

The first two scenarios are quite straightforward, but the third merits further discussion. In Scenario three, different populations of bacteria might inhabit different micro-niches, making it rare for a single genotype to sweep through all populations. How likely is such a scenario? In the coastal water column, at least 15 ecologically distinct subpopulations of *Vibrio splendidus* co-exist (Hunt et al, 2008a). Because the coastal ocean is relatively well-mixed, there might be even more opportunity for resource partitioning and niche subdivision in other (e.g. terrestrial or host-associated) environments (e.g. (Koeppel et al, 2008)). Therefore, a meta-population model, such as the one described by Majewski and Cohan (Majewski & Cohan, 1999), could explain the distance-dependent decay of LD. Their model requires two or more populations, each adapted to a different niche and these niches might only be very slightly different or even transient niches. Each population experiences independent selective sweeps, purging diversity within each sub-population, yet genetic diversity remains high when summed over all populations. Occasionally, a globally adaptive mutant occurs in one of the populations, but cannot sweep through other populations because its genotype as a whole is relatively unfit in the other micro-niches. But if the globally adaptive mutation is recombined into the 'native' background of another population, it can confer a fitness advantage and rise in frequency.

Eventually, globally adapted mutations can purge diversity locally in the genome without disrupting unlinked genomic diversity. Thus, Majewski and Cohan's model could explain the observed distance-dependent decay of LD without invoking a high rate of recombination. Because niche partitioning in the microbial world likely occurs at a very fine spatial scale (Hunt et al, 2008a, Thompson et al, 2005), it is probable that many bacterial population samples actually encompass multiple subpopulations, connected by rare recombination of globally adaptive alleles.

0.7 When is recombination an adaptive event?

Under Scenarios one and two above, neutral recombinational events occur faster than selection, whereas under Scenario three recombinant genotypes are driven to high frequency by selection. So how much of recombination is adaptive? On the one hand, recombination across wide phylogenetic distances (by horizontal gene transfer), followed by conservation of the foreign DNA in the recipient, in itself provides compelling evidence for adaptive evolution. But the picture is not as clear for homologous recombination between closely-related strains.

Recent work by Lefebure and Stanhope helps clarify the relationship between recombination and positive selection (Lefebure & Stanhope, 2007). Recombination and positive selection were both quantified in the *Streptococcus* core genome, and it was concluded that genes under positive selection are frequently recombined, a result recently supported in a study of *Listeria* genomes (Orsi et al, 2008). Specifically, 78% of genes under positive selection in the *S. pyogenes* core genome were also inferred to be recombinant (Lefebure & Stanhope, 2007). Yet, of the genes identified as recombinant within this species, only a small fraction experienced positive selection (Figure 0.3). Therefore, although positively selected genes are frequently recombined, a substantial amount of within-species recombination shows no evidence of direct adaptive value. In other words, recombination within a species could be largely neutral. But this is not the case for recombination between species (from *S. agalactiae* to *S. pyogenes*, or vice versa). Comparison of between-species and within-species evolutionary events yields a surprising insight not highlighted in the original work: 81% of all genes recombined between species also experienced positive selection, whereas only 4% of genes recombined within species also experienced positive selection (Figure 0.3).

This striking result has several implications. First, it provides empirical confirmation of Scenario three above and Milkman's hypothesis that in order for a horizontally transferred gene to be fixed (at least across species), it must enjoy a "considerable selective advantage" (Milkman et al, 2003). Second, it furnishes evidence that short-distance recombination events are likely to be neutral and unlikely to be

under positive selection (Scenarios one and two). Third, it provides inspiration for a new class of tests for positive selection in bacterial populations: identify positively selected genes in bacterial populations as recombinant sequences transferred across population boundaries. Fourth, it suggests a pragmatic solution to the long-standing challenge of defining bacterial species. Sexual eukaryotic species undergo neutral recombination in each generation. By analogy, a group of bacteria that undergo frequent neutral recombination could also constitute a discrete species. Recombination between species is not precluded in this species definition, but would require positive selection for maintenance of the introduced recombinant sequences (Figure 0.4). While previous studies have shown that clusters of closely-related strains (or putative species, with more frequent neutral recombination within than between species) can theoretically emerge and be maintained in the absence of selection, these same studies suggest that neutral recombination alone is insufficient to explain fine-scale genetic differentiation actually observed among clusters, supporting the idea that speciation might require population structure (*e.g.* microepidemics) or positive selection (Fraser et al, 2005, Hanage et al, 2006, Fraser et al, 2007).

0.8 Detecting selection among species and higher-order groups

Over millions to billions of years of evolution, populations of bacteria have diverged to form distinct species (although there is considerable controversy over exactly when two populations can be called independent species (Gevers et al, 2005, Doolittle & Papke, 2006)). This process might occur by restricted gene flow, followed by neutral drift, or by niche partitioning and natural selection (Hanage et al, 2006, Fraser et al, 2007). Recombination between distant species is relatively rare and often have great functional consequences (usually deleterious, but potentially positively-selected, as discussed above), and potentially straightforward to detect using phylogenetic methods (e.g. http://almlab.org/AnGST). While most distant recombination events will introduce deleterious substitutions, only adaptive events will be selectively retained, resulting in a greater number of these events being observed.

Substitution patterns indicative of positive selection at long time scales (between rather than within populations) could be detected using codon-based (*dN/dS*) and other relative rates-based methods (e.g. selective signatures, M-K test; see Table 0.1). The ratio of nonsynonymous to synonymous substitution rates has been widely used in genome-wide scans for positive selection in bacteria, often providing evidence for function- or gene-specific selection (Lefebure & Stanhope, 2007, Chen et al, 2006). Yet *dN/dS* is inappropriate when comparing either very distantly-related strains (*dS* saturated with multiple substitutions), or very closely-related strains, within which *dN/dS* is inflated by segregating nonsynonymous polymorphism (Rocha et al, 2006, Kryazhimskiy & Plotkin, 2008).

Metrics that explicitly measure deviations from the expected pattern of amino acid substitution (relative to a within-species near-neutral expectation as in the M-K test, or to a between-species expectation as in selective signatures) are perhaps better suited to detecting sequence-level changes associated with changes in ecological preferences. Recently, such deviations from a protein's expected rate of evolution (based on the genome and protein family to which it belongs) were quantified as its selective signature, identifying substitutions with potential ecological relevance ((Shapiro & Alm, 2008); described in Chapter 2). This approach can yield insights into the cellular functions and pathways that contribute to niche adaptation. For example, selective signatures showed that glycolysis and phenylalanine metabolism genes evolve unusually rapidly in *Idiomarina loihiensis*, mirroring this lineage's shift in carbon source preference from sugars to amino acids.

Nearly every rate-based test (Table 0.1) has the potential to mistake recombination for positive selection (e.g. (Yang & Nielsen, 2002, Anisimova et al, 2003)), but it is possible to control for recombination by ensuring that the correct gene phylogeny is being used while testing for selection. Certain implementations of the M-K test, for example, assume that all genes in the genomes being compared diverged at the same time (Bustamante et al, 2005) – an assumption that is violated when genes have different histories of recombination. Thus, if care is not taken to control for recombination before testing for selection, these two evolutionary events – both potentially interesting and with adaptive merit – may easily be confused.

Finally, there is mounting evidence that taxonomic units broader than individual species indeed have ecological meaning, and thus show similar patterns of selection. For example, clades of bacteria in the same Family or Order tend to have similar habitat preferences (Mering et al, 2007). In comparisons of obese (enriched in Firmicutes) and lean (enriched in Bacteroidetes) gut microbiomes, habitat preference was observed at the level of Division (Ley et al, 2006, Turnbaugh et al, 2006). The genetic basis of these higher-order habitat preferences is only just beginning to be elucidated, and likely involves both genome content and sequence-level variation.

0.9 Concluding remarks and future directions

Identifying the signature of natural selection in microbial genomes can help to shed light on the hidden world of microbes. Which techniques can be used to identify positive selection depends on the rates and bounds of recombination in microbial populations. The first step in any study of natural selection in bacteria is to quantify the extent of recombination within a population before moving on to sequence-based tests. Once recombinant portions of the genome are identified, they can be tested for evidence of

positive selection using diversity-based methods. Meanwhile, the non-recombinant clonal frame can be identified (e.g. using the ClonalFrame program (Didelot & Falush, 2007) or STARRInIGHTS, described in Chapter 1), and tested for convergent evolution or excessive rates of functional substitutions (Figure 0.4). If possible, all tests should be performed genome-wide to estimate and control for demographic effects (Table 0.1) that might otherwise provide spurious evidence of positive selection.

In his "Difficulties on Theory" chapter in *On the Origin of Species*, Darwin wrote: "We are profoundly ignorant of the causes producing slight and unimportant variations [...]"(Darwin, 1859). On the sesquicentennial of its original publication, we know that random mutation and recombination are the causes of heritable fitness variations, yet we remain largely ignorant of the selective pressures that cause advantageous variations to be favored and maintained. Even within our own species, the list of uncontroversial cases of selective pressures leading to genetic adaptations is not long. Yet the list of candidate adaptive variations has grown much longer since genome-wide scans for selection became viable in humans (Sabeti et al, 2006), and we are beginning to see the same happen for microbes. Metazoans and microbes will soon be on similar footing; with a list of candidate genes in hand, the challenge will be to translate this list into a meaningful genome-wide map of selection, linking genetic variation to phenotype and ecology. With such genome-wide maps, we are optimistic that evolutionary adaptations will be revealed, even at the finest resolutions - for example, among closely-related, yet ecologically differentiated subpopulations of *Vibrio splendidus* in the coastal ocean (Hunt et al, 2008a, Thompson et al, 2005). To encourage efforts in this direction, we and others are establishing a database of nearly 100 complete genome sequences of marine *Vibrio* strains encompassing multiple ecologically distinct populations, which will be freely available to the microbial ecology and population genetics communities. Darwin was right in saying that many variations are slight (e.g. a single nucleotide mutation that subtly alters protein structure or expression), but cumulatively they leave a trail of footprints, which, given the right set of population genomic tools, will ultimately lead us to a better understanding of the microbial world.

0.10 Glossary

Positive/diversifying selection: The evolutionary force causing novel alleles conferring a fitness advantage to rise in frequency in a population. This leads to reduced genetic variation at the selected locus within the population, but increased genetic variation between populations.

Negative/purifying/stabilizing selection: The evolutionary force selecting against deleterious mutations and promoting conservation of the ancestral state.

Neutral drift: The process by which mutations with negligible effects on fitness become stochastically fixed in a population of finite size.

Selective sweep: The process of a positively-selected allele rising in frequency and ultimately becoming fixed in a population. In the absence of recombination, a single clone will sweep through the population, purging genetic diversity genome-wide. In the presence of recombination, diversity may be retained at genomic loci that are unlinked to the selected allele.

Homologous recombination: The exchange of identical or similar (homologous) stretches of DNA, either between two homologous chromosomes, or between a chromosome and a fragment of DNA taken up directly into the cell (transformation), DNA introduced by a phage (transduction), or by a conjugative plasmid (conjugation). The process results in 'allelic replacement' of a stretch of the recipient genome.

Illegitimate recombination (horizontal transfer): The acquisition of DNA that previously had no homolog in the acceptor genome. The process may be mediated by transposons, phage or other mobile genetic elements, and may result in the acquisition of entirely new genes or operons from very distant relatives.

Acceptor lineage: In a recombination or horizontal transfer event, the acceptor lineage is the recipient of a stretch of novel DNA.

Restricted gene flow: The reduction or prevention of recombination between bacterial lineages owing to physical or ecological barriers, or DNA sequence divergence.

Panmictic population: A population undergoing frequent recombination (e.g. a sexual population in which recombination occurs every generation). Qualitatively, this results in random association between loci. More quantitatively, panmixis may be defined as when a single nucleotide change is more likely to have resulted from a recombination event than a mutation event ($r/m > 1$).

Sympatric speciation: The process through which new species arise in the absence of physical barriers to gene flow between them.

Clonal population: A population that never (or extremely rarely) undergoes recombination. All loci in the genome are thus in complete linkage, meaning that a selective sweep will affect diversity in the entire genome, not just at a selected locus.

Globally adaptive mutation: In a meta-population model, a globally adaptive mutation confers a fitness advantage in all of the sub-populations making up the meta-population. If the mutation is recombined into a sub-population, it will purge genetic diversity only in the recombined portion of the genome.

Niche partitioning: The process whereby different organisms co-exist in a community rather than competing for resources. Niches can be partitioned when lineages avoid competition by using different resources, or restricting their activity to different physical spaces, seasons, times of day, etc.

Tajima's *D*: A statistic to measure deviations from allele frequencies expected in a population evolving under a neutral model. Deviations may indicate purifying selection ($D < 0$), diversifying selection or population subdivision ($D > 0$), or a recent selective sweep or population bottleneck ($D < 0$).

Long-range haplotypes: A test to detect positively-selected alleles that have risen to high frequency in a population in a short period of time, so that recombination has not had time to break down linkage to distant hitchhiking mutations. The test exploits the genome-wide distribution of allele frequencies and haplotype lengths to detect outlying haplotypes that are at unusually high frequency for their length.

McDonald-Kreitman (M-K) test: A test for selection on protein-coding nucleotide sequences that measures an unusually high between-species dN/dS, relative to a near-neutral standard of within-population dN/dS.

Selective signatures: A measure of selection that can be applied to nucleotide or protein sequences from distantly related species. Selective signatures quantify the extent to which a gene deviates from the evolutionary rate (number of substitutions per site) predicted by the gene family and genome it belongs to. Such deviations suggest gene-specific, species-specific changes in the selective pressures on a gene.

Phylogenetic incongruence: If a gene has experienced horizontal transfer, duplication, and/or loss in some lineages, this will often result in the gene's phylogeny (gene tree) having a different topology from the species' phylogeny. Phylogenetic incongruence is often used as evidence for horizontal transfer.

Convergence: The independent fixation of the same mutation in two or more independent (distantly-related) lineages, also called homoplasy, is often used as evidence for positive selection. Because it generates phylogenetic incongruence (see above), it may also be used as evidence for recombination. Recombination may be ruled out if the convergence is restricted to a single mutation, rather than a long stretch of mutations. If the convergence consists of different nucleotide-level mutations that result in convergence to the same amino acid, this also supports positive selection as a more likely explanation that recombination.

Box 1. Key challenges in bacterial population genetics

Compared to eukaryotic systems, our limited understanding of microbial population genetics can be summarized by several key challenges in studying environmental and host-associated bacterial communities:

- Limited understanding of gene flow patterns and population boundaries in bacteria. Arguably, a universally accepted species definition is lacking even for animal taxa (de Queiroz, 2005, Hey, 2006); however, the problem is exacerbated in bacteria where the rates and bounds (genetic and/or ecological) of gene flow are not known. Complicating matters further, different natural populations likely occupy a continuum of recombinational rates from clonal to panmictic.
- Lack of approaches to detect recent positive selection. A wealth of statistical tools based on allele frequencies or haplotype structure are available for detecting the signature of natural selection in sexual eukaryotes, but which if any of these tests can be adapted for use in bacteria has not been studied.
- The unknown role of the 'peripheral' genome. A large fraction of the genetic diversity within microbial lineages is contained within the 'peripheral' or 'flexible' genome -- strains that are nearly identical in nucleotide sequence at orthologous loci can differ by genomic islands containing megabases of strain-specific DNA. The extent to which this extraordinary diversity contributes to adaptive evolution is not known.

Figure legends

Figure 0.1. Population diversity following a selective sweep with varying selection and recombination rate.
(A) We simulated a selective sweep in an initially diverse population of size $N=10^7$, mutation rate $m=10^{-10}$ per bp per generation and a range of selection coefficients (σ) and recombination rates (ρ). The genome contained only two loci: the foreground locus, which could contain a beneficial allele, and a neutral background locus. After the beneficial allele has swept through the population, we computed the population's Clonality as Σp_i^2, where p_i is the frequency of the i^{th} allele at the background locus. Populations with small ρ and large σ are dominated by few genotypes, while populations with large ρ and small σ consist of many genotypes following fixation of a beneficial allele.
(B) Effect of population size on diversity following a selective sweep. $\rho=Nr$ and $\sigma=Ns$ are the 'natural' variables of the system, such that populations of different sizes with same values of ρ and σ (at fixed m) have similar structure following a selective sweep. Eight combinations of ρ and σ are compared using simulations with different population sizes. For each population size, we plot the clonality index vs. that observed for a population of size $N=10^7$. Deviations from the dashed line (y=x), are relatively small indicating that the results hold over a range of population sizes.

Figure 0.2. Illustration of distance-dependent decay of LD in the *E. coli* genome.
We gathered 1672 core orthologs present in each of 24 *E. coli* strains, as described in (Shapiro & Alm, 2008). Each unique allele at a given locus was assigned a unique allele number. We then chose pairs of loci separated by increasing distances in the *E. coli* K12 reference genome. Pairs of loci on the same operon and neighboring loci on the same strand were excluded. LD was estimated using the D_A' metric, which provides a summary measure of LD between 2 loci, each containing an arbitrary number of alleles (Kalinowski & Hedrick, 2001). When $D_A' = 1$, linkage is at its theoretical maximum. For pairs of loci separated by increasing genetic distance (kb, on a $\log_{10}$ scale), the proportion of pairs in full linkage (# of pairs with $D_A' = 1$ / total # of pairs in that distance bin) is plotted on the y-axis. Inset: Distances of 0-100 kb shown on a linear scale (red points).

Figure 0.3. Intersections of sets of recombining and positively selected genes in *Streptococcus* spp.
Overall, Lefebure and Stanhope (Lefebure & Stanhope, 2007) observed recombination in 753 genes and positive selection in 217 genes; 72 genes experienced both. The overall dataset (top) was segmented into a genus-wide core-genome (between species, bottom left) and a species-specific set of genomes (within species, bottom right). 53 genes were found to recombine between species of which 43 experienced positive selection. By contrast, 700 genes were found to be recombinant within species, but only 29 of these experienced positive selection.

Figure 0.4. Flow chart of methods to identify positively selected loci in bacteria.
Tajima's *D* and Fay and Wu's *H* tests measure unusually high or low allele frequencies within a population. They require a sample of allele sequences, preferably genome-wide to help quantify recombination and demographics, representing polymorphism within a population. The tests differ in that the *H* test also requires an outgroup species to distinguish derived and ancestral mutations, allowing it to distinguish positive and negative selection (Zeng et al, 2006). The M-K test also requires a sample of alleles from a population, and at least one outgroup species, but differs from the diversity-based tests in that it is restricted to protein-coding genes. The important assumption of the M-K test is that in the absence of selection, the dN/dS ratio should remain constant over time, and thus be the same for fixed substitutions (between outgroup and ingroup) as for segregating polymorphism (within the ingroup). When the ratio of fixed:polymorphic dN/dS exceeds 1, this provides strong evidence that positive selection has played a role in the divergence the of outgroup and ingroup (McDonald & Kreitman, 1991). AnGST is a phylogeny-based approach to detecting recombination (http://almlab.org/AnGST). It

identifies ancestral recombinations and specifies donors and acceptor lineages. See the main text and glossary for brief descriptions of other methods.

Table legend

Table 0.1. Overview of methods for identifying loci affected by positive selection

[a] Method cannot distinguish between these events, unless the test is performed genome-wide in order to account for demographic effects.

[b] Provided there is no recent recombination with the outgroup at this locus, and that the correct gene phylogeny is used.

[c] Provided that at least some synonymous and nonsynonymous substitutions have occurred, that *dS* is not saturated with multiple substitutions per site, and that time scales are not so short that *dN* is dominated by slightly-deleterious polymorphism segregregating within a population.

[d] Requires at least one outgroup species.

[e] Complete linkage of all loci on the chromosome results in selective sweeps that purge diversity genome-wide, preventing selected loci from being identified.

[f] Will require evaluating the degree of recombination between sites.

[g] Recombination might not be sufficient to disrupt the clonal frame, therefore no distance-dependent decay of LD.

[h] If gene conversion fragments are large enough so that the pattern of LD is similar to that generated by crossing-over.

Figure 0.1

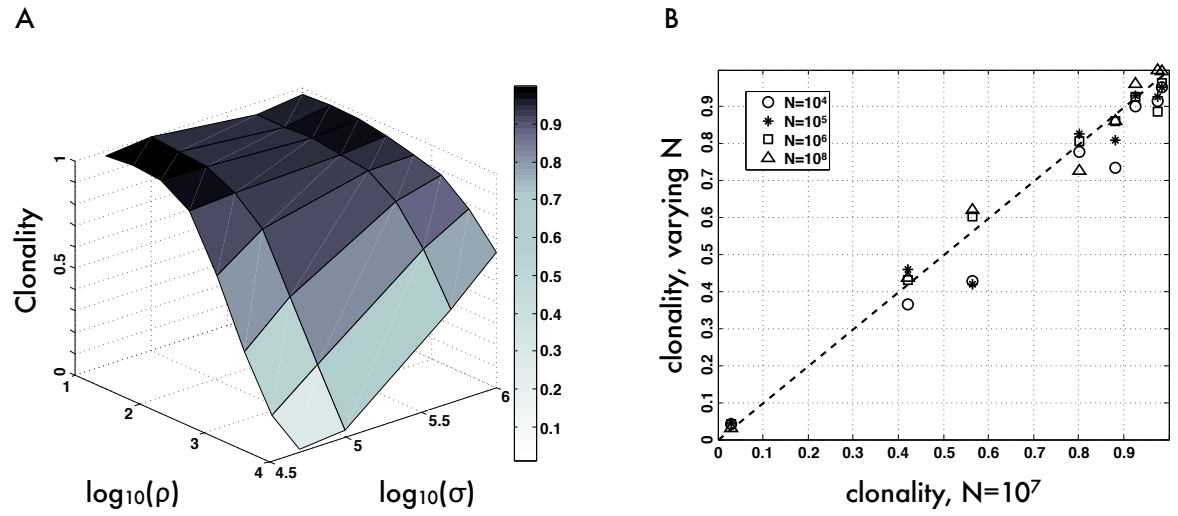


Figure 0.2

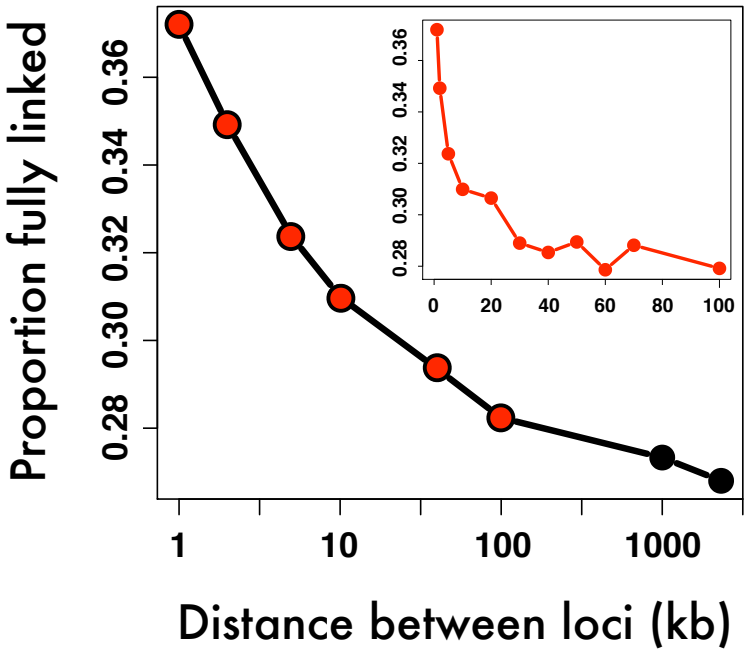


Figure 0.3

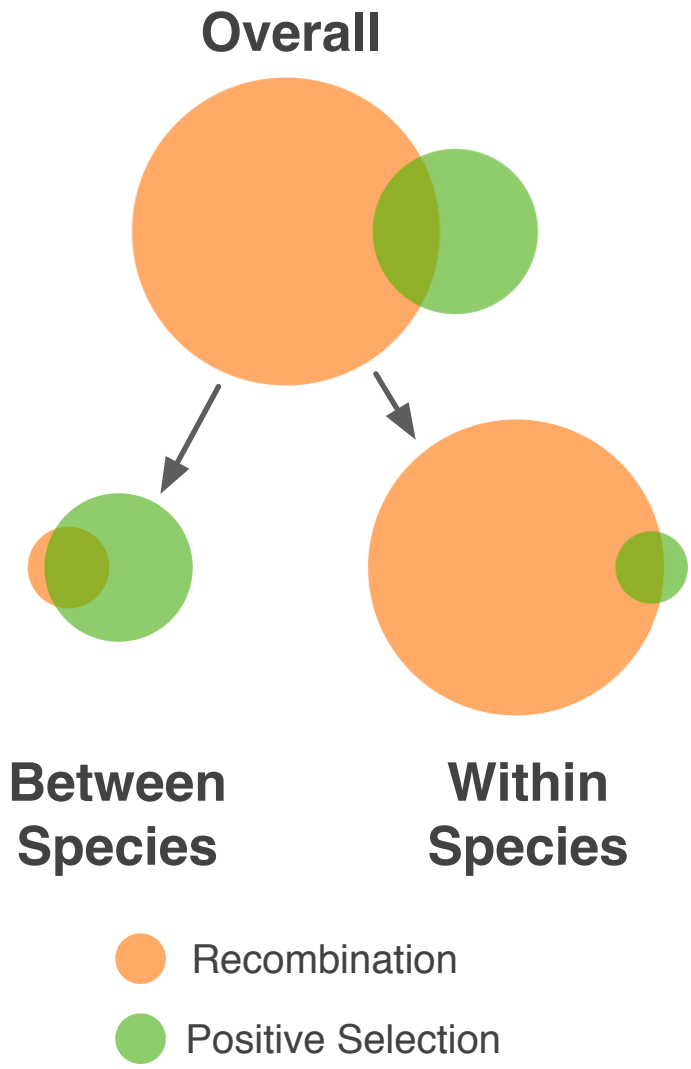

Overall
Between Species
Within Species
Recombination
Positive Selection

Figure 0.4

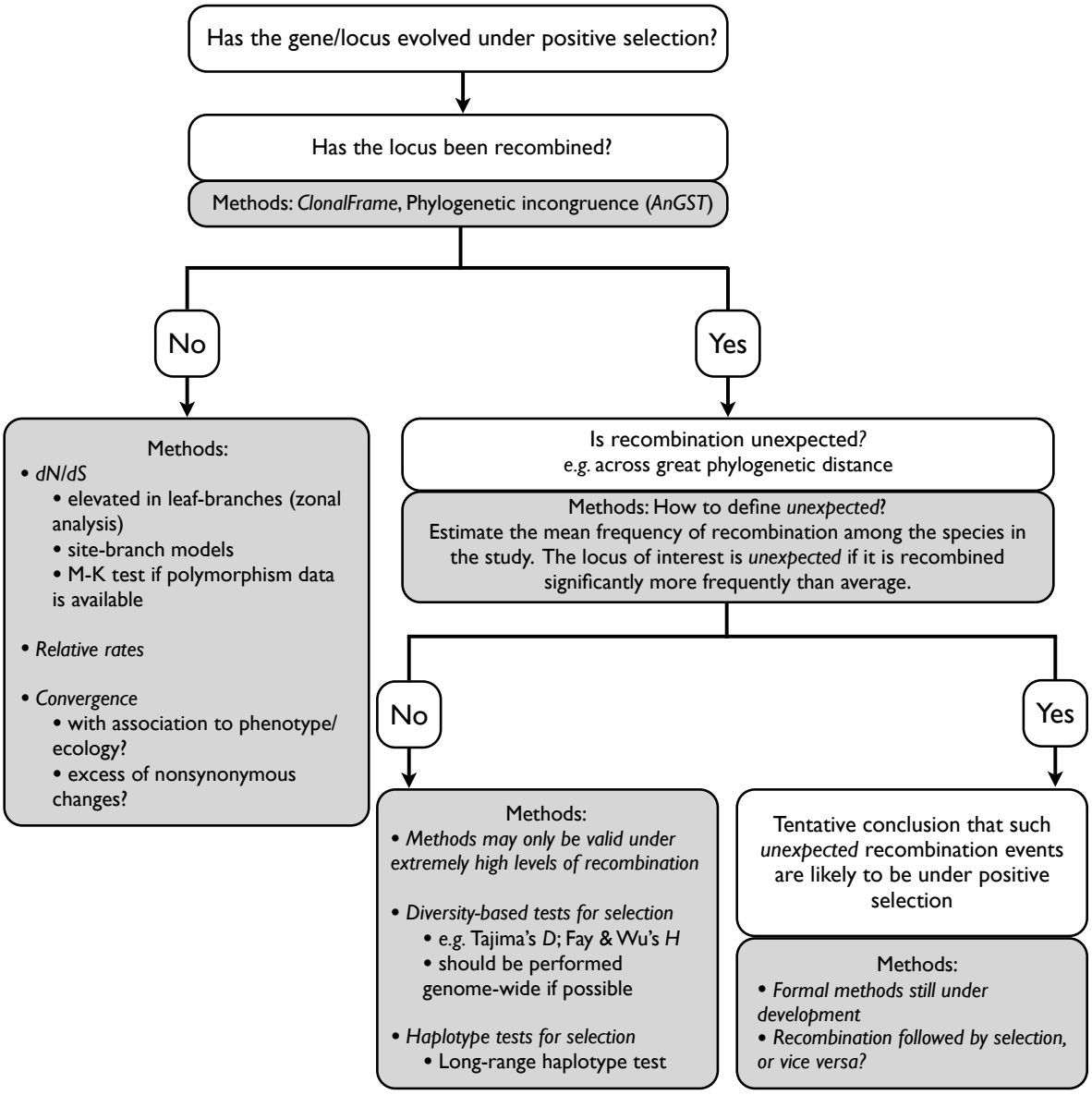

Has the gene/locus evolved under positive selection?
Has the locus been recombined?
Methods: *ClonalFrame*, Phylogenetic incongruence (*AnGST*)
No
Yes
Methods:
• *dN/dS*
• elevated in leaf-branches (zonal analysis)
• site-branch models
• M-K test if polymorphism data is available
• *Relative rates*
• *Convergence*
• with association to phenotype/ ecology?
• excess of nonsynonymous changes?
Is recombination unexpected?
e.g. across great phylogenetic distance
Methods: How to define *unexpected*?
Estimate the mean frequency of recombination among the species in the study. The locus of interest is *unexpected* if it is recombined significantly more frequently than average.
No
Yes
Methods:
• *Methods may only be valid under extremely high levels of recombination*
• *Diversity-based tests for selection*
• e.g. Tajima's *D*; Fay & Wu's *H*
• should be performed genome-wide if possible
• *Haplotype tests for selection*
• Long-range haplotype test
Tentative conclusion that such *unexpected* recombination events are likely to be under positive selection
Methods:
• *Formal methods still under development*
• *Recombination followed by selection, or vice versa?*

Table 0.1

Method	Basis	Time Range	Events detected	Effective for:			Refs
				$r < s$ (clonal)	$r \approx s$	$r > s$ (sexual)	
RATE OF FUNCTIONAL CHANGE							
Relative rates	excess in amino acid substitution rate, relative to outgroup(s)	long	positive[a], purifying, or relaxed purifying[a] selection	yes	yes	yes[b]	Shapiro & Alm 2008
dN/dS ratio	ratio of amino-acid replacement vs. silent substitution rates	intermediate[c]	positive[a] or purifying selection; demographic changes[a]	yes	yes	yes[b]	Yang & Nielsen 2002
McDonald-Kreitman (M-K test)	*dN/dS* between vs. within species	long[d]	positive[a] or purifying selection; demographic changes[a]	yes	yes	yes[b]	McDonald & Kreitman 1991
Zonal analysis	excess amino acid substitutions in the tips of a phylogeny; convergence	short	positive[a] or relaxed purifying[a] selection	yes	yes	yes[b]	Sokurenko et al. 2004
CONVERGENCE							
Convergence test	phylogenetically incongruent substitutions; often involving amino acid changes; often associated with a phenotype or environment	flexible	positive selection or recombination	yes	yes	yes	Holt et al. 2008
DIVERSITY-BASED							
Tajima's *D*	excess of low-frequency vs. intermediate frequency alleles	short	positive[a] or purifying[a] selection; demographic changes[a]	no[e]	maybe[f]	yes	Tajima 1989
Fay & Wu's *H* (and related tests)	high-frequency derived alleles[d]	short	positive selection[a], population subdivision[a]	no[e]	maybe[f]	yes	Zeng et al. 2006
Population differentiation (F_{ST})	within vs. between population heterogeneity	short	positive[a] or purifying[a] selection, population subdivision[a]	no[e]	maybe[f]	yes	Lewontin & Krakauer 1973
HAPLOTYPE-BASED							
Long-range haplotype (LRH) test	rise in frequency of a selected mutation, along with an extended haplotype of linked mutations	very short	positive selection	no[e]	maybe[g]	yes[h]	Sabeti et al. 2002, 2007

Chapter 1:

Recombination in the core and flexible genome drives ecological differentiation in sympatric ocean microbes

Shapiro BJ, Friedman J, Cordero OX, Preheim SP, Timberlake S, Polz MF and Alm EJ (2010) Recombination in the core and flexible genome drives ecological differentiation in sympatric ocean microbes. *in prep.*

Abstract

Microbes adapt to changing selective pressures in their natural environments, leading to a dynamic process of ecological specialization and speciation, even over short evolutionary time spans. Yet little is known about the microevolutionary processes leading to ecological differentiation of microbial populations in the wild. In particular, it is unclear how speciation proceeds in a sympatric setting, where there are no barriers to the homogenizing force of recombination. Speciation has rarely been observed in the wild; even more rarely in microbes, and in the near absence of genome sequence data.

We sequenced and analyzed complete genomes from 8 closely-related strains of *Vibrio splendidus*, representing two nascent populations, that appear to have recently diversified ecologically: 3 strains found on small particles, and 5 strains found attached to zooplankton in the coastal ocean. We developed a comprehensive method to assess mutation and gene flow within and between habitats, and inferred ~2000 homologous recombination breakpoints. Although gene-flow between populations in the two habitats is common, we observe a significant excess of very recent recombination within habitats. Barriers to gene flow between habitats may therefore be emerging, and are likely driven by a few highly divergent habitat-specific alleles, including genes involved in stress response (rpoS) and chitin metabolism. Adaptation through gain and loss of 'flexible' DNA is also extensive (each genome contains ~100-300 kb of strain-specific DNA), and a few recently acquired genes may provide habitat-specific adaptive value. For example, a suite of genes involved in O-antigen and mannose-sensitive hemagglutinin (MSHA) biosynthesis are absent in small-particle strains but present in zooplankton-associated strains, perhaps promoting preferential attachment to zooplankton.

Taken together, these results support a model in which ecological differentiation is initiated by the acquisition of a small number of habitat-specific alleles, gene flow between habitats is gradually restricted, and populations resembling biological species – with distinct but not disjoint gene pools – begin to emerge.

1.1 Introduction

Microbial species outnumber multicellular species by at least an order of magnitude, partly due to their continued ability to adapt to new environments and diversify both ecologically and genetically (Cohan & Koeppel, 2008). The identification of functional units – adaptive genes or species – within this diversity is a key step towards understanding microbial ecology. In the clinical realm, a key question is whether diagnosis and treatment should be based on presence or absence of particular pathogenic *species* of bacteria, or on particular virulence *genes* or *alleles*. An obvious example of the importance of genes over species identity is *Staphylococcus aureus* infection, the clinical outcome of which is heavily dependent on antibiotic resistance genes (Weigel et al, 2003). Antibiotic resistance has also been observed to evolve through an ordered series of mutations acquired over the course of a single patient's *S. aureus* infection, with no evidence for wholesale gain or loss of genes (Mwangi et al, 2007). Whether due to point mutation or gene transfer, ecological differentiation (in this case from a state of antibiotic sensitivity to resistance) is clearly possible on evolutionary time scales much more fine-grained than named species. Sub-populations may emerge into full-fledged species via the process of ecological differentiation; a fundamental question is whether sub-populations are produced by point mutations and clonal expansions – the so-called 'ecotype' model (Cohan, 2001), or by a more complex interplay of vertical (clonal) and horizontal transmission of DNA and natural selection. The different models have distinct evolutionary and epidemiological dynamics (Maiden, 2008, Fraser et al, 2009), and also differ in the genomic signatures of positive selections they leave in their wake (Shapiro et al, 2009).

Distinguishing between these models of ecological differentiation requires simultaneously measuring rates of mutation, homologous recombination, and illegitimate recombination (horizontal transfer) within carefully chosen populations. In microbial ecology, the focus has tended toward the role of illegitimate recombination in shaping the 'flexible genome', as in the example of antibiotic resistance cassettes, but also in adaptation to more multi-faceted niches (Kettler et al, 2007, Wilmes et al, 2009, Mandel et al, 2009). Mutation plays a clear role in adaptation of clonal organisms such as *Salmonella enterica* serovar Typhi (Holt et al, 2008) and also in numerous experimental evolution experiments in the laboratory (Bantinaki et al, 2007, Barrick et al, 2009). Homologous recombination within the 'core' genome (common to all strains in the population) has generally been considered as a selectively neutral, cohesive force maintaining genetic homogeneity within a population (Sheppard et al, 2008, Simmons et al, 2008). Generally underappreciated is the potential for homologous recombination to drive adaptation, as has been documented repeatedly in animal populations (Turner et al, 2005, Papa et al, 2008, Schluter & Conte, 2009, Michel et al, 2010). Here we describe a comprehensive and generalizeable pipeline for quantifying each of these evolutionary forces and their relative contributions to ecological differentiation.

The pipeline begins with two or more closely related microbial populations that appear to be specializing ecologically (*e.g.* strains of *S. aureus* with different clinical phenotypes). Whole genome sequences are obtained from a sample of strains in each population, genome sequences are aligned and divided into regions that are present in all strains (core), and those that are specific to a single strain or a subset (flexible). The flexible genome is leveraged to assess rates of illegitimate recombination, while the core genome is tested for evidence of homologous recombination either among the sampled strains, or with an unsampled relative. Population-specific genes (in the flexible genome) or alleles (in the core genome), if present, are identified in an unbiased way and tested for evidence of positive selection promoting ecological differentiation. Perhaps most importantly, the pipeline tests whether the ecological parameter under study (*e.g.* virulence, temperature sensitivity, habitat preference etc.) is a major driver of genome-wide divergence or not.

In this study, we investigate the genomewide evolutionary forces underlying a putative ecological differentiation between sympatric marine populations of *Vibrionaceae*. In an earlier study, vibrios were isolated from different compartments of the coastal water column; some strains were found to be free-living, while others were associated with particulate matter or zooplankton (Hunt et al, 2008a). Each compartment, or microhabitat, favors distinct ecological strategies: free living vibrios exploit dissolved, evenly distributed nutrients at low concentrations, particle-associated strains degrade nutrient-rich but patchily distributed detritus (Polz et al, 2006), while zooplankton-associated strains attach to and metabolize chitinous exoskeletons (Hunt et al, 2008b). In many cases, ecological populations mapped to relatively deep phylogenetic groups, corresponding to named species. More recent habitat switches were also observed, potentially catching the dynamic process of ecological differentiation – a key first step toward speciation – in action. In one such case, a zooplankton-associated population of *Vibrio splendidus* appeared to have given rise to a small particle-associated population. The populations differed by just a single nucleotide substitution (SNP) in a region of the marker gene *hsp60*, suggesting a very recent time scale for the ecological switch. Here we ask whether this recent ecological switch drives genomewide divergence, and deploy our pipeline to uncover the genomic changes (due to mutation, homologous or illegitimate recombination) that promote, or hinder, ecological differentiation.

Results

1.2 The overwhelming signal of genomewide divergence follows ecological lines

Genome sequences were obtained for three predicted small-particle-associated strains (SP) and five zooplankton-associated strains (Z). The initial hypothesis that the SP population was derived from the Z population is not supported by the genomewide data: based on 3.6Mb of aligned 'core' genome (Methods), the SP and Z strains form two distinct, monophyletic sister clades (Figure 1.1A). The consensus phylogenetic signal in the genome is dominated by SNPs which distinguish SP from Z strains (henceforth 'ecoSNPs'; *e.g.* a site where all SP strains have an 'A' and all Z strains have a 'G'), with the ancestral SNP occurring with roughly equal frequency in either the SP or Z population. This phylogenetic pattern, coupled with a modest genomewide inflation of *dN/dS* between habitats (Table 1.S5; Supplementary note 2), is consistent with significant genetic divergence in both habitats. Therefore, the small particle/zooplankton ecological axis does indeed appear to be a major correlate and potential driver of genomewide divergence.

1.3 Evidence for frequent homologous recombination

Next, we asked whether the signal of divergence along ecological lines is supported throughout the genome, consistent with a clonal expansion model, or whether divergence is driven by a few (potentially recombinant) loci rich in ecoSNPs. While clonal expansions can arise from neutral processes such as founder effects and bottlenecks, cross-species (or cross-population) recombination events are unlikely to be successful in the absence of selection (Shapiro et al, 2009). Distinguishing between these alternatives requires a genomewide analysis of two classes of homologous recombination events: (i) involving exchange of DNA among the 8 genomes studied, resulting in rearrangements of the tree topology, or (ii) exchange with more distant relative, resulting in a contiguous region of sequence with a high rate of substitutions, and possible topology changes. Existing methods consider one class of recombination event or the other (Didelot & Falush, 2007, Mau et al, 2006); we developed a model that consider both. Our model, **S**train-based **T**ree **A**nalysis and **R**ecombinant **R**egion **In**ference **I**n **G**enomes from **H**igh-**T**hroughput **S**equencing-projects, or STARRInIGHTS allows the genome to be divided into recombinant 'blocks,' each with its own local substitution frequency, and each supporting its own, maximum-parsimony phylogeny. We then use dynamic programming to solve for the optimal number and allocation of blocks in the genome, followed by Expectation-Maximization to estimate the most likely recombination rate (Methods). In mostly clonal lineages, the genome is expected to consist of a single block, with the vast majority of polymorphisms supporting a single phylogeny, and only a few phylogenetically inconsistent polymorphisms (homoplasies) scattered throughout the genome. As

expected, this is the case in a population of 19 *Salmonella enterica* serovar Typhi genomes, previously shown to evolve essentially by mutation alone (Holt et al, 2008), with negligible contribution from recombination (Table 1.1).

In contrast, and consistent with previous estimates of recombination rates in vibrios (Vos & Didelot, 2009), the most likely model to account for patterns of polymorphism in the *V. splendidus* genomes involves a relatively high rate of recombination. The probability of a recombination breakpoint was estimated to be 4.90×10^{-4} per site per branch (about one expected recombination event, involving any set of strains or nodes on the tree, every 50 kb), resulting in ~1000-2000 total recombination events in the core genome (Table 1.1). The lower bound assumes that one half of the recombinant blocks are in the 'clonal frame', consistent with a single phylogeny of vertical descent, whereas the upper bound assumes no such majority-rule phylogeny. Only 199 (~10%) of recombination breakpoints were due to a change in substitution frequency, rather than a change in tree topology, between adjacent blocks. This implies that the majority (>90%) of recombination events in the core genome involved an exchange within the SP/Z populations rather than with a more distant relative.

Given the substantial role of recombination in shaping the *V. splendidus* genomes, we wondered what fraction of the genome contributed to the observed ecological divergence. Surprisingly, of the 1995 inferred blocks, only 234 (<12%) supported an 'ecological split' (ECO split, defined as a tree dividing strains into distinct, monophyletic habitat groupings). Moreover, the remainder of the genome is not simply ecologically uninformative – 1372 blocks (69%) support trees *actively rejecting* the ECO split. This implies that most of the genome experiences 'mixing' between habitats: either the genome contains polymorphism that pre-dates the habitat split, there has been extensive gene-flow between habitats, or both. As illustration of this mixing, we considered all 10,395 possible unrooted tree topologies for the 8 strains and ranked them according to their popularity among blocks of the core genome (Figure 1.1B). Strikingly, the top-ranked tree only accounts for 110 Kbp (3% of the core genome), 55 trees are required to account for 50%, and 438 trees to account for the entire core genome (Figure 1.S4). This highlights the lack of a clear 'clonal frame' phylogeny due to rampant recombination. Moreover, the majority of trees are inconsistent with the ECO split (grey bars in Figure 1.1B), suggesting that only a few habitat-specific alleles are responsible for the genomewide signal of ecological divergence.

1.4. Habitat-specific alleles in the core genome

What are the habitat-specific alleles, and where are they located in the genome? Most of the 2971 ecoSNPs are clustered in 12 regions of the core genome (Table 1.S1), with just 3 regions accounting for >50% of ecoSNPs. (Table 1.2; Figure 1.1C)

The first ecologically differentiated region, located on the small chromosome, contains 2 genes: a predicted RNA polymerase stress-response sigma factor, RpoS, and an RTX (repeat in toxin) family protein. The second region consists of an operon involved in chitin metabolism. The third region flanks the highly variable O-antigen locus (part of the flexible genome, not the core) and contains genes coding for phosphoglyceromutase (EC 5.4.2.1), methylated-DNA-protein-cysteine methyltransferase (EC 2.1.1.63), NormM (a putative Na^+-driven multidrug efflux pump), and the chaperone protein GroEL (*hsp60*) initially used to define SP and Z groups (Hunt et al a). Further details on the genes in these regions can be found in Supplementary note 1.

While it is unclear which, if any, of these ECO regions provide habitat-specific adaptive value, they represent promising targets for future investigation. As initial validation of the habitat-specificity of these regions, we sequenced portions of them in an additional 6 SP strains and 3 Z strains. As a control, we also sequenced portions of 3 probable 'housekeeping' loci (unlikely to be involved in adaptation) in the same strains. As expected, the housekeeping loci were not diverged along ecological lines (Figure 1.1C; Figure 1.S2). In contrast, SP and Z strains formed separate monophyletic groups in ecological regions 1 and 3. Intriguingly, the SP strains were polyphyletic at the chitin-related locus (region 2), suggesting that this locus may be fitness-neutral in the SP habitat. Yet the Z strains are paraphyletic (neither monophyletic nor polyphyletic, but still forming a coherent group) at the chitin locus, suggesting that a zooplankton-associated lifestyle may impose certain constraints on the type of chitin metabolism alleles required for growth. Alternatively, a small-particle-adapted allele of the chitin locus may be in the process of sweeping through the SP population, but the sweep is still incomplete. Overall, these results support, but do not unequivocally confirm, the idea that these 3 regions contribute to habitat adaptation.

1.5. Rapid turnover and habitat-specific genes in the flexible genome

Even over the short evolutionary time span separating the SP and Z strains, the flexible genome is incredibly dynamic. Consistent with previous observations (Thompson et al, 2005), each of the 8 *V. splendidus* genomes contains ~130-360 kb, or 300-600 blocks, of strain-unique DNA (Figure 1.S13). Despite this rapid turnover, there are substantial amounts of habitat-specific DNA unique to either the SP or Z habitat (Table 1.S2). Of particular interest, a set of 5 loci involved in O-antigen and mannose-sensitive hemaglutinin (MSHA) biosynthesis are absent in small-particle strains but present in

zooplankton-associated strains, perhaps promoting preferential attachment to zooplankton (Meibom et al, 2004). We confirmed this result by PCR of the 5 loci in an additional 3 Z strains and 4 SP strains (Table 1.S3).

1.6. Quantifying gene flow within and between habitats

The data presented thus far suggest that, with the exception of a few habitat-specific alleles, the SP and Z populations share a history of extensive gene flow between habitats. Yet a shared *history* of gene flow in the *distant* past does not preclude restricted gene flow in the more *recent* past and present, as might be expected in newly emerging biological species.

To assess the extent of only the most recent gene flow within and between habitats, we considered only recombination events confidently predicted as recent: events that separate the most recently diverged pair of strains (1F111, 1F273) in our analysis. This pair of genomes, henceforth 'sister strains', have more shared polymorphism than any other group (7708 SNPs), and represent the two shortest leaf branches in the tree (343 and 217 leaf-specific 'singleton' SNPs in 1F111 and 1F273, respectively, compared to an average of 6576 singletons in the other 6 strains). Recombination events that split up the sister strains are rare; only 83 blocks in the genome reject their grouping together. We used these rare blocks, involving recent recombination events, to ask whether within-habitat recombination events (those involving one of the sister strains and the remaining member of the SP group) were more common than between-habitat events. To ensure we were only considering recent recombination, we restricted our analysis to events involving only pairs of leafs (extant, thus recent, nodes on the tree): one of the sister strains, and one of the other 6 leafs on the tree (5 Z strains and 1 SP strain). After excluding blocks whose maximum-likelihood trees included probable recombination events with distant relatives (leaf branches accounting for more than 50% of the total tree length), which could result in long-branch attraction and spurious disruption of the sister strain pairing, and merging adjacent blocks attributable to the same recombination event, we were left with 27 high-confidence events. Despite the small number of events, there was a marginally significant excess of recombination within habitats (Figure 1.2A; observed 10 within-habitat events, expected 4.5; Binomial test $p = 0.009$). Barriers to gene flow between habitats may thus be in the process of emerging, despite a history of ancestral gene flow between habitats. Yet the gene pools are far from closed: 17 out of 27 recent recombination events involved exchanges between habitats.

Barriers to gene flow between habitats is a feature of the flexible genome as well as the core. There is significantly more flexible DNA shared within habitats than between them, and SP and Z habitats consistently form separate groups (100% bootstrap support) when strains were clustered according to

shared blocks of flexible DNA (Figure 1.2B). Considering the rapid turnover of flexible DNA (Figure 1.S13), the similarity of gene pools within habitats is not to be taken for granted. For example, the sister strains differ by only 560 substitutions (or ~8.1e-5 substitutions/site, based on singleton SNPs, and excluding likely recombinant blocks that break up the sister strains), but each contains ~300 kb of strain-specific DNA not present in the other sister (321 kb (distributed over 440 blocks) in 1F111; 287 kb (over 590 blocks) in 1F273). This translates to an average gain or loss of ~1 kb (or 1-2 blocks) of DNA for every single nucleotide change in the core genome – a surprisingly rapid turnover. Thus, similar flexible gene pools are maintained within habitats despite rapid turnover and limited vertical inheritance. Distinct habitat-specific gene pools may be maintained in part by selection (*e.g.* MSHA genes) and in part by higher encounter rates and exchange of neutral/selfish elements (*e.g.* phage and transposons) within habitats (Table 1.S2).

1.7 Discussion

We have presented the first study of a recent habitat switch exploiting ecological data and whole genome sequencing of very closely related strains. An ecological switch is an important first step for speciation, a process where an ancestral population splits into new populations with either or both of the following characteristics: (1) distinct ecological niches, thus allowing coexistence of both daughter species ('ecotypes') without competitive exclusion, (2) restricted gene flow between populations, either due to physical barriers (allopatry), or not (sympatry). Understanding the relative contributions of these mechanisms to the speciation process requires surveying diversity across the entire genome of nascent species, but the discourse on microbial speciation has to date proceeded largely in the absence of genome sequence data.

Our results support the idea that the initial stages of ecological differentiation may be driven by a few genomic islands of habitat-specific genes (in the flexible genome) or alleles (in the core genome), despite the prevailing homogenizing force of recombination acting on most of the genome. We propose a sympatric-like ancestral population, in which extensive gene flow results in many recombination blocks supporting many different phylogenies, mostly inconsistent with divergence along ecological lines (Figure 1.3A). The acquisition of habitat-specific alleles, either by mutation or gene flow from another population initiates (and later may maintain) the divergence between strains specialized to one habitat or the other (Figure 1.3B). Finally, gene flow becomes more common within than between habitats (Figure 1.3C), possibly due to reduced opportunity for encounters between strains in different habitats (partial allopatry). While we have apparently caught a snapshot of speciation in action, the fate of the nascent species is completely unknown. Speciation rates may be high, but extinction rates may be equally high,

especially when habitats (*e.g.* particular types of zooplankton or particles in the marine water column) are transient or frequently disturbed (Polz et al, 2006).

The chitin operon presents an intriguing candidate for ecological adaptation since zooplankton exoskeletons are composed largely of chitin, while small particles might contain chitin at lower concentration, in more degraded forms, or not at all (Hunt et al, 2008b, Velde & Kiekens, 2004). Strains adapted to these different habitats may have experienced selective pressures to fine-tune their chitin utilization strategies in different ways. Although the chitin operon contains a few transport proteins, most of the habitat-specific divergence is confined to 4 enzymes further downstream in the pathway of chitin uptake and breakdown (Hunt et al, 2008b). These enzymes contain a relatively large number of habitat-specific amino acid changes, although evidence for positive selection on these changes is inconclusive (Table 1.2; Supplementary Note 1). Nevertheless, the extreme levels of habitat-specific divergence in the chitin operon, in contrast with most loci being freely recombined across habitats, is inconsistent with purely neutral processes. Divergence in the chitin operon also makes sense ecologically in the context of the flexible genome: MSHA genes, present in Z but not SP genomes, promote attachment to chitinous particles such as zooplankton (Meibom et al, 2004). Zooplankton-associated strains may therefore by adapted to attach to and specifically metabolize zooplankton-derived chitin, while small particle-associated strains may rely less on chitin, or prefer more degraded forms.

High levels of synonymous divergence in the ecologically differentiated regions (Table 1.2) suggest that they were likely acquired via gene flow from a relative outside the SP/Z populations. Many ECO blocks, including those in regions 2 and 3, appear to have been acquired early, perhaps helping initiate the divergence process (the median divergence time in ECO blocks is 6.98e5 generations since the ecological split, not much less than the estimated number of generations since the most recent common ancestor of all strains, 7.15e5 generations (Table 1.1). However, certain regions, including the RpoS/RTX locus, have very little within-habitat polymorphism (Table 1.2; Figure 1.S10), indicating their more recent acquisition.

The ability of microbes to acquire novel, potentially niche-specifying genes by homologous or illegitimate recombination may explain their ability to speciate rapidly and invade new niches (Cohan & Koeppel, 2008, Doolittle & Papke, 2006). Over evolutionary time scales much more ancient than those investigated here, genes that cross species boundaries by horizontal gene transfer tend to be enriched in functions that promote rapid exploitation of new niches (*e.g.* cell-surface and pathogenicity-related genes), while 'housekeeping' genes are transferred only rarely (Nakamura et al, 2004, Pal et al, 2005).

Although we did not find a strong functional bias in habitat-specific alleles (those in genomic blocks supporting the ecological split), this is probably due to the relatively recent ecological differentiation, and that divergence is mostly due to just a few loci (Figure 1.S3).

In sexual organisms, the feasibility of sympatric speciation hinges on the number of adaptive loci required to distinguish between nascent species: when only a few loci (<10) are sufficient for niche adaptation, distinct species can be maintained even with high recombination rates (Kondrashov & Mina, 1986, Kondrashov, 1986). As more loci are required to attain an optimal niche-specific phenotype, speciation is impeded because it becomes more likely that recombination will degrade this optimal allele combination. The necessity of few adaptive loci may also be a feature of microbial sympatric speciation, but for a different reason: even with relatively high recombination rates, the waiting time for all the optimal alleles to become fixed in a single genome may be very long (Shapiro & Alm, in prep.). Our observation that only 3 loci account for over 50% of the divergence between SP and Z habitats (Table 1.2) is consistent with the prediction that sympatric speciation is only possible when just a few adaptive loci are involved, but further theoretical work is needed to fully understand why.

This study has provided a first glimpse of the early stages of ecological differentiation at the genomic scale. Even in highly panmictic, sympatric bacteria, ecological differentiation appears to be driven by a few habitat-specific loci, potentially leading to restricted gene flow between habitats and the emergence of distinct populations resembling biological species. The generality of this process, the frequency of speciation with respect to extinction, and the exact number and function of the adaptive loci required for speciation all remain to be verified in other *Vibrio* clades and across the microbial tree of life. We have proposed a general pipeline for population genomic analysis, amenable to identifying the mechanisms of ecological adaptation across a variety of microbial taxa.

Methods

1.8. Assembly, alignment and definition of core and flexible genome

Genomic DNA was extracted from isolated *Vibrio* strains from our strain collection and sequenced at an average coverage of 30X per bp using the Illumina platform. Genome sequences were assembled *de novo* using Velvet (Schmidt et al, 2009). We used Mauve (Darling et al, 2004) to generate alignments (locally colinear blocks; LCBs) of Velvet contigs. Default parameters were chosen as they produced longer LCBs compared to other parameter settings. The core genome consisted of LCBs containing all 8 SP/Z strains, as well as all 3 outgroup strains (12B01, 12F01 and 13B01), while the flexible genome included blocks of minimum length 500 bp present in only a subset of genomes. Gaps in the core genome (any site with <5X coverage in any genome) were excluded, and the alignment was mapped to the *Vibrio splendidus* 12B01 reference genome (GenBank AAMR00000000), resulting in 14 'core' contigs of total length 3,583,079 bp.

1.9. Inference of mutation rates and recombination breakpoints using STARRInIGHTS

STARRInIGHTS combines aspects of the two major classes of methods to detect homologous recombination events in microbial genomes: 'substitution distribution methods' (*e.g.* ClonalFrame (Didelot & Falush, 2007)) and 'phylogenetic methods' (*e.g.* (Mau et al, 2006)). Like substitution distribution methods, STARRInIGHTS allows the mutation rate to vary along the genome, allowing for the detection of recombination events that import a large number of substitutions simultaneously into a stretch of the genome. Like phylogenetic methods, STARRInIGHTS also explicitly models the tree topology for each stretch of the genome, allowing for detection of recombination events that change the tree topology without necessarily importing a large number of new substitutions.

The input to STARRInIGHTS is an aligned contig (or set of contigs) of genomic sequences from the strains of interest, from which we wish to infer recombination breakpoints and relative rates of mutation and recombination. Our goal was to infer recombination events using phylogenetically informative single nucleotide polymorphisms (SNPs) in the aligned contigs – in this case, the 'core' genome described above. We propose that the core genome, consisting of G contigs each of length L_G bp, can be divided into B+1 blocks divided by B recombination breakpoints. Due to recombination between blocks, each block may have its own phylogeny and substitution rates (branch lengths). We assume each block has evolved according to its maximum-likelihood (ML) phylogeny. ML trees were inferred using phyML v. 2.4.5 (Guindon & Gascuel 2003) with a JC69 substitution model, a BIONJ starting tree, and two gamma-distributed evolutionary rate categories. The gamma distribution shape parameter was set to 0.031, the median value estimated by phyML in 5 kb windows along the core genome (the choice of window size

was ranged from 500 bp to 20 kb without affecting the estimate). An example of the STARRInIGHTS procedure is shown in Figure 1.S5. Due to the high sequencing coverage, we do not expect a considerable amount of sequencing error. Any remaining errors are not expected to introduce spurious breakpoints but rather to be accounted for by an increased local mutation rate. The only way that sequence errors could cause spurious breakpoints is if they introduced spatially clustered groups of polymorphisms all supporting a single phylogeny, an unlikely scenario to occur by chance.

To find the optimal number of breakpoints (B) and their locations in the genome, we define a cost function C, where both mutation events (on an ML tree within a block) and recombination breakpoints between blocks contribute to the cost incurred by a stretch of DNA from base i to base j ($i \leq j$).

$$\begin{aligned} C(i,j) &= c_b \cdot b_{ij} \\ &+ c_{nb} \cdot (l_{ij} - b_{ij}) \\ &+ c_{Tree(i,j)} \end{aligned} \qquad \text{Equation 1.}$$

where l_{ij} is the length, in base pairs, from i to j, c_b is the per-site cost of adding a breakpoint, c_{nb} is the cost for not adding a breakpoint, b_{ij} is the number of breakpoints between i and j, and $c_{Tree(i,j)}$ is the cost of the ML tree topology and branch lengths (mutation events) estimated for the alignment between i and j. The costs are in fact negative log probabilities:

$$\begin{aligned} c_b &= -\log P(b) \\ c_{nb} &= -\log(1 - P(b)) \\ c_{Tree(i,j)} &= -\log P(\tau, \nu, \theta \mid A_{i,j}) \end{aligned} \qquad \text{Equations 2a-d.}$$

where $P(b)$ is the probability of a breakpoint and $P(\tau, \nu, \theta \mid A_{i,j})$ is the probability, estimated by phyML, of the tree topology τ, branch lengths ν and substitution model θ given the alignment A from i to j. We then minimize C over the whole genome using the dynamic programming recursion:

$$M_j = \min_{\{i,j;\ 1 \leq i \leq j\}} \left(M_{i-1} + C(i,j) \right) \qquad \text{Equation 3.}$$

where M_j is the minimum cost for the first j bp, and setting $M_0 = 0$.

The number of breakpoints (*B*) and their locations will of course depend on the value of *P(b)*. The probability of a breakpoint, *P(b)*, is estimated from the data using expectation maximization (E-M) (Durbin et al, 1998). The E-M steps are as follows:

1. Initialize *P(b)* with a value between 0 and 0.5. (In practice, try 10 different values and check for convergence).
2. Using the current value of *P(b)*, solve for the optimal number and location of breakpoints using dynamic programming.
3. Compute the log likelihood of observing all SNPs in all *G* contigs of the genome:

$$\log L = \sum_{g=1}^{G} \log P(b) \cdot B_g + \log(1 - P(b)) \cdot (L_g - B_g) + \sum_{k=1}^{B_g+1} \log P(\tau_k, \nu_k, \theta \mid A_k) \qquad \text{Equation 4.}$$

where L_g is the length in bp of contig *g*, B_g is the number of inferred breakpoints and B_g+1 the number of blocks in contig *g*, and τ_k, ν_k and A_k are respectively the ML tree topology, branch lengths and alignment within block *k*.

4. Update the value of *P(b)* such that:

$$P(b) = \sum_{g=1}^{G} \frac{B_g}{L_g} \qquad \text{Equation 5.}$$

5. Iterate through steps 2-4 using the updated value of *P(b)* and continuing until convergence: $\log L_t - \log L_{t-1} \approx 0$, where *t* is the number of iterations.

1.10. Benchmarking on simulated contig sequences and correction for model complexity.

We tested the sensitivity and specificity of breakpoint detection by applying STARRInIGHTS to simulated contig sequences with predetermined recombination events. Simulated contig sequences of length 2500 bp were generated using seq-gen (Rambaut & Grassly, 1997), assuming the HKY85 substitution model, a transition/transversion ratio of 2, equal base frequencies, and a tree of 8 genomes. A

contig consisted of 5 blocks of 500 bp, each generated by an independent simulation run. In all cases, blocks 1, 3 and 5 constituted the 'clonal frame', all sharing the same tree topology and branch lengths (sampled at random from the distribution of trees observed across subsequences of the core genome). We investigated two different scenarios: (A) true recombination events: blocks 2 and 4 simulated with a different tree topology than 1, 3 and 5 (also sampled at random from the observed distribution of trees); (B) no event: blocks 1, 2, 3, 4 and 5 all simulated along the same tree. The results of 100 simulations for each scenario are shown in Figure 1.S6A.

As can be seen from this figure, the ML method yields unacceptably high levels of false positive breakpoints. In Scenario B, where no breakpoints should be inferred, the 100 simulated contigs collectively contained 593 breakpoints, which translates to a false-positive probability of 0.0024 per site (593 breakpoints / (2500 sites * 100 simulations). When spurious breakpoints in Scenario A (any breakpoints in excess of the four expected) are also counted, the false-positive probability increases to 0.0032 per site. For a core genome of 3.5 Mbp, this would lead to an expected 8,400-11,200 false-positive breakpoints – clearly an unreasonably low specificity.

One reason for spurious breakpoint calls might be lack of correction for model complexity in STARRInIGHTS. Every time a new breakpoint is included, a new tree topology and branch lengths are added as additional parameters to the model. Unless this is corrected for, many false breakpoints might be added in order to increase the likelihood of the model. For example, consider a hypothetical subsequence of 100 bp containing 10 phylogenetically-informative SNPs. For simplicity, assume that all 10 SNPs support the exact same tree topology, partitioning the strains into two groups. If by chance 5 of the SNPs fell within the first 40 bp and the other 5 SNPs in the last 60 bp, the likeliest model might result in two blocks, each supporting the same tree topology, but with a longer branch length in the 40 bp block (5 SNPs / 40 bp $\approx$ 0.125 subs/site) than the 60 bp block (4 SNPs / 60 bp $\approx$ 0.067 subs/site). To quantify the contribution of this effect, we simulated sequences ranging in length from l = 10 bp to l = 211 kb (in binned increments each spanning ~10% of the observed core genome subsequences), and ranging in mutation rate from $\lambda = 0.001$ to $\lambda = 1$, where λ is the number of SNPs per site. For each combination of λ and l we simulated 100 sequences using seq-gen with a tree chosen at random from the distribution of trees observed across subsequences of the core genome, and estimated the likelihood of a model with no breakpoints (*L0*) and a model with exactly one breakpoint (*L1*), placed optimally in the sequence to maximize the likelihood. Note that sequences were simulated using a single tree, so breakpoints introduced in the *L1* model are necessarily due to increased model complexity rather than actual recombination events. The maximum values of the log (*L1*/*L0*) ratio observed in 100 simulated contigs

for each combination of l and λ are shown in Figure 1.S7. This distribution was smoothed and used as an empirical correction for model complexity. STARRInIGHTS was modified to include an appropriate correction factor for the values of l and λ in the subsequence being considered, and we again benchmarked on contigs simulated under Scenarios A and B. The correction for model complexity was accomplished by adding a penalty, $pen(i,j) = \log (L1/L0)_{\lambda^*,l^*}$, to the cost function $C(i,j)$ from Equation 1, where $\log (L1/L0)_{\lambda^*,l^*}$ is the observed maximum log ratio from 100 simulations with a given $\lambda = \lambda^*$ and $l = l^*$. The designations λ^* and l^* designate discrete, percentile-incremented parameter values used in the simulations that most closely match the observed $\lambda(k,j)$ and $l(k,j)$, where k = traceback(i-1) and the subsequence (k,j) consists of two flanking breakpoints at k and j, with a third breakpoint in between at i. By adding the appropriate penalty, we are correcting for the probability that the breakpoint at i due to model overfitting (*e.g.* due to the effect described in the hypothetical subsequence of 100 bp with unevenly distributed SNPs). Running STARRInIGHTS with the correction resulted in a noticeable improvement in specificity without a substantial decrease in sensitivity to identify true breakpoints (Figure 1.S6B). The probability of a false-positive breakpoint improved from 0.0032 per site without the correction for model complexity to 0.00042 per site with the correction. Only considering Scenario B, where no true breakpoints are present, the false-positive probability decreased from 0.0024 to 0.000024 per site. Using a strict measure of sensitivity, only counting breakpoints in simulations that yielded ≤ 4 breakpoints in Scenario A, the sensitivity went from 6% (in the uncorrected model; Figure 1.S6A) to 16.5% of correctly identified breakpoints in the corrected model. Using a less strict measure of sensitivity (including simulations with >4 breakpoints, indicating some false positives), the uncorrected model had a sensitivity of 95% and the corrected model of 88.5%.

Finally, we identified breakpoint probabilities for which specificity was near perfect, either under a strict definition of specificity (zero false-positive breakpoints in 100 Scenario A simulations (no instances of >4 breakpoints), and zero false-positives in 100 Scenario B simulations), or a more relaxed definition (zero breakpoints in Scenario B). Under the strict specificity requirement, we found log $P(b)$ = -97.20, and under the relaxed requirement log $P(b)$ = -10.35 as the maximum log $P(b)$ values that achieved the desired specificity levels. We proceeded to apply the model complexity-corrected STARRInIGHTS algorithm to the *Vibrio* core genome alignment. Estimating the breakpoint probability by E-M yielded log $P(b)$ = -6.79 and 4206 inferred breakpoints (compared to log $P(b)$ = -6.38 and 6463 breakpoints in the uncorrected ML model). Setting log $P(b)$ = -10.35 or log $P(b)$ = -97.20 yielded 3227 or 663 inferred breakpoints, respectively.

1.11. Parsimony approximation

A variant of STARRInIGHTS using parsimony trees, which are less computationally intensive to infer than ML trees, was also evaluated. The parsimony-based method requires the assumption of equal branch lengths throughout the tree, and the assumption that mutation events occur only once per site per branch. This allows us to pool all branches together and treat mutation events as a Poisson process, with homoplasies (convergent mutation to the same base in two or more branches) being more likely as the Poisson rate parameter (a proxy of the mutation rate) increases. In the transition from one recombinant block to the next, clusters of homoplasies inconsistent with the first block's tree will be encountered, and if these homoplasies are sufficiently dense, a model with a breakpoint separating them into their own parsimony tree becomes increasingly likely. The model is seriously flawed when the assumptions are violated; take for example a tree of 8 genomes with a single long internal branch, and all other branch lengths equal to zero. The mutation rate (Poisson rate parameter) will be estimated as non-zero based on the single long branch, yielding a non-zero probability of observing a homoplasy. Yet homoplasies will *never* be observed in such a tree because it effectively has only two branches, a scenario in which homoplasy is not defined. It is unclear how the parsimony model will perform when such trees are encountered. It seems likely that the model will be overly conservative in calling breakpoints because when stretches of DNA with no homoplasies and a non-zero probability of homoplasy are encountered, these stretches will be extended, without calling breakpoints, until regions that do contain breakpoints are included. In practice, the parsimony model does indeed appear to be conservative, calling zero false-positive breakpoints in the Scenario B simulation (Figure 1.S6C), but with a reduced sensitivity compared to the ML model.

Applied to the *Vibrio* core genome, STARRInIGHTS using parsimony infers 1981 breakpoints. Of these breakpoints, over half (1033) are also identified exactly in the ML+correct model (described above, yielding a total of 4206 breakpoints). When the criterion for breakpoint overlap is relaxed to a 100 bp window, 1684 of the parsimony breakpoints are found by ML+correct, and all parsimony breakpoints are found within a 1 kb window. A full analysis of the *Vibrio* genome using the ML+correct model is underway, but the data presented here use the parsimony approximation. Because of the good overlap of the breakpoints called by both methods, we expect most of the features of the analysis to be preserved across methods.

The parsimony method is conceptually similar to the ML method, but we assume a block has evolved according to its maximum-parsimony phylogeny (minimizing the number of inconsistent/homoplasic

substitutions in the tree). Parsimony trees were inferred using the the DNACOMP program in the phylip package (Felsenstein, 1993). The cost function described in Equation 1 is replaced with:

$$\begin{aligned} C(i,j) = & c_b \cdot b_{ij} \\ & + c_{nb} \cdot (l_{ij} - b_{ij}) \\ & + c_h \cdot h_{ij} \\ & + c_{nh} \cdot (l_{ij} - h_{ij}) \end{aligned}$$

Equation 6.

where l_{ij} is the length, in base pairs, from i to j, c_b is the per-site cost of adding a breakpoint, c_{nb} is the cost for not adding a breakpoint, c_h is the cost of a homoplasic mutation, c_{nh} is the cost of no homoplasy at a site, and b_{ij} and h_{ij} are respectively the numbers of breakpoints and homoplasies between i and j. The costs are in fact negative log probabilities of breakpoints (b), homoplasic mutations (h), or the absence of these events:

$$\begin{aligned} c_b &= -\log P(b) \\ c_{nb} &= -\log(1 - P(b)) \\ c_h &= -\log P(h) \\ c_{nh} &= -\log(1 - P(h)) \end{aligned}$$

Equations 7a-d.

We then minimize C over the whole genome using the dynamic programming recursion described in Equation 3.

The number of breakpoints (B) and their locations will depend on the values of both *P(b)* and *P(h)*. The probability of a breakpoint, *P(b)*, is estimated from the data using expectation maximization (E-M) as described above, but replacing Equation 4 with:

$$\begin{aligned} \log L = & \sum_{g=1}^{G} \log P(b) \cdot B_g \\ & + \log(1 - P(b)) \cdot (L_g - B_g) \\ & + \sum_{k=1}^{B_g+1} \log P(h) \cdot h_k + \log(1 - P(h)) \cdot (l_k - h_k) \end{aligned}$$

Equation 8.

where L_g is the length in bp of contig g, B_g is the number of inferred breakpoints and B_g+1 the number of blocks in contig g, h_k is the number of homoplasies and l_k is the length of block k.

Meanwhile, the probability of homoplasy, *P(h)* can be estimated based on the observed mutation rate in a stretch of DNA from base *i* to base *j*. Let us begin by defining λ_x the per-bp rate of mutation from an ancestral base *y* to a new base *x* ($x \neq y$) on a maximum-parsimony tree, assuming that branch lengths are equal and non-zero, and that all types of mutations are equally likely:

$$\lambda_x = \frac{1}{3} \cdot \frac{s}{l_{ij}} \qquad \text{Equation 9.}$$

where *s* is the number of single-base substitution events observed in the parsimony tree for the stretch of DNA (*i*,*j*) with length $l_{ij} = j\text{-}i\text{+}1$. The factor of 1/3 is included because there are only 3 available bases *x* that differ from the ancestral base *y*.

Mutations at a single site can be modeled as a Poisson process $P(n; \lambda_x)$, where λ_x is mutation rate to base *x* and *n* is the number of mutation events to base *x* at the site. Homoplasies are defined as $n \geq 2$, meaning that two or more branches have undergone convergent mutation to base *x* at the same site. Repeated mutations at the same site in the same branch are assumed not to occur. Therefore the probability of homoplasy *x* in the region (*i*,*j*) is:

$$\begin{aligned} P(h_x \mid s, l_{ij}) &= Poisson(n \geq 2 \mid \lambda_x) \\ &= 1 - e^{-\lambda_x} - \lambda_x e^{-\lambda_x} \end{aligned} \qquad \text{Equation 10.}$$

Given that:

$$P(h \mid s, l_{ij}) = 1 - \left[1 - P\left(h_x \mid s, l_{ij}\right)\right]^3 \qquad \text{Equation 11.}$$

and substituting in Equation 10, we have

$$P(h \mid s, l_{ij}) = 1 - \left[e^{-\lambda_x} + \lambda_x e^{-\lambda_x}\right]^3 \qquad \text{Equation 12.}$$

We can now use $P(h|s,l_{ij})$ as our estimate of *P(h)* in Equations 7 and 8 above. This effectively allows a local estimate of the mutation rate in any stretch (*i*,*j*) or block of the genome. In regions with higher mutation rates, homoplasies will be relatively likely due to repeated mutation events at a site. Conversely,

in regions with low mutation rates, homoplasies are less likely to occur by mutation, and unparsimonious SNPs are more likely to be explained by invoking a recombination breakpoint.

To account for the possibility that some sites might be *truly* invariant (not just *observed* as invariant, but unchangeable for biological reasons), and thus must be considered separately from the Poission mutation model, we added the parameter a to our model, where a = the fraction of truly invariant sites in the genome, and modified Equation 12 accordingly:

$$P(h \mid s, l_{ij}) = (1-a) \cdot \left(1 - \left[-e^{-\lambda_x} - \lambda_x e^{-\lambda_x}\right]^3\right) \qquad \text{Equation 13.}$$

We tried a range of values of a between 0 and 1, and conducted the dynamic-programming/E-M procedure with each, eventually choosing the a yielding the maximum likelihood. We repeated this procedure, narrowing in on a smaller and smaller range of a until the optimum was found.

1.12. Pre-filtering for regions of phylogenetic incongruence

The cost functions and dynamic programming described above rely on ML trees for each possible subsequence (i,j) of the genome. This requires precomputing a large number ($\sim L^2$) of trees. Specifically, $N(T)$, the number of ML trees to be inferred is:

$$N(T) = \sum_{g=1}^{G} \frac{L_g \cdot (L_g - 1)}{2} \qquad \text{Equation 14.}$$

where L_g is the length in bp of contig g , and there are G contigs in the core genome. To reduce the computational burden of building $\sim O(L^2)$ ML trees, we perform a pre-filtering step to avoid building trees for subsequences (i,j) that almost certainly contain at least one breakpoint. To search for these clear cases of phylogenetic incongruence, we slide a 150 bp window along the sequence of informative SNPs and calculate, for each window, the probability that the SNPs on the left side and the right side of the window come from the same distribution. This is done by registering the frequency of the different SNPs observed in the window in a n-row, 2-column contingency table, where n is the number of different SNPs observed in the window and the columns correspond to the left and right sides of the window. We can then use a χ^2 test to calculate the significance that the observed SNPs are unevenly distributed over the window. This gives us a statistical criterion to split the alignment into smaller blocks tractable by the downstream dynamic-programming algorithm.

In addition, we distinguish cases of significant unevenness caused by incongruent phylogenetic topologies from those caused by long branch lengths, by explicitly measuring the average percentage of conflicting SNPs between all pairs of topologies found in the window:

$$\overline{F}_{dis} = \left\langle \frac{\min(SNP_{T1}, SNP_{T2}) - SNP_{T1 \cap T2}}{SNP_{T1} + SNP_{T2} - SNP_{T1 \cap T2}} \right\rangle \qquad \text{Equation 15.}$$

where SNP_{T1}, SNP_{T2} and $SNP_{T1 \cap T2}$ are the number of SNPs supporting topologies *T1*, *T2* or both *T1* and *T2*. The average in Equation 7 runs over all pairs of topologies found in the window.

For the analysis presented in the main text, we split the alignment in smaller blocks at all positions with chi-square *p*-value < 1e-6 (corresponding to ~1 false positive breakpoint inserted every 1 Mbp, or < 4 total in the core genome) and a $\overline{F}_{dis}$ of at least 15% discordance between topologies (Figure 1.S8).

1.13. Population genetics and phylogenetic analysis

i. McDonald-Kreitman (M-K) test

Gene calls were obtained from the *Vibrio splendidus* 12B01 genome annotation (MicrobesOnline taxon ID 314291) and aligned to core genome contigs with the Smith-Waterman global alignment program *water*, part of the EMBOSS 6.2.0 package (Rice et al, 2000). Gaps in the codon sequence, potentially due low sequencing coverage, were skipped, but a stop codon in any strain always ended the alignment. Polymorphic substitutions (within ingroup strains) and divergent substitutions (fixed between ingroup and outgroup) were counted, and assigned to synonymous or nonsynonymous categories, as previously described (McDonald & Kreitman, 1991b). Only codons for which there were no more than two states were retained for analysis, and we always chose the pathway between codons that minimized the number of nonsynonymous changes. We calculated the Fixation Index, FI = (FN/FS)/(PN/PS), where FN and FS are respectively the numbers of fixed nonsynonymous and synonymous sites, and PN and PS are respectively the number of polymorphic nonsynonymous and synonymous sites. FN, FS, PN and PS were corrected for multiple substitutions using a Jukes-Cantor correction:

$$d = -\frac{3}{4}\ln\left(1 - \left(\frac{4}{3} \cdot \frac{c}{s}\right)\right) \qquad \text{Equation 16.}$$

where d is the corrected value of FN, FS, PN or PS, c is the number of observed substitutions, s is the number of sites (synonymous or nonsynonymous, as appropriate).

A fixation index greater than one suggests positive selection at the protein level between ingroup and outgroup, whereas a value less than one suggests negative selection between ingroup and outgroup, or segregating deleterious polymorphism or balancing selection in the ingroup population. The fixation index is an Odds Ratio statistic; significant deviations from 1 are evaluated with a Fisher exact test.

Computing FI genomewide requires pooling Fisher test contingency tables for multiple genes, potentially resulting in artificial inflation or deflation of the genomewide FI due to Simpson's paradox. To control for this effect, we calculated the expected genomewide FI by permuting the contingency table for each gene (*e.g.* generating a random, 'neutral' contingency table with row and column sums equal to those observed), calculating FI, and repeating this procedure 1000 times to obtain a distribution of expected FI (Shapiro et al, 2007). In the pooled genomewide analysis, we used 3417 coding genes. In rare cases where the same codon is used in multiple genes, the codon was only counted once.

ii. Population differentiation (F_{ST})

F_{ST} was calculated within each inferred recombinant block as:

$$F_{ST} = \frac{\hat{k}_{XY} - \left(\frac{\hat{k}_X - \hat{k}_Y}{2} \right)}{\hat{k}_{XY}} \qquad \text{Equation 17.}$$

where $\hat{k} = \sum_{i<j} k_{ij} \Big/ \binom{n}{2}$, and k_{ij} is the number of single nucleotide differences between sequence (block) i and j of the n total sequences. $\hat{k}_X$ is only calculated between pairs of sequences coming from habitat X, $\hat{k}_Y$ between pairs from habitat Y, and $\hat{k}_{XY}$ between all pairs.

iii. Phylogenetics

An approximate maximum-likelihood tree based on a concatenation of the 14 core genome contigs (3.6 Mbp) was constructed with FastTree (settings: -topm 2 –boot 100 –slow –refresh 0.9 –nni 100) (Price et al, 2010). For shorter sequences, including individual blocks and regions of interest re-sequenced in additional strains, maximum likelihood trees were constructed with PhyML (Guindon & Gascuel, 2003),

using the HKY substitution model, 4 gamma-distributed rate categories, transition/transversion ratio and shape parameter estimated from the data, and 100 bootstraps (settings: 100 HKY e e 4 e BIONJ). Parsimony trees for subsequences of the core genome were inferred using the DNACOMP program of the phylip package. The total number of substitutions along the most parsimonious tree were recorded, as were the number consistent with the parsimony tree (occurring only once), and the number of homoplasic substitutions (occurring more than once in the tree).

1.14. PCR and sequencing

Primers were designed to test for putative 'niche-specifying' alleles in an additional 9 strains: 3 predicted members of the Z population (ZF28, ZF30 and ZF205), 4 predicted members of the SP population (1F97, 1F124, 1F127 and1F175), and 2 free-living strains falling within the SP clade based on *hsp60* similarity (FF160 and FF274).

The following loci in the core genome were targeted for sequencing: two regions of the chitin operon (parts of the N-acetylglucosamine kinase (VIMSS2681958) and N-acetyl-hexosaminidase (VIMSS2681959) genes; primers coreG and coreA, respectively), part of *rpoS* (VIMSS2678244*;* primer coreB), part of *pgm* (VIMSS2679313; primer coreC), an intergenic region near lactate dehydrogenase (VIMSS2678827; primer coreD), part of *gyrB* (VIMSS2682261; primer coreI), and part of formate dehydrogenase (VIMSS2677858; primer coreJ). Five loci in the flexible genome were also targeted by PCR to test for presence/absence in each strain: MSHA biogenesis proteins *mshN* and *mshF* (primers flex1 and flex2), probable maltose O-acetyltransferase (primer flex3), a probable glycosyltransferase (primer flex4), and a putative intercellular adhesion protein (primer flex5). Primer sequences are shown in Table 1.S4. Core PCR reactions consisted of of a denaturing step (30s at 98^{o}C), followed by 30 cycles of denaturing (30s at 98^{o}C), annealing (30s at 50^{o}C), and extension (15s at 72^{o}C). Core PCR products were Sanger sequenced at the MIT Biopolymers laboratory. Sequences were aligned, trimmed and filtered for any regions of low-quality sequence using CLC DNA Workbench 5. This resulted in 372 bp of high-quality sequence for N-acetyl-hexosaminidase, 558 bp for N-acetylglucosamine kinase, 674 bp for *rpoS*, 621 bp for *pgm*, 425 bp near lactate dehyrogenase, 515 bp for *gyrB* and 701 bp for formate dehydrogenase.

Flexible PCR reactions consisted of a denaturing step (30s at 98^{o}C), followed by 30 cycles of denaturing (30s at 98^{o}C), annealing (30s at the temperature shown in Table 1.S4), and extension (15s at 72^{o}C). Presence/absence of a PCR product of the expected size was checked on a gel, with previously published *hsp60* primers as a control (Hunt et al, 2008a).

Supplementary Note 1. Details of genes in 3 ecologically-divergent regions.

Two general features are evident in these 3 regions of high ecological divergence: (1) genetic novelty is more likely acquired by recombination than mutation, and (2) genetic hitchhiking between linked genes may obscure the identities of the true targets of habitat-specific selection. Support for the first point - that habitat-specific genes are commonly acquired by recombination with distant relatives - comes from the observation that most genes in the ecological regions have very high rates of synonymous substitutions between habitats (Table 1.2). Given that, on average, the strains in this study are ~98% genetically identical (Figure 1.1A; Table 1.1), such high synonymous divergence (dS) is best explained by recombination with relatives beyond the 8 genomes considered here. One consequence of such high dS is that, despite relatively high nonsynonymous divergence (dN), traditional tests for positive selection at the protein level (such as *dN/dS* and the McDonald-Kreitman test) suffer a substantial loss of power. For example, three chitin-related genes in region 2 with high *dN* (beta-hexosaminidase, N-acetylglucosamine kinase and endoglucanase) also have high *dS*, obscuring any signal of positive selection in the M-K test. Yet many of the amino acid changes between habitats may have adaptive value and their potential contribution to speciation should not be ignored.

The second point, concerning genetic hitchhiking, is exemplified by the GroEL gene (*hsp60*). The 10 habitat-specific substitutions in GroEL were used in the initial identification of SP- and Z-specific strains (Hunt et al, 2008a), yet it seems unlikely that this housekeeping gene plays a large part in habitat adaptation. More likely, another gene in this cluster, such as DNA methyltransferase or NormM, confers the habitat-specific adaptive value while GroEL has simply hitchhiked along with it on the same linked piece of DNA. Therefore, it was somewhat serendipitous to have chosen GroEL as a putative neutral marker gene, when in fact it may be linked to genes with habitat-specific fitness effects. The majority of loci in the genome show little or no support for the partitioning of strains into habitats; had one of these loci been chosen, the distinction between SP and Z strains would never have been discovered in the first place. In the future, use of putatively fitness-neutral marker genes (*e.g.* MLSA genes) should be used with caution: depending on the degree of recombination in the population of interest, markers could be linked to recombined, selected loci whose evolutionary history does not reflect the population's history of clonal descent.

The first region, located on the small chromosome, contains 2 genes: a predicted RNA polymerase stress-response sigma factor, RpoS, and an RTX (repeat in toxin) family protein. The sigma factor may be a *Vibrio*-specific second copy of RpoS (henceforth "RpoS2"), the first copy of which is located on the large chromosome (Figure 1.S1). RpoS2 contains 26 fixed amino acid (aa) changes between SF and Z

strains, 3 of which occur in predicted DNA binding domains: positions 141 (SP=His, Z=Arg), 153 (SP=Ser, Z=Asn) and 161 (SP=Thr, Z=Lys) of the *T. thermophilus* reference sequence (Lee & Gralla, 2002). An additional 2 DNA binding residues differ between RpoS2 and the canonical RpoS, but are identical in SP and Z strains: positions 140 (RpoS=Phe, RpoS2=Tyr) and 156 (RpoS=Arg, RpoS2=Ala). These observations suggest, first, that RpoS2 may target significantly different DNA binding sites than the canonical stress-response sigma factor RpoS, and second, that RpoS2 may have experienced functional modifications between SP and Z strains - potentially promoting transcriptional activation of different sets of genes in each habitat.

The RTX family protein adjacent to RpoS2 has similarity to *V. cholerae* RTX, an excreted cytotoxic protein (Lin et al, 1999). RTX is highly diverged between habitats, with >200 fixed aa changes (Table 1.2) and significant domain reorganization (the RTX homolog in SP strains is 1505 aa in length, but only 786 aa in Z strains and the 12B01 outgroup).

The second ecologically differentiated region consists of an 8kb operon involved in chitin metabolism. This locus presents an intriguing candidate for ecological adaptation since zooplankton exoskeletons are composed largely of chitin, while small particles might contain chitin at lower concentration, in more degraded forms, or not at all. Strains adapted to these different habitats may have experienced selective pressures to fine-tune their chitin utilization strategies in different ways. Although the chitin operon contains a few transport proteins, most of the habitat-specific divergence is confined to 4 enzymes further downstream in the pathway of chitin uptake and breakdown (Table 1.1). These enzymes are involved in producing N-acetyl-glucosamine (GlcNAc) monomers from $(GlcNAc)_{2,3}$ oligomers, either in the periplasm by N-acetyl-hexosaminidase or in the cytosol by diacetylchitobiose phosphorylase, and in breaking down deacetylated chitin dimers $(GlcN)_2$ to monomers (by the endoglucanase) and phosphorylating them (by glucosamine kinase) to be fed into the pentose phosphate pathway (Hunt et al, 2008b). Proteins in the upstream portion of chitin metabolism (chitin transporters and chitinases) show no differences between SP and Z strains. Thus, it may be that different selective pressures between habitats are primarily due to 'downstream' differences in chitin quality (long polymers vs. shorter oligomers) and acetlyation levels, rather than 'upstream' differences in local chitin concentration. These 'downstream' differences have the potential to vary widely depending on the chitin source (*e.g.* shrimp, crab or squid shell, whether it is mechanically broken, treated with acids or bases, etc.; (Velde & Kiekens, 2004)).

The third region contains genes coding for phosphoglyceromutase (EC 5.4.2.1), methylated-DNA-protein-cysteine methyltransferase (EC 2.1.1.63), NormM (a putative Na^+-driven multidrug efflux pump),

and the chaperone proteins GroES and GroEL (Table 1.2). Although the functional significance of habitat-specific evolution of these proteins is unclear, it is worth noting that DNA methyltransferase plays an important role in repairing damaged DNA - specifically by reversing O-6-methylation of guanine and preventing G-A transitions from accumulating (Myers et al, 1993). Loss of DNA repair capabilities may be an important mechanism in generating 'hypermutator' phenotypes, which may promote rapid adaptation under strong natural selection (Barrick et al, 2009, Visser,J Arjan G M de, 2002, Hazen et al, 2009). The hypermutator state may eventually be reversed by re-acquisition of the lost DNA repair gene(s), thus rescuing the lineage from eventual mutational meltdown. Our data reveal 25 fixed aa changes in DNA methyltransferase between SP and Z habitat-associated strains. Both SP and Z strains are fairly diverged from the 12B01 outgroup, with 14 and 23 aa changes, respectively. Such a high density of substitutions in a relatively small protein (132 aa) strongly suggests that DNA methyltransferase was recombined independently into SP and Z strains from distant yet unidentified relatives (because 12B01 is always the top BLAST hit, but with relatively low identity). Acquisition of DNA methyltransferases from distant relatives could be explained if hypermutator strains emerge frequently in nature, and eventually must revert to a normal DNA repair phenotype by re-gaining the necessary genes, potentially from any available donor. In the case of our ecologically-associated *Vibrio* strains, we hypothesized that the Z strains might have hypermutator phenotypes, given that they contain a DNA methyltransferase homolog with more changes relative to the outgroup than SP strains. If DNA methyltransferase activity were impaired in Z strains, we would expect an increased incidence of G-A transitions in the recent evolution of Z strains relative to SP strains. Indeed, we found that Z strains (leaves on the phylogeny, recent mutations) had a transition:transversion ratio (ti:tv) of 2.06, moderately higher than the ti:tv of 1.93 in SP strains (Fisher test: Odds Ratio = 1.1, $p = 0.002$). Consistent with a Z strain-specific lesion in DNA methyltransferase, 43% of recent mutations in Z strains involved G-A transitions - significantly more than the 33% observed in SP strains (Odds Ratio = 1.55, $p < 2.2e\text{-}16$). More generally, there is some evidence for 'bursts' of evolution (unusually long branches in the phylogeny) to involve elevated ti:tv ratios, potentially due to hypermutation events. We tentatively conclude that rapid gain and loss of hypermutator phenotypes may be common in natural *Vibrio* populations, possibly contributing to rapid adaptation to new habitats.

Supplementary Note 2. Genomewide M-K test

Although the M-K test is not appropriate when ingroup/outgroup are not well defined, as is the case in the majority of genomic blocks that reject the partitioning of genomes along ecological lines, it still reveals a modest genomewide inflation in *dN/dS* between habitats (*dN/dS* = 0.063), relative to within the Z habitat (*pN/pS* = 0.060; *p* = 0.008; Table 1.S5). This would be expected as slightly-deleterious nonsynonymous mutations hitchhike to fixation along with the genome targeted by a selective sweep. Alternatively, the excess of nonsynonymous changes between habitats could have been driven by recurrent selective sweeps targeting the amino acid changes themselves. These data are also consistent with an ecotype model of speciation involving genetic drift rather than selection. The drift model would require that SP and Z strains undergo population bottlenecks independently of one another. A bottleneck may indeed have occurred recently in the SP strains, which show a significant excess of segregating nonsynonymous polymorphism (*pN/pS* = 0.068; *p* = 0.001; Table 1.S5), likely due to genomewide fixation of slightly deleterious nonsynonymous mutations.

Figure legends

Figure 1.1. Phylogeny follows ecology at just a few habitat-specific genomic loci.
(A) The SP (filled green squares) and Z strains (open red circles) form two distinct monophyletic clades in an approximate maximum-likelihood tree based on 3.6Mb of aligned 'core' genome. Top tree: support for tree partitions based 100 bootstraps (branch lengths not to scale); Bottom tree: branch lengths (substitutions/site);
(B) Many different tree topologies are required to account for most of the core genome. All 10,395 different unique, unrooted trees with 8 leafs were considered. Each recombinant block of the genome was allowed to 'vote' on any tree consistent with the block's informative SNPs. The tree with the most votes (accounting for the most bases in the genome) was chosen as the top ranked tree, and removed from further iterations along with its recruited blocks. The voting process was repeated iteratively, greedily recruiting blocks to the remaining tree topologies until all blocks were accounted for (requiring 438 unique trees, the top 100 of which are shown; see Figure 1.S4 for all 438 and Figure 1.S9 for the top 20 ranked tree topologies).
(C) Regions of the genome supporting (black line) or rejecting (grey line) the partitioning of strains into distinct habitats. Core genome contigs are shown as alternating white/dark grey lines along the x-axis. The right-most (white) contig corresponds to the small chromosome. Tick marks below the x-axis represent the inferred recombination breakpoints. Black peaks represent uninterrupted stretches of support for the ecological split, with peak height showing the number of SNPs supporting the split (*e.g.* all SP strains have base 'A' while all Z strains have base 'C') in that stretch. Numbered peaks correspond to the regions described in Table 1.2 and 1.S1. Grey peaks represent uninterrupted stretches of SNPs inconsistent with the ecological split. Trees for 'ecological' and 'housekeeping' loci sequenced in additional strains are shown above and below the plot, respectively (see Figure 1.S2 and Table 1.S4).

Figure 1.2. Recent gene flow is more common within than between habitats.
(A) Core genome recombination events that split up the two most recently diverged sister strains either by recombination within the SP habitat (green bar), or between habitats (Z strains shown as grey bars).
(B) Distance matrix based on the amount of shared flexible genomic blocks between strains. Similarity between a pair of strains *i* and *j* was defined as 2*(# blocks shared between *i* and *j*)/(# blocks in *i* + # blocks in *j*). The heatmap represents UPGMA clustering of the distance matrix (distance = 1 – similarity). The tree on the left was generated using 1000 bootstraps resamplings (with replacement) of the flexible blocks and building a Neighbour-Joining tree with the distance matrix. The tree shown represents an 80% consensus across bootstrap replicates, with percentage support shown above nodes in the tree.

Figure 1.3. Schematic of microbial speciation with gene flow.
Lineages of clonal (vertical) descent are shown as black lines, which become red or green upon acquisition of habitat-specific alleles (red or green arrows). Recombination events are depicted by vertical arrows, with grey arrows representing recombination of fitness-neutral alleles and colored arrows representing habitat-specific alleles.

Table legends

Table 1.1. Recombination and mutation in *Vibrio* and *Salmonella* population genomes.
All values and parameters were inferred using STARRInIGHTS. The mutation rate (# substitutions per site per branch) for *V. splendidus* represents the median value across all 1995 blocks. The number of generations to the most recent common ancestor (MRCA) was estimated as (# subs./site/branch) * (# sites in core genome) * (300 generations/substitution in genome) after Drake ((Drake, 1991)), with substitution counts extracted from the block's parsimony tree.

Table 1.2. Divergence and polymorphism in top 3 regions supporting the ecological split.
The degree of population differentiation, *Fst* between SP and Z, was calculated as described in Methods. FN, FS, PN, PS and FI are as described in Table 1.S5. Only genes with divergence between habitats (FN and/or FS > 0) are shown.

Supplementary figure legends

Figure 1.S1. 'RpoS2' is a *Vibrio*-specific second copy of RpoS (sigma 38).
The protein sequences of *V. splendidus* 12B01 RpoS (VIMSS2680752), RpoD (VIMSS2682712), and putative second copy of RpoS ('RpoS2'; VIMSS2678244) were obtained from MicrobesOnline and BLASTed against the NCBI nr database to obtain homologs in other species (E-value < 1e-20). Homologous sequences were aligned with MUSCLE (Edgar, 2004), and used to generate a maximum-likelihood PhyML tree. The SP and Z strains used in this study are highlighted in green and red, respectively, on the tree. RpoS2 appears to form a distinct clade on the tree, consisting entirely of vibrios.

Figure 1.S2. Trees of 'ecological' and housekeeping genes resequenced in additional strains.
Regions of 4 'ecological' genes and 3 housekeeping genes were amplified from additional strains (indicated with asterisks) using PCR primers described in Table 1.S4, and sequenced as described in Methods. Trees were built using PhyML, rooted with 12B01 and 12F01 as outgroups. SP strains = filled green boxes; Z strains = red outlined boxes.

Figure 1.S3. Recombination across COG functional gene categories.
(A) Recombination *blocks* were classed as either supporting the ecological partition of strains (upper panel), or rejecting the ecological partition (lower panel). The relative fraction of genes within a given function that either support (upper) or reject (lower) the ecological split is shown (normalized to the mean fraction across all functions). Enrichments or depletions of each COG function within each class (ECO-support or –reject) were assessed using a hypergeometric test. No significant enrichments were found after multiple-hypothesis correction; individually significant, uncorrected *p*-values are shown (**p* < 0.05; ***p* < 0.01).
(B) Recombination *breakpoints* that split between an ECO-supporting and an ECO-rejecting block ('between-habitat breaks'; upper panel) and those that split between two ECO-supporting blocks ('within-habitat breaks'; lower panel) were counted. We then measured the mean distance from the edges of each coding gene to a breakpoint of either class (*e.g.* (distance from start codon to upstream breakpoint + distance from stop codon to downstream breakpoint) / 2). The y-axis shows (1 - the relative mean distance to a breakpoint). Distances are relative to the mean across all functions. Higher values on the y-axis denote COG functions that tend to be closer to breakpoints; lower values denote COG functions that tend to be further from breakpoints. Differences between each COG function and the mean distance across functions were assessed using a Wilcoxon test. No functions were significantly different from the mean after multiple-hypothesis correction; individually significant, uncorrected *p*-values are shown (**p* < 0.05; ***p* < 0.01).

Figure 1.S4. All 438 different tree topologies required to cover the entire core genome.
See figure legend 1.1B in the main text.

Figure 1.S5. Example STARRInIGHTS calculations and workflow.
See methods for a detailed description.

Figure 1.S6. STARRInIGHTS benchmarked on simulated sequence.

Frequency of breakpoint calls in simulated contigs using (A) the ML STARRInIGHTS method, (B) the ML method corrected for model complexity, or (C) the Parsimony method. Contigs of length 2500 bp were simulated under either Scenario A (two recombination events, yielding four breakpoints) or Scenario B (no recombination events), as described in Methods. The number of breakpoints at each site, summed over 100 simulations, is shown as the height of vertical bars along the contig length (x-axis). Boundaries of simulated 'true' recombination events are indicated with blue shading.

Figure 1.S7. Empirical estimates of model complexity penalties.
The log-likelihood ratio of a model with a single breakpoint to a model with no breakpoints is plotted as a heatmap for different simulated sequence lengths (y-axis) and mutation rates (x-axis). As described in Methods, high values of log (*L1/L0*) indicate that introducing a breakpoint in the sequence significantly improves the likelihood, even though no recombination event actually occurred. Likelihoods were computed using phyML. The shade of each square in the grid represents the maximum value of log (*L1/L0*) observed in 100 simulated sequences with the given length and mutation rate, and each with a unique, randomly chosen tree topology.

Figure 1.S8. Example phylogenetic incongruence filter applied to contig 56.
The phylogenetic discordance metric, described in Methods, is plotted in the upper panel. The middle panels displays informative, dimorphic SNPs in the 8 strains (*e.g.* strains with the same base shown in white; strains with the alternative base shown in blue). The bottom panel shows the χ^2 *p*-value for stretches of significantly discordant SNPs. The yellow highlighted regions show significantly discordant stretches in which at least one breakpoint must be present (discordance metric > 0.15 and $p < 1e\text{-}6$). The contig is broken in these stretches such that four different, partially overlapping subcontigs are considered in the downstream STARRInIGHTS algorithm (horizontal black lines in lower panel). Within each subcontig, all subsequences (*i,j*) are used to build parsimony trees and infer further breakpoints.

Figure 1.S9. Top 20 ranked tree topologies.
Trees are ranked as described in Figure 1.1B. SP strains are shown as filled green boxes. Trees that support the 'ECO split' (separate monophyletic clusters for SP and Z strains) are indicated with filled number boxes. All trees are unrooted and branch lengths are not to scale.

Figure 1.S10. More divergent ECO blocks also tend to be highly polymorphic.
Each point represents one of the 234 ECO blocks, with blocks within the top 3 densest clusters of ecological divergence indicated in red, green and blue, respectively. Region 1, consisting of just one block, is a clear outlier: it is highly divergent between habitats, but contains relatively little polymorphism, suggesting a relatively recent transfer into one or both habitats.

Figure 1.S11. Schematic of very recent core genome recombination events.
(A) Core genome recombination events that split up the two most recently diverged sister strains (green squares with bold outline) either by recombination within the SP habitat (green square with light outline), or between habitats (Z strains shown as red circles). Events are depicted as arrows (within habitat = green, between habitat = grey), with line width proportional to the number of events.

Figure 1.S12. Distribution of F_{ST} is skewed toward high values in top 3 ECO regions.
Histogram of F_{ST} across all 1995 recombinant blocks (empty bars), or restricted to 41 blocks within the top 3 ECO regions described in Figure 1.1C (black lines).

Figure 1.S13. Flexible genome sizes.
Flexible genome is defined as any aligned block of DNA not present in all 8 SP/Z strains and the 12B01/12F01 outgroup. Strain-unique DNA is defined as a block present in only a single genome. Flexible genome sizes of each strain shown (A) in amount of DNA (Mb) or (B) number of blocks.

Supplementary table legends

Table 1.S1. Divergence and polymorphism in top 12 regions supporting the ecological split.
See Table 1.2 in main text for description.

Table 1.S2. List of habitat-specific flexible genome blocks.
Flexible genome blocks were identified as described in Methods. Habitat was 'Z-unique' if the block was found in all Z strains and no SP strains, and 'SP-unique' if the inverse was true. Annotations of genes in each block were obtained by blastx against the NCBI nr protein database. Alternating shaded/unshaded rows distinguish between genes on different blocks, and genes shown in italic were tested for presence/absence in additional strains using the primer set indicated.

Table 1.S3. PCR assay for presence/absence of flexible genome blocks.
PCR products from the indicated template strain and primer combinations were run on an agarose gel, and presence/absence of a band of the expected length was scored as +/-.

Table 1.S4. PCR primers used in this study.

Table 1.S5. Genomewide divergence and polymorphism in coding regions.
FN = number of fixed nonsynonymous (ns) sites; FS = number of fixed synonymous (syn) sites; *dN/dS* = ratio of fixed ns substitutions per ns sites to fixed syn substitutions per syn site, corrected for multiple substitutions; PN = number of polymorphic ns sites; PS = number of polymorphic syn sites; *pN/pS* = ratio of polymorphic ns substitutions per ns sites to polymorphic syn substitutions per syn site, corrected for multiple substitutions; FI obs = observed fixation index, (FN/FS)/(PN/PS), corrected for multiple substitutions; FI exp = expected fixation index based on 1000 permuted resamplings of 2x2 McDonald-Kreitman (M-K) contingency tables (Methods).

Figure 1.1

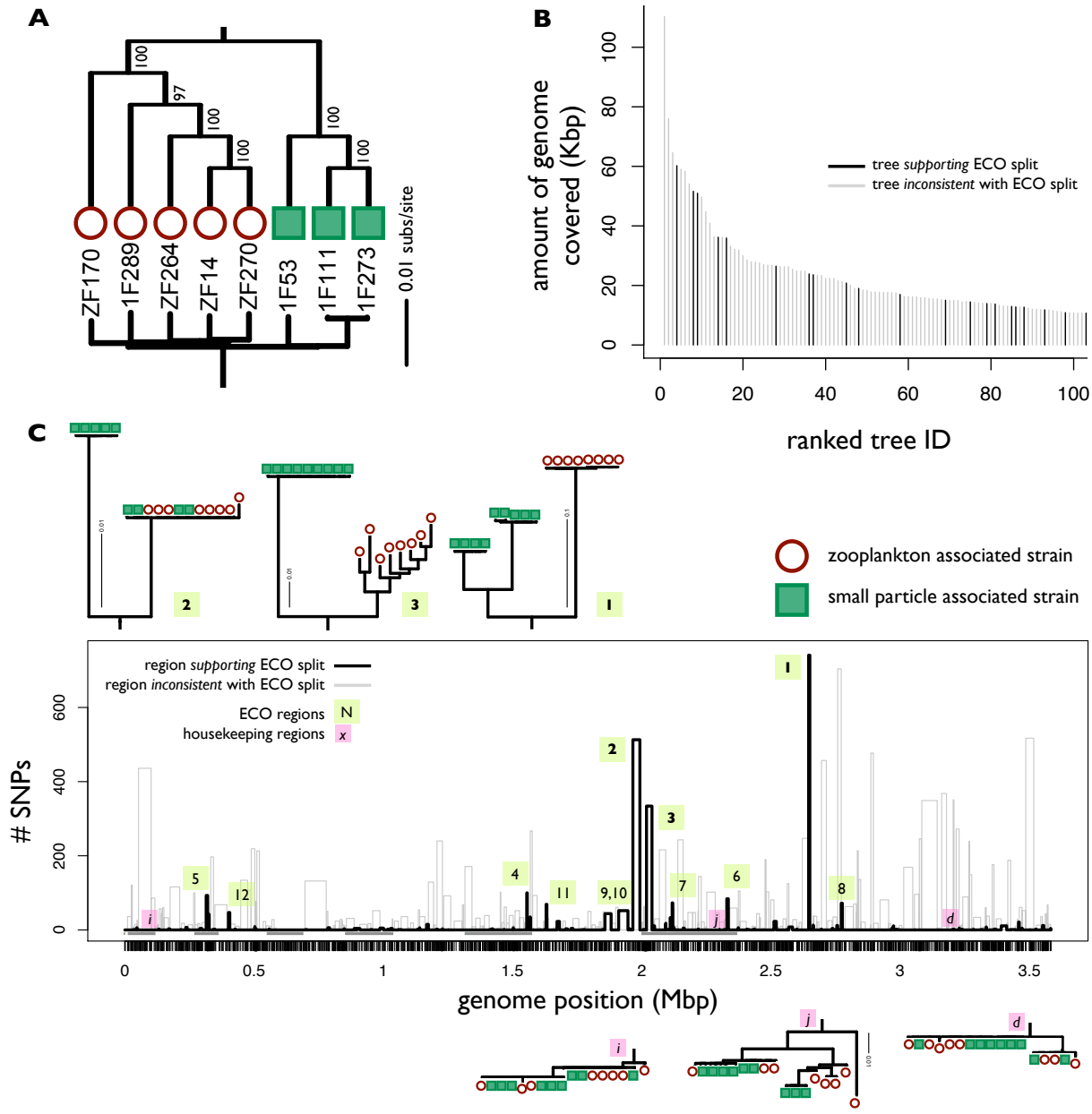

A
100
97
100
100
100
100
ZF170
1F289
ZF264
ZF14
ZF270
1F53
1F111
1F273
0.01 subs/site
B
amount of genome covered (Kbp)
0
20
40
60
80
100
tree supporting ECO split
tree inconsistent with ECO split
ranked tree ID
C
zooplankton associated strain
small particle associated strain
region supporting ECO split
region inconsistent with ECO split
ECO regions N
housekeeping regions x
SNPs
genome position (Mbp)
0
0.5
1
1.5
2
2.5
3
3.5
200
400
600
1
2
3
4
5
6
7
8
9,10
11
12
i
j
d

Figure 1.2

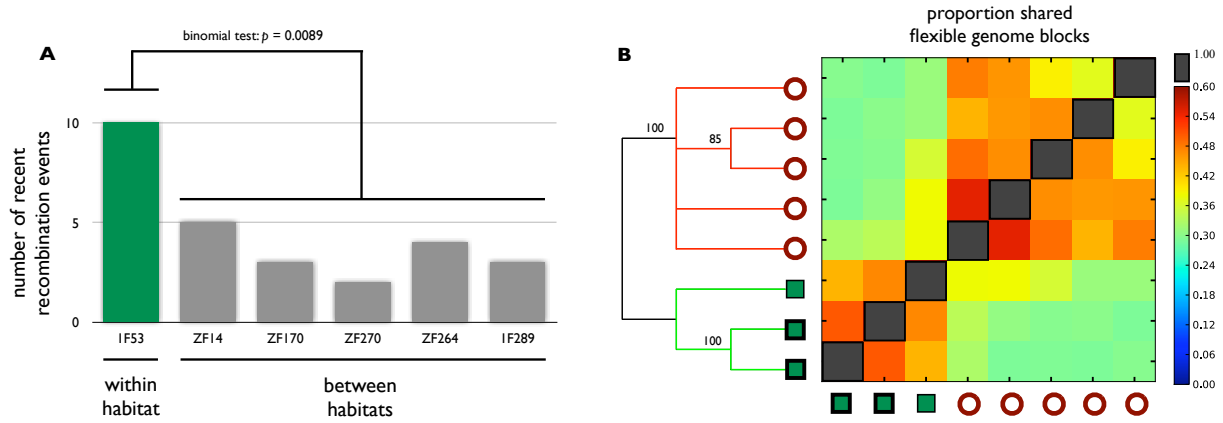


Figure 1.3

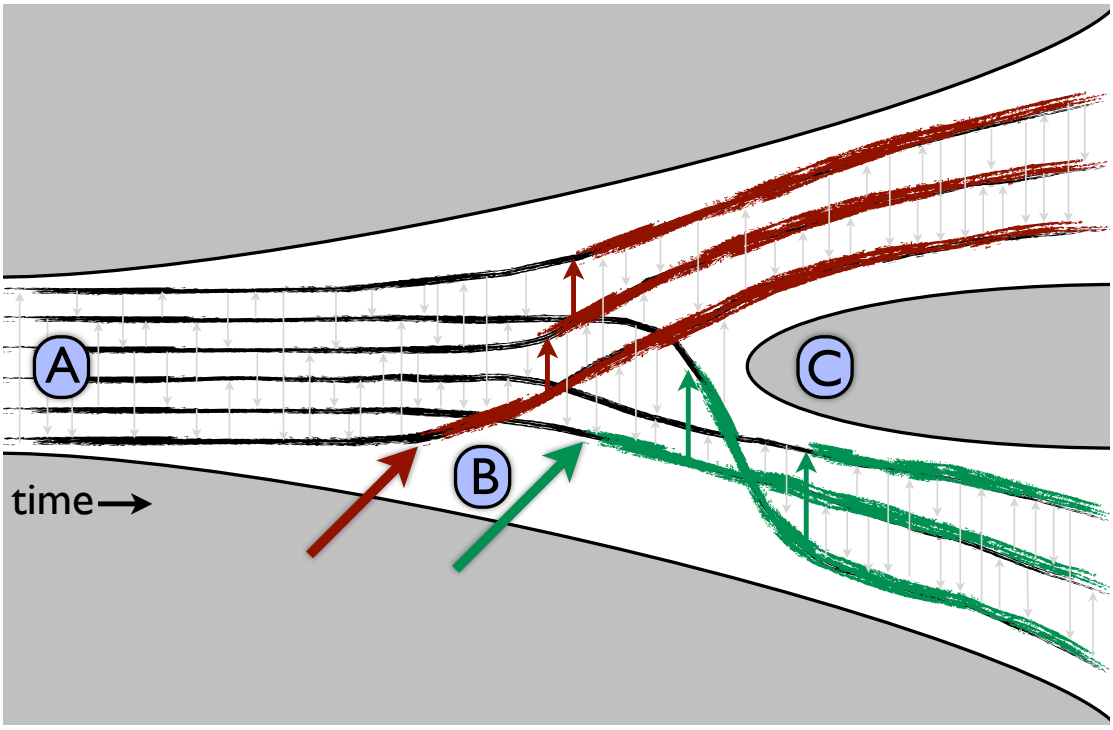


Table 1.1

			Polymorphic sites			Recombination		Mutation (within blocks)	
organism	# genomes	core size (Mbp)	informative	non-inform.	# homoplasies	# events	ρ(Break)	subs/site/branch	gen.'s since MRCA
V. splendidus	8	3.58	42,873	38,459	1,375	991 - 1982	4.90E-04	6.65E-04	7.15E+05
S. enterica	19	5.81	473	1,313	13	0	2.10E-07	1.05E-05	1.51E+04

Table 1.2

Position					Number of mutations			Fixed between habitats			Polymorphism within habitat				M-K test			
											SP		Z					
region	contig	start	end	length (Kbp)	total	# ecoSNPs (%)	*Fst*	N	S	dN/dS	pN	pS	pN	pS	FI	*p*	# codons	gene
1	60	272305	276907	4.6	764	741 (24.9%)	0.99	26	103	0.46	1	0	0	1	Inf	1	262	RNA polymerase sigma factor RpoS
								223	175	0.43	0	1	1	4	5.1	0.176	784	RTX family protein
2	58	387879	414935	27.1	815	513 (17.3%)	0.71	1	2	1.46	2	0	4	3	0.41	1.00	428	D-alanyl-D-alanine carboxypeptidase (penicillin-binding protein 4)
								0	44	0.00	0	0	2	0	Inf	1.00	259	Phosphomannomutase
								2	138	0.04	0	0	9	60	0.09	0.00	792	Cellobiose phosphorylase
								20	97	0.55	1	0	21	68	0.63	0.22	642	N-acetyl-beta-hexosaminidase
								12	58	0.51	0	0	8	22	0.49	0.17	294	Predicted N-acetylglucosamine kinase
								19	64	0.84	0	0	6	39	1.82	0.26	573	Endoglucanase-related protein)
								1	23	0.12	0	0	0	0	0.00	1.00	331	ABC-type transport system, ATPase component
								0	2	0.00	0	0	0	21	0.00	1.00	327	ABC-type transport system, ATPase component
								0	3	0.00	1	0	0	15	0.00	1.00	335	ABC-type transport systems, permease components
								0	1	0.00	0	0	0	7	0.00	1.00	328	ABC-type transport systems, permease components
								0	1	0.00	0	0	1	38	0.00	1.00	555	ABC-type transport system, periplasmic component
								0	1	0.00	0	0	4	31	0.00	1.00	1129	Signal transduction histidine kinase
3	59	18462	39387	20.9	578	334 (11.2%)	0.81	14	30	1.29	0	0	2	41	9.71	0.00	506	Phosphoglyceromutase
								26	58	0.82	0	0	2	7	0.94	1.00	145	Methylated DNA-protein cysteine methyltransferase
								20	99	0.45	0	0	0	33	Inf	0.02	408	Na+-driven multidrug efflux pump
								0	0	n/a	0	0	0	0	0	1	96	Co-chaperonin GroES (HSP10)
								2	8	0.67	0	1	0	26	Inf	0.07	548	Chaperonin GroEL (HSP60 family)
								0	4	0	0	1	1	5	0	1	219	3-hydroxyisobutyrate dehydrogenase
								1	7	0.37	0	2	1	3	0.46	1	185	Uncharacterized low-complexity protein
								1	3	0.89	0	0	1	4	1.29	1	132	Uncharacterized membrane protein
								0	7	0	0	0	2	11	0	0.52	340	Lysine 2,3-aminomutase
								0	1	0	0	0	0	0	0	1	188	Translation elongation factor P (EF-P)
								0	2	0	0	0	0	1	0	1	128	Fumarate reductase subunit D
								0	2	0	0	0	1	3	0	1	127	Fumarate reductase subunit C
								1	5	0.58	0	0	1	22	4.11	0.38	607	Succinate dehydrogenase, flavoprotein subunit
								0	2	0	0	0	6	14	0	1	323	Truncated, possibly inactive, lysyl-tRNA synthetase (class II)
								1	3	0.92	1	0	1	9	2.74	0.51	459	Di- and tricarboxylate transporters
								0	3	0	0	0	1	6	0	1	304	Phosphatidylserine decarboxylase

Figure 1.S1

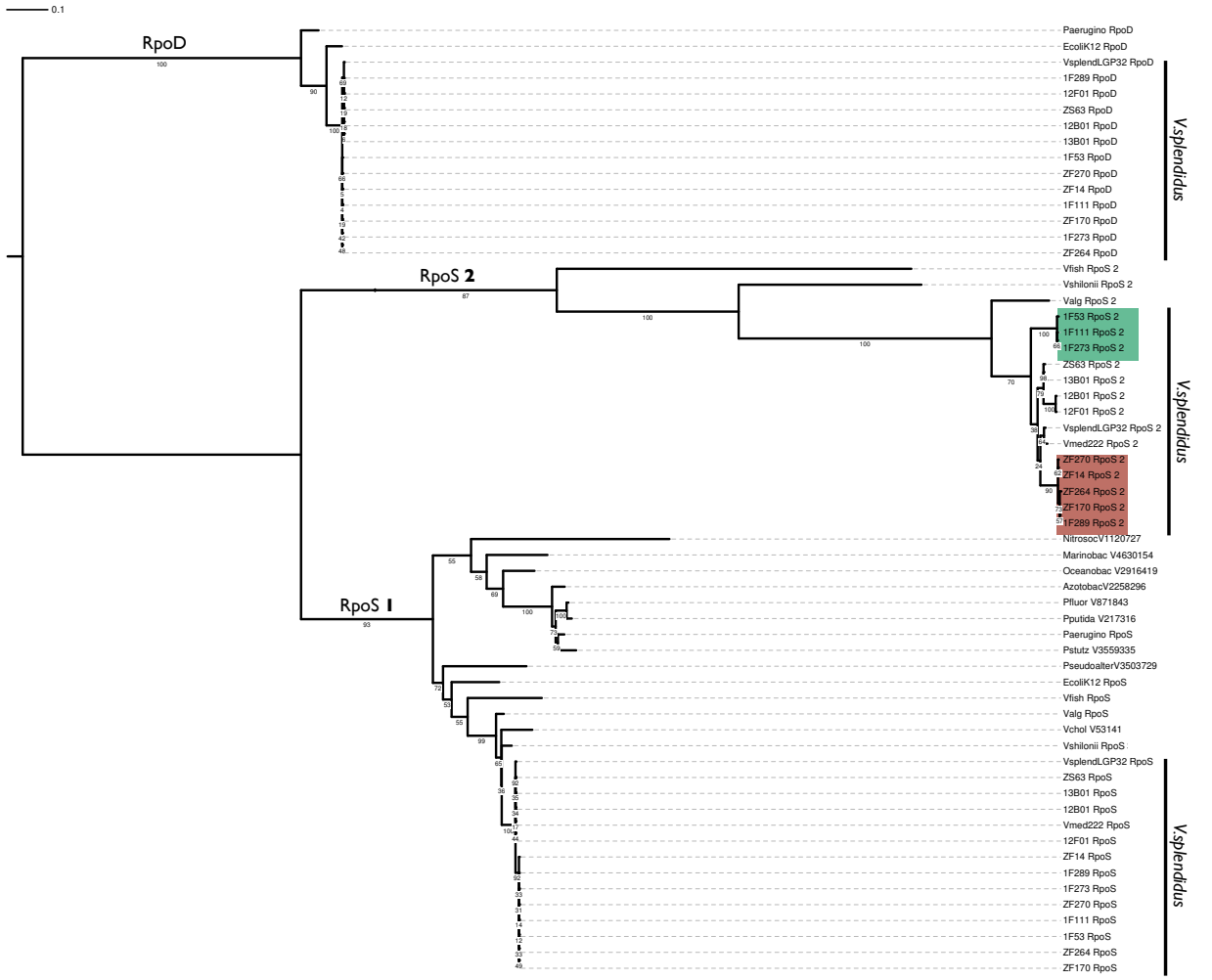

0.1
RpoD
RpoS 2
RpoS 1
Paerugino RpoD
EcoliK12 RpoD
VsplendLGP32 RpoD
1F289 RpoD
12F01 RpoD
ZS63 RpoD
12B01 RpoD
13B01 RpoD
1F53 RpoD
ZF270 RpoD
ZF14 RpoD
1F111 RpoD
ZF170 RpoD
1F273 RpoD
ZF264 RpoD
Vfish RpoS 2
Vshilonii RpoS 2
Valg RpoS 2
1F53 RpoS 2
1F111 RpoS 2
1F273 RpoS 2
ZS63 RpoS 2
13B01 RpoS 2
12B01 RpoS 2
12F01 RpoS 2
VsplendLGP32 RpoS 2
Vmed222 RpoS 2
ZF270 RpoS 2
ZF14 RpoS 2
ZF264 RpoS 2
ZF170 RpoS 2
1F289 RpoS 2
NitrosocV1120727
Marinobac V4630154
Oceanobac V2916419
AzotobacV2258296
Pfluor V871843
Pputida V217316
Paerugino RpoS
Pstutz V3559335
PseudoalterV3503729
EcoliK12 RpoS
Vfish RpoS
Valg RpoS
Vchol V53141
Vshilonii RpoS
VsplendLGP32 RpoS
ZS63 RpoS
13B01 RpoS
12B01 RpoS
Vmed222 RpoS
12F01 RpoS
ZF14 RpoS
1F289 RpoS
1F273 RpoS
ZF270 RpoS
1F111 RpoS
1F53 RpoS
ZF264 RpoS
ZF170 RpoS
V.splendidus
V.splendidus
V.splendidus

Figure 1.S2

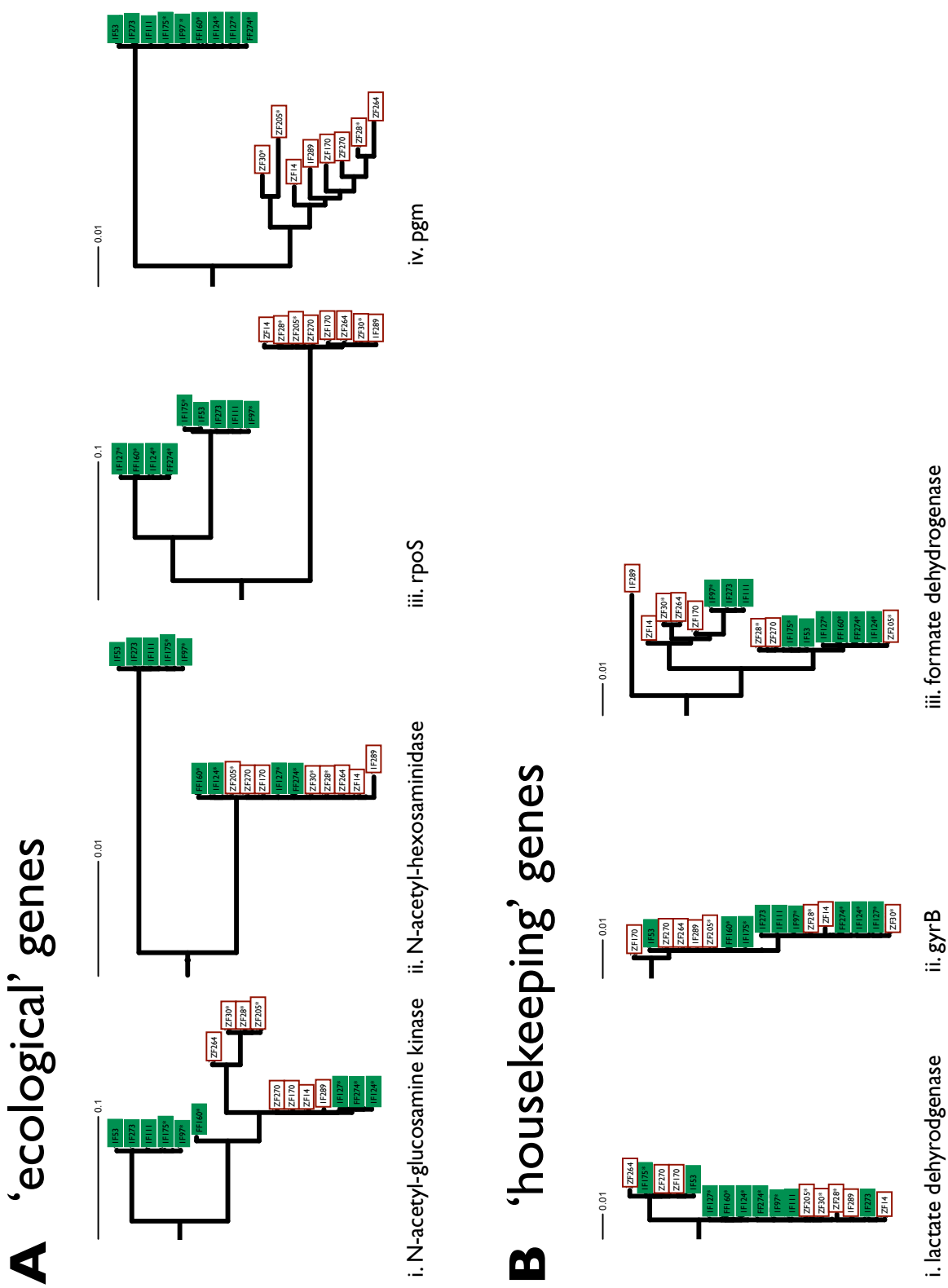
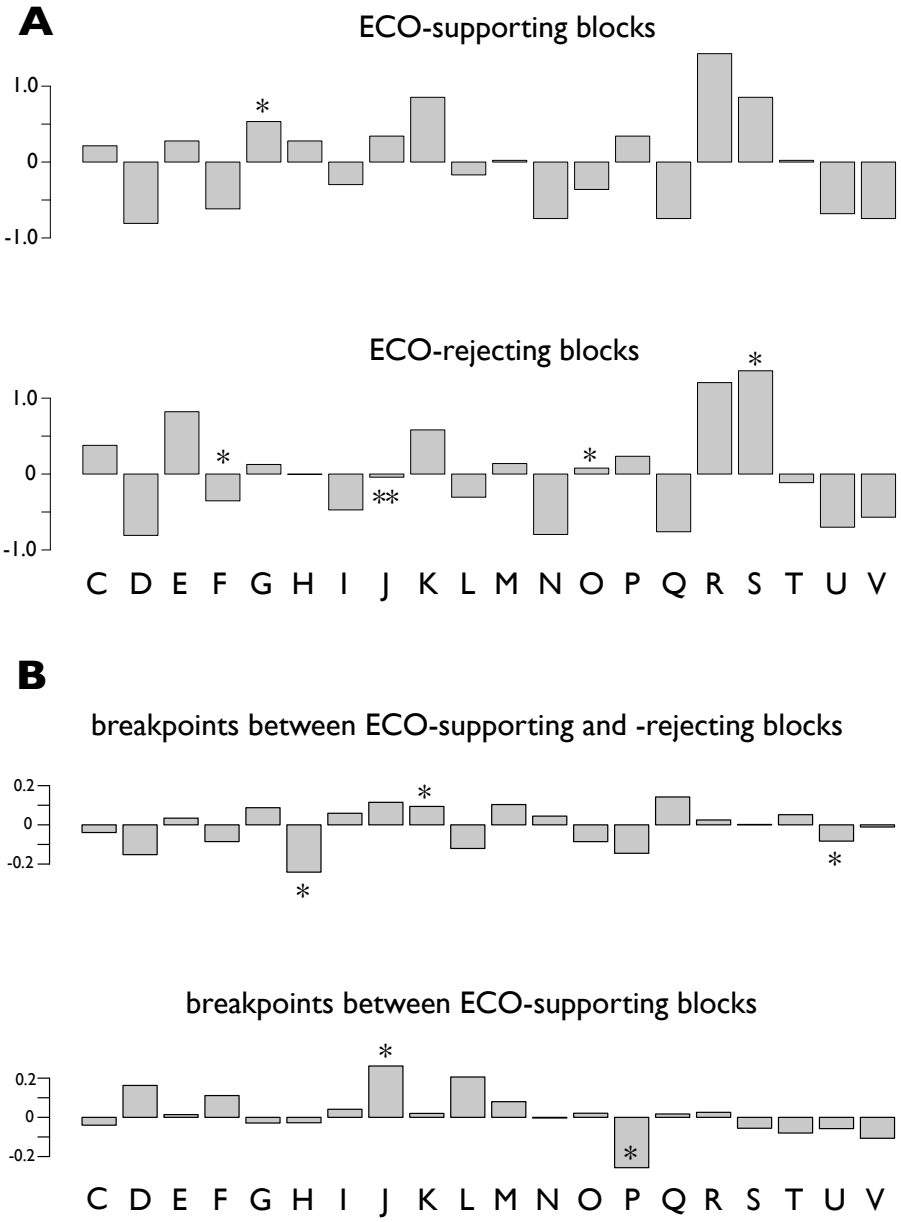

A 'ecological' genes
i. N-acetyl-glucosamine kinase
ii. N-acetyl-hexosaminidase
iii. rpoS
iv. pgm
B 'housekeeping' genes
i. lactate dehyrodgenase (intergenic)
ii. gyrB
iii. formate dehydrogenase

Figure 1.S3



COG functions:

C: Energy production and conversion
D: Cell cycle control, cell division, chromosome partitioning
E: Amino acid transport & metabolism
F: Nucleotide transport & metabolism
G: Carbohydrate transport & metabolism
H: Coenzyme transport & metabolism
I: Lipid transport & metabolism
J: Translation, ribosomal structure and biogenesis
K: Transcription
L: Replication, recombination & repair
M: Cell wall/membrane/envelope biogenesis
N: Cell motility
O: Posttranslational modification, protein turnover, chaperones
P: Inorganic ion transport & metabolism
Q: Secondary metabolites biosynthesis, transport & catabolism
R: General function prediction
S: Function unknown
T: Signal transduction mechanisms
U: Intracellular trafficking, secretion & vesicular transport
V: Defense mechanisms

Figure 1.S4

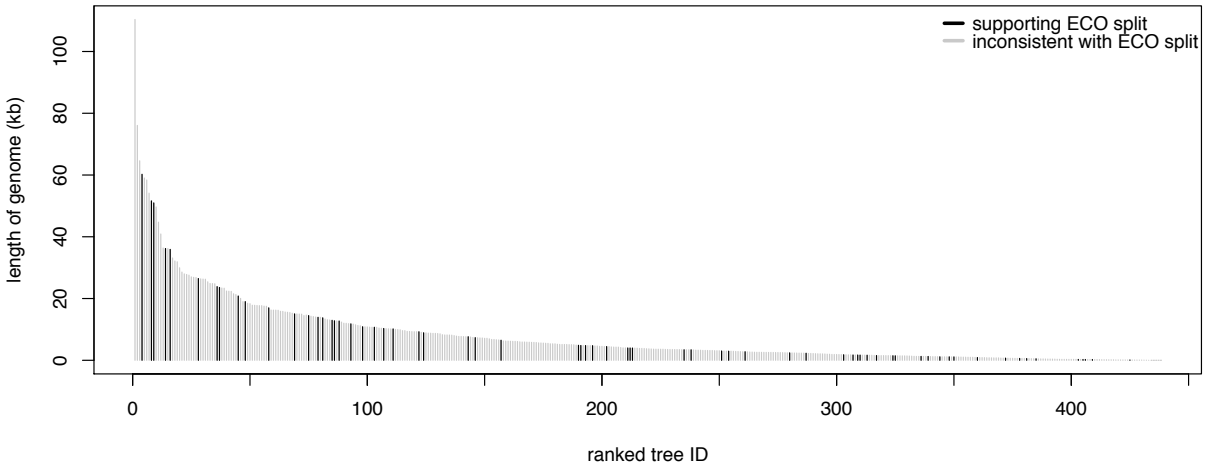
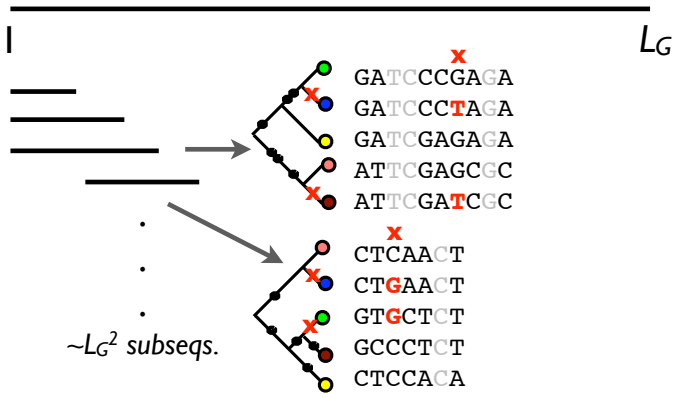
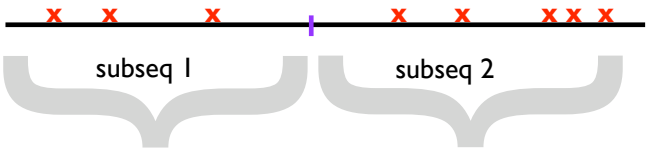

supporting ECO split
inconsistent with ECO split
length of genome (kb)
0
20
40
60
80
100
0
100
200
300
400
ranked tree ID

Figure 1.S5



e.g. *1*. combine 2 subseqs with a breakpoint in between:



$C(1,L_G) = c_b(1) + c_{nb}(L_G-1) + c_{Tree\ 1} + c_{Tree\ 2}$

e.g. 2. combine 4 subseqs with 3 breakpoints in between:

$C(1,L_G) = c_b(3) + c_{nb}(L_G-3) + c_{Tree1} + c_{Tree2} + c_{Tree3} + c_{Tree4}$

... consider further e.g.'s and choose the best by DP

1. Consider all subsequences of the core genome (~L^2).

2. Each subsequence gets an ML tree.

x = homoplasic / unparsimonious site

3. Define a cost function for recomb. breakpoints (b) and trees in intervening sequences.

$$C(i,j) = c_b \cdot b_{ij} + c_{nb} \cdot (l_{ij} - b_{ij}) + c_{Tree(i,j)}$$

-log probabilities

events (breakpoint or not)

4. Find optimal breakpoint locations by dynamic programming (DP). Estimate c_b by Expectation-Maximization.

Figure 1.S6

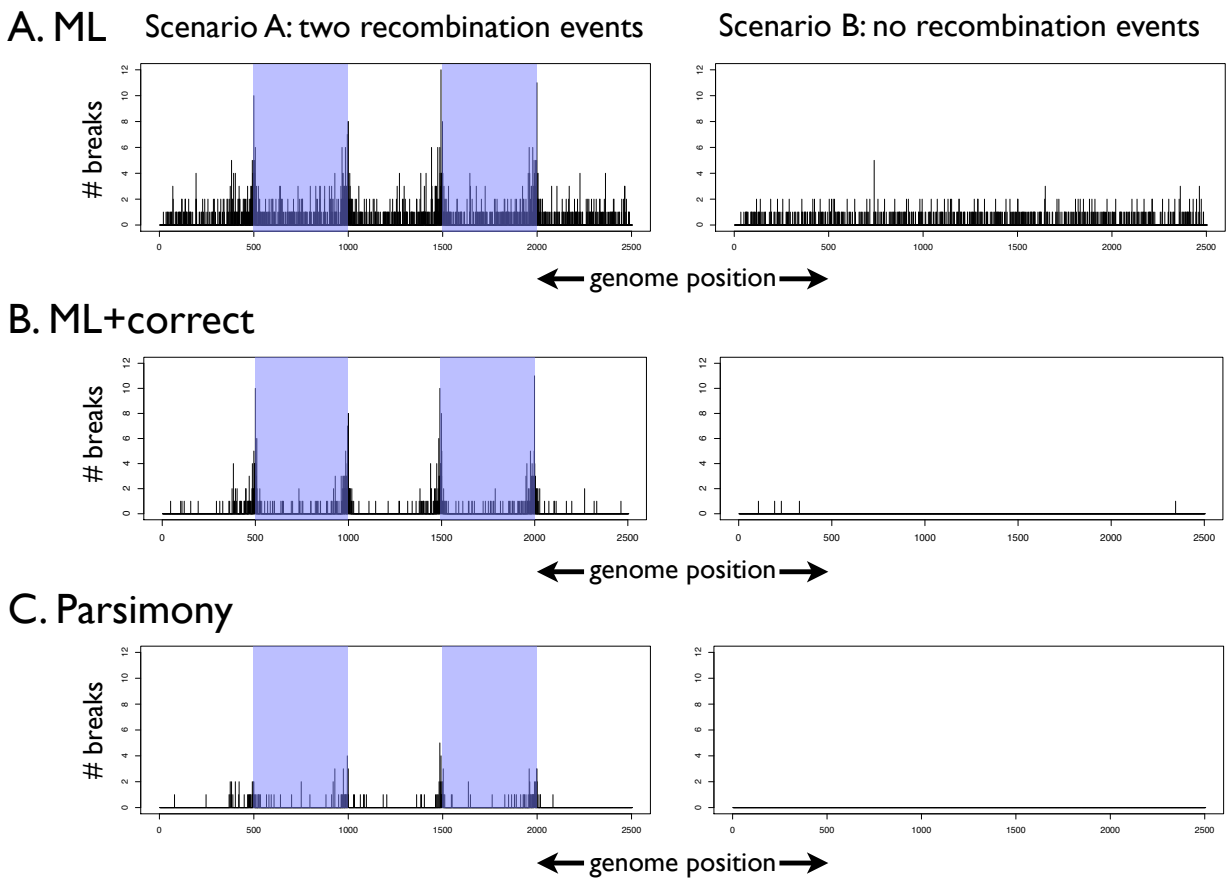

A. ML
Scenario A: two recombination events
Scenario B: no recombination events
breaks
genome position
B. ML+correct
breaks
genome position
C. Parsimony
breaks
genome position

Figure 1.S7

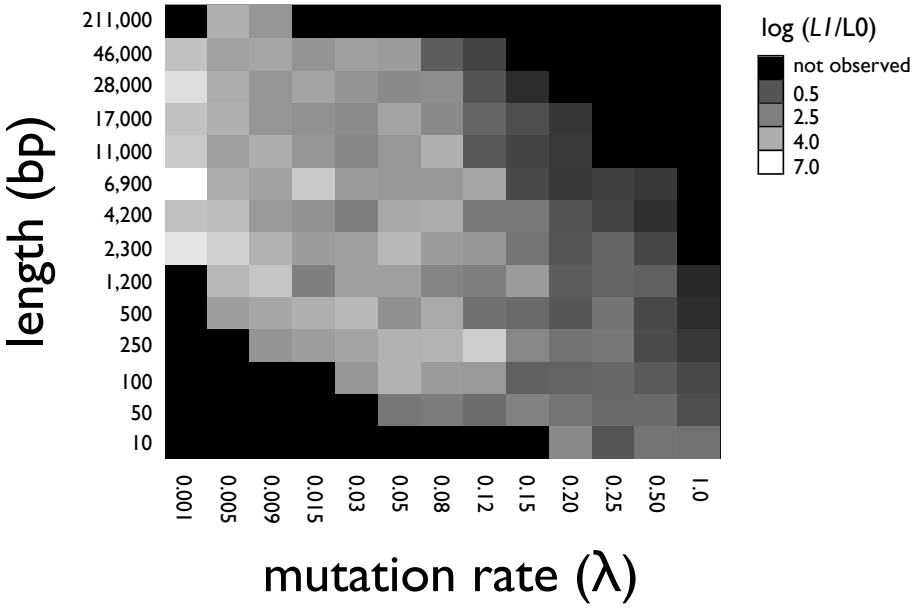

211,000
46,000
28,000
17,000
11,000
6,900
4,200
2,300
1,200
500
250
100
50
10
length (bp)
0.001
0.005
0.009
0.015
0.03
0.05
0.08
0.12
0.15
0.20
0.25
0.50
1.0
mutation rate (λ)
log (L1/L0)
not observed
0.5
2.5
4.0
7.0

Figure 1.S8

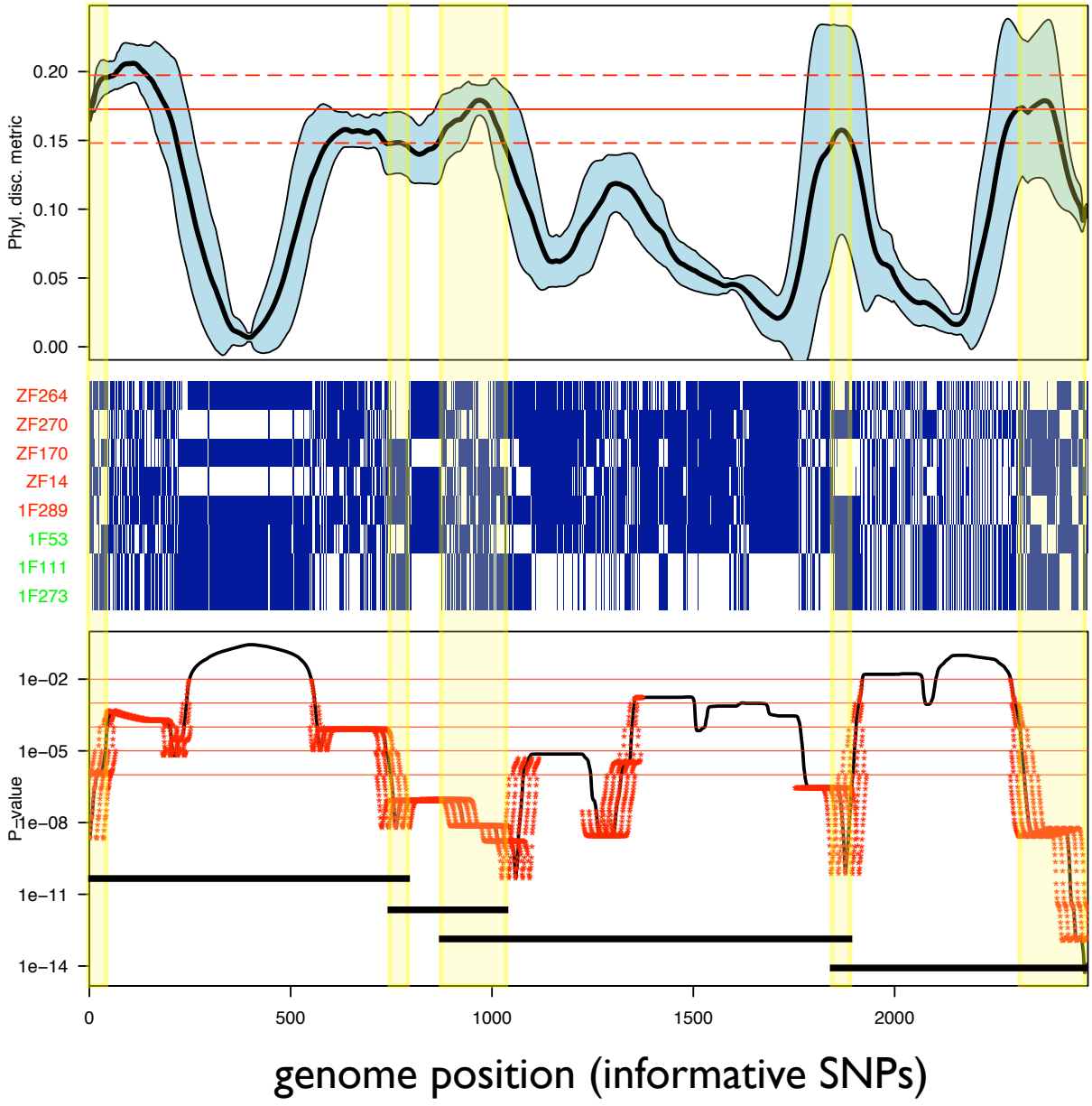

Phyl. disc. metric
0.20
0.15
0.10
0.05
0.00
ZF264
ZF270
ZF170
ZF14
1F289
1F53
1F111
1F273
P-value
1e−02
1e−05
1e−08
1e−11
1e−14
0
500
1000
1500
2000
genome position (informative SNPs)

Figure 1.S9

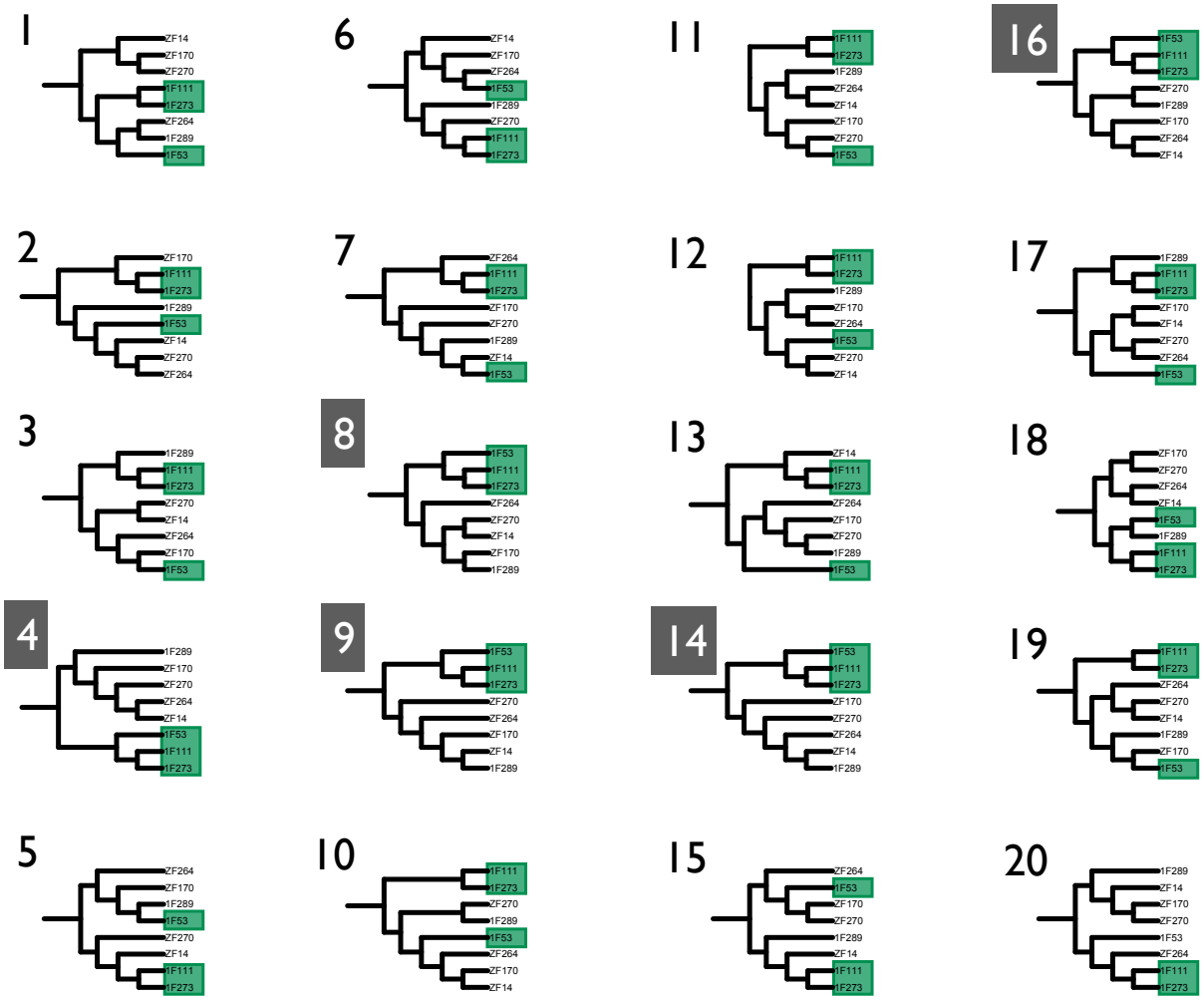

1
ZF14
ZF170
ZF270
1F111
1F273
ZF264
1F289
1F53
6
ZF14
ZF170
ZF264
1F53
1F289
ZF270
1F111
1F273
11
1F111
1F273
1F289
ZF264
ZF14
ZF170
ZF270
1F53
16
1F53
1F111
1F273
ZF270
1F289
ZF170
ZF264
ZF14
2
ZF170
1F111
1F273
1F289
1F53
ZF14
ZF270
ZF264
7
ZF264
1F111
1F273
ZF170
ZF270
1F289
ZF14
1F53
12
1F111
1F273
1F289
ZF170
ZF264
1F53
ZF270
ZF14
17
1F289
1F111
1F273
ZF170
ZF14
ZF270
ZF264
1F53
3
1F289
1F111
1F273
ZF270
ZF14
ZF264
ZF170
1F53
8
1F53
1F111
1F273
ZF264
ZF270
ZF14
ZF170
1F289
13
ZF14
1F111
1F273
ZF264
ZF170
ZF270
1F289
1F53
18
ZF170
ZF270
ZF264
ZF14
1F53
1F289
1F111
1F273
4
1F289
ZF170
ZF270
ZF264
ZF14
1F53
1F111
1F273
9
1F53
1F111
1F273
ZF270
ZF264
ZF170
ZF14
1F289
14
1F53
1F111
1F273
ZF170
ZF270
ZF264
ZF14
1F289
19
1F111
1F273
ZF264
ZF270
ZF14
1F289
ZF170
1F53
5
ZF264
ZF170
1F289
1F53
ZF270
ZF14
1F111
1F273
10
1F111
1F273
ZF270
1F289
1F53
ZF264
ZF170
ZF14
15
ZF264
1F53
ZF170
ZF270
1F289
ZF14
1F111
1F273
20
1F289
ZF14
ZF170
ZF270
1F53
ZF264
1F111
1F273

Figure 1.S10

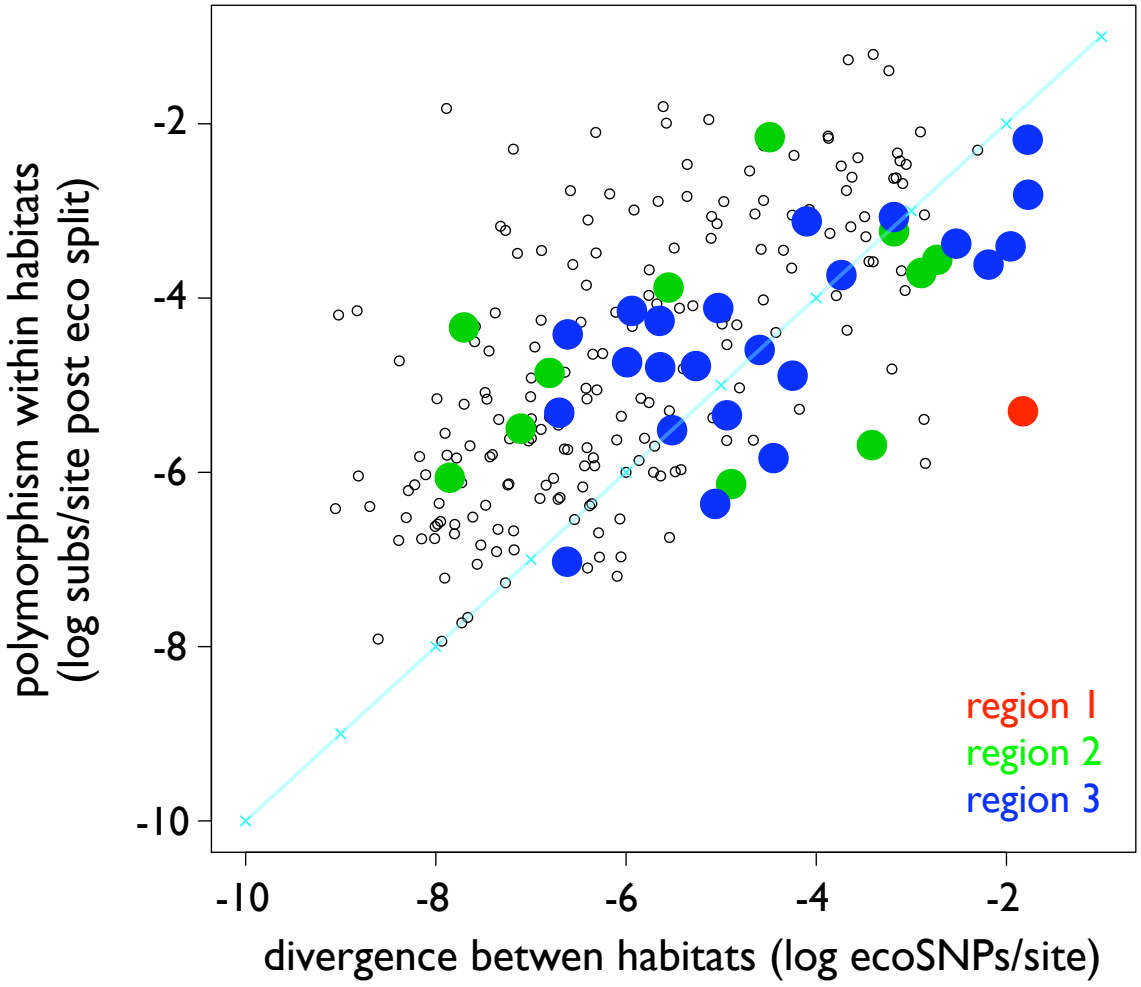

polymorphism within habitats
(log subs/site post eco split)
-2
-4
-6
-8
-10
-10
-8
-6
-4
-2
divergence betwen habitats (log ecoSNPs/site)
region 1
region 2
region 3

Figure 1.S11

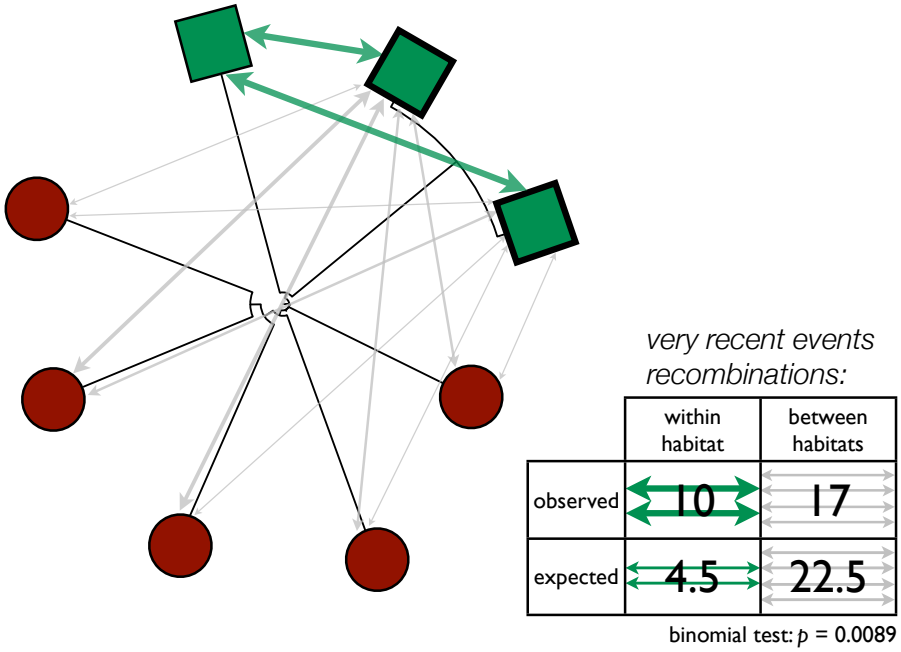


Figure 1.S12

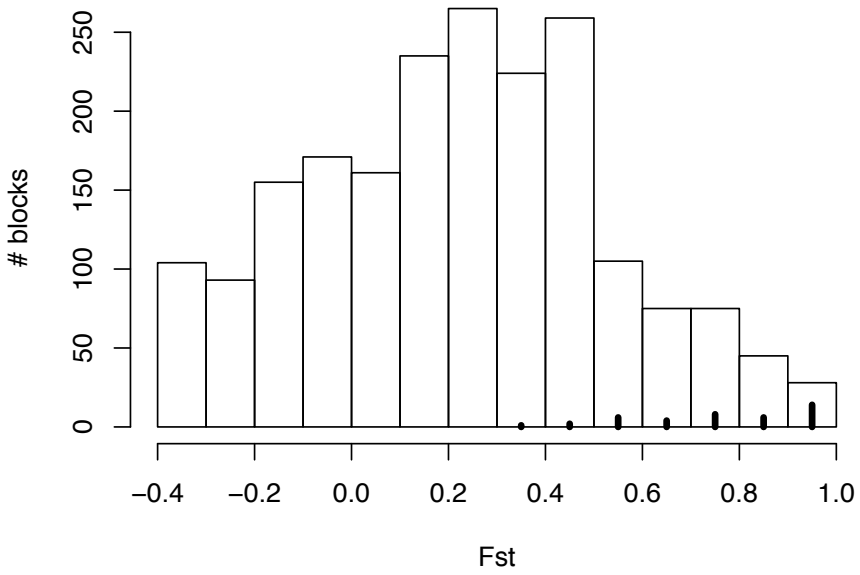


Figure 1.S13

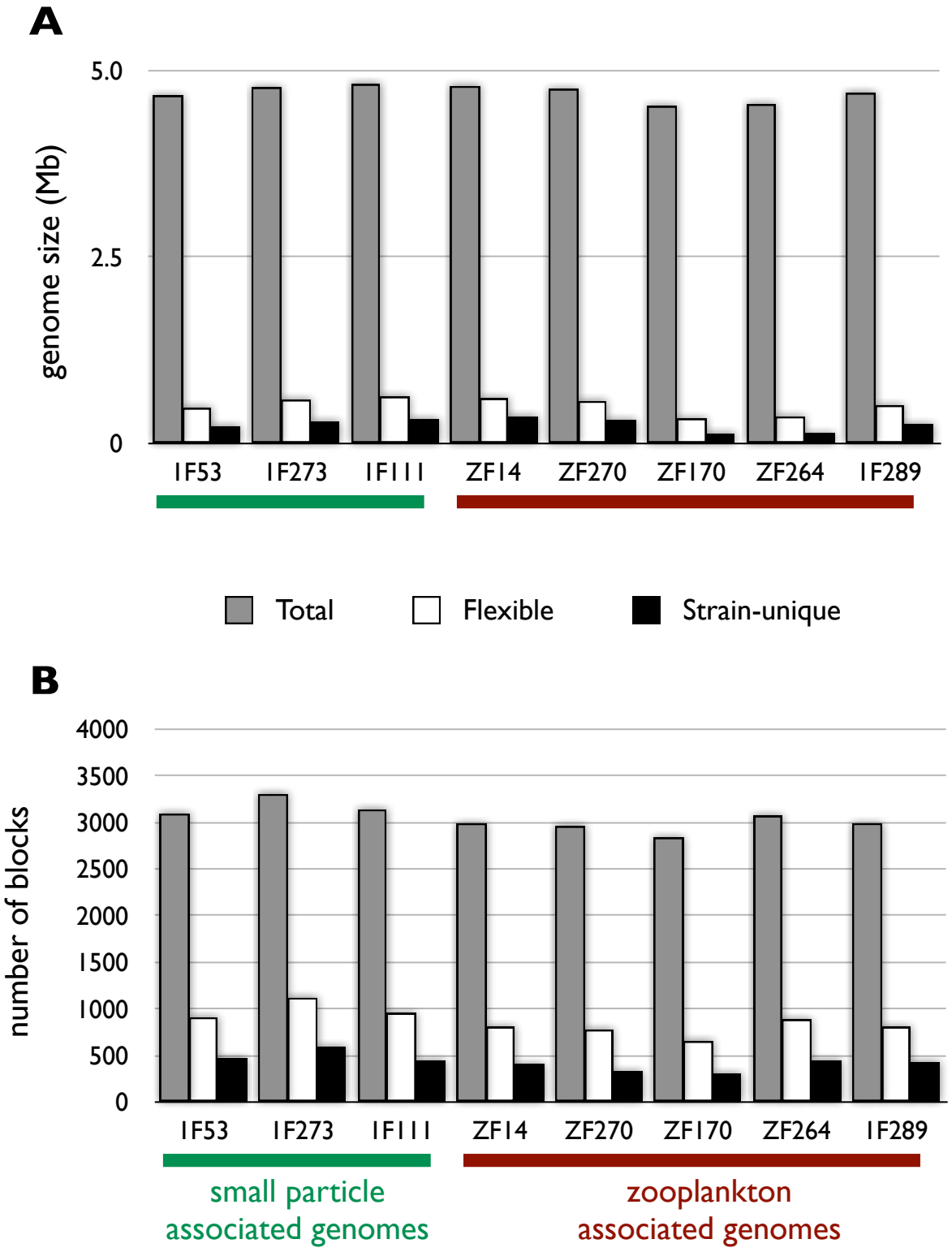

A
5.0
2.5
0
genome size (Mb)
1F53
1F273
1F111
ZF14
ZF270
ZF170
ZF264
1F289
Total
Flexible
Strain-unique
B
4000
3500
3000
2500
2000
1500
1000
500
0
number of blocks
1F53
1F273
1F111
ZF14
ZF270
ZF170
ZF264
1F289
small particle associated genomes
zooplankton associated genomes

Table 1.S1

Position						Number of mutations			Fixed between habitats			Polymorphism within habitat				M-K test			
												SP		Z					
region	contig	# blocks	start	end	length (Kbp)	total	# ecoSNPs	*Fst*	N	S	dN/dS	pN	pS	pN	pS	FI	*p*	# codons	gene
1	60	1	272305	276907	4.60	764	741	0.99	26	103	0.46	1	0	0	1	Inf	1	262	RNA polymerase sigma factor RpoS
									223	175	0.43	0	1	1	4	5.1	0.176	784	RTX family protein
2	58	14	387879	414935	27.06	815	513	0.71	1	2	1.46	2	0	4	3	0.41	1.00	428	D-alanyl-D-alanine carboxypeptidase (penicillin-binding protein 4)
									0	44	0.00	0	0	2	0	Inf	1.00	259	Phosphomannomutase
									2	138	0.04	0	0	9	60	0.09	0.00	792	Cellobiose phosphorylase
									20	97	0.55	1	0	21	68	0.63	0.22	642	N-acetyl-beta-hexosaminidase
									12	58	0.51	0	0	8	22	0.49	0.17	294	Predicted N-acetylglucosamine kinase
									19	64	0.84	0	0	6	39	1.82	0.26	573	Endoglucanase-related protein (NCBI)
									1	23	0.12	0	0	0	0	0.00	1.00	331	ABC-type dipeptide/oligopeptide/nickel transport system, ATPase component
									0	2	0.00	0	0	0	21	0.00	1.00	327	ABC-type dipeptide/oligopeptide/nickel transport system, ATPase component
									0	3	0.00	1	0	0	15	0.00	1.00	335	ABC-type dipeptide/oligopeptide/nickel transport systems, permease components
									0	1	0.00	0	0	0	7	0.00	1.00	328	ABC-type dipeptide/oligopeptide/nickel transport systems, permease components
									0	1	0.00	0	0	1	38	0.00	1.00	555	ABC-type dipeptide transport system, periplasmic component
									0	1	0.00	0	0	4	31	0.00	1.00	1129	Signal transduction histidine kinase
3	59	26	18462	39387	20.93	578	334	0.81	14	30	1.29	0	0	2	41	9.71	0.00	506	Phosphoglyceromutase
									26	58	0.82	0	0	2	7	0.94	1.00	145	Methylated DNA-protein cysteine methyltransferase
									20	99	0.45	0	0	0	33	Inf	0.02	408	Na+-driven multidrug efflux pump
									0	0	n/a	0	0	0	0	0	1	96	Co-chaperonin GroES (HSP10)
									2	8	0.67	0	1	0	26	Inf	0.07	548	Chaperonin GroEL (HSP60 family)
									0	4	0	0	1	1	5	0	1	219	3-hydroxyisobutyrate dehydrogenase and related beta-hydroxyacid dehydrogenases
									1	7	0.37	0	2	1	3	0.46	1	185	Uncharacterized low-complexity protein (NCBI)
									1	3	0.89	0	0	1	4	1.29	1	132	Uncharacterized membrane protein
									0	7	0	0	0	2	11	0	0.52	340	Lysine 2,3-aminomutase
									0	1	0	0	0	0	0	0	1	188	Translation elongation factor P (EF-P)/translation initiation factor 5A (eIF-5A)
									0	2	0	0	0	0	1	0	1	128	Fumarate reductase subunit D
									0	2	0	0	0	1	3	0	1	127	Fumarate reductase subunit C
									1	5	0.58	0	0	1	22	4.11	0.38	607	Succinate dehydrogenase/fumarate reductase, flavoprotein subunit
									0	2	0	0	0	6	14	0	1	323	Truncated, possibly inactive, lysyl-tRNA synthetase (class II)
									1	3	0.92	1	0	1	9	2.74	0.51	459	Di- and tricarboxylate transporters
									0	3	0	0	0	1	6	0	1	304	Phosphatidylserine decarboxylase
4	57	1	236326	238083	1.76	108	100	0.98	0	2	0	3	8	2	1	0	0.4	215	*Cobalamin-5-phosphate synthase*
									0	0	n/a	0	0	0	0	0	1	203	*Adenosyl cobinamide kinase/adenosyl cobinamide phosphate guanylyltransferase*
									41	56	1.70	1	1	0	1	Inf	1	205	*Fructose-2,6-bisphosphatase*
5	51	12	43135	48889	5.76	258	93	0.67	0	48	0	0	1	4	21	0	0.01	290	*Polyprenyltransferase (cytochrome oxidase assembly factor)*
									0	5	0	1	0	0	13	0	1	103	*Heme/copper-type cytochrome/quinol oxidase, subunit 4*
									0	7	0	1	0	2	43	0	1	205	*Heme/copper-type cytochrome/quinol oxidase, subunit 3*
									0	31	0	3	7	5	106	0	0.58	647	*Heme/copper-type cytochrome/quinol oxidases, subunit 1*
6	59	1	333272	334727	1.46	88	84	0.99	8	76	0.19	2	0	1	3	0.19	0.24	199	*Response regulator containing CheY-like receiver, AAA-type ATPase, and DNA-binding domains*
7	59	2	117515	119704	2.19	107	73	0.87	1	45	0.06	0	31	0	39	Inf	1	515	*Histidine ammonia-lyase*
8	60	1	398408	402633	4.23	286	72	0.69	2	4	1.56	14	105	10	57	3.19	0.22	408	*putative glycosyltransferase (NCBI)*
9	58	8	332814	368802	35.99	143	52	0.68	0	1	0	0	0	0	2	0	1	360	*Flp pilus assembly protein TadG*
									0	1	0	0	0	0	2	0	1	404	*Flp pilus assembly protein, secretin CpaC*
									2	0	n/a	1	0	5	0	0	1	431	*hypothetical protein (NCBI)*
									2	0	n/a	0	0	2	2	Inf	0.47	246	*AraC-type DNA-binding domain-containing proteins*
									0	1	0	0	0	1	1	0	1	25	*hypothetical protein (NCBI)*
									0	1	0	0	0	0	0	0	1	61	*hypothetical protein (NCBI)*
									0	2	0	0	0	0	1	0	1	678	*Translation elongation factors (GTPases)*
									1	1	2.85	0	0	2	2	1	1	397	*Phosphopentomutase*
									1	2	1.40	0	0	2	1	0.32	1	442	*Thymidine phosphorylase*
									0	1	0	0	0	0	2	0	1	420	*Nucleoside permease*
									0	1	0	0	0	3	2	0	1	270	*Mg-dependent DNase*
									0	4	0	0	0	0	25	0	1	679	*Predicted signal transduction protein containing a membrane domain, an EAL and a GGDEF domain*
									0	26	0	1	67	0	2	0	1	337	*Collagenase and related proteases*
10	58	10	279763	303637	23.88	220	44	0.62	1	22	0.14	1	1	13	108	0.47	0.70	561	Glycosidases
									1	1	2.81	0	0	2	4	1.83	1	366	3,4-dihydroxy-2-butanone 4-phosphate synthase
									1	0	n/a	0	0	0	0	0	1	156	Riboflavin synthase beta-chain
									0	2	0	0	0	0	0	0	1	319	Thiamine monophosphate kinase
									0	1	0	1	0	2	2	0	1	146	Phosphatidylglycerophosphatase A and related proteins
									0	5	0	0	0	5	27	0	1	627	Deoxyxylulose-5-phosphate synthase
									2	0	n/a	0	1	0	4	Inf	0.066667	294	Geranylgeranyl pyrophosphate synthase
11	58	2	52153	55170	3.02	192	69	0.72	5	61	0.21	18	114	3	4	0.11	0.02	543	Predicted NADH:ubiquinone oxidoreductase, subunit RnfC
12	52	2	34363	38695	4.33	118	47	0.88	13	18	1.98	0	0	9	70	6.76	0.00	332	AMP-(fatty) acid ligase
									1	9	0.29	0	0	0	2	Inf	1	185	Predicted membrane protein

Table 1.S2

block length (bp)	habitat	primers	Annotation
716	Z-unique		No hits found
3340	Z-unique		hypothetical protein V12B01_02745 [Vibrio splendidus 12B01]
3340	Z-unique		Predicted amino acid racemase [Vibrio splendidus 12B01]
3340	Z-unique		hypothetical protein V12B01_02735 [Vibrio splendidus 12B01]
581	Z-unique		ThiJ/PfpI family protein [Vibrio sp. MED222]
761	Z-unique		MSHA pilin protein MshB [Vibrio splendidus 12B01]
5221	Z-unique		MSHA biogenesis protein MshM [Vibrio splendidus 12B01]
5221	*Z-unique*	*flex1*	*MSHA biogenesis protein MshN [Vibrio splendidus 12B01]*
5221	Z-unique		Type II secretory pathway [Vibrio splendidus 12B01]
5221	Z-unique		Type II secretory pathway, component PulF [Vibrio splendidus 12B01]
1220	Z-unique		Sialic acid synthase [Vibrio splendidus 12B01]
1220	Z-unique		hypothetical protein V12B01_22960 [Vibrio splendidus 12B01]
862	Z-unique		hypothetical protein MED222_04835 [Vibrio sp. MED222]
958	Z-unique		hypothetical protein MED222_04835 [Vibrio sp. MED222]
836	*Z-unique*	*flex4*	*Glycosyltransferase [Vibrionales bacterium SWAT-3]*
3348	Z-unique		multidrug resistance protein [Chromobacterium violaceum ATCC 12472]
3348	Z-unique		predicted protein [Aspergillus terreus NIH2624]
737	Z-unique		hypothetical protein V12B01_12555 [Vibrio splendidus 12B01]
4866	Z-unique		hypothetical protein VS_II0856 [Vibrio splendidus LGP32]
2043	*Z-unique*	*flex5*	*putative intercellular adhesion protein A [Vibrio splendidus 12B01]*
2043	Z-unique		putative glycosyltransferase [Vibrio splendidus 12B01]
778	Z-unique		periplasmic protein involved in polysaccharide export [Vibrio sp. MED222]
1320	Z-unique		hypothetical protein VS_II0856 [Vibrio splendidus LGP32]
1066	Z-unique		Outer membrane protein [Vibrio splendidus 12B01]
1066	Z-unique		periplasmic protein involved in polysaccharide export [Vibrio splendidus 12B01]
591	Z-unique		hypothetical protein V12B01_22960 [Vibrio splendidus 12B01]
5268	Z-unique		putative polysaccharide export protein [Vibrio splendidus 12B01]
5268	Z-unique		glycosyltransferase [Vibrio splendidus 12B01]
5268	*Z-unique*	*flex3*	*probable maltose O-acetyltransferase [Vibrio splendidus 12B01]*
5268	Z-unique		putative membrane protein of ExoQ family [Vibrio splendidus 12B01]
5268	Z-unique		putative hexosyltransferase [Vibrio splendidus 12B01]
1160	Z-unique		hypothetical protein MED222_04835 [Vibrio sp. MED222]
676	*Z-unique*	*flex2*	*MSHA biogenesis protein MshF [Vibrio splendidus 12B01]*
580	Z-unique		periplasmic protein involved in polysaccharide export [Vibrio splendidus 12B01]
517	SP-unique		No hits found
1787	SP-unique		permease of the major facilitator superfamily [Vibrio sp. Ex25]
1931	SP-unique		hypothetical protein V12G01_06656 [Vibrio alginolyticus 12G01]
760	SP-unique		hypothetical protein VPMS16_1245 [Vibrio parahaemolyticus 16]
636	SP-unique		No hits found
772	SP-unique		No hits found
505	SP-unique		No hits found
1029	SP-unique		No hits found
2237	SP-unique		SH3 domain protein [Vibrio cholerae 1587]
1279	SP-unique		ISSod13, transposase [Vibrio parahaemolyticus]
907	SP-unique		No hits found
5412	SP-unique		ISPsy5 [Marinobacter sp. ELB17]
5412	SP-unique		PAS/PAC sensor signal transduction histidine kinase [Saccharophagus degradans 2-40]
627	SP-unique		transposase [Alteromonas macleodii 'Deep ecotype']
4473	SP-unique		No hits found
565	SP-unique		No hits found
1466	SP-unique		hypothetical protein V12B01_00967 [Vibrio splendidus 12B01]
2405	SP-unique		No hits found
604	SP-unique		No hits found
876	SP-unique		No hits found
536	SP-unique		No hits found
7413	SP-unique		hypothetical protein PXO_01414 [Xanthomonas oryzae pv. oryzae PXO99A]
7413	SP-unique		putative pore-forming cytotoxin integrase [Shewanella sediminis HAW-EB3]
7413	SP-unique		hypothetical protein VP2137 [Vibrio parahaemolyticus RIMD 2210633]
7413	SP-unique		hypothetical protein VSWAT3_17015 [Vibrionales bacterium SWAT-3]
7413	SP-unique		hypothetical protein VSWAT3_17020 [Vibrionales bacterium SWAT-3]
7413	SP-unique		hypothetical protein VSWAT3_17025 [Vibrionales bacterium SWAT-3]
2764	SP-unique		FG-GAP/YD repeat domain protein [Vibrio splendidus 12B01]
620	SP-unique		transposase [Vibrio splendidus 12B01]
955	SP-unique		No hits found
1386	SP-unique		2`,3`-cyclic-nucleotide 2`-phosphodiesterase, putative [Vibrio splendidus 12B01]
1012	SP-unique		transposase [Vibrionales bacterium SWAT-3]
2518	SP-unique		No hits found
3092	SP-unique		hypothetical protein VSWAT3_06691 [Vibrionales bacterium SWAT-3]
1683	SP-unique		methyl-accepting chemotaxis protein [Vibrio vulnificus YJ016]
1347	SP-unique		hypothetical protein V12B01_09036 [Vibrio splendidus 12B01]
766	SP-unique		hypothetical protein VSWAT3_12067 [Vibrionales bacterium SWAT-3]
2814	SP-unique		hypothetical protein V12B01_12660 [Vibrio splendidus 12B01]

Table 1.S3

	primers					
strain	**flex1**	**flex2**	**flex3**	**flex4**	**flex5**	***hsp60***
ZF28	+	+	+	+	+	+
ZF30	+	+	+	+	+	+
ZF205	+	+	+	+	+	+
1F97	-	-	-	-	-	+
1F127	-	-	-	-	-	+
1F124	-	-	-	-	-	+
1F175	-	-	-	-	-	+
FF160	+	+	-	-	-	+
FF274	-	-	-	-	-	+

Table 1.S4

name	**direction**	**# bases**	**annealing temp.**	**sequence**
coreA	F	22	50	TTC CTT TTT AAC AAA CTG CAT C
	R	23		TAT CAA GAC AGC AAA GAC TTA CA
coreG	F	18	50	GCG CCT TCR ATA GCG TCA
	R	18		GGC AGT GCC AAC ATC CTT
coreB	F	18	50	TAA ATA AGG ACA AGG AAA
	R	18		TTM GAM ACA TTY ACA TCC
coreC	F	16	50	AGC AGT GAT AAG CAG T
	R	16		TGA AAT GGC AGC AGA A
coreD	F	18	50	GCG GTT GTT GTA GTG AGT
	R	18		GCA GAG GGT GTA TTC GGT
coreI	F	19	50	CAT CAT ACG GCG AGT TTC A
	R	19		AGA AAG GTA AGC AAG AGC A
coreJ	F	18	50	GTA AAG CGT AAG GAG TCT
	R	18		GTT TGA CCT GGA GTA CTG
flex1	F	18	54.6	GTC AAA ACA CGC TCT CCA
	R	18		TCG ATC TCG CAC AAA ACT
flex2	F	19	48.2	TTT TGT TAT TTG GAG TGT G
	R	19		CTT TTT CGA GCA ATA TAT C
flex3	F	18	51.2	GCA TGA TCA CCT CCA CGA
	R	18		ACC TAC CCC ACA AAT ACT
flex4	F	18	52.2	ACA CCT TAC ATC GGA TCA
	R	19		GTA CAA CCA CAC TTT TCC T
flex5	F	19	55	CGC TTC GCA CTC TTT AAC A
	R	18		CCA CGC AGC ATA TCC AAT

Table 1.S5

		Divergence (in vs. outgroup)				**Polymorphism (within ingroup)**				**Bootstapped M-K test**				**# sites**	
ingroup	**outgroup**	**FN**	**FS**	**FN/FS**	**dN/dS**	**PN**	**PS**	**PN/PS**	**pN/pS**	**FI obs**		**FI exp**	***p***	**synonymous**	**nonsynon.**
SP+Z	12B01+F01	40,762	336,510	0.12	0.042	7,595	52,238	0.15	0.050	0.83	<	1.075	<0.001	700,089	2,039,508
SP	Z	369	2,008	0.18	0.063	4,752	24,136	0.20	0.068	0.93	<	0.989	0.001	701,156	2,042,458
Z	SP	369	2,008	0.18	0.063	8,104	46,594	0.17	0.060	1.06	>	0.989	0.008	701,156	2,042,458

Chapter 2:

Comparing Patterns of Natural Selection Across Species Using Selective Signatures

Shapiro BJ and Alm EJ (2008) Comparing Patterns of Natural Selection across Species Using Selective Signatures. *PLoS Genetics* **4**:e23

Abstract

Comparing gene expression profiles over many different conditions has led to insights that were not obvious from single experiments. In the same way, comparing patterns of natural selection across a set of ecologically distinct species may extend what can be learned from individual genome-wide surveys. Toward this end, we show how variation in protein evolutionary rates, after correcting for genome-wide effects such as mutation rate and demographic factors, can be used to estimate the level and types of natural selection acting on genes across different species. We identify unusually rapidly and slowly evolving genes, relative to empirically derived genome-wide and gene family-specific background rates for 744 core protein families in 30 γ-proteobacterial species. We describe the pattern of fast or slow evolution across species as the 'selective signature' of a gene. Selective signatures represent a profile of selection across species that is predictive of gene function: pairs of genes with correlated selective signatures are more likely to share the same cellular function, and genes in the same pathway can evolve in concert. For example, glycolysis and phenylalanine metabolism genes evolve rapidly in *Idiomarina loihiensis*, mirroring an ecological shift in carbon source from sugars to amino acids. In a broader context, our results suggest that the genomic landscape is organized into functional modules even at the level of natural selection, and thus it may be easier than expected to understand the complex evolutionary pressures on a cell.

2.1 Introduction

An enormous genetic diversity exists on earth, particularly in the microbial domains of life - yet how much diversity is functional, and what are the important adaptations that serve to partition species into different niches? Adaptive differences can be identified in genes subject to positive Darwinian selection - the evolutionary force that causes advantageous genetic traits to spread in populations, allowing species to diverge ecologically. Natural selection acts not just on individual proteins, but on the complex assemblage of proteins specified by an organism's genome. Thus, looking for natural selection across the entire genome is valuable for two reasons. First, it allows us to identify systems-level patterns of adaptation - for example, selection on consecutive enzymes in a metabolic pathway. Secondly, it provides a built-in empirical distribution against which outliers (candidates for selection) can be evaluated. In addition, by simultaneously considering multiple genomes, we can compare relative amounts of selection on a gene in different species subject to different ecological constraints.

Much recent work has focused on genome-wide scans for positive selection in human (Chimpanzee Sequencing and Analysis Consortium, 2005, Consortium, 2005) and other eukaryotic species (e.g. *Drosophila*, *Plasmodium* (Shapiro et al, 2007, Volkman et al, 2007)). Many of these scans rely on skews in polymorphism patterns at a genomic locus as a selectively favored allele increases in frequency and becomes fixed in the population (Thornton et al, 2007). To identify such selected loci requires that their polymorphism patterns be unlinked from the rest of the genome by recombination, making them stand out as regions of reduced variation, or unexpectedly long haplotypes (Sabeti et al, 2006). It is thus unclear whether any of these 'diversity-based' tests (e.g. Tajima's D (Tajima, 1989), Fay & Wu's *H* (Fay & Wu, 2000)) for positive selection on sexual genomes - which rely on the assumption that recombination occurs between genomic loci - will be amenable to bacteria, in which recombination is decoupled from reproduction, and thus may occur very rarely, or across species boundaries (due to horizontal gene transfer; HGT).

Alternative 'rate-based' approaches to detecting positive selection (in both sexual and asexual species) include finding genes with high rates of amino acid substitution - relative to (i) the rate of evolution in other lineages (relative rates), or (ii) the number of silent substitutions in the gene (nonsynonymous : synonymous substitution ratio; dN/dS) (Anisimova & Liberles, 2007). These approaches may lack power when positive selection only affects a small number of sites (Sabeti et al, 2006, Hughes, 2007), and the latter may be inappropriate as dS becomes saturated with multiple substitutions over very long time scales. Both approaches may have difficulty distinguishing between positive selection (fixation of beneficial mutations) and relaxed purifying selection (loss of constraint, fixation of neutral or deleterious

mutations, for example during population bottlenecks). These two types of selection can, however, be better distinguished by normalizing out species-wide bottleneck effects, and when polymorphism data is available, using independent methods such as the McDonald-Kreitman (MK) test, which compares the rate of synonymous and nonsynonymous substitutions within and between groups (McDonald & Kreitman, 1991).

In this study, we focus on relative evolutionary rates because our model system, the γ-proteobacteria, span a considerable evolutionary time period over which synonymous substitution rates are saturated in many branches, and because polymorphism data from *Escherichia coli* provide an independent means to estimate the relative contributions of positive selection and relaxed negative selection to elevated evolutionary rates. Nonetheless, we show results from dN/dS profiling for comparison.

The biological factors driving protein evolutionary rates are complex and widely debated (Fraser et al, 2002, Rocha & Danchin, 2004, Drummond et al, 2005, Wall et al, 2005, Saeed & Deane, 2006) (for recent reviews see (McInerney, 2006, Rocha, 2006)). In addition, selection may lead to subtle lineage-specific variation in evolutionary rates. To identify potentially important rate variation from the background of gene family and genome-specific rates, we factor evolutionary rates into three components that contribute to the total evolutionary distance (amino acid substitutions per site) as defined in Equation 1 (where r is the total evolutionary rate, and t is time):

$$\text{evolutionary distance} = r \cdot t = \rho(\text{gene family}) \cdot \beta(\text{genome}) \cdot \nu(\text{gene,genome}) \cdot t \quad (1)$$

The first and most significant background component (ρ in Eq. 1) is related to the protein family: for example, the ribosomal machinery is known to evolve slowly across all sequenced microbes, while surface-exposed proteins often evolve rapidly to avoid recurrent predation and antibiotic recognition. The second major contribution (β in Eq. 1) is the background rate of evolution that results from the 'molecular clock' associated with each lineage, perhaps due to between-species differences in population size, generation time, constraint on codon usage, or environmental factors such as UV light exposure (Mering et al, 2007). For example, due to such demographic factors, genes from the intracellular parasites of the *Buchnera* genus evolve more rapidly than those in other Enterobacteria. This may be due to frequent population bottlenecks, allowing more frequent fixation of neutral or slightly deleterious alleles, or an increased mutation rate (Moran, 1996, Itoh et al, 2002). Of course, ρ and β are not always independent, and are expected to interact, resulting in evolutionary rate-variation that is both gene-specific and species-specific (ν in Eq. 1). When a gene evolves at the rate predicted by its gene family and genome, ν will be

equal to one. However, when ν deviates from one, this may represent natural selection acting to favor different functionality in different genomic/ecological milieus,

Deviations from the 'expected' rate of protein evolution can be used to detect positive selection and functional diversification between orthologous proteins (Sarich & Wilson, 1967, Muse & Weir, 1992, Jordan et al, 2001), and the 'expected' background is best estimated empirically, by measuring rates across the entire genome. A recent study demonstrated global differences in evolutionary rate between environments (Mering et al, 2007), but did not attempt to identify patterns of natural selection on genes in different genomes. The growing number of organisms with fully sequenced genomes provides an opportunity to look for patterns of selection on genomes, and to begin to address a question of fundamental interest: to what extent does differentiation in core, 'housekeeping' genes drive functional divergence between species across the tree of life? And can we identify genes under selection in different species, and make predictions about their biological/ecological significance?

2.2 Results

Using a well-sampled sub-tree of γ-proteobacterial genomes, we detected deviations from the 'expected' rate of evolution (controlling for ρ and β, as described in the Methods and Figure 2.S1), by estimating ν (Eq. 1) for each of 744 'core' proteins present in single-copy in the majority of species. Of these protein families, 718 (97%) reject a single molecular clock for all species (Likelihood ratio test, $P < 0.05$), indicating substantial species-specific rate variation over the long time scales considered here. As recently shown to be the case among species of fruit flies and fungi (Rasmussen & Kellis, 2007), protein-family and genome-wide effects account for most (80%) of the variation in evolutionary distances among orthologous proteins (Figure 2.1; Pearson correlation = 0.89, $P < 2.2e-16$); we used the residual variation on each branch as an estimate of ν, and calculated a Z-score (ratio of the mean of ν to its standard deviation over bootstrap-resamplings from the sequence data) to assess confidence in any deviation from ν=1. As expected, ν correlates well with dN (Pearson's correlation = 0.44, $P < 2.2e-16$), and the correlation is improved substantially once dN is also normalized for protein-family and molecular effects (Pearson's correlation = 0.78, $P < 2.2e-16$). Interestingly, ν correlates less well with normalized dN/dS (Pearson's correlation = 0.11, $P < 2.2e-16$), perhaps due to dS becoming saturated over the long time scales considered, or simply because dN/dS and relative rates (ν) detect different types and magnitudes of selection, thus predicting different sets of selected genes.

When relative rates are overlaid onto the species tree (Letunic & Bork, 2007), patterns of selection across both genes and species become apparent. For example, genes involved in flagellar biosynthesis (*e.g. flgN*, *flgA* and *fliS*) are unusually fast-evolving in species of enterobacteria, while genes putatively involved in sulfur oxidation (*yheL* and *yheM*) are unusually slow-evolving in species of *Buchnera* (Figure 2.2). As described below, genes involved in the same biological function (*e.g.*, flagellar biosynthesis or sulfur oxidation) tend to have a similar 'selective signature' (pattern of fast or slow evolution across species). In other words, they evolve in a manner more similar to each other than to genes of a different function. This similarity could be due to genes of the same function being encoded on the same operon (as is the case for *flgA*/*flgN* and *yheL*/*yheM*, respectively). Yet *fliS*, which is encoded on a different operon than *flgA*/*flgN*, has a selective signature similar to as the other flagellar genes (Figure 2.2), suggesting selection on gene function.

2.3 Selection acts coherently at the level of function.

In addition to the anecdotal cases described above, we examined more generally whether genes of

common function tend to experience similar regimes of selection. Indeed, in our overall data set, pairs of genes sharing the same COG (clusters of orthologous groups (Tatusov et al, 1997)) functional annotation have significantly more correlated selective signatures (the vector of ν across all species) than pairs with different functions (Kolmogorov-Smirnov (KS) test, D=0.12, P<2.2e-16; Figure 2.3A). This indicates that selection can act coherently at the level of function, and across levels of organization larger than single genes. Considering each functional category in isolation, we find that most functions (11 of 16 COG function categories, excluding 'general' and 'unknown' categories) contribute significantly to this effect. Thus, selective signatures are a surprisingly good predictor of common function – a feature that could be useful in the annotation of genes of unknown function, provided that they have orthologs in several species. Correlation in ν is also a significantly better predictor of function than correlation in dN/dS (Figure 2.3A), or raw evolutionary distance, and the predictive power remains strong even after removing genes used to construct the species tree or genes on the same operon (Figure 2.S3). When dN/dS is normalized by its median for each ortholog and genome to produce a 'relative' dN/dS measure, it correlates much better with function, almost equal to ν, highlighting the generality of the empirical multi-species approach used in this study.

Our data set of 744 genes is enriched in highly conserved 'housekeeping' genes (median dN/dS = 0.047, with 70% of dN/dS values (within 1 standard deviation on a $\log_2$ scale) ranging from 0.005 to 0.26). Despite this uniformly low range of dN/dS, the subtle rate variation captured by selective signatures is able to identify co-dependencies between genes of related functions. We explicitly tested the ability to detect co-dependencies between genes by simulating codon data for 30 species under 36 different models of evolution, half of which allowed dN/dS to vary on different branches, chosen at random. All models allowed dN/dS to vary among sites. However, for any site, dN/dS was only allowed to range within 1 standard deviation of the mean of the observed data (0.005 to 0.26). For each of the 36 models, 5 replicate data sets were generated, and we treated replicates as genes with known evolutionary co-dependence. We computed ν for each of the resulting 180 simulated genes, and found that in models with branch variation in dN/dS, replicates of the same model had significantly more correlated ν across species than expected (KS test versus all models, D=0.58, P<2.2e-16; Figure 2.3B). Thus, when at least some branch variation is present, selective signatures are able to uncover genes with similar evolutionary patterns, even amidst a strong background of purifying selection.

2.4 Patterns of selection reflect ecology.

The relationship between selective signatures and gene function is borne out in several genomes in our study. For example, evolution of flagellar proteins appears to be most rapid in some species of

Enterobacteria, perhaps reflecting diversifying (positive) selection from 'arms races' with hosts or predators. In contrast, ion transport/metabolism proteins, especially those involving sulfur, are slowest evolving in *Buchnera aphidicola APS* (Tables 2.S3), indicating the importance of these proteins in the lifestyle of this intracellular symbiont.

A deep-sea bacterium that lives at the periphery of hydrothermal vents, *Idiomarina loihiensis*, presents a particularly interesting case study. Having lost many genes essential for sugar metabolism, it relies instead on amino acids as its primary source of energy and carbon (Hou et al, 2004). Consistent with disuse of sugar metabolism, we find that glycolysis genes, as well as an upstream phosphotransferase system component (COG2190) have some of the highest values of ν in the *Idiomarina* genome, suggesting relaxed negative selection on this pathway (Figure 2.4). Moreover, carbohydrate transporters and key glycolytic enzymes in the pentose phosphate and Entner-Doudoroff pathways have been lost in *Idiomarina*, and two of these relatively rapidly-evolving enzymes have been lost (COG166 and COG2190) in *Colwellia*, the most closely related sister-taxon of *Idiomarina* in our study. Taken together, these results suggest the relaxation of purifying (negative) selection on this pathway resulting from the disuse of sugars as a carbon source. By contrast, the relatively rapid evolution of amino acid metabolic enzymes in *Idiomarina* might reflect adaptation to growth on amino acids, particularly phenylalanine (Figure 2.4). Further supporting the idea of a species-specific adaptation in *Idiomarina*, none of the rapidly-evolving phenylalanine metabolism genes are also rapidly-evolving in *Colwellia*, nor have they been lost in this sister species. The 7 glycolysis genes and 3 phenylalanine biosynthesis genes were also analyzed in PAML (Yang & Nielsen, 2002, Yang, 2000), using models allowing dN/dS to vary among sites and branches, or branches only (Table 2.S4). In the branch-only models, none of these genes had significantly high average dN/dS in *Idiomarina*, but the branch-site models found evidence for a few sites in each gene with unusually high dN/dS in *Idiomarina*. While selective signatures cannot distinguish positive from relaxed negative selection on these genes, the known ecology and genome dynamics suggest positive selection on phenylalanine metabolism and relaxed negative selection on sugar metabolism. Although the true patterns of selection may be more complex, our results paint a broad picture of how the *Idiomarina* core metabolism has been optimized for a diet of amino acids rather than sugars, and lay a path for more targeted follow-up studies.

2.5 Contributions of purifying and positive Darwinian selection.

For the cases above, we used biological intuition to discriminate the roles of positive and negative selection on gene evolutionary rates. In general though, natural selection may act to accelerate changes in

a protein's sequence (positive selection; $\nu > 1$) or to slow down and constrain its rate of change (negative selection; $\nu < 1$). Alternatively, when negative selection is relaxed, the apparent rate of evolution may increase due to fixation of slightly deleterious mutations (relaxed negative selection; $\nu > 1$). Because these scenarios cannot be distinguished by relative rates methods alone, we employed an independent test for selection (the McDonald-Kreitman (MK) test (McDonald & Kreitman, 1991)) using polymorphism data from 473 genes from 24 fully sequenced *E. coli* strains, with *Salmonella enterica* as an outgroup. In the MK test, rather than normalizing according to a sample of distantly-related species (as in the selective signatures approach), we normalize according to the expected dN/dS from a within-species polymorphism sample. Specifically, the ratio of synonymous (S) and nonsynonymous (NS) changes at polymorphic sites (within the 24 strains) is compared to the ratio at (non-polymorphic) divergent sites (comparing *E. coli* to *S. enterica*). The Fixation Index is calculated as FI = (divergent NS / S) / (polymorphic NS / S) [3]. Under neutral evolution, FI is expected to equal 1; under positive selection it may exceed 1, and under negative selection it may be less than 1. We compared the FI values of the 473 genes to their corresponding selective signatures (ν) in *E. coli* and found a significant positive correlation (Pearson's correlation = 0.23, P = 6.5e-7). Although relaxation of negative selection in either the *E. coli* or *S. enterica* lineage could generate high values of FI, at least some of the genes with the highest values of FI are expected to be under positive selection (Charlesworth & Eyre-Walker, 2006). This demonstrates that relative rate acceleration is often associated with positive selection, and deceleration with purifying selection (for a complete list of selected genes identified by both methods, see Table 2.S5). The correlation between ν and FI is striking because, although the same set of gene families were used to calculate relative rates and the FI, the former used protein sequence while the latter used DNA, and the alignments were performed independently using different sets of species. These results imply that many genes have experienced selective changes since the divergence of *E. coli* and *Salmonella*, despite low overall values of dN/dS.

When the distributions of FI values are compared between genes with fast ($\nu > 2$) versus slow ($\nu < 0.5$) relative rates (Figure 2.5A), the difference is very clear. Fast-evolving genes have significantly higher FI values than slow-evolving genes (one-sided KS test; D = 0.43, P = 4.1e-6). The fast and slow subsets are also both significantly different from the mid-range ($0.5 < \nu < 2$) subset of genes (one-sided KS tests: D = 0.17; P = 0.04, and D = 0.30; P = 2.7e-5, respectively for fast and slow). Moreover, the distribution of FI values for fast-evolving genes has a broad shoulder with mean slightly less than 1, and a sharper peak with mean greater than 1 (note the $\log_2$ scale in the figure). The simplest interpretation of these results is that increased relative rate reflects both relaxed negative selection and positive selection. Interestingly, the two hypothesized distributions appear to contain a similar number of genes, suggesting that positive selection is about as common as relaxed negative selection as a cause for acceleration of evolutionary

rate. This result is largely in agreement with the previous finding that ~50% of amino acid substitutions between *E. coli* and *S. enterica* were fixed by positive selection (Charlesworth & Eyre-Walker, 2006), with the remaining substitutions due to relaxed negative selection, or hitchhiking with positively selected mutations (discussed below).

Unusually slowly evolving genes ($\nu < 0.5$), on the other hand, show greater levels of negative selection (low FI) than normal genes ($0.5 < \nu < 2$). While these results may seem unsurprising at first, it is important to note that our evolutionary rates have been normalized for gene family-specific effects, thus even the fastest evolving genes (in terms of 'raw' rate) will appear 'slow-evolving' ($\nu < 1$) in about half of the genomes. Conversely, the slowest evolving genes (*e.g.*, the ribosomal machinery) will appear to be 'fast-evolving' ($\nu > 1$) in about half of the genomes.

To further investigate the role of negative selection, we used gene deletions within a clade as evidence of relaxed negative selection, with the expectation that genes under relaxed selective constraint are lost more frequently. Consistent with a significant role for negative selection in constraining rate variation, genes evolving much more slowly than expected ($\nu < 0.25$) were less likely to have undergone deletion in a sister clade (Figure 2.5B). Conversely, genes evolving much faster than expected ($\nu > 4.0$) were more likely to have lost their ortholog in a sister clade, pointing toward relaxed negative selection.

2.6 Evidence for genetic hitchhiking in bacteria.

In sexually recombining organisms, positively-selected mutations are thought to sweep rapidly through the population, lowering effective population size and decreasing the effectiveness of negative selection at linked loci. When sweeps occur faster than recombination can separate the beneficial allele from 'hitchhikers', clusters of rapidly-evolving genes (*i.e.*, one gene under positive selection, and linked genes under relaxed negative selection) can arise (Sabeti et al, 2006). Perhaps unexpectedly for an asexual species, selective sweeps and genetic hitchhiking between linked (~30 kb apart), but not unlinked loci, have been documented in *E. coli* (Guttman & Dykhuizen, 1994). Theoretically, there exist regimes of selective sweeps and recombination in asexual prokaryotes that would be able to produce a pattern of genetic hitchhiking (Majewski & Cohan, 1999). Early work on variation across ~1700 strains of *E. coli* showed genetic linkage between loci separated by ~45 kb (Whittam et al, 1983) - an estimate largely supported by recent whole-genome scans, which find recombinational segments of up to 100 kb (Mau et al, 2006). To determine whether genetic hitchhiking was detectable among fast-evolving genes in this study, we examined proximal pairs of genes (separated on the chromosome by 0-5 genes) and asked whether they showed a tendency to co-evolve - either both 'fast' ($\nu > 1$), or both 'slow' ($\nu < 1$). Proximal

genes are frequently encoded on the same operon, and are thus expected to be under similar selective pressures due to co-expression and common function. Indeed, we find that pairs of genes predicted to be on the same operon (Price et al, 2005) co-evolve in the same direction (either both genes with $\nu > 1$, or both with $\nu < 1$, Z-score > 1; Fisher's Exact Test: Odds Ratio = 3.1, P < 2.2e-16). In fact, selective signature (correlation in ν across species) is a better predictor of operons than dN/dS, and about as accurate as a small compendium of gene expression data from *E.coli* under different experimental conditions (Figure 2.S2). Because these operon effects could confound the detection of hitchhiking, we restricted our analysis to pairs of genes on different operons, or separated by at least one gene on the opposite strand of DNA. In this operon-free data set, we observe a slight but statistically significant tendency for fast-evolving genes ($\nu > 1$), but not slow-evolving genes ($\nu < 1$), to cluster together in a genome, not only at distances of 0-5 intervening genes, but even as far as 20-100 genes apart (Figure 2.5C). Assuming an average gene length of ~1 kb in prokaryotes (Xu et al, 2006), clustering of fast-evolving genes up to 100 genes apart (Figure 2.5C) is very much consistent with earlier predictions (Majewski & Cohan, 1999, Mau et al, 2006, Guttman & Dykhuizen, 1994, Whittam et al, 1983). Alternatively, genomic mutational hotspots might explain the observed clustering, but this hypothesis is currently difficult to test. Therefore, we tentatively conclude that selective sweeps are occurring in a significant fraction of the 30 species analyzed in this study, and that these sweeps leave a detectable signal in the form of accelerated evolutionary rates.

Taken together, the observed correlations between ν and the Fixation Index (MK test), deletion frequency, and 'hitchhiking' lead us to conclude that ν is reflective of both positive and negative natural selection on core genes.

2.7 Discussion

We have described an approach to detecting selection across genes and genomes. By applying a simple, empirical normalization, we have identified unusually fast- and slow-evolving genes in a phylogeny of 30 bacterial species. Many of these genes are likely targets of natural selection, and are thus among the most important in shaping phenotypic and ecological divergence among species. As genome sequencing outpaces phenotypic and functional characterization, efforts to identify the genetic basis underlying ecological differentiation will rely increasingly on sequence-based approaches. Our approach is widely applicable across the tree of life, as it requires only a set of sequenced genomes with common orthologs. Selective signatures have the advantage of detecting subtle gene- and lineage-specific variation in evolutionary rates, but the disadvantage of being limited to core orthologs with representatives in several genomes. For this reason, the timescale and resolution of our approach will depend on the set of species

included in the analysis. This study was restricted to extant species (terminal branches), but could easily be extended to include ancestral species (internal branches), providing insight into ancient selective pressures and adaptations.

Relative rates provide information about which genes are evolving unusually rapidly or slowly, but not about what type of natural selection is responsible. We have complemented our between-species relative evolutionary rate estimations with within-species polymorphism data from *E. coli* to show that relative rates are a reasonable and easily-estimated predictor of positive and negative selection. In the absence of polymorphism data (available for well-studied species such as *E. coli*, but lacking for most others), relative rates can still yield high-quality predictions of selected genes, which should be followed up with further experimentation to test their functional significance.

2.8 Selective signatures as a measure of Natural Selection, or of niche-specific *changes* in selection

Even for detecting selection in single genomes, the selective signatures approach can be powerful because it can identify positive (or relaxed negative) selection for genes with low values of dN/dS, while in some other cases selection is more easily detected using dN/dS with a variable branch or branch-site model. To illustrate this, we simulated codon data for 180 genes families under different models of natural selection across our tree of 30 γ-proteobacteria, and calculated dN/dS and ν in each branch (Methods). In cases with elevated dN/dS in all branches (Model 1 in Figure 2.6), PAML is able to correctly identify all branches under selection. Because there is very little variation among branches, ν is uninformative, despite positive selection in all lineages. When branch variation is present, and selection is strong in some branches but not others (Model 2 in Figure 2.6), both ν and dN/dS are able to correctly identify the species under selection. Yet when branch variation is present but the branch under selection is only weakly selected (few sites and dN/dS only slightly higher than background), it is identified correctly by ν but not dN/dS (Model 3 in Figure 2.6). Therefore, ν is well-suited to detect subtle cases of species-specific selection, but is powerless to detect uniform positive selection in all species. This is further demonstrated in an example from a gene family in our data set: *PstC* (COG573), which encodes a permease involved in phosphate transport. This gene is highly conserved across 18 species, with dN/dS near zero in most species except *Xylella fastidiosa* and *Xanthomonas campestris*, which have among the highest genome-wide average dN/dS, suggesting the high dN/dS of *PstC* may be due in part to demographic effects. Despite the lack of information from dN/dS, this gene shows substantial variation in ν across species (Figure 2.6), which may be related to species-specific ecological factors.

Like the Fixation Index computed in the MK test, but unlike dN/dS, selective signatures measure

selection relative to a baseline. While the MK identifies selection relative to a baseline of within-population polymorphism, selective signatures test for selection relative to a baseline established by comparing to related species. Despite their contrasting and independent normalization procedures, the two measures tend to overlap significantly in their predictions of natural selection. Moreover, the positive association between them (Figure 2.7; Odds ratio > 1) persists at high, intermediate, and low levels of dN/dS. The association may be slightly stronger when dN/dS is very high, due to correct identification of strong positive selection by all three methods. Yet even when *absolute* dN/dS is low, the FI and ν often agree that evolutionary rate is *relatively* fast, suggesting positive or relaxed negative selection (or strong negative selection, when both FI and ν are low), perhaps on just a few sites. While the MK test may wrongly predict selection after a population bottleneck, leading to between-species fixation of slightly deleterious mutations (Hughes, 2007), selective signatures explicitly normalize out such genome-wide effects. On the other hand, if demographic effects are not significant, the MK test has the advantage of distinguishing positive selection from relaxed negative selection, which is not possible with selective signatures. In addition, HGT (*e.g.*, from *Salmonella enterica* to *E. coli*) is expected to reduce the observed divergence, lowering ν without affecting FI or dN/dS. Thus, the intersection of genes predicted by both high FI and ν (Table 2.S5) provides additional confidence in inferring selective events.

Because selective signatures are also lineage-specific, they represent a measure of niche-specific changes in selection, and have the advantage of being sensitive to substitutions in just a few amino acid sites, provided these are unexpected relative to the gene-family and genome-specific background rates. For example, we identified several *Idiomarina* genes with high values of ν, which corresponded to only a few sites with high dN/dS, while average dN/dS across each gene was low (Table 2.S4). Even if rate acceleration is due to relaxed negative selection rather than positive selection, the change in selection detected by ν is both gene- and lineage-specific, and thus may be relevant to ecological differentiation among species. Genes with similar values of ν in the same species may be part of a co-evolving functional module, and correlations in ν are able to identify such sets of genes (Figures 2.3 & 2.4, Figure 2.S2).

2.9 Genome evolution through horizontal transfer and changes in core genes.

Can horizontal transfer alter effective protein evolutionary rates, thereby affecting selective signatures? HGT is prevalent in prokaryotes (Susko et al, 2006, Gogarten et al, 2002), especially among closely-related taxa (Alm et al, 2006). For example, we suspect that homologous recombination (or HGT between close relatives) within 'species' contributes to the observed clustering of rapidly evolving genes (Figure 2.5C). HGT can also complicate inferred evolutionary rates in two qualitatively different ways: (i) transfer from distant lineages (or replacement with paralogs) can make distances to sister taxa appear long

(and disrupt tree topology); and (ii) transfer between sister taxa does not affect tree topology, but can shorten observed distances. Thus, some of our observed rate variation is likely due to lateral gene flow. We investigated the extent to which HGT affects our results by repeating our analyses with a set of genes more likely to include horizontal gene flow, and concluded that our main findings are not easily attributable to artifacts of HGT (Figures 2.S4-S6). Moreover, our main findings are supported by methods not directly biased by HGT (MK and dN/dS tests). Even if the selective signatures of core genes are not directly confounded by HGT of the core genes themselves, they may be indirectly affected by gain and loss of other non-core genes in the genome. A discussion of the potential for such epistastic interactions to exert selective pressures on core genes is presented in Supplementary Note 2.

2.10 Summary

Species are believed to diverge only when they gain the ability to exploit a new ecological niche (Cohan, 2001), and this may come about through mutations in existing (core) genes, or acquisition of new genes. It is gaining widespread acceptance that the latter is responsible for many, if not most adaptations (Gogarten et al, 2002, Lerat et al, 2005), and possibly ensuing speciation events. Yet, as we demonstrate, core genes are also subject to selection, and likely contribute strongly to differentiation between species over long time spans. Much of this selection is positive, leading to novel adaptations in core genes. Thus, core genes, which are by definition retained in genomes over long periods of time, may be quite dynamic in terms of functional change. The coherence of selective patterns across genes of similar function (those with the same operon, functional annotation, or in the same pathway) is exciting because it suggests that the genomic landscape is organized into functional modules even at the level of natural selection. Thus, it may be easier than anticipated to understand the complex evolutionary pressures acting on genomes. Correlations in selective signatures could be used to identify fitness co-dependencies among genes in much the same way that correlated mRNA expression profiles are used to identify genes connected in the physical or regulatory networks of the cell.

Methods.

2.11 Estimation of relative evolutionary rates (ν).

To calculate relative evolutionary rates (ν), normalized to remove protein-specific 'scaffold' constraints (ρ) and species-specific 'molecular clock' (β) effects, we first constructed a 'species tree' for 30 species of γ-proteobacteria (see Table 2.S2 for species names and taxonomy IDs). Our tree is based on a concatenation of amino acid sequences for 80 housekeeping genes that occur in single-copy in each genome (Table 2.S1), and have previously been shown to be orthologous and consistent with a single organismal phylogeny (Lerat et al, 2005). Gene trees were then constructed for 977 putative 'core' gene families (members of the same cluster of orthologous genes (Tatusov et al, 1997), retrieved from the MicrobesOnline database (Alm et al, 2005)), each occurring as a single copy in at least 16 of the 30 genomes. Multiple sequence alignments (MSAs) were performed using MUSCLE (Edgar, 2004), and all gaps were removed, along with one flanking residue on either side. Gene trees were constructed from the resulting MSAs using Tree-Puzzle (Schmidt et al, 2002)with a JTT amino acid substitution model (Jones et al, 1992) and 8 γ-distributed rate categories. Estimation of ν proved to be independent of the substitution model used (see Figure 2.S7 for comparison with WAG model (Whelan & Goldman, 2001)). Gene trees were screened to remove genes that may have resulted from horizontal transfer by excluding all gene families with topologies that conflicted with the species tree topology according to a Kishino-Hasegawa (K-H) test (Kishino & Hasegawa, 1989) ($p<0.05$). Of the remaining 744 'core' gene families, 99% of the top BLAST hits were to a member of the same Genus, or to a neighboring branch on the species tree. For the 744 gene families consistent with the species tree phylogeny, trees were re-built using the consensus 'species tree' topology, but with branch lengths estimated separately for each gene. These gene trees were first normalized to remove gene family-specific contributions (ρ) by re-scaling each tree such that the sum of all branch lengths in the tree matched that expected by the species tree (considering only those branches of the species tree that are present in the gene tree). Gene trees were further normalized to remove 'molecular clock'-type effects ($\beta \cdot t$) by dividing each branch by the corresponding branch length in the species tree (Figure 2.S1). Only terminal branches (those leading directly to extant species) were used in this study, and branches with near-zero sequence changes were excluded from the analysis. Finally, the resulting relative rates were median centered within each genome, leaving an estimate of ν in which values greater than 1.0 indicate faster than expected evolution (*e.g.*, due to positive or relaxed negative selection), and values smaller than 1.0 indicate slower than expected evolution (*e.g.*, due to increased negative selection). To estimate the significance of the deviation from 1.0 (no unusual selective pressures), we computed 100 replicates of our estimate for ν by non-parametric sequence bootstrapping, and computed a 'Z-score' as the ratio of the observed $\log_2(\nu)$ to the square root of

its variance over the bootstrap replicates.

2.12 Estimation of synonymous and non-synonymous substitution rates (dS and dN).

We used the *codeml* program from the PAML 4.0 package (Yang, 2000) to estimate dN and dS, allowing their ratio to vary freely along branches of the species tree ('free-ratio' model). Estimates of dN, dS and dN/dS were made for each of the 744 core orthologs described above. To generate 'relative' values of dN, dS and dN/dS, each of these values was first normalized by its median value for each genome, then by the median for each ortholog. Note the order of normalization steps is reversed from that for relative rates, because there is no prior expectation that dN/dS values across the tree are proportional to evolutionary time/distance.

2.13 Simulation of genes under different models of selection.

We used the *evolver* program from the PAML 4.0 package (Yang, 2000) to simulate gene families of 300 codons in 30 species, using the γ-proteobacteria species tree topology. In the first set of simulations (Figure 2.3B), we used two classes of sites (occurring at frequency 0.1 and 0.9, respectively), each with a different value of dN/dS, randomly chosen from within ±1 standard deviation of the mean of the observed distribution of dN/dS in our data set of 744 genes across 30 species. In 18 of the models, dN/dS was not allowed to vary among branches; in the remaining 18 a different dN/dS value was chosen at random for each site class and each branch. For each model, we generated 5 replicate codon sequences in 5 independent runs of *evolver*. In the second set of simulations (Figure 2.6), we used either 2 or 3 classes of sites (with frequency chosen within the range of 0.1 to 0.9), each with dN/dS of either 2.0, 1.5, 1.1, 1.0, 0.5 or 0. We generated 180 different models, 45 of which did not allow branch variation, and the remaining 135 with 1 to 5 branches under selection, with one site class having a higher dN/dS than the other branches. We generated 12 replicate sequences for each model. For both sets of simulations, we translated the codons to amino acid sequence in order to calculate ν, treating each replicate of each model as a protein family. We also estimated dN/dS in each branch using the free-ratio model in PAML.

2.14 McDonald-Kreitman tests.

Gene families were retrieved from 24 strains of *E. coli* (including some strains of *Shigella*; see Table 2.S2b), and an outgroup, *Salmonella enterica*. Each gene had exactly one representative in each strain. Genes were assigned to orthologous families using OrthoMCL (Li et al, 2003). Only the 473 gene families corresponding to COGs in the relative rates data set, and not violating the K-H test, retained for analysis. We tried excluding genes with a large number of divergent sites relative to polymorphic sites, which might reflect HGT from closely-related species, but this did not significantly affect results.

Nucleotide sequences were aligned and trimmed using MUSCLE, as described above. Polymorphic substitutions (within the 24 strains of *E. coli*) and divergent substitutions (fixed between *E. coli* and *Salmonella*) were counted, and assigned to synonymous or nonsynonymous categories, as previously described (McDonald & Kreitman, 1991). Only codons for which there were no more than two states were retained for analysis, and we always chose the pathway between codons that minimized the number of nonsynonymous changes. An Odds Ratio statistic, the Fixation Index (FI), was then calculated as described in the main text.

2.15 Simulation of ν.

To ensure that the patterns of selection across species were not simply due to an artifact of the species tree topology and normalization procedure, we simulated sequence data for 744 genes over the tree of 30 species. First, gene presence/absence in the 30 species studied was sampled from the distribution observed for the 744 gene families. Then branch lengths ($\beta \cdot t$) were chosen for each branch by multiplying the corresponding branch in the species tree by randomly chosen values from the observed distributions of ν and ρ, respectively. Sequences were then generated based on these trees using Seq-Gen (Rambaut & Grassly, 1997), with a JTT amino acid substitution model and γ-distributed rates and analyzed as for the actual gene sequences.

2.16 Divergence time estimation.

We used the 'multidivtime' software package (Thorne et al, 1998, Battistuzzi et al, 2004) to estimate divergence times in the species tree, using a relaxed molecular clock model. The tree of 30 γ-proteobacteria was rooted using *Neisseria meningitidis* MC58 (taxonomy ID 122586) and *Nitrosomas europaea* (taxonomy ID 228410) as outgroups. The 30 γ-proteobacteria and outgroup species all shared the 80 single-copy orthologs used to construct the tree. The most likely evolutionary rates and divergence times were estimated using the Bayesian Markov chain Monte Carlo procedure, as implemented in multidivtime using the following priors: Mean tree length (root to tips) = 1.75 By (standard deviation = 1.0 By) and rate at the root = 3.9 substitutions/site/By (chosen to equal the median rate at the tips, with standard deviation = 3.9). Time bounds were set based on the divergence of *E. coli* and *S. typhimurium* between 57 Mya (upper bound) and 176 Mya (lower bound) (Ochman & Wilson, 1987, Ochman et al, 1999). Three independent runs were performed with these settings, and did not differ in the estimated divergence times.

Supplementary Note 1. Analysis of Horizontally and Vertically-inherited gene sets.

We repeated many of our analyses with a set of genes more likely to include horizontal gene flow (those 173 genes previously excluded because they rejected the species tree according to a K-H test, see Methods), and a set of 203 ubiquitous housekeeping genes (of which 161 are present in our data set) previously reported to be primarily vertically inherited (though the previous study used fewer genomes within this clade (Lerat et al, 2005)). The results of these analyses (shown in Figures 2.S4-2.S6) suggest that most of the effects we report cannot be simply explained by HGT: similar results to those shown in Figures 3 and 5B&C were observed for the non-HGT set, whereas less striking results are observed for the HGT set. In addition, the dN/dS results reported in Figure 2.3 (main text), which should be less sensitive to the effects of HGT, confirm our finding that patterns of selection correlate with function. Thus, we conclude that HGT likely plays a significant role in shaping evolutionary rates, especially among groups more closely related than our study resolves, but that long-range transfer is not solely responsible for the patterns observed.

Supplementary Note 2. Patterns of selection across genomes.

To better appreciate the timescale for rate variation across species, we compared the values of ν between orthologous genes in different species. Figure 2.S8 (upper diagonal) shows the correlation in ν over all genes shared by each pair of species. When species are ordered according to their phylogeny, there is a striking trend: genes in closely-related species appear to be under similar selective pressures (high pairwise ν-correlations are close to the diagonal). The observed pattern is not due to artefacts of the normalization procedure or phylogenetic inference, as simulated data produce no pattern (lower diagonal; see Methods). Remarkably, strong positive correlations are observed within relatively ancient clades, dating as far back as ~900 million years as estimated by a Bayesian relaxed molecular clock model (Thorne et al, 1998) (e.g., the *Colwellia*/*Idiomarina*/*Shewanella* clade in Figure 2.S8; all correlations significant with $P < 1e\text{-}4$). The species within this clade have qualitatively different lifestyles, and differ greatly in their respiratory capabilities and preferred temperatures ranges. On the other hand, we observe no finer-scale patterns *within* these closely-related, co-evolving clades (e.g., *Pseudomonas* or *Vibrio* species; see Figure 2.S10). This suggests that, on shorter time scales, evolutionary rate variation does not correlate with phylogenetic relatedness, but perhaps with fine-scale niche adaptation, or neutral drift. Indeed, the constancy of (relative) rates within related clades has been recently exploited as a powerful feature for ortholog identification in several eukaryotic groups (Rasmussen & Kellis, 2007).

Is ecological similarity responsible for the apparent similarity in core gene evolution among close relatives? Correlations in ν are generally observed between species within the same order, consistent with

the observation that species related at the level of order have more similar habitats than expected by chance (Mering et al, 2007). If ecological similarity leads to similar regimes of selection, we would also predict some degree of ecological similarity between the outlying pairs of distantly-related species with highly correlated ν across their genomes. Indeed, the biggest outliers to the trend that 'rate variation recapitulates phylogeny' all involve *Coxiella*, *Xylella*, and members of the *Buchnera*/*Wigglesworthia* clade - all of which are ecologically similar in that they are fastidious, intracellular, host-associated species (Tamas et al, 2002).

Core genes are highly conserved, and might be affected only indirectly by selection imposed from the outside environment. Adaptation to a new ecological niche often involves acquisition of new modules of genes rather than mutational changes in existing genes. However, epistatic interactions between genes might require bacteria to fine-tune their 'core' genomes to accommodate a new or altered module, and maintain fitness. On average, epistatic interactions are likely to exist between most genes (as has been reported for *E. coli* (Kishony & Leibler, 2003)). Thus, large changes such as the addition or deletion of genes *via* HGT may lead to changes in the intracellular environment, and ensuing selection on core genes. Therefore the relationship between natural selection on core genes and phylogenetic similarity could be driven by shared gene content.

What are the relative contributions of these factors in modulating the rate of core gene evolution? As discussed above, similar ecology may contribute to co-evolution of core genes in different species. Other contributions may come from genetic 'fine-tuning' in response to dynamic gene content (quantifiable as the proportion of shared orthologs between species), and phylogenetic closeness (quantifiable as the evolutionary distance in substitutions/site or divergence time between species). We compared these quantitative measures for their ability to predict how strongly values of ν are correlated between pairs of species. Even after controlling for evolutionary distance or divergence time, we found gene-content similarity to be a better predictor than either evolutionary distance or time alone (Figure 2.S9). Thus, gene-content appears to be a prime determinant of selection on core genes, while phylogenetic closeness does not contribute as strongly once gene-content is controlled for. As a result, processes that affect gene-content, such as horizontal gene transfer, duplication and gene loss, may play an under-appreciated role in the evolution of vertically inherited 'core' genes.

Selection on core genes may be imposed either directly from the environment, or to accommodate newly acquired functions that might otherwise incur a fitness cost. HGT among microbes, analogous to recombination among viral strains, might reduce fitness by disrupting co-evolved functional modules of

genes (Martin et al, 2005). Although horizontally transferred genes tend to attach to the periphery of the cell's metabolic network so as not to disrupt major housekeeping functions (Pal et al, 2005), they still interact (directly or epistatically (Segre et al, 2005)) with downstream and parallel pathways, perhaps leading to subtle fitness effects. To compensate for this changing fitness landscape, bacteria may 'fine-tune' the rest of their genomes accordingly. As we have shown, such selective fine-tuning on core genes is common enough to be detectable, and is at least partially explained by gene-content variation. Of course, gene-content similarity may co-vary substantially with ecological similarity, and the lack of sufficient data or methods to quantify differences between ecological niches prevent us from disentangling these factors.

Figure legends

Figure 2.1. Evolutionary rate deviations as evidence of natural selection.

Observed branch length is plotted against the branch length predicted from gene-specific (ρ) and species-specific (β) effects (see Methods). A total of 16,681 points are plotted, corresponding to 744 orthologous proteins present in 16-30 species. Amino acid substitutions per site are shown on a $\log_2$ scale. The grey line corresponds to y=x.

Figure 2.2. Genes of common function have similar selective signatures.

Relative rates of evolution are shown for 5 genes across 30 species. Fast-evolving genes ($\log_2 \nu > 0$) are shown as red bars; slow-evolving genes ($\log_2 \nu < 0$) as blue bars; genes absent in a given species are not shown. The time scale for the phylogeny was estimated using a Bayesian relaxed molecular clock model (Thorne et al, 1998). Flagellar genes: *flgN* (COG 3418; Flagellar biosynthesis/type III secretory pathway chaperone), *flgA* (COG 1261; Flagellar basal body P-ring biosynthesis protein), *fliS* (COG 1516; Flagellin-specific chaperone). Sulfur metabolism genes: *yheL* (COG 2168; Uncharacterized conserved protein involved in oxidation of intracellular sulfur), *yheM* (COG 2923; Uncharacterized conserved protein involved in oxidation of intracellular sulfur).

Figure 2.3. (A) Selection acts coherently on cellular functions.

Correlations in ν, dN/dS and relative dN/dS (normalized as described in Methods) were obtained for the 109,405 gene-pairs with a COG functional category annotation (16 categories, excluding 'general' or 'unknown' function). Of these pairs, 10,377 have the same COG function, accounting for a proportion of ~0.09 of the total (plotted as a solid grey line). Pairs were binned according to correlation-percentile in groups of 10 percentile points except for the last three (90-95%, 95-99%, 99-100%). Shown is the fraction with common function in each bin. To avoid potential bias, percentiles were calculated separately for genes present in different numbers of species (15 bins ranging from 16-30 species).

(B) Gene families under the same model of evolution have highly correlated selective signatures. Correlations in ν were obtained for all pairs of simulated gene families, with or without branch variation in dN/dS, and with dN/dS chosen randomly from within ±1 standard deviation of the mean of the observed dN/dS values (range: 0.005 to 0.26). The distribution of correlations is shown for pairs of gene families with branch variation in dN/dS, and that are replicates of the same evolutionary model (light blue). The distribution of all pairwise correlations – including gene families with or without branch variation, and pairs from the same or different models – is also shown (grey).

Figure 2.4. Rapidly-evolving pathways in *Idiomarina loihiensis.*

Simplified schematic of glycolysis and phenylalanine metabolism in *Idiomarina loihiensis*. Metabolic intermediates are denoted by white circles; enzymes by arrows. 'Fast-evolving' enzymes, depicted as red arrows, are defined as those with ν in the top 10% of genes in the *Idiomarina loihiensis* genome. The names of genes encoding fast-evolving enzymes are shown, highlighted in light blue or orange, respectively for glycolysis or phenylalanine metabolism. Non-functional pathways (those with many key enzymes or transporters missing) are shown in grey. Of the 'present' enzymes shown in black, only one is slow-evolving ($\nu < 1$) in *Idiomarina*: COG 191, encoding the enzyme fructose bisphosphate aldolase, which interconverts F1,6P and GA3P. Abbreviations for metabolic intermediates: PEP: phoshphenolpyruvate, E4P: erythrose-4-phosphate, DAHP: 7P-2-dehydro-3-deoxy-arabinoheptonate, DHQ: 3-dehydroquinate; DHS: 3-dehydroshikimate, prCat: protocatechuate, shik: shikimate, shik-3P:

shikimate-3-phosphate, CVPS: 5-O-(1-carboxyvinyl)-3-phosphoshikimate, chor: chorismate, prePh: prephenate, phPy: phenylpyruvate, Phe: phenylalanine, G6P: glucose-6-phosphate, F6P: fructose-6-phosphate, F1,6P: fructose-1,6-bisphosphate, GA3P: glyceraldehyde-3-phosphate, DHAP: dihydroxyacetone phosphate, G1,3P: glyerate-1,3-bisphospate, G3P: glycerate-3-phosphate, G2P: glycerate-2-phosphate. COG and EC numbers of fast-evolving genes: *AroB*: COG337, EC4.2.3.4, *AroQ*: COG757, EC4.2.1.10, *AroE*: COG169, EC1.1.1.25, *PheA*: COG77, EC4.2.1.51, *Pgi*: COG166, EC5.3.1.9, *Fbp*: COG158, EC3.1.3.11, *Pfk*: COG205, EC2.7.1.11, *TpiA*: COG149, EC5.3.1.1, *Eno*: COG148, EC4.2.1.11.

Figure 2.5. (A) Comparison of relative rates (ν) and Fixation Index.

Histograms show the frequency (probability density) distribution of FI values for fast-evolving ($\nu > 2$; dark red; N=69) and slow-evolving ($\nu < 0.5$; light blue; N=63) genes. Bins are labelled with the FI value corresponding to their midpoint, on a $\log_2$ scale. FI was calculated by counting fixed and polymorphic substitutions at synonymous and nonsynonymous sites, in a sample of 473 COGs (all present in the relatives rates data set, and passing the K-H test) in 24 *E. coli* strains, using *Salmonella enterica* as an outgroup.

(B) Purifying selection and gene deletions. Fast-evolvers (or slow-evolvers) were defined as those genes evolving 4 times faster (or slower) than expected ($\nu > 4.0$ or $\nu < 0.25$, respectively for fast and slow, with a Z-score > 1.0). For the fast and slow sets of genes, we counted the number with lost orthologs in the closest sister clade in the species tree. When the sister clade contains multiple species, loss indicates the gene was absent from all species in the clade. Frequency of loss among the fast and slow sets was significantly different than the average over all other genes: higher in the fast-evolving set (Fisher's exact test: Odds Ratio = 3.1, P = 2.4e-7), and lower in the slow-evolving set (Fisher's exact test: Odds Ratio = 0.55, P = 0.01).

(C) Evidence for genetic hitchhiking. A binomial test was used to determine whether fast (or slow) evolving genes tend to be clustered in the genome near other fast (or slow) evolving genes across all 30 species combined ($\nu > 1$ or $\nu < 1$, respectively for fast and slow, with a Z-score > 1.0). Log *p*-values are plotted for pairs separated by distance-windows of 0-5 genes, 6-20 genes, 21-100 genes, 101-200 genes, and 201-300 genes (points shown indicate the maximum separation). The grey line represents $p = 0.05$.

Figure 2.6. Detection of positive selection by dN/dS and ν under different evolutionary models.

Values of dN/dS and ν (mean over 12 replicates of each model) are shown for 3 simulation models. Model 1: dN/dS = 2 at 3/10 of sites and dN/dS = 1 at 7/10 of sites, in all species (shown in red). Model 2: dN/dS = 2 at 3/10 of sites and dN/dS = 1 at 7/10 of sites, respectively, for the species shown in red. All other branches had dN/dS = 0 at all sites. Model 3: dN/dS = 2, dN/dS = 1 and dN/dS = 0 at 1/10, 7/10 and 2/10 of sites, respectively, in the species shown in red. All other branches had dN/dS = 1 and dN/dS = 0 at 8/10 and 2/10 of sites, respectively. Values of dN/dS and ν are also shown, as estimated for a real protein family from our data set of 744 protein families in 30 species.

Figure 2.7. Positive association of selective signatures (ν) and Fixation Index, independent of dN/dS. We counted *E. coli* genes with FI > 1.2 or FI < 0.6 as 'high' and 'low', and with $\log_2 \nu > 0.5$ ($\nu > 1.4$) or $\log_2 \nu < -0.5$ ($\nu < 0.7$) as 'high' and 'low'. The genes were divided into sets with relatively high dN/dS (> 0.06), medium (0.02 < dN/dS < 0.06), or low dN/dS (< 0.02). Within each set, counts were binned in 2 X 2 contingency tables to calculate the Odds Ratio statistic, with Odds Ratio > 1 indicating positive association between ν and FI. One-sided *P*-values of Fisher's exact test are shown.

Supplementary figures legends.

Figure 2.S1. Example of tree normalization and calculation of ν.

The normalization procedure is illustrated for two example protein families (columns). We begin with a gene tree of three species ((A,B),C), with branch length ((substitutions / site) x 10^{-2}) equal to total evolutionary distance (top row). The gene tree is first normalized so that terminal branches all sum to 1. The resulting gene tree, normalized to remove gene-family effects (ρ) is shown in the second row. Each branch in the normalized gene-tree is then divided by the corresponding branch in the normalized species-tree (β, shown in the third row) to yield an estimate of ν for each branch, shown in the bottom row. Values of $\nu > 1$ reflect faster-than-expected evolution; $\nu < 1$ reflect slower-than-expected evolution.

Figure 2.S2. Operon prediction by correlation in ν, dN, dS, and dN/dS.

Receiver Operating Characteristic curve of several methods for operon prediction. Pairs of genes predicted to be on the same operon are considered 'true positives'; pairs on different operons as 'false positives'. Correlations were computed between the 157,612 pairs of genes for which both gene expression (from *E. coli* microarrays under 14-17 different experimental conditions) and relative rate (ν) data was available, and for which the pair is present in at least 16 of 30 species. Of these pairs, 898 are predicted to fall on the same operon in *E. coli* (Price et al, 2005). To avoid systematic biases in correlations (pairs present in fewer species might achieve higher correlations by chance), the correlations were percentile-ranked together with other genes present in the same number of species. For each level of percent-ranked correlation (in either ν, expression level or raw/normalized dN, dS, or dN/dS, estimated as described in the Methods of the main text), the percentage of pairs above this level in the different-operon set is plotted against the percentage in the same-operon set. The solid grey line represents random prediction ($y = x$). We also assessed the ability of similarity in Fixation Index (FI) to predict *E. coli* genes on the same operon. For each pair of COGs present in *E. coli*, we calculated 'delta FI' as the absolute difference between the FI of each COG. Delta-FI was percentile-ranked and plotted on the operon-predicting ROC curve. Similarity in FI is a rather poor predictor of operons, perhaps because it a scalar value from one species, rather than a correlation across many species. The *E. coli* gene expression data was obtained from microbesonline.org. Correlations in expression level (mean log-ratio of experimental condition to control) were taken between genes for which expression had been measured in at least 14 of 17 experimental conditions, including heat shock (experiment ID 12; (Gutierrez-Rios et al, 2003)), low pH (experiment ID 40; Blattner lab), UV exposure (experiment ID 46, 7 time points; GEO accession GSE9), and tryptophan exposure/starvation (experiment ID 47, 8 time points; Geo accession GSE7).

Figure 2.S3. Effect of normalization procedure (A) and COGs used in species-tree construction (B) on rate-function correlation.

Methods as described in Figure 2.3 of the main text. (A) Comparing normalized rates (ν) to non-normalized 'raw' evolutionary distance. Red lines: Correlation in ν between gene-pairs. Black lines: Correlation in raw evolutionary distance (non-normalized gene-tree branch lengths) between gene-pairs. Filled circles: including all gene pairs; Open circles: excluding pairs on the same operon. The distribution of ν-correlations among gene-pairs of common function is significantly more biased toward high correlation than the distribution of raw-distance correlations between gene-pairs of common function (KS test, D=0.12 and D=0.10, respectively for total and same-operon excluded data sets, both $P<2.2e-16$), indicating that ν-correlation is a significantly better predictor of function than raw distance, even when same-operon pairs are not considered. The grey line denotes the mean fraction with shared function, over all genes. (B) Comparing the full set of normalized rates (shown in red) with the set excluding the 80 COGs used to construct the species tree (shown in blue). The mean fraction of genes with shared function for each of these data sets is shown in red or blue, respectively. The distribution of ν-correlations among gene-pairs of common function does not differ significantly between these data sets (KS test, D=0.02 $P>0.05$).

Figure 2.S4. Effect of topology violation (putative HGT) on rate-function correlation.

ν-correlations between genes were obtained for gene-pairs falling into 3 categories: (1) Overall: those among the 744 used throughout this work (red), (2) non-HGT :those among the 161 high-quality orthologs of Lerat, Daubin & Moran (Lerat et al, 2005) (blue), or (3) HGT: those among the 173 orthologs found to significantly violate the species tree topology by the K-H test (green). Only gene-pairs with a COG functional category annotation (16 categories, excluding 'general' or 'unknown' function), and which occur in at least 16 of 30 species, were used. (A) All pairs, including genes on the same or different operons; (B) Excluding gene-pairs on the same operon. For each case, we binned the pairs according to ν-correlation-percentile (as described in Figure 2 of the main text), and plot the fraction with common function in each bin. ν-correlations are expressed as percentiles to control for possible bias in computing correlation for different vector lengths (ranging from 16-30 species). Solid horizontal lines represent the mean fraction of gene-pairs with shared function, averaged over all values of ν-correlation. For both (A) and (B), the distribution of ν-correlations differed between same-function and different-function pairs only for the Overall and non-HGT sets, but not the HGT set (KS tests: $P<2.2e-16$ and NS, respectively for Overall/non-HGT and HGT set). The distribution of Overall ν-correlations among gene-pairs of common function is significantly different from the HGT distribution for gene-pairs of common function (KS test, $D>0.12$, $P<1e-4$, for both operon-excluded or total data set).

Figure 2.S5. Frequency of lost orthologs in sister taxa for HGT and non-HGT data sets.

Within each of 3 data sets (topology violators (failed K-H test, putative HGT), high-confidence (putative non-HGT from Lerat, Daubin & Moran, 2003), or Overall (744 genes, passed K-H test)), fast-evolvers (or slow-evolvers) were defined as those genes evolving 4 times faster (or slower) than expected ($\nu > 4.0$ or $\nu < 0.25$, respectively for fast and slow, with a Z-score > 1.0). For the fast and slow sets of genes, we counted the number with lost orthologs in the closest sister clade in the species tree. When the sister clade contained multiple species, the loss was only counted if all species in the clade had lost the ortholog. Frequency of loss among the fast (red) and slow (blue) sets was compared to the frequency of loss over all genes (grey) using Fisher's exact test; asterices denote significant differences ($P < 0.01$). In each of the 3 data sets, the proportion of lost orthologs was significantly different between fast-evolving and slow-evolving genes. The proportion of fast-evolving genes with lost orthologs did not differ significantly between the HGT and Overall set, nor between the high-confidence non-HGT and Overall set. However, the proportion of slow-evolving genes with lost orthologs was significantly higher in the HGT set than the Overall set (Fisher's exact test, Odds Ratio = 3.2, $P = 0.009$), but not between the non-HGT and Overall set.

Figure 2.S6. Clustering of 'Fast' genes in HGT and non-HGT data sets.

(A) High confidence data set (161 probable non-HGT genes; Lerat, Daubin & Moran, 2003), (B) Topology-violating data set (173 probable HGT genes; failed K-H test). For each data set, a binomial test was used to determine whether fast (or slow) evolving genes tend to be clustered near other fast (or slow) evolving genes on the chromosome across all 30 species ($\nu > 1$ or $\nu < 1$, respectively for fast and slow, with a Z-score > 1.0). The log P-values from the binomial test are plotted against genomic distance between genes. All pairs were on different operons. The P-values are plotted for pairs separated by distance-windows of 0-5 genes, 6-20 genes, 20-100 genes, 100-200 genes, and 200-300 genes. Points shown indicate the maximum separation in each non-overlapping window. The grey line represents a P-value of 0.05.

Figure 2.S7. Agreement of values of ν estimated using JTT and WAG substitution models.

For each of 744 genes across 30 species, the value of ν estimated under a JTT substitution model is plotted against the value estimated under a WAG substitution model. The species tree topology did not differ significantly between JTT and WAG models implemented in Tree-Puzzle (K-H and S-H tests, $P > 0.05$), therefore the same topology was used for both models, but with branch lengths estimated separately to compute ν under each model. The models produce near-identical values of ν (Pearson's correlation = 0.987, $R^2 = 0.974$, $P < 2.2e-16$, N = 16,681).

Figure 2.S8. Patterns of selection across genomes.

ABOVE DIAGONAL: As a global measure of similarity in selection over all genes between two species, Pearson's correlations were computed between each pair of species, using the vectors of ν for all shared orthologs of the pair (always in the range of 250-650 orthologs). The pairwise correlations between species are illustrated, ordered according to the species topology. The time scale for the phylogeny was estimated using a Bayesian relaxed molecular clock model. BELOW DIAGONAL: Random values of ν simulated along the same species tree show no pattern of correlation between species (see Methods).

Figure 2.S9. Full and partial correlations between co-evolution (ν-correlation between species) and gene content, raw evolutionary distance, and divergence time.

Pearson's correlations and partial correlations were computed between ν-correlation (over all genes) and (i) shared gene-content, (ii) distance, and (iii) time, for all pairs of species. Gene-content varies positively with ν-correlation between species, while distance and time both vary negatively. For simplicity, absolute values of correlations are shown. All correlations and partial correlations were significant with $P < 0.0001$, except for ν x distance | time, for which $P = 0.007$. The full correlation of co-evolution (ν) with gene content was stronger than the correlation with either distance or time, with 95% confidence. ν-correlation between each species-pair was computed as for Figure 2.S8. Shared gene-content between a species-pair was defined as the number of shared orthologs (same COG), divided by the geometric mean of the total number of COGs in each pair of genomes, excluding multi-copy COGs and genes not assigned to a COG. Distance was defined as total evolutionary change (amino acid substitutions per site) estimated to have occurred since the last common ancestor of the pair.

Figure 2.S10. No phylogenetic pattern of ν-correlation is observed within shorter time scales.

ν was calculated as described in Methods, using either a species tree consisting of only the 4 *Pseudomonas* species (left) or the 5 *Vibrio/Photobacterium* species (right). 828 or 805 shared single-copy orthologs were used to build the respective species trees. ν was calculated for each ortholog in each species, provided that the gene-tree did not violate the species topology. As in Figure 2.S8, Pearson's correlations were computed between each pair of species, using the vectors of ν for all shared orthologs of the pair. The pairwise correlations between species are illustrated, ordered according to the species topology.

Supplementary table legends.

Table 2.S1. List of genes used to construct the species tree.

Table 2.S2a. Taxonomy IDs of species used in this study.

Table 2.S2b. Taxonomy IDs of strains used in McDonald-Kreitman tests.

Table 2.S3a. Enrichment of COG functions in top 10% fast- or slow-evolving sets of genes .

Fast and slow-evolving sets of genes were defined as those with the highest or lowest 10% values of ν in a genome, with a Z-score threshold of 1.0. A hypergeometric test was performed to test whether each functional category was enriched in the fast or slow set, relative to the expected fraction of that category in the whole genome. *p*-values are Bonferonni-corrected for multiple tests on 16 COG categories.

Table 2.S3b. List of fast-evolving flagellar genes in 3 species of enterobacteria.

The COG function "Motility & Secretion" is enriched in the fast-evolving set of genes in 3 species of Enterobacteria: *Escherichia coli K12, Photorhabdus luminescens*, and *Yersinia pestis*. The genes responsible for this trend are all involved in flagellar functions. Different genes are fast-evolving in each species: these are denoted by asterices in the relevant column.

Table 2.S4. Evidence for site-specific changes in dN/dS in *Idiomarina* genes.

To support the high values of ν observed in glycolysis and phenylalanine biosynthesis in *Idiomarina*, we analyzed these genes in PAML under 5 different evolutionary models: (1) No branch variation (single value of dN/dS for the whole tree); no site variation, (2) Different dN/dS allowed in the *Idiomarina* branch (two values of dN/dS in the tree); no site variation, (3) No branch variation; variable selective pressure among sites allowed (3 categories of sites, each allowed a different dN/dS), (4) Different dN/dS allowed in *Idiomarina*; variable selective pressure among sites allowed (3 categories of sites for *Idiomarina* and 3 for the rest of the tree, each allowed a different dN/dS), and (5) Same as Model 4, but fixing dN/dS not to exceed 1. In Test A, Model 2 is compared with Model 1 in a likelihood ratio test (compare $2(L_2 - L_1)$ to χ^2 distribution with degrees of freedom = # parameters in Model 2 – Model 1). In effect, this tests whether *Idiomarina* has a different dN/dS, averaged over all sites in the gene, than other lineages. It should be noted that for all genes in the table above, dS is saturated (> 50 substitutions per site), perhaps explaining the relatively low average values of dN/dS. In Test B, Model 4 is compared to Model 3 in a likelihood ratio test (compare $2(L_4 - L_3)$ to χ^2 distribution with degrees of freedom = # parameters in Model 4 – Model 3). For all 10 genes, this test supports sites under different selection in *Idiomarina*. In all cases, there is at least on site category with higher dN/dS in *Idiomarina* than other lineages. We performed a final test to determine whether there were some sites in *Idiomarina* with dN/dS > 1. In this test, we compared Model 4 to Model 5 in a likelihood ratio test (compare $2(L_4 - L_5)$ to χ^2 distribution with degrees of freedom = # parameters in Model 4 – Model 5). For genes with significant evidence ($P < 0.05$) for sites with dN/dS > 1, the proportion of sites with elevated dN/dS in *Idiomarina* (far right column) is shown in bold.

Table 2.S5. Genes predicted under selection in *E. coli* by selective signatures and the MK test.

Genes are listed only when the MK test and selective signatures agree, using the same criteria as Figure 7 in the main text ($\log_2$FI > 0.26 and $\log_2\nu > 0.5$, or $\log_2$FI < -0.74 and $\log_2\nu < -0.5$). For ν, * denotes Z score > 1, ** Z > 2. For FI, * denotes $\chi^2 > 3.84$ ($P < 0.05$ with 1 d.f.).

COG functional category key:

C	Energy Production	K	Transcription
D	Cell Division	L	DNA replication, repair & modification
E	Amino Acid Metabolism	M	Cell Envelope
F	Nucleic Acid Metabolism	N	Motility & Secretion
G	Carbohydrate Metabolism	O	Protein Modification & degradation
H	Coenzyme Metabolism	P	Ion Transport & Metabolism
I	Lipid Metabolism	Q	Secondary Metabolism
J	Ribosome & Translation	R	General Function
S	Unknown	T	Signal Transduction

Figure 2.1

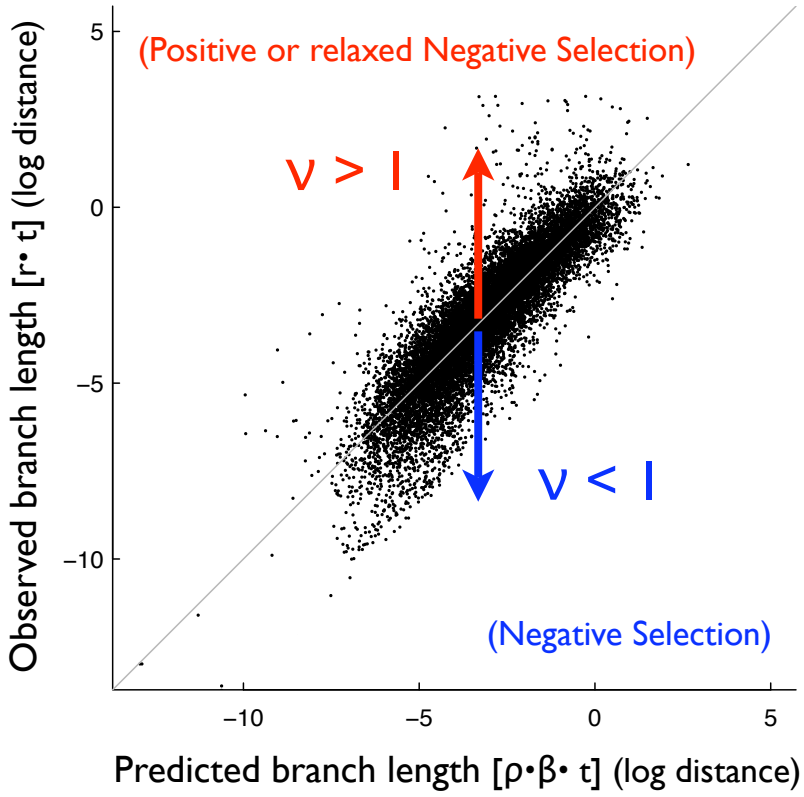

(Positive or relaxed Negative Selection)
ν > 1
ν < 1
(Negative Selection)
Observed branch length [r• t] (log distance)
Predicted branch length [ρ•β• t] (log distance)
5
0
−5
−10
−10
−5
0
5

Figure 2.2

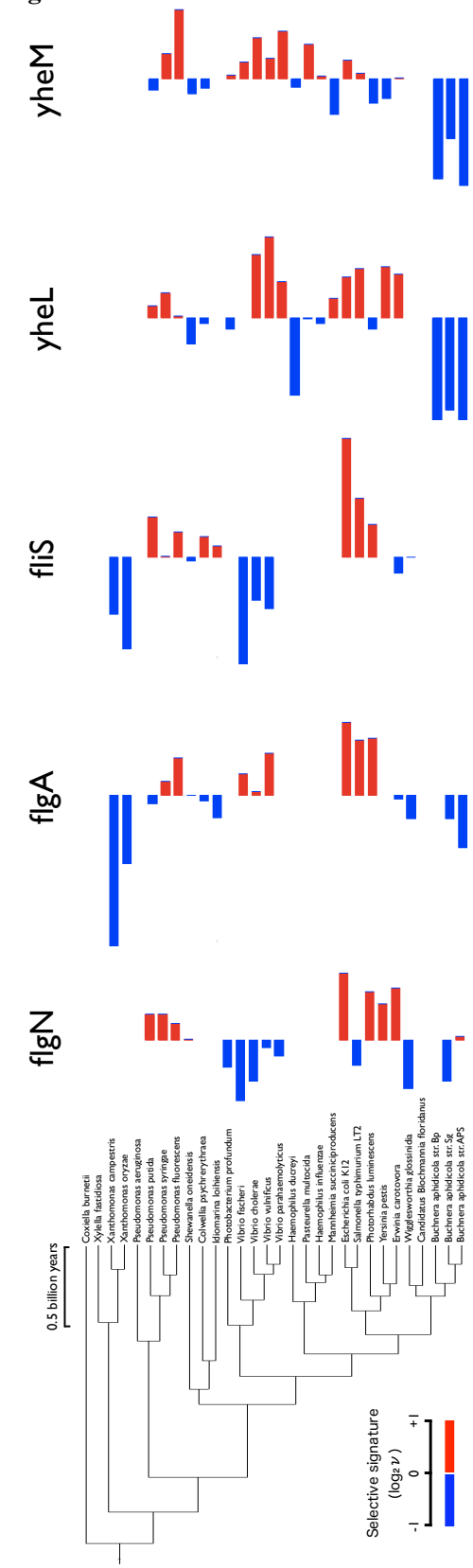

yheM
yheL
fliS
flgA
flgN
0.5 billion years
Coxiella burnetii
Xylella fastidiosa
Xanthomonas campestris
Xanthomonas oryzae
Pseudomonas aeruginosa
Pseudomonas putida
Pseudomonas syringae
Pseudomonas fluorescens
Shewanella oneidensis
Colwellia psychrerythraea
Idiomarina loihiensis
Photobacterium profundum
Vibrio fischeri
Vibrio cholerae
Vibrio vulnificus
Vibrio parahaemolyticus
Haemophilus ducreyi
Pasteurella multocida
Haemophilus influenzae
Mannheimia succiniciproducens
Escherichia coli K12
Salmonella typhimurium LT2
Photorhabdus luminescens
Yersinia pestis
Erwinia carotovora
Wigglesworthia glossinida
Candidatus Blochmannia floridanus
Buchnera aphidicola str. Bp
Buchnera aphidicola str. Sg
Buchnera aphidicola str. APS
Selective signature
(log2 ω)
-1
0
+1

Figure 2.3

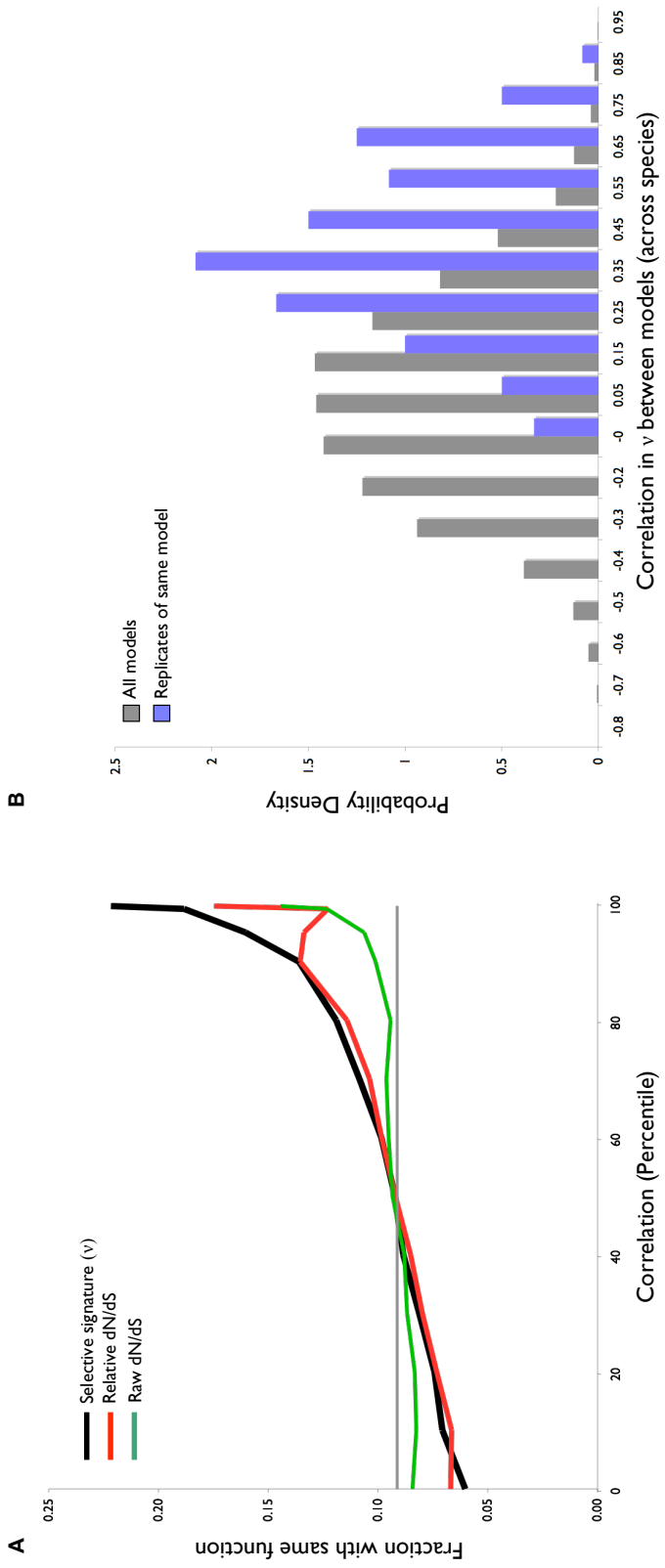

A
Fraction with same function
0.25
0.20
0.15
0.10
0.05
0.00
0
20
40
60
80
100
Correlation (Percentile)
Selective signature (ν)
Relative dN/dS
Raw dN/dS
B
Probability Density
2.5
2
1.5
1
0.5
0
All models
Replicates of same model
-0.8
-0.7
-0.6
-0.5
-0.4
-0.3
-0.2
-0
0.05
0.15
0.25
0.35
0.45
0.55
0.65
0.75
0.85
0.95
Correlation in ν between models (across species)

Figure 2.4

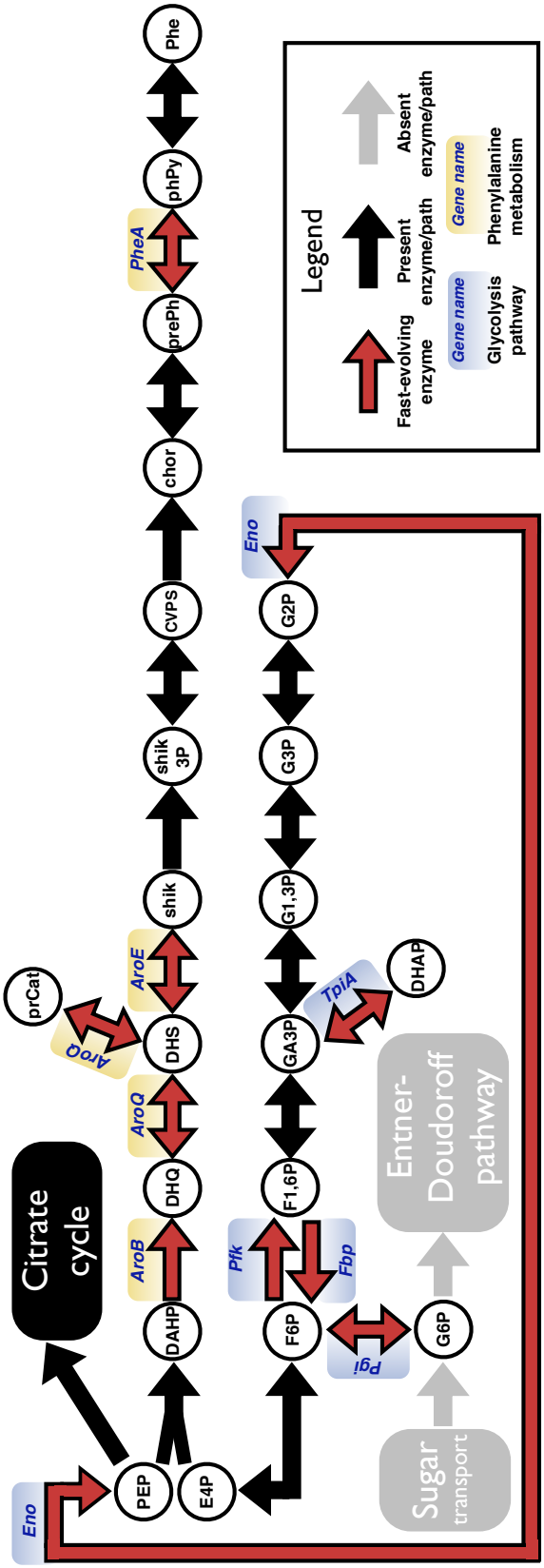

Phe
phPy
PheA
prePh
chor
CVPS
shik 3P
shik
AroE
prCat
AroQ
DHS
AroQ
DHQ
AroB
DAHP
PEP
E4P
Eno
Citrate cycle
F6P
Pfk
Fbp
F1,6P
GA3P
TpiA
DHAP
G1,3P
G3P
G2P
Eno
Pgi
G6P
Entner-Doudoroff pathway
Sugar transport
Legend
Fast-evolving enzyme
Present enzyme/path
Absent enzyme/path
Gene name
Glycolysis pathway
Gene name
Phenylalanine metabolism

Figure 2.5

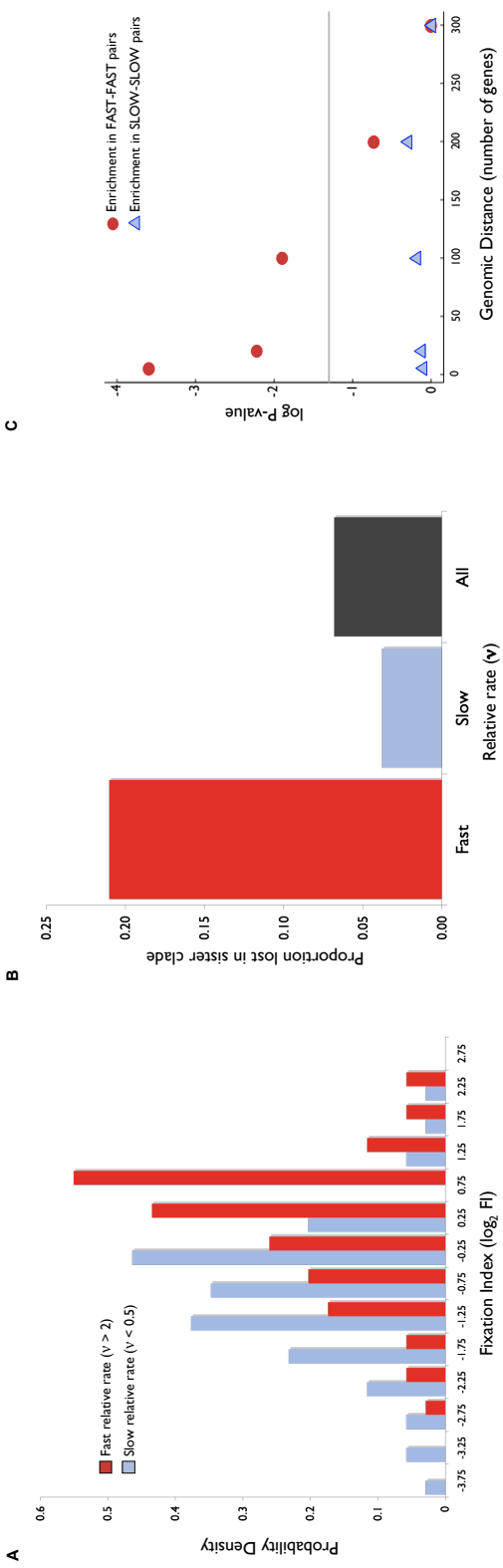

A
Probability Density
Fast relative rate (v > 2)
Slow relative rate (v < 0.5)
Fixation Index (log2 FI)
B
Proportion lost in sister clade
Fast
Slow
All
Relative rate (v)
C
log P-value
Enrichment in FAST-FAST pairs
Enrichment in SLOW-SLOW pairs
Genomic Distance (number of genes)

Figure 2.6

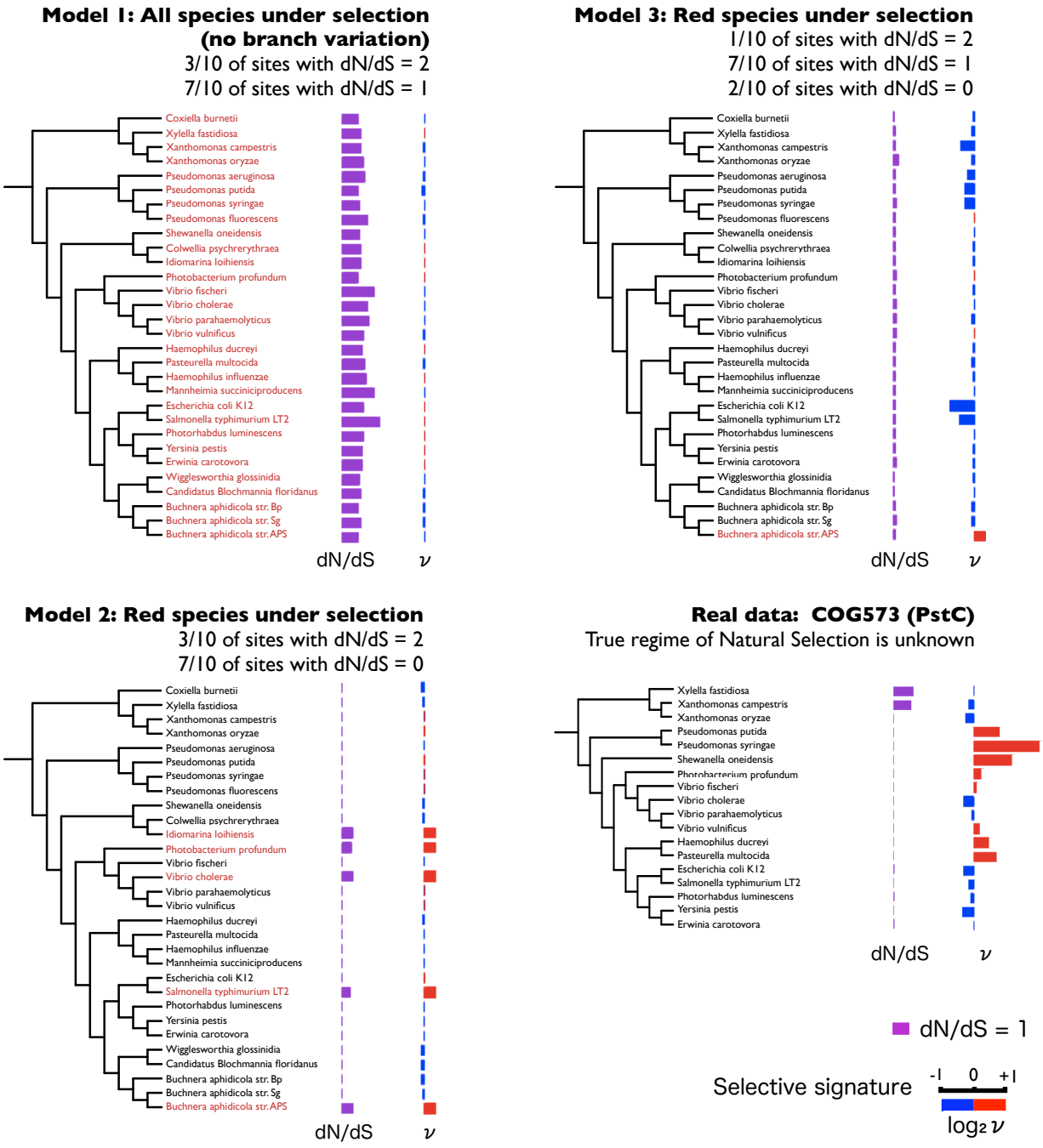

Model 1: All species under selection
(no branch variation)
3/10 of sites with dN/dS = 2
7/10 of sites with dN/dS = 1
Coxiella burnetii
Xylella fastidiosa
Xanthomonas campestris
Xanthomonas oryzae
Pseudomonas aeruginosa
Pseudomonas putida
Pseudomonas syringae
Pseudomonas fluorescens
Shewanella oneidensis
Colwellia psychrerythraea
Idiomarina loihiensis
Photobacterium profundum
Vibrio fischeri
Vibrio cholerae
Vibrio parahaemolyticus
Vibrio vulnificus
Haemophilus ducreyi
Pasteurella multocida
Haemophilus influenzae
Mannheimia succiniciproducens
Escherichia coli K12
Salmonella typhimurium LT2
Photorhabdus luminescens
Yersinia pestis
Erwinia carotovora
Wigglesworthia glossinidia
Candidatus Blochmannia floridanus
Buchnera aphidicola str. Bp
Buchnera aphidicola str. Sg
Buchnera aphidicola str. APS
dN/dS
ν
Model 3: Red species under selection
1/10 of sites with dN/dS = 2
7/10 of sites with dN/dS = 1
2/10 of sites with dN/dS = 0
Coxiella burnetii
Xylella fastidiosa
Xanthomonas campestris
Xanthomonas oryzae
Pseudomonas aeruginosa
Pseudomonas putida
Pseudomonas syringae
Pseudomonas fluorescens
Shewanella oneidensis
Colwellia psychrerythraea
Idiomarina loihiensis
Photobacterium profundum
Vibrio fischeri
Vibrio cholerae
Vibrio parahaemolyticus
Vibrio vulnificus
Haemophilus ducreyi
Pasteurella multocida
Haemophilus influenzae
Mannheimia succiniciproducens
Escherichia coli K12
Salmonella typhimurium LT2
Photorhabdus luminescens
Yersinia pestis
Erwinia carotovora
Wigglesworthia glossinidia
Candidatus Blochmannia floridanus
Buchnera aphidicola str. Bp
Buchnera aphidicola str. Sg
Buchnera aphidicola str. APS
dN/dS
ν
Model 2: Red species under selection
3/10 of sites with dN/dS = 2
7/10 of sites with dN/dS = 0
Coxiella burnetii
Xylella fastidiosa
Xanthomonas campestris
Xanthomonas oryzae
Pseudomonas aeruginosa
Pseudomonas putida
Pseudomonas syringae
Pseudomonas fluorescens
Shewanella oneidensis
Colwellia psychrerythraea
Idiomarina loihiensis
Photobacterium profundum
Vibrio fischeri
Vibrio cholerae
Vibrio parahaemolyticus
Vibrio vulnificus
Haemophilus ducreyi
Pasteurella multocida
Haemophilus influenzae
Mannheimia succiniciproducens
Escherichia coli K12
Salmonella typhimurium LT2
Photorhabdus luminescens
Yersinia pestis
Erwinia carotovora
Wigglesworthia glossinidia
Candidatus Blochmannia floridanus
Buchnera aphidicola str. Bp
Buchnera aphidicola str. Sg
Buchnera aphidicola str. APS
dN/dS
ν
Real data: COG573 (PstC)
True regime of Natural Selection is unknown
Xylella fastidiosa
Xanthomonas campestris
Xanthomonas oryzae
Pseudomonas putida
Pseudomonas syringae
Shewanella oneidensis
Photobacterium profundum
Vibrio fischeri
Vibrio cholerae
Vibrio parahaemolyticus
Vibrio vulnificus
Haemophilus ducreyi
Pasteurella multocida
Escherichia coli K12
Salmonella typhimurium LT2
Photorhabdus luminescens
Yersinia pestis
Erwinia carotovora
dN/dS
ν
dN/dS = 1
Selective signature
-1
0
+1
log2 ν

Figure 2.7

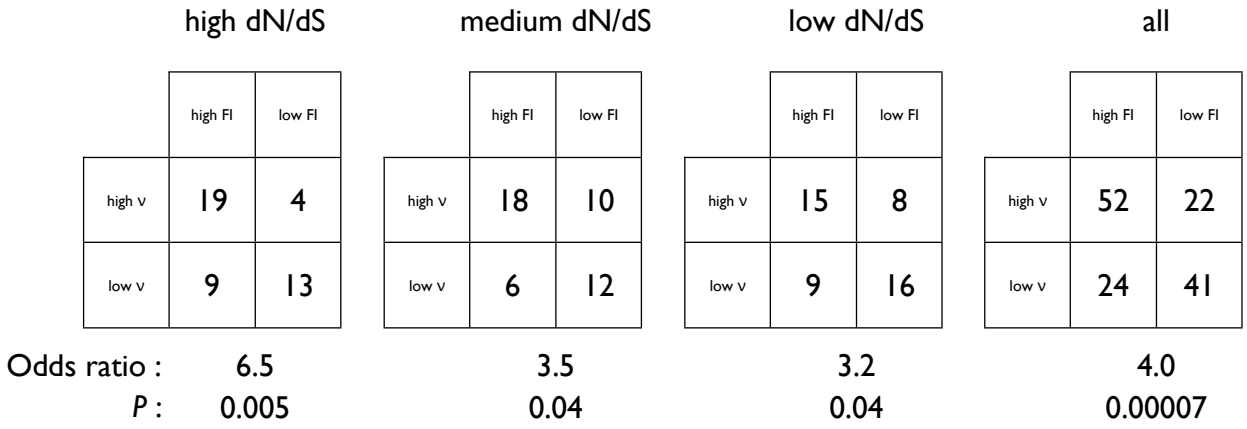

high dN/dS
medium dN/dS
low dN/dS
all
high FI
low FI
high v
low v
19
4
9
13
18
10
6
12
15
8
9
16
52
22
24
41
Odds ratio :
P :
6.5
0.005
3.5
0.04
3.2
0.04
4.0
0.00007

Figure 2.S1

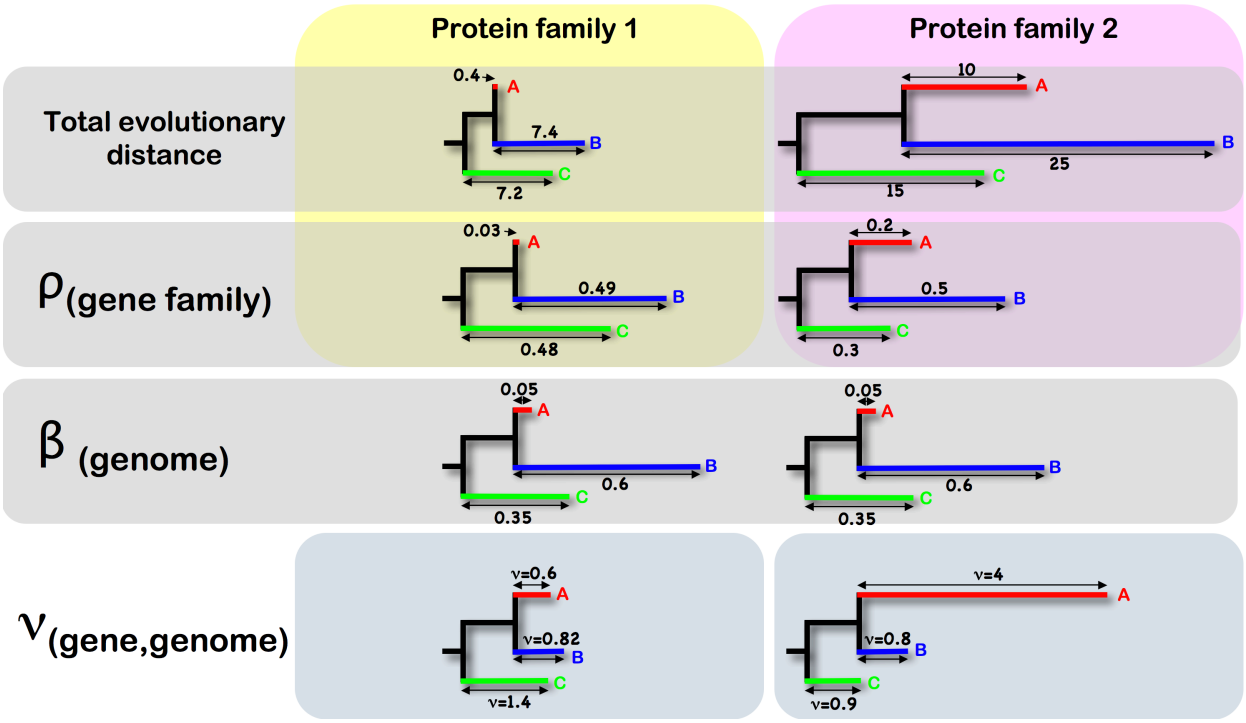

Protein family 1
Protein family 2
Total evolutionary distance
0.4
A
7.4
B
7.2
C
10
A
25
B
15
C
ρ(gene family)
0.03
A
0.49
B
0.48
C
0.2
A
0.5
B
0.3
C
β (genome)
0.05
A
0.6
B
0.35
C
0.05
A
0.6
B
0.35
C
ν(gene,genome)
ν=0.6
A
ν=0.82
B
ν=1.4
C
ν=4
A
ν=0.8
B
ν=0.9
C

Figure 2.S2

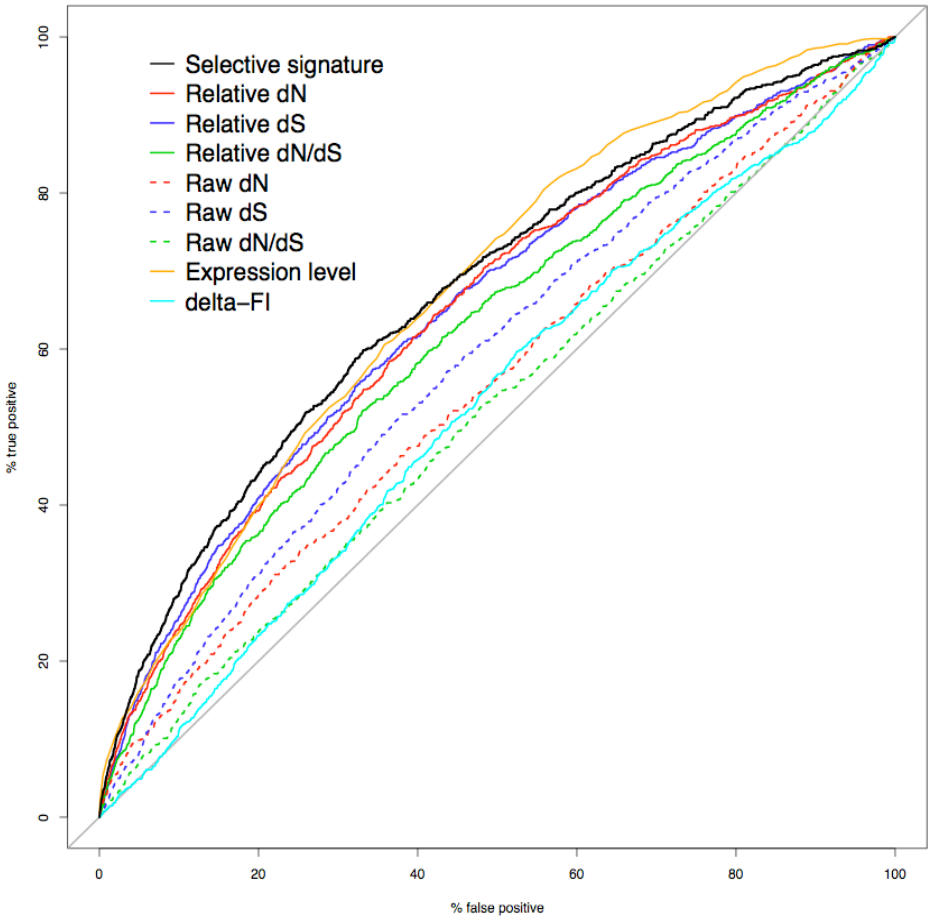

Selective signature
Relative dN
Relative dS
Relative dN/dS
Raw dN
Raw dS
Raw dN/dS
Expression level
delta−FI
% true positive
% false positive
0
20
40
60
80
100

Figure 2.S3

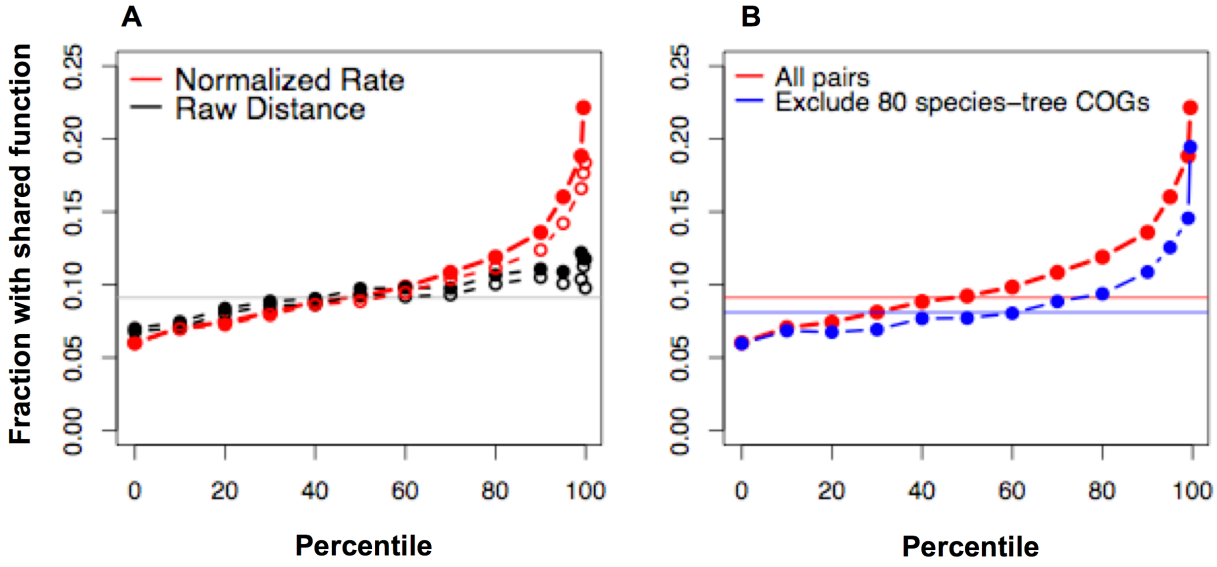


Figure 2.S4

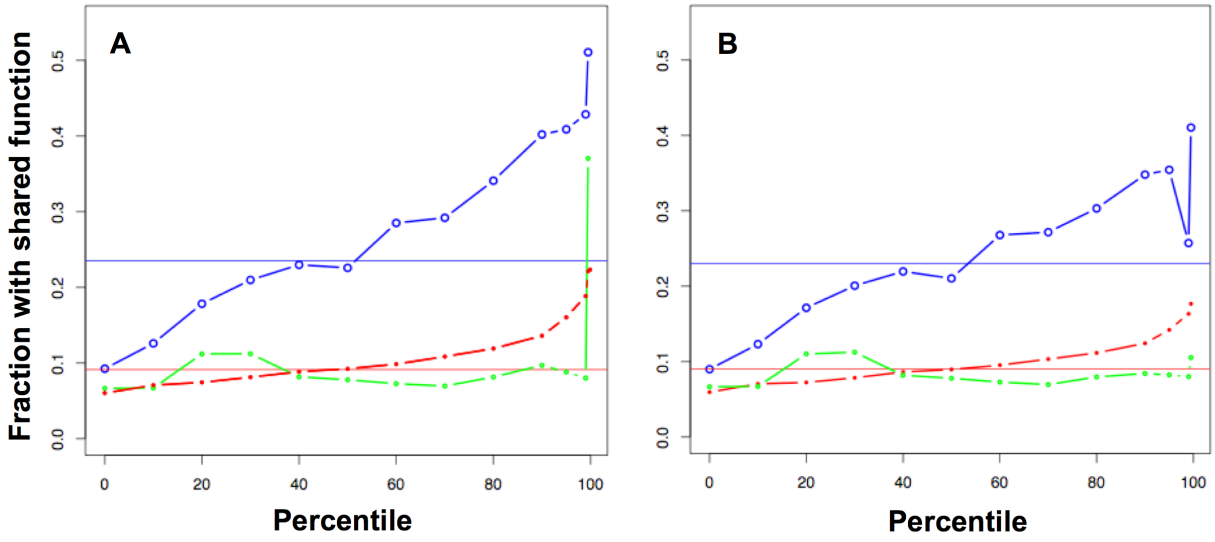


Figure 2.S5

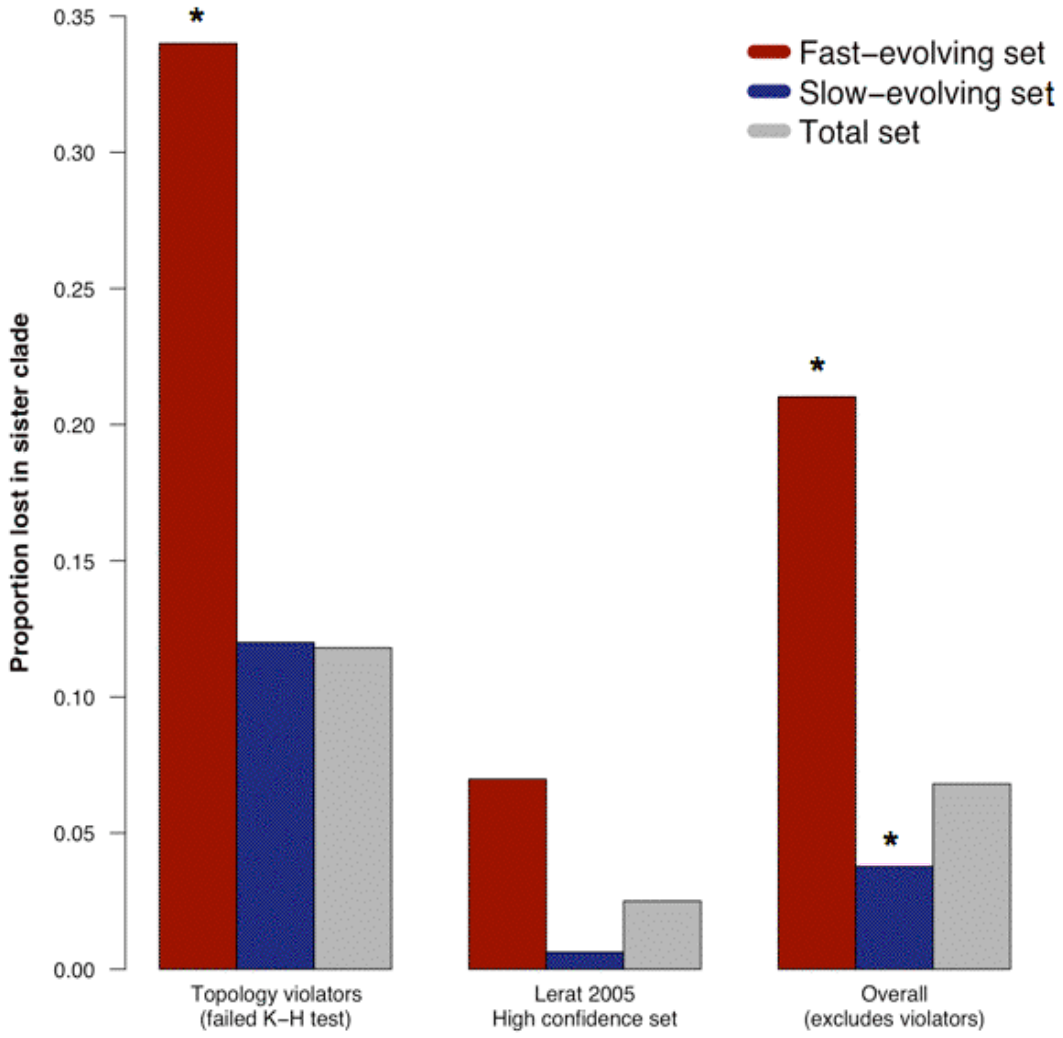

Fast−evolving set
Slow−evolving set
Total set
Proportion lost in sister clade
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
*
*
*
Topology violators
(failed K−H test)
Lerat 2005
High confidence set
Overall
(excludes violators)

Figure 2.S6

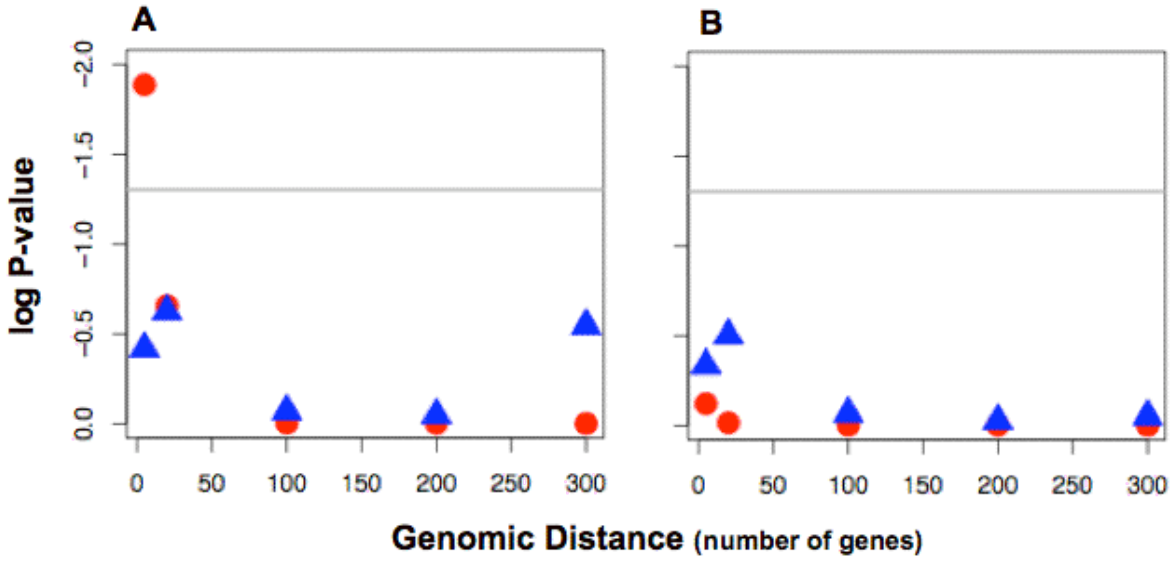


Figure 2.S7

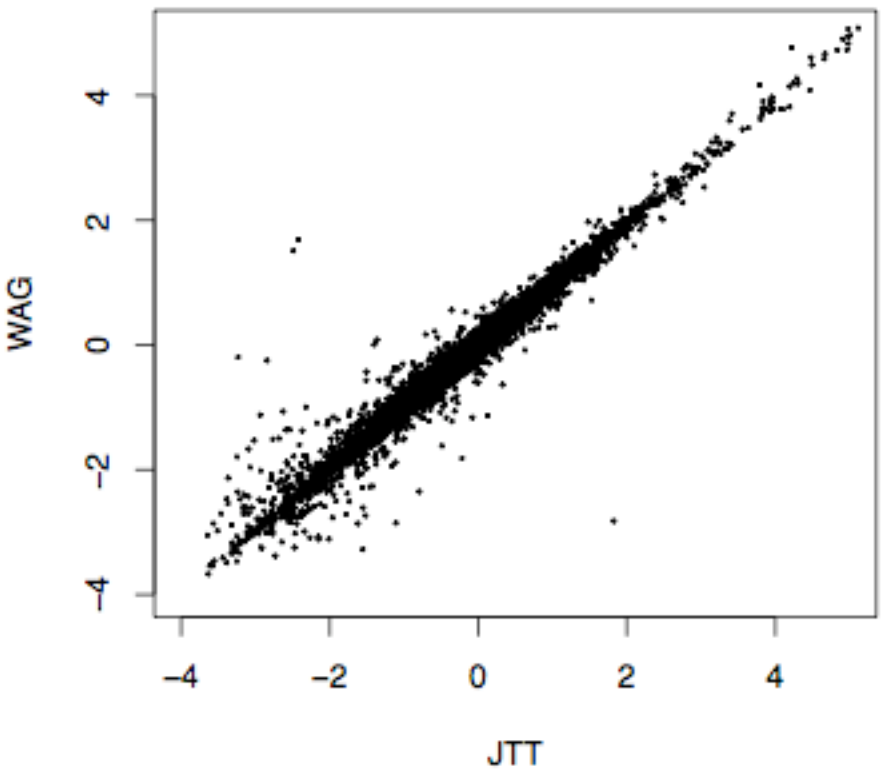


Figure 2.S8

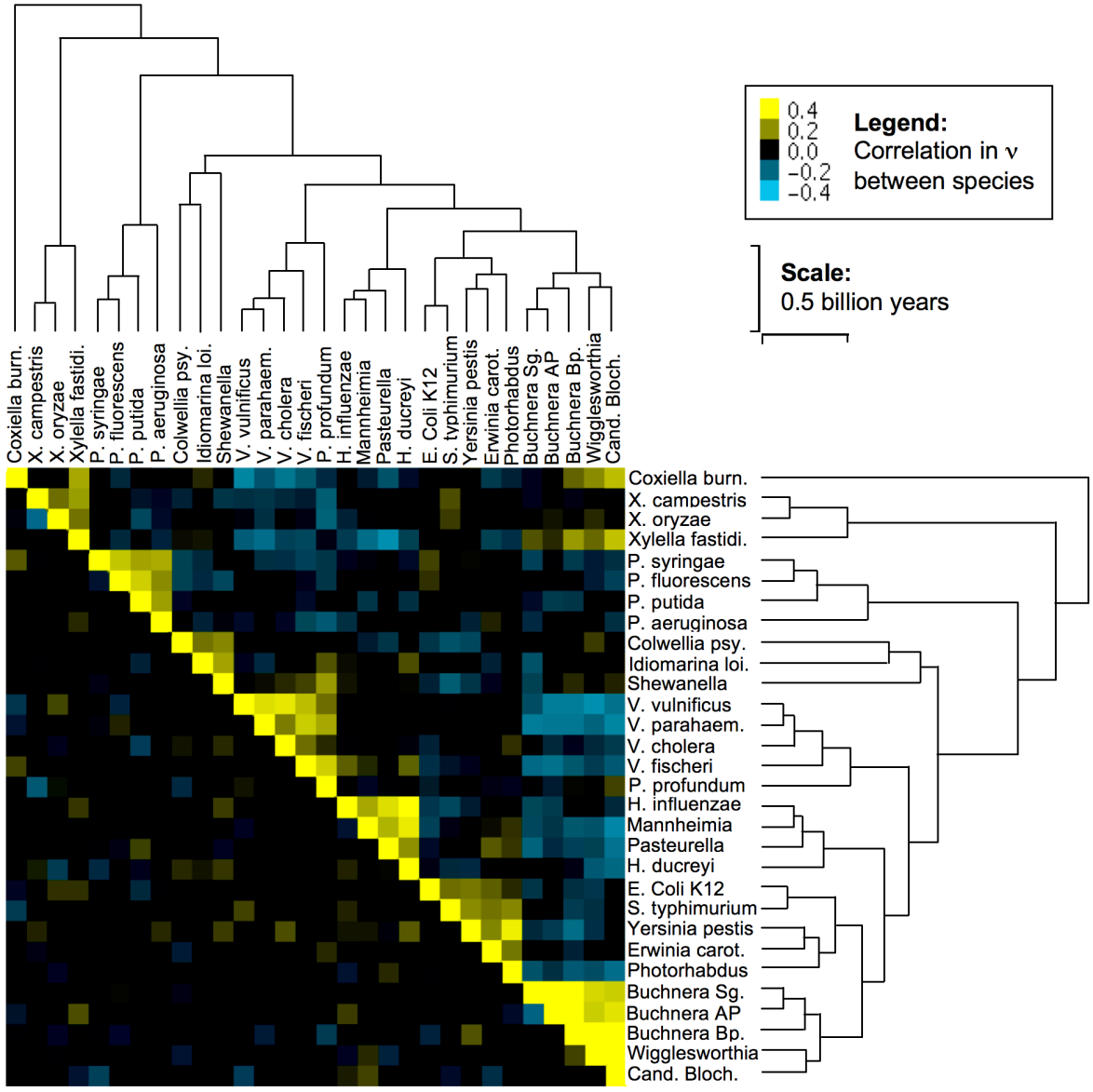

0.4
0.2
0.0
−0.2
−0.4
Legend:
Correlation in ν
between species
Scale:
0.5 billion years
Coxiella burn.
X. campestris
X. oryzae
Xylella fastidi.
P. syringae
P. fluorescens
P. putida
P. aeruginosa
Colwellia psy.
Idiomarina loi.
Shewanella
V. vulnificus
V. parahaem.
V. cholera
V. fischeri
P. profundum
H. influenzae
Mannheimia
Pasteurella
H. ducreyi
E. Coli K12
S. typhimurium
Yersinia pestis
Erwinia carot.
Photorhabdus
Buchnera Sg.
Buchnera AP
Buchnera Bp.
Wigglesworthia
Cand. Bloch.

Figure 2.S9

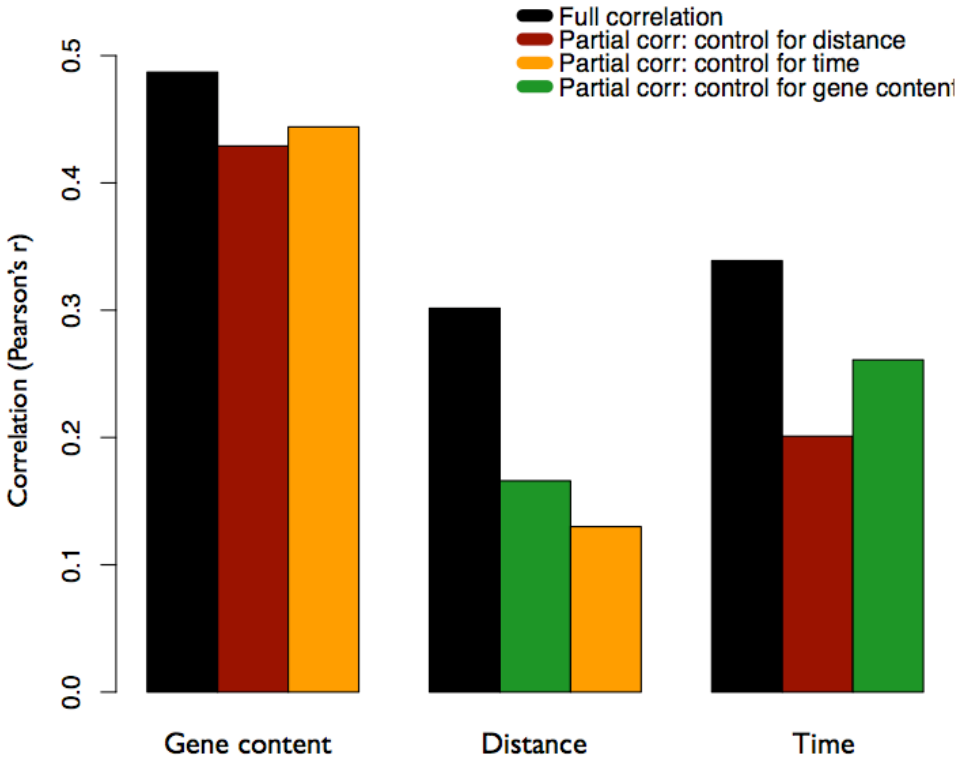


Figure 2.S10

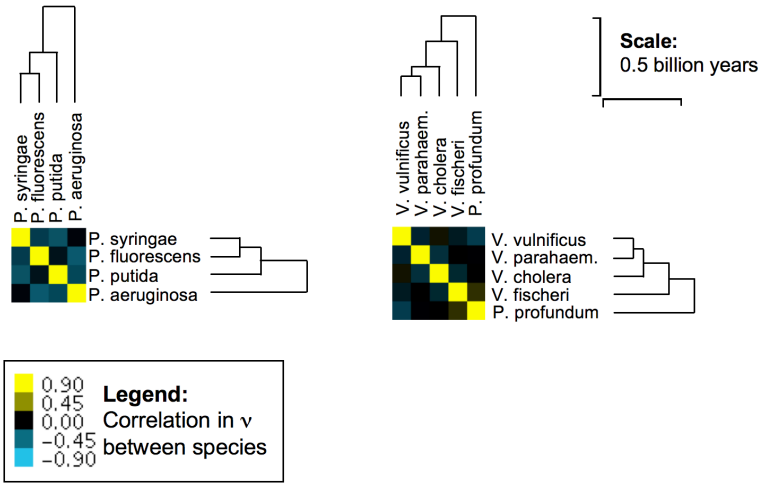


Table 2.S1

COG	Name	Functional Category	Description
20	UppS	I	Undecaprenyl pyrophosphate synthase [Lipid metabolism]
30	KsgA	J	Dimethyladenosine transferase (rRNA methylation) [Translation, ribosomal structure and biogenesis]
48	RpsL	J	Ribosomal protein S12 [Translation, ribosomal structure and biogenesis]
51	RpsJ	J	Ribosomal protein S10 [Translation, ribosomal structure and biogenesis]
52	RpsB	J	Ribosomal protein S2 [Translation, ribosomal structure and biogenesis]
60	IleS	J	Isoleucyl-tRNA synthetase [Translation, ribosomal structure and biogenesis]
80	RplK	J	Ribosomal protein L11 [Translation, ribosomal structure and biogenesis]
86	RpoC	K	DNA-directed RNA polymerase, beta' subunit/160 kD subunit [Transcription]
87	RplC	J	Ribosomal protein L3 [Translation, ribosomal structure and biogenesis]
88	RplD	J	Ribosomal protein L4 [Translation, ribosomal structure and biogenesis]
89	RplW	J	Ribosomal protein L23 [Translation, ribosomal structure and biogenesis]
90	RplB	J	Ribosomal protein L2 [Translation, ribosomal structure and biogenesis]
92	RpsC	J	Ribosomal protein S3 [Translation, ribosomal structure and biogenesis]
97	RplF	J	Ribosomal protein L6P/L9E [Translation, ribosomal structure and biogenesis]
98	RpsE	J	Ribosomal protein S5 [Translation, ribosomal structure and biogenesis]
102	RplM	J	Ribosomal protein L13 [Translation, ribosomal structure and biogenesis]
103	RpsI	J	Ribosomal protein S9 [Translation, ribosomal structure and biogenesis]
124	HisS	J	Histidyl-tRNA synthetase [Translation, ribosomal structure and biogenesis]
172	SerS	J	Seryl-tRNA synthetase [Translation, ribosomal structure and biogenesis]
177	Nth	L	Predicted EndoIII-related endonuclease [DNA replication, recombination, and repair]
193	Pth	J	Peptidyl-tRNA hydrolase [Translation, ribosomal structure and biogenesis]
194	Gmk	F	Guanylate kinase [Nucleotide transport and metabolism]
195	NusA	K	Transcription elongation factor [Transcription]
197	RplP	J	Ribosomal protein L16/L10E [Translation, ribosomal structure and biogenesis]
198	RplX	J	Ribosomal protein L24 [Translation, ribosomal structure and biogenesis]
199	RpsN	J	Ribosomal protein S14 [Translation, ribosomal structure and biogenesis]
200	RplO	J	Ribosomal protein L15 [Translation, ribosomal structure and biogenesis]
201	SecY	U	Preprotein translocase subunit SecY [Intracellular trafficking and secretion]
203	RplQ	J	Ribosomal protein L17 [Translation, ribosomal structure and biogenesis]
211	RpmA	J	Ribosomal protein L27 [Translation, ribosomal structure and biogenesis]
216	PrfA	J	Protein chain release factor A [Translation, ribosomal structure and biogenesis]
222	RplL	J	Ribosomal protein L7/L12 [Translation, ribosomal structure and biogenesis]
228	RpsP	J	Ribosomal protein S16 [Translation, ribosomal structure and biogenesis]
238	RpsR	J	Ribosomal protein S18 [Translation, ribosomal structure and biogenesis]
244	RplJ	J	Ribosomal protein L10 [Translation, ribosomal structure and biogenesis]
256	RplR	J	Ribosomal protein L18 [Translation, ribosomal structure and biogenesis]
268	RpsT	J	Ribosomal protein S20 [Translation, ribosomal structure and biogenesis]
275	MraW	M	Predicted S-adenosylmethionine-dependent methyltransferase involved in cell envelope biogenesis
285	FolC	H	Folylpolyglutamate synthase [Coenzyme metabolism]
292	RplT	J	Ribosomal protein L20 [Translation, ribosomal structure and biogenesis]
313	YraL	R	Predicted methyltransferases [General function prediction only]
319	YbeY	R	Predicted metal-dependent hydrolase [General function prediction only]
335	RplS	J	Ribosomal protein L19 [Translation, ribosomal structure and biogenesis]
336	TrmD	J	tRNA-(guanine-N1)-methyltransferase [Translation, ribosomal structure and biogenesis]
354	YgfZ	R	Predicted aminomethyltransferase related to GcvT [General function prediction only]
359	RplI	J	Ribosomal protein L9 [Translation, ribosomal structure and biogenesis]
360	RpsF	J	Ribosomal protein S6 [Translation, ribosomal structure and biogenesis]
361	InfA	J	Translation initiation factor 1 (IF-1) [Translation, ribosomal structure and biogenesis]
445	GidA	D	NAD/FAD-utilizing enzyme apparently involved in cell division [Cell division and chromosome partitioning
481	LepA	M	Membrane GTPase LepA [Cell envelope biogenesis, outer membrane]
482	TrmU	J	Predicted tRNA(5-methylaminomethyl-2-thiouridylate) methyltransferase [Translation, ribosomal struct.]
486	ThdF	R	Predicted GTPase [General function prediction only]
495	LeuS	J	Leucyl-tRNA synthetase [Translation, ribosomal structure and biogenesis]

504	PyrG	F	CTP synthase (UTP-ammonia lyase) [Nucleotide transport and metabolism]
522	RpsD	J	Ribosomal protein S4 and related proteins [Translation, ribosomal structure and biogenesis]
525	ValS	J	Valyl-tRNA synthetase [Translation, ribosomal structure and biogenesis]
532	InfB	J	Translation initiation factor 2 (IF-2; GTPase) [Translation, ribosomal structure and biogenesis]
533	QRI7	O	Metal-dependent proteases with possible chaperone activity [Posttranslational modification, chaperones]
536	Obg	R	Predicted GTPase [General function prediction only]
539	RpsA	J	Ribosomal protein S1 [Translation, ribosomal structure and biogenesis]
571	Rnc	K	dsRNA-specific ribonuclease [Transcription]
592	DnaN	L	DNA polymerase sliding clamp subunit (PCNA homolog) [DNA replication, recombination, and repair]
691	SmpB	O	tmRNA-binding protein [Posttranslational modification, protein turnover, chaperones]
707	MurG	M	UDP-N-acetylglucosamine:LPS N-acetylglucosamine transferase [Cell envelope biogenesis]
751	GlyS	J	Glycyl-tRNA synthetase, beta subunit [Translation, ribosomal structure and biogenesis]
771	MurD	M	UDP-N-acetylmuramoylalanine-D-glutamate ligase [Cell envelope biogenesis, outer membrane]
806	RimM	J	RimM protein, required for 16S rRNA processing [Translation, ribosomal structure and biogenesis]
812	MurB	M	UDP-N-acetylmuramate dehydrogenase [Cell envelope biogenesis, outer membrane]
816	YqgF	L	Predicted endonuclease involved in recombination (possible Holliday junction resolvase) [DNA recomb.]
849	FtsA	D	Actin-like ATPase involved in cell division [Cell division and chromosome partitioning]
858	RbfA	J	Ribosome-binding factor A [Translation, ribosomal structure and biogenesis]
1185	Pnp	J	Polyribonucleotide nucleotidyltransferase (polynucleotide phosphorylase) [Translation, ribosomal]
1207	GlmU	M	N-acetylglucosamine-1-phosphate uridyltransferase [Cell envelope biogenesis, outer membrane]
1214	YeaZ	O	Inactive homolog of metal-dependent proteases, putative molecular chaperone [Posttrans./chaperones]
1825	RplY	J	Ribosomal protein L25 (general stress protein Ctc) [Translation, ribosomal structure and biogenesis]
1841	RpmD	J	Ribosomal protein L30/L7E [Translation, ribosomal structure and biogenesis]
1862	YajC	U	Preprotein translocase subunit YajC [Intracellular trafficking and secretion]
1932	SerC	HE	Phosphoserine aminotransferase [Coenzyme metabolism / Amino acid transport and metabolism]
1949	Orn	A	Oligoribonuclease (3'->5' exoribonuclease) [RNA processing and modification]
2924	YggX	S	Uncharacterized protein conserved in bacteria [Function unknown]

Table 2.S2a

Taxonomy ID	Species Name
107806	*Buchnera aphidicola str. APS*
224915	*Buchnera aphidicola str. Bp*
198804	*Buchnera aphidicola str. Sg*
203907	*Candidatus Blochmannia floridanus*
167879	*Colwellia psychrerythraea 34h*
227377	*Coxiella burnetii RSA*
218491	*Erwinia carotovora SCRI1043*
83333	*Escherichia coli K12*
233412	*Haemophilus ducreyi35000HP*
71421	*Haemophilus influenzae Rd KW20*
283942	*Idiomarina loihiensis L2TR*
221988	*Mannheimia succiniciproducens MBEL55E*
747	*Pasteurella multocida*
298386	*Photobacterium profundum SS9*
243265	*Photorhabdus luminescens TTO1*
208964	*Pseudomonas aeruginosa PAO1*
220664	*Pseudomonas fluorescens Pf-5*
160488	*Pseudomonas putida KT2440*
223283	*Pseudomonas syringae*
99287	*Salmonella typhimurium LT2*
211586	*Shewanella oneidensis MR-1*
666	*Vibrio cholerae*
312309	*Vibrio fischeri ES114*
223926	*Vibrio parahaemolyticus RIMD2210633*
216895	*Vibrio vulnificus CMCP6*
36870	*Wigglesworthia glossinidia*
190485	*Xanthomonas campestris ATCC33913*
291331	*Xanthomonas oryzae pv. oryzae KACC103*
160492	*Xylella fastidiosa 9a5c*
273123	*Yersinia pestis IP32953*

Table 2.S2b

Taxonomy ID	Species Name
358709	*Escherichia coli 101-1*
362663	*Escherichia coli 536*
405955	*Escherichia coli APEC O1*
37762	*Escherichia coli B*
344601	*Escherichia coli B171*
340184	*Escherichia coli B7A*
199310	*Escherichia coli CFT073*
340186	*Escherichia coli E110019*
340185	*Escherichia coli E22*
331111	*Escherichia coli E24377A*
340197	*Escherichia coli F11*
331112	*Escherichia coli HS*
83333	*Escherichia coli K12*
155864	*Escherichia coli O157:H7 EDL933*
386585	*Escherichia coli O157:H7 str. Sakai*
364106	*Escherichia coli UTI89*
316407	*Escherichia coli W3110*
344609	*Shigella boydii BS512*
300268	*Shigella boydii Sb227*
300267	*Shigella dysenteriae Sd197*
198215	*Shigella flexneri 2a str. 2457T*
198214	*Shigella flexneri 2a str. 301*
373384	*Shigella flexneri 5 str. 8401*
300269	*Shigella sonnei Ss046*
	outgroup:
99287	*Salmonella typhimurium LT2*

Table 2.S3a

Species	COG function	Enrichment	Bonferonni-corrected P-value
Haemophilus ducreyi	Ribosome & Translation	slow	<0.0001
C. Blochmannia floridanus	Ribosome & Translation	fast	<0.0001
Xylella fastidiosa	Ribosome & Translation	fast	<0.0001
Escherichia coli K12	Motlity & Secretion	fast	<0.001
Pasteurella multocida	Ribosome & Translation	slow	<0.001
Photorhabdus luminescens	Motlity & Secretion	fast	<0.001
Mannheimia succiniciproducens	Ribosome & Translation	slow	<0.001
Photorhabdus luminescens	Ribosome & Translation	slow	<0.001
Vibrio parahaemolyticus	Amino acid metabolism	slow	<0.01
Haemophilus influenzae	Ribosome & Translation	slow	<0.01
Buchnera aphidicola str. APS	Ion transport & metabolism	slow	<0.01
Wigglesworthia glossinidia	Coenzyme metabolism	slow	<0.01
Haemophilus ducreyi	Cell Division	fast	<0.05
Vibrio vulnificus	Nucleic acid metabolism	slow	<0.05
Yersinia pestis	Motlity & Secretion	fast	<0.05
Idiomarina loihiensis	Carbohydrate metabolism	fast	<0.05
Xylella fastidiosa	Energy production	slow	<0.05
Idiomarina loihiensis	Amino acid metabolism	fast	<0.05
Photobacterium profundum	Cell Division	fast	<0.05

Table 2.S3b

COG	Name	*E. coli K12*	*P. luminescens*	*Y. pestis*	Description
1377	FlhB	*	*		Flagellar biosynthesis pathway
1684	FliR	*	*		Flagellar biosynthesis pathway
3418	FlgN	*	*	*	Flagellar biosynthesis/type III secretory pathway chaperone
4787	FlgF	*	*		Flagellar basal body rod protein
1261	FlgA	*	*		Flagellar basal body P-ring biosynthesis protein
3190	FliO	*			Flagellar biosynthesis protein
4967	PilV	*		*	Tfp pilus assembly protein
1815	FlgB	*			Flagellar basal body protein
1345	FliD	*			Flagellar capping protein
1516	FliS	*	*		Flagellin-specific chaperone
4786	FlgG		*		Flagellar basal body rod protein
1706	FlgI		*		Flagellar basal body P-ring protein
1677	FliE		*		Flagellar hook basal body protein
2805	PilT		*	*	Tfp pilus assembly protein, pilus retraction ATPase
4969	PilA			*	Tfp pilus assembly protein, major pilin
1989	PulO			*	Type II secretory pathway, prepilin signal peptidase PulO and related peptidases

Table 2.S4

			Test A: Selection on *Idiomarina*		Test B: Site-specific selection on *Idiomarina*	
Pathway	**Gene**	**COG**	**Different dN/dS in *Idiomarina* ?**	**dN/dS in *Idiomarina* (relative to other lineages)**	**Sites with higher dN/dS in *Idiomarina*?**	**Proportion of sites with higher dN/dS in *Idiomarina* (dN/dS > 1 in bold)**
Glycolysis	*Pgi*	166	yes*	0.0066 (lower)	yes***	0.40
	Fbp	158	yes**	0.0022 (lower)	yes***	0.18
	Eno	148	yes***	0.0017 (lower)	yes***	**0.08**
	TpiA	149	yes*	0.0056 (lower)	yes***	**0.26**
	Pfk	205	yes*	0.0038 (lower)	yes***	0.30
	NagE	2190	no	0.0185	yes***	0.75
Phenylalanine biosynthesis	*AroQ*	757	yes*	0.0028 (lower)	yes*	0.07
	AroB	337	no	0.0044	yes***	0.23
	AroE	169	no	0.0055	yes***	**0.16**
	PheA	77	no	0.0115	yes**	0.45

Likelihood ratio test: * P <0.05, ** P < 0.005, *** P < 0.00001

Table 2.S5

COG	$\log_2$ v		$\log_2$ FI		dN/dS	Func	Description
112	2.13	**	2.47	*	0.12	E	serine hydroxymethyltransferase
118	2.02	**	0.82		0.07	E	imidazole glycerol phosphate synthase subunit HisH
150	1.96	**	0.69		0.03	F	phosphoribosylaminoimidazole synthetase
2973	1.86	*	0.41		0.00	K	Trp operon repressor
157	1.86	**	0.68		0.02	H	nicotinate-nucleotide pyrophosphorylase
2980	1.73	**	0.43		0.07	M	minor lipoprotein
704	1.54	*	0.84		3.21	P	negative regulator of PhoR/PhoB two-component regulator
357	1.51	**	0.61		0.02	M	glucose-inhibited division protein B
99	1.48	*	0.74		0.24	J	30S ribosomal protein S13
649	1.48	*	1.87	*	0.02	C	NADH:ubiquinone oxidoreductase, chain C,D
196	1.33	**	0.54		0.03	H	hypothetical protein
414	1.30	*	2.02	*	0.04	H	pantoate--beta-alanine ligase
2109	1.28	*	0.50		0.00	H	cob(I)yrinic acid a,c-diamide adenosyltransferase
3123	1.28	*	0.94		0.02	S	hypothetical protein
1214	1.27	**	0.94		0.05	O	predicted peptidase
3159	1.21	*	0.90		0.09	S	hypothetical protein
850	1.19	**	0.37		0.06	D	septum formation inhibitor
221	1.16		1.03		0.09	C	inorganic pyrophosphatase
2861	1.15	*	0.85		0.63	S	orf, hypothetical protein
1983	1.12	*	0.93		0.06	KT	DNA-binding transcriptional activator
529	1.11		0.55		0.05	P	adenylylsulfate kinase
149	1.07	*	2.08		0.08	G	triosephosphate isomerase
205	1.05	*	0.89		0.03	G	6-phosphofructokinase
3076	1.03		0.30		0.02	S	hypothetical protein
229	0.97	*	0.67		0.02	O	methionine sulfoxide reductase B
540	0.96	*	1.68		0.02	F	aspartate carbamoyltransferase catalytic subunit
1005	0.95		1.82	*	0.16	C	NADH dehydrogenase subunit H
751	0.88	*	2.09	*	0.05	J	glycyl-tRNA synthetase subunit beta
17	0.86		1.00		0.05	J	asparaginyl-tRNA synthetase
481	0.85		0.94		0.02	M	GTP-binding protein LepA
1666	0.84		2.52	*	0.07	S	orf, hypothetical protein
344	0.82	*	1.74		0.08	S	hypothetical protein
682	0.80	*	1.93	*	2.50	M	prolipoprotein diacylglyceryl transferase
2377	0.80	**	0.37		0.01	O	anhydro-N-acetylmuramic acid kinase
554	0.74		0.27		0.03	C	glycerol kinase
151	0.73		0.29		4.33	F	phosphoribosylamine--glycine ligase
233	0.73		0.95		0.07	J	ribosome releasing factor
249	0.69	*	1.51	*	0.02	L	DNA mismatch repair protein
66	0.69		0.32		0.08	E	isopropylmalate isomerase small subunit
212	0.68	*	0.61		0.01	H	putative ligase
1207	0.64	*	0.91		0.06	M	bifunctional N-acetylglucosamine-1-phosphate uridyltransferase/glucosamine-1-phosphate acetyltransferase
2908	0.64	*	0.44		0.06	S	UDP-2,3-diacylglucosamine hydrolase
165	0.63		0.37		0.09	E	argininosuccinate lyase
468	0.63		0.67		0.01	L	recombinase A
3004	0.62		0.66		0.00	P	pH-dependent sodium/proton antiporter
2360	0.62	*	0.56		0.04	O	leucyl/phenylalanyl-tRNA--protein transferase
1485	0.60	*	0.37		0.04	R	conserved protein with nucleoside triphosphate hydrolase domain
106	0.58		0.32		0.02	E	N-(5'-phospho-L-ribosyl-formimino)-5-amino-1-(5'- phosphoribosyl)-4-imidazolecarboxamide isomerase
323	0.58	*	0.60		0.02	L	DNA mismatch repair protein
418	0.58	*	0.30		0.02	F	dihydroorotase
629	0.56		1.41		0.06	L	single-strand DNA-binding protein

2968	0.52		1.34		0.02	S	hypothetical protein
519	-0.51	*	-1.10		0.01	F	bifunctional GMP synthase/glutamine amidotransferase protein
2917	-0.52		-1.13		0.01	D	intracellular septation protein A
2925	-0.52	*	-1.45	*	0.01	L	exonuclease I
262	-0.57	*	-0.91		0.77	H	dihydrofolate reductase
1188	-0.59		-1.46		2.06	J	ribosome-associated heat shock protein Hsp15
275	-0.65	*	-0.98		0.01	M	S-adenosyl-methyltransferase
1314	-0.71	*	-0.87		0.09	U	protein-export membrane protein
4121	-0.73	*	-0.76	*	0.08	S	putative peptidase
324	-0.74		-1.90	*	0.05	J	tRNA delta(2)-isopentenylpyrophosphate transferase
2976	-0.76	*	-1.58		0.04	S	hypothetical protein
93	-0.81		-2.32		0.15	J	50S ribosomal protein L14
1212	-0.83	*	-1.61	*	0.00	M	3-deoxy-manno-octulosonate cytidylyltransferase
777	-0.99	*	-2.35	*	0.05	I	acetyl-CoA carboxylase subunit beta
774	-1.04	*	-1.26		0.04	M	UDP-3-O-[3-hydroxymyristoyl] N-acetylglucosamine deacetylase
5633	-1.05	*	-2.65	*	0.08	R	hypothetical protein
2190	-1.11	*	-0.99		0.03	G	glucose-specific PTS system enzyme IIA component
4657	-1.13	*	-2.93	*	0.03	C	Na(+)-translocating NADH-quinone reductase subunit E
360	-1.15	*	-3.00	*	0.06	J	30S ribosomal protein S6
563	-1.25	*	-0.85		0.10	F	adenylate kinase
466	-1.28	*	-3.55	*	0.01	O	DNA-binding ATP-dependent protease La
3151	-1.37	*	-2.46		0.02	S	predicted dehydrogenase
4785	-1.37	**	-0.99		0.08	R	hypothetical protein
101	-1.41	*	-1.14		1.40	J	tRNA pseudouridine synthase A
821	-1.50	*	-0.81		0.01	I	4-hydroxy-3-methylbut-2-en-1-yl diphosphate synthase
2983	-1.52	*	-1.84	*	0.00	S	orf, hypothetical protein
509	-1.58	*	-2.62	*	0.01	E	glycine cleavage system protein H
287	-1.58	*	-0.88		1.62	E	fused chorismate mutase T/prephenate dehydrogenase
788	-1.63	*	-1.34		1.28	F	formyltetrahydrofolate deformylase
634	-1.69	*	-1.68		0.02	F	hypoxanthine phosphoribosyltransferase
283	-1.69	*	-1.75		0.01	F	cytidylate kinase
131	-1.75	*	-1.71	*	0.02	E	imidazoleglycerolphosphate dehydratase and histidinol-phosphate phosphatase
452	-2.01	*	-1.05		0.03	H	flavoprotein affecting synthesis of DNA and pantothenate metabolism
41	-2.43	*	-1.23		0.04	F	phosphoribosylaminoimidazole carboxylase catalytic subunit
3317	-2.51	*	-1.52	*	0.03	M	lipoprotein-34
575	-2.60	*	-1.77	*	1.38	I	CDP-diglyceride synthetase
779	-2.62	*	-1.15		0.03	S	orf, hypothetical protein
539	-2.71	*	-1.32		0.02	J	30S ribosomal protein S1
3117	-2.77	*	-1.72	*	0.00	S	hypothetical protein
103	-2.78	*	-1.42		0.06	J	30S ribosomal protein S9
55	-2.86	*	-1.16		0.01	C	F0F1 ATP synthase subunit beta
1385	-2.89	*	-1.68	*	1.34	S	orf, hypothetical protein

Chapter 3:

The Slow:Fast substitution ratio reveals changing patterns of natural selection in γ-proteobacterial genomes

Shapiro BJ and Alm E (2009) The slow:fast substitution ratio reveals changing patterns of natural selection in gamma-proteobacterial genomes. *ISME Journal* **10**:1180-1192

Abstract

Different microbial species are thought to occupy distinct ecological niches, subjecting each species to unique selective constraints, which may leave a recognizable signal in their genomes. Thus, it may be possible to extract insight into the genetic basis of ecological differences among lineages by identifying unusual patterns of substitutions in orthologous gene or protein sequences. We use the ratio of substitutions in slow versus fast-evolving sites (nucleotides in DNA, or amino acids in protein sequence) to quantify deviations from the typical pattern of selective constraint observed across bacterial lineages. We propose that elevated S:F in one branch (an excess of slow-site substitutions) can indicate a functionally-relevant change, due to either positive selection or relaxed evolutionary constraint. In a genome-wide comparative study of γ-proteobacterial proteins, we find that cell-surface proteins involved with motility and secretion functions often have high S:F ratios, while information-processing genes do not. Change in evolutionary constraints in some species is evidenced by increased S:F ratios within functionally-related sets of genes (*e.g.* energy production in *Pseudomonas fluorescens*), while other species apparently evolve mostly by drift (*e.g.* uniformly elevated S:F across most genes in *Buchnera* spp.). Overall, S:F reveals several species-specific, protein-level changes with potential functional/ecological importance. As microbial genome projects yield more species-rich gene-trees, the S:F ratio will become an increasingly powerful tool for uncovering functional genetic differences among species.

3.1 Introduction

Natural selection is an evolutionary force that promotes the spread of beneficial alleles in a population (positive/diversifying selection), and impedes the spread of deleterious alleles (negative/purifying selection). Selection is intimately tied to ecology: depending on the ecological niche of an organism (e.g. its source of carbon and nutrients, interactions with predators, hosts, competitors, etc.), different mutations will be favoured by selection. Genome-wide scans for natural selection have the potential to identify ecologically-relevant genetic adaptations, even when the adaptive traits themselves remain obscure (Li et al, 2008). Such genome-wide approaches, sometimes referred to as 'reverse ecology', thus have great potential to elucidate the 'hidden world' of microbial ecology. Reverse ecology requires a sampling of related genomes to quantify genetic differences and similarities within or between species, and an appropriate genome-wide test for selection. Most tests for selection have been developed with sexual eukaryotes in mind, and may not always be amenable to microbes (Shapiro et al, 2009). Evidence for selection can be detected over relatively recent time scales by studying allele frequencies within populations (Sabeti et al, 2002, Zeng et al, 2006), or over longer time scales by studying protein evolution between species (Shapiro & Alm, 2008, Jordan et al, 2001, Yang, 1998).

At the protein sequence level, natural selection is often quantified using dN/dS (substitutions per nonsynonymous site/substitutions per synonymous site). The theoretical foundation of dN/dS can be traced back in the development of the neutral theory of molecular evolution, when Kimura made an important observation: while not all synonymous mutations are necessarily functionally neutral, "the possibility is very high that, on average, synonymous changes are subject to natural selection very much less than the mis-sense mutations" (Kimura, 1977). When an excess of mis-sense mutations is observed relative to nearly-neutral silent mutations ($dN/dS > 1$), this provides strong evidence for positive selection on a protein or portion thereof. Meanwhile, dN/dS close to zero indicates strong selective constraint, and $dN/dS \approx 1$ indicates low or 'relaxed' selective constraint (e.g. pseudogenes). Relaxed constraint amounts to reduced efficacy of purifying selection in purging deleterious mutations from a population. It has long been recognized that dN/dS loses power to detect positive selection over long divergence times because dS becomes 'saturated' with multiple substitutions. More recently, it has been recognized that dN/dS may also be unreliable over very short divergence times between species (Nozawa et al, 2009) or when it is applied within a single population (Kryazhimskiy & Plotkin, 2008). These issues may be particularly acute in studies of microbial genomes, where populations may be ill-defined, or divergences times may be ancient (on the scale of millions to billions of years).

In this work, we introduce the slow:fast substitution ratio (S:F) as a metric for detecting variation in natural selection on biological sequences - either nucleotides or amino acids – and apply it to detect variation in natural selection among bacterial species that are sufficiently diverged that most synonymous sites have undergone multiple substitutions (saturated dS). The logic underlying this new method is analogous to the logic of dN/dS (see Supplementary Note 3): sequences with an excess of substitutions in sites (positions of a nucleotide or protein sequence alignment) of probable functional importance (slow-evolving), relative to the nearly-neutral standard of substitutions in sites of less functional importance (fast-evolving), are candidate targets of positive or relaxed negative selection. Unlike dN/dS, which defines site categories based on the genetic code, the S:F ratio instead relies on each site's observed substitution rate in a phylogeny of many species – and is thus applicable not just to codon sequences, but also to noncoding or protein sequences. Substitutions are first counted in each branch of the species phylogeny. For any extant or ancestral branch, we define S as the number of substitutions per 'slow-evolving' site, F as the number of substitutions per 'fast-evolving' site, and S:F as their ratio (Figure 3.1). Along with S:F, we calculate an odds ratio and *p*-value to assess the significance of the branch's deviation from the S:F observed for that gene in the rest of the gene-tree (Methods).

A potential limitation of the S:F method is that it condenses all the complexity of a gene sequence into a single number. Gene sequences contain a multitude of sites, some of which may be under strong purifying selection, some selectively neutral, and others under strong positive selection at any moment in time (Hughes, 2007, Hughes & Nei, 1988). Such an intricate pattern of selection across sites cannot be adequately captured by a simple summary statistic (prompting the development of site-specific models of dN/dS (Yang & Nielsen, 2002, Massingham & Goldman, 2005), although these model-based methods may suffer from false-positive and false-negative adaptive site predictions (Nozawa et al, 2009)). However, S:F is not designed to summarize the complex pattern of selection across the gene, but instead to quantify the extent of *change* in that pattern. Thus, S:F is a simple statistic quantifying changes in the regime of selection, while still acknowledging that this regime may be complex (*e.g.* sites and lineages with different selective constraints).

Even if a lineage is found to have a significantly high value of S:F, this may result from either adaptation (positive natural selection favoring novel mutations), or relaxed selective constraint (accumulation of neutral or mildly-deleterious mutations). In cases of S:F > 1, positive selection is a likely explanation, but such cases are expected to be rare. In more subtle cases where S:F is excessively high (but still < 1) in

one lineage, either positive or relaxed negative selection may be responsible, and S:F alone cannot be used to discriminate between these possibilities.

In this study, we calculated S:F for ~1000 protein families from 30 species of γ-proteobacteria, an ancient and ecologically diverse group, with evidence for species-specific positive selection on many of their core genes (Shapiro & Alm, 2008). We aimed to identify which genes – or functional modules of genes – contribute to species-specific adaptations. More specifically, we tested the hypotheses that (1) selective constraint, as measured with S:F, varies with protein function, and (2) that ecologically-distinct species experience different regimes of selection on proteins of different functions. We describe several examples of elevated S:F in proteins with functions relevant to species ecology, suggesting ecological adaptation at the protein level.

Results

3.2 Performance of S:F under simulated evolution

To understand the response of S:F to changing regimes of selection, we generated simulated gene sequences along the γ-proteobacteria species tree (Supplementary Methods). The simulated sequences, 300 codons long, contained 3 default site-classes: (1) dN/dS = 0.2 at 70% of sites, (2) dN/dS = 1.0 at 20% of sites, and (3) dN/dS = 0.1 at 10% of sites. In the 'baseline' scenario, all branches in the tree evolved according to these site-classes. Under 'selection' scenarios, dN/dS was increased to 0.5, 2.0, or 5.0 in site-class 3 ('slow' sites) for the designated 'target' branch(es). The resulting codon sequences were translated to amino acids, and S:F was calculated for each branch, with 'slow' and 'fast' defined such that sites ranking among the slowest 33% were considered 'slow' ($k = 0.33$).

In simulations mimicking species-specific positive selection, a single 'target' species was assigned an accelerated nonsynonymous substitution rate in the 'slow' 10% of codons (site-class 3), which remained slow-evolving in the other branches (dN/dS = 0.1). Even with a moderately elevated dN/dS in 'slow' sites (from dN/dS = 0.1 to 0.5), the median S:F ratio in the target branch increased significantly from the 'baseline' scenario in 100 replicate simulations (Kolmogorov-Smirnov test: $D = 0.22$; $p = 0.016$). We then modeled more dramatic branch-specific positive selection, keeping dN/dS within the range previously observed (Yang & Nielsen, 2002). Dramatic increases in dN/dS in the target branch's slow-sites led to a monotonic increase in S:F (Figure 3.2, light-grey bars). However, when many species (8/30 species, interspersed over the tree) were targeted by selection, S:F became less sensitive to detect selection (Figure 3.2, black bars). With so many species experiencing substitutions in 'slow' sites, they could no longer be classed as 'slow', and thus did not result in high S:F. This illustrates how S:F is sensitive to species-specific changes in selective pressure, yet relatively blind to positive selection on many/all branches. In an intermediate scenario (4 species under selection; Figure 3.2, dark-grey bars), S:F behaves similarly to the 1-target-species case for dN/dS ≤ 2, but plateaus around dN/dS = 5.

3.3 Delineating 'fast' and 'slow' sites

The S:F approach relies on empirical definitions of 'slow' and 'fast' sites, necessitating an optimal cutoff (k) between 'slow' and 'fast' sites. We evaluated both methods for choosing k based on mutual consistency, and consistency with dN/dS. Applied to codon data, dN/dS correlates best with S:F when $k = 0.75$ (Pearson's correlation = 0.92, $p < 2.2e\text{-}16$; Figure 3.S1A). While *minSD* allows each gene to have a different k, its average estimates of S:F also correlate best with $k = 0.75$. This agrees with *minFDR*, which

finds the minimum false-positive rate at $k = 0.75$.

Applied to protein sequence, *minFDR* converges on $k = 0.55$ (FDR = 0.18 for $p<0.05$ and FDR=0.055 for $p<0.005$; Figure 3.S1B). However, the *minSD* method chooses values of k that are on average lower (k = 0.30; Figure 3.S1B). Thus. *minSD* is 'stricter', allocating fewer sites into the 'slow' category. Nevertheless, the two methods agree fairly well with one another (Pearson's correlation = 0.47, $p <$ 2.2e-16).

Henceforth, we use *minSD* estimates of S:F as these generally provide a stricter definition of 'slow' sites, and the method makes fewer assumptions about the signature of selection in the data, but we report *minFDR* estimates for comparison.

3.4 Regimes of natural selection on different protein functions

We set out to quantify variation in selective pressures on 917 gene families in 30 species of γ-proteobacteria, gathered as described in Chapter 2 (Shapiro & Alm, 2008). For each branch of each gene tree, we computed S:F (using amino acid and codon sequences), and estimated dN/dS (using codon sequences). When applied to codon sequences with an appropriate cutoff ($k = 0.67$, approximating the expected proportion of nonsynonymous sites), S:F closely tracks dN/dS (Pearson's correlation = 0.91, $p <$ 2.2e-16; Supplementary Note 1; Figure 3.S2). Yet applied to amino acid sequences, S:F behaves differently than dN/dS (Table 3.S1; correlations in range 0.2-0.3, $p <$ 2.2e-16). The poor correlation between the amino acid-based S:F measure and dN/dS may be due to saturation of dS over the relatively long time scales investigated.

To test the hypothesis that different cellular functions are under different regimes of selection, we compared S:F ratios among proteins annotated with different biological functions (from the Clusters of Orthologous Groups (COG) database (Tatusov et al, 1997)). We picked proteins with values of S:F in the top 10% of their respective branch, and pooled together all branches into a 'high-S:F' subset (Figure 3.3A). We then used a Hypergeometric test to determine if any functional categories were over- or under-represented in the high-S:F subset, relative to the entire set of proteins. We used a percentile cutoff for S:F values within a genome to control for any genome-wide inflations or deflations of S:F in a particular lineage (*e.g.* inflation in *Buchnera* likely due to relaxed negative selection). The high-S:F protein set should therefore reflect protein-specific variation in S:F, rather than genome-wide variations in mutation rate, generation time, or effective population size. We also applied an additional p-value cutoff, reducing the size of the data set while preserving its main features (Supplementary Note 2 and Figure 3.S3).

Most noticeably, genes involved in motility and secretion (function N) are significantly over-represented in the high-S:F subset (Figure 3.3A). This is consistent with the notion that these genes, which often code for surface proteins targeted by immune systems, predators or phage, are frequent targets of positive selection, as has been documented previously in the γ-proteobacteria, most notably in plant and enteric pathogens (Shapiro & Alm, 2008, Guttman et al, 2006, Weber & Koebnik, 2006, Ma et al, 2006). Nonetheless, this result is not necessarily anticipated because S:F cannot detect genes that are under positive selection in *all* lineages (Figure 3.2). Thus, not only are motility/secretion genes subject to strong positive selection, but selection must frequently apply to different genes, or different sets of amino acids, in each lineage. Elevated S:F ratios in function N are observed using both amino acid (AA) and codon (DNA) sequences, both estimators of k, and dN/dS (Figure 3.3A), providing evidence for recurrent diversifying selection spanning ancient to more recent time scales. Motility/secretion genes are also significantly enriched among the set of genes with dN/dS > 1 (Hypergeometric test, $p = 0.005$), supporting the hypothesis of frequent positive selection on these genes, rather than relaxed negative selection.

Genes involved in cell envelope biosynthesis (M), ion transport and metabolism (P), and signal transduction (T) also tend to have high S:F, although with less statistical significance. Nevertheless, these functions may be common targets of lineage-specific positive or relaxed negative selection, constituting a more 'adaptable', less constrained, component of these genomes.

In contrast, positive and relaxed negative selection are much less frequent among genes involved in information-processing and central metabolism (functions C, E, F, G and J). These COG functions are all under-represented among the 'high-S:F' component of genomes (Figure 3.3A), and are likely under similar regimes of mostly purifying selection.

3.5 Species-specific, function-specific variation in selection

We next investigated to what extent function-specific selection may also be species-specific. In other words, does the set of cellular functions with unusually high S:F differ among branches of the γ-proteobacteria species tree? To address this, we again looked for enrichment/depletion of COG functions in the highest 10% of S:F values in each branch, this time on a branch-by-branch basis. By visual inspection, branches clearly differ in the set of COG functions with unusually high or low S:F ratios (Figure 3.3B). This difference is statistically significant: when choosing pairs of genes from the pooled high-S:F set, pairs from the same branch are more likely to have the same function than pairs from

different branches (Fisher test: Odds Ratio = 1.33, $p < 2.2e-16$). For example, the tendency toward high S:F in motility/secretion genes is attributable mostly to enterobacteria and members of the *Vibrio* clade (Figure 3.3B), perhaps due to unusually strong diversifying selection on cell-surface proteins in these species.

Certain lineages have globally skewed rates of evolution across all their genes, due to species-specific differences in effective population size, mutation rates, or generation times (Moran, 1996, Ochman et al, 1999). The *Buchnera* clade of aphid endosymbionts is a classic example: *Buchnera* experience population bottlenecks in each transmission cycle, reducing the efficacy of purifying selection, and allowing frequent fixation of deleterious mutations (Herbeck et al, 2003, Itoh et al, 2002, Fry & Wernegreen, 2005). This is recapitulated in the genome-wide S:F distributions for *Buchnera*, as well as the *Wigglesworthia* and *Candidatus Blochmannia* species of insect endosymbionts, which are all biased toward high S:F ratios (Figure 3.3B). The bias applies across all gene functions: *Buchnera* show little functional enrichment or depletion among their high-S:F genes, consistent with reduced efficacy of selection relative to genetic drift.

In addition to the insect endosymbionts, several other branches are shifted toward high values of S:F. For example, internal branches often have high S:F (and even higher values of dN/dS), perhaps due to ancestral sequence reconstruction errors (Table 3.S2). Moreover, short branches have slightly higher dN/dS than longer branches (Figure 3.S4), because purifying selection has had less time to purge deleterious mutations from the population (Rocha et al, 2006). The short-branch effect, like sequencing error, is only expected to influence dN/dS or S:F in leaf-branches, because the same deleterious mutation (or sequencing error) would have to occur twice independently in order to be incorporated into an internal branch. Because S:F is not inflated in short leaf-branches (Figure 3.S4), it appears that neither sequencing errors nor unpurged deleterious polymorphisms present a major source of bias in our results. Yet errors in ancestral reconstruction may significantly bias S:F estimates in internal branches, and although they may be less biased than dN/dS (Table 3.S2), S:F in internal branches should still be interpreted cautiously.

3.6 Selection example I: Redox metabolism in pseudomonads

Proteins involved in energy production (function C) tend to have low S:F in most species (Figure 3.3), consistent with uniform purifying selection. The only exception is the *Pseudomonas* clade, notably *P. fluorescens*, which has an excess of high-S:F energy production genes (Figure 3.3B). Many of these genes are co-expressed on the same operon (Figure 3.4) and tend to have elevated S:F in *Pseudomonas* but rarely in other clades (Figure 3.4). This pattern of selection is discernible at the protein level (both *minSD*

and *minFDR* methods), but is weaker at the codon level, partially due to saturation of dS (Figure 3.4, bottom panel). For example, the pyruvate dehydrogenase E1 component (AceE; COG 2609) has high S:F at the protein level in *P. putida*, but codon-level selection is not detectable with dN/dS. Consistent with species-specific, protein-level adaptation, significant structural differences are known to have occurred in AceE between *P. putida* and *E. coli* (Arjunan et al, 2002). Moreover, pseudomonads often inhabit oxygen-limited biofilms, where they produce alternative electron acceptors such as phenazines to maintain redox homeostasis (Price-Whelan et al, 2006, Price-Whelan et al, 2007). Phenazines may interact with AceE: inhibiting it by generating superoxides, or promoting its activity by re-oxidizing one of its products, NADH (Price-Whelan et al, 2007). These potentially *Pseudomonas*-specific biochemical interactions may impose lineage-specific selective pressures on AceE and other redox metabolism genes.

In another example, we found both transmembrane subunits of the succinate:ubiquinone dehydrogenase complex, SdhC and SdhD, among the high-S:F subset of *P. fluorescens* genes (Figure 3.4). This complex shuttles electrons from succinate to ubiquinone as part of the electron transport chain. SdhC has high S:F in two *Pseudomonas* species, but no other lineages (Figure 3.4), suggesting a lineage-specific evolutionary change. SdhC is in the 1% highest values of S:F in the *P. fluorescens* genome, due to 3 slow-site substitutions (S:F = 1.5, $p<0.05$, *minSD*; Table 3.1). We mapped these substitutions onto the *E. coli* Sdh protein structure (Yankovskaya et al, 2003) and discovered that one substitution, Phe→Tyr58, is in contact with a bound cardiolipin phospholipid (Yankovskaya et al, 2003), while the other two, Ala→Gly24 and Ile→Phe28, fall in the path of electron transport between the 3Fe-4S cluster and ubiquinone (Figure 3.5A). The latter site, Ile28, makes up part of the ubiquinone binding site, and is perfectly conserved across species in this study except *P. fluorescens*. Further confirming the species-specificity of this substitution, *P. fluorescens* Pf-5 (the strain used in this study) and *P. fluorescens* PfO-1 (the only other *P. fluorescens* genome in MicrobesOnline) are the only 2 strains harboring the Ile→Phe28 substitution, of 16 total *Pseudomonas* strains with SdhC orthologs in the database. This also serves as tentative confirmation that the substitution is fixed in *P. fluorescens*, and is not simply a slightly-deleterious polymorphism segregating in the population (Hughes et al, 2008). The substituted side-chain (Phe) is substantially larger than Ile, and would clash directly with ubiquinone unless there were some local modification of the protein structure (Figure 3.5A). Moreover, mutations at the equivalent site in human Sdh cause disease (Astuti et al, 2001), and result in oxidative stress in nematodes (Ishii et al, 1998) due to electron leakage (Yankovskaya et al, 2003). Often associated with superoxide-producing plants, *P. fluorescens* has a number of mechanisms for coping with oxidative stress (Paulsen et al, 2005). The Ile→Phe28 substitution might therefore be tolerated by *P. fluorescens* due to relaxed negative selection against free radical production. However, the occurrence of another nearby substitution in the path of

electron transport (Ala→Gly24) suggests an adaptive change. Given the diversity and ecological importance of secondary 'respiratory pigments' produced by pseudomonads (Price-Whelan et al, 2006, Mavrodi et al, 2006), it is not unreasonable to speculate that central metabolic respiratory pathways involving redox balance may be under positive selection to better interface with these secondary pathways.

3.7 Selection example II: Outer membrane in *V. cholerae*

Another potential ecological adaptation is presented by the outer membrane protein OmpW (COG 3047) of the human pathogen *Vibrio cholerae*. Low DNA-S:F and dN/dS show that OmpW is highly conserved in *V. cholerae*, with few amino acid-altering substitutions relative to silent substitutions (Table 3.2). Yet of these few amino acid changes, an unexpectedly high number occur in slow-evolving sites, suggesting lineage-specific positive or relaxed negative selection (S:F = 1.38, $p<0.05$, *minSD*; Table 3.2). We focus on this protein because it is present in all known *V. cholerae* strains, is highly immunogenic, suggesting it may be subject to immune selection (Das et al, 1998), and is up-regulated in related vibrios under high-NaCl stress (Xu et al, 2005), suggesting a role in osmoregulation Of the 12 substitutions inferred in *V. cholerae* using the *minFDR* method, the 6 in slow-sites cluster slightly more closely with one another in the 3D structure (Hong et al, 2006) than do the 6 in fast-sites (Mean pairwise Euclidean distance between C_α atoms = 21.5 Å for slow-sites; 27.2 Å for fast-sites; Two-sample one-sided Wilcoxon test: W=75, p=0.06), suggesting that the slow-site substitutions may represent structurally-coordinated adaptive changes. Indeed, the substitutions in the two most highly conserved sites, Leu→Val55 and Leu→Phe83, are adjacent in the protein structure, despite being distant in the linear protein sequence (Figure 3.5B). They are localized just below the putative exit channel, where a small molecule may exit the hydrophobic barrel and enter the outer membrane (Hong et al, 2006). The substitutions might thus alter substrate specificity or transport kinetics of the channel. All six slow-site substitutions are present in the additional 6 *V. cholerae* strains (V51, V52, RC385, O395, MO10 and 2740-80) with sequences in MicrobesOnline, consistent with functional significance of these substitutions, and confirming that they are not slightly-deleterious polymorphisms or sequencing errors. However, these substitutions are not all unique to *V. cholerae*: Leu→Val55 and Leu→Phe83 both occur in *V. splendidus 12B01* and *Photobacterium profundum* 3TCK (not in *P. profundum* SS9, which appears to have lost COG3047). Horizontal gene transfer could be responsible for this phylogenetically incongruence, but would require two separate transfer events because *V. cholerae* contains an insertion of the sequence SGGELG between residues 67 and 68, which is not present in either potential donor, *P. profundum* or *V. splendidus* (Figure 3.5B). Therefore, convergent evolution is the more parsimonious explanation for this covarying pair of

substitutions, and this likely implies positive selection (Holt et al, 2008, Sokurenko et al, 2004, Falush & Bowden, 2006).

Discussion

3.8 S:F as a method to detect changes in the regime of selection

We have described a method for detecting selection at the protein or DNA level that is conceptually similar to dN/dS, but is more general, relying on empirical definitions of 'slow' and 'fast' sites rather than pre-defined non-synonymous/synonymous sites. In general, S:F identifies deviations from a sequence's 'usual' regime of selection, whether that regime is neutral, involves strong purifying or diversifying selection, or some complex combination of these regimes. An advantage of S:F over dN/dS is its suitability to anciently-diverged clades, such as the γ-proteobacteria, in which synonymous sites are often saturated with multiple substitutions. Applied to more closely-related strains, it may lack power due to paucity of substitutions, but should still be more conservative (e.g. fewer false-positives) than branch- and site-based models of dN/dS (Nozawa et al, 2009).

As an empirical method, S:F exploits the availability of species-rich protein families, made possible by whole-genome sequencing of related species. Depending on the diversity and breadth of species included, S:F will identify different sets of slow- and fast-evolving sites. The method is therefore flexible, and potentially sensitive to selection at different time scales. In this study, we investigated the relatively broad hypothesis that patterns of function-specific natural selection vary among ecologically distinct species. The method also lends itself well to more specific hypotheses, aimed at particular groups of interest.

3.9 Distinguishing adaptive evolution

Elevated S:F may be attributed to either positive selection, or species-specific relaxation of negative selection. Both scenarios have the potential to be biologically informative, and may suggest ecological adaptation. For example, the Ile→Phe28 substitution in *P. fluorescens* SdhC may have been 'passively tolerated' by relaxed selective constraint on this residue, or 'actively' pushed to fixation by positive selection for a novel or improved function. Without within-population sampling (*e.g.* McDonald-Kreitman tests; (Shapiro et al, 2009, Li et al, 2008)), it is difficult to distinguish between these scenarios Yet the substitution is lineage-specific (Figures 3.4 and 3.5A), strongly suggesting some sort of functional re-wiring of redox metabolism and electron transport in *P. fluorescens*. The substitution is also gene-specific: SdhC has an S:F ratio in the top 1% of the *P. fluorescens* genome (Table 3.1) and therefore cannot be attributed it to a genome-wide shift in substitution rates, or possible biases in S:F due to branch length. By further accounting for *P. fluorescens*' ecology – a phenazine-producing, plant-associated

organism with a high metabolic capacity – we gain confidence in the adaptive value of substitutions in an electron-transport protein. Similar lines of evidence lend support to the hypothesis that OmpW has acquired ecologically adaptive substitutions in *V. cholerae*. In both examples, further experimental work is needed to fully understand and validate the predictions of our 'reverse ecology' approach.

3.10 Conclusions

In our analysis of adaptive protein evolution across 30 γ-proteobacteria, we were able to glean several insights, both global and specific. Globally, we found that proteins localized to the cell surface (functioning in motility/secretion or cell envelope biosynthesis) are frequent targets of positive or relaxed negative selection, showing elevated S:F ratios across many species, especially those involved in host-pathogen or host-symbiont interactions. Meanwhile, proteins involved in 'housekeeping' roles tend to be under purifying selection, which we observe as low S:F ratios. Yet there are exceptions to this rule: we observe instances of species- or clade-specific reversals of purifying selection, for example the unusually high S:F ratios observed in a suite of energy metabolism proteins in pseudomonads.

The method we describe is a flexible, empirical approach for detecting varying regimes of natural selection. It can be applied to study selection on protein-coding sequences, or non-coding genomic sequences, such as promoters and non-coding RNAs. In this work, we showed how S:F can be applied over evolutionary time scales beyond the reach of dN/dS. Discriminating between positive and relaxed negative selection remains a challenge, but we reason that both scenarios are ecologically informative. As we accumulate whole-genome sequences for more and more ecologically diverse species, the S:F method will be useful in detecting the protein-level adaptations that functionally distinguish between them.

Methods

3.11 Data set

A set of 917 gene families (members of the same Cluster of Orthologous Groups (Tatusov et al, 1997)), each represented by a single copy in at least 16 of the 30 genomes in this study, was retrieved from the MicrobesOnline database as previously described (Alm et al, 2005). Maximum-likelihood (ML) gene trees, and a consensus species tree topology were constructed using PhyML ((Guindon & Gascuel, 2003); Supplementary Methods).

3.12 Calculation of Slow:Fast substitution ratio (S:F)

We performed joint reconstruction of ancestral sequences (Pupko et al, 2000), implemented in PAML: Phylogenetic Analysis by Maximum Likelihood 4.0 (Yang, 2000) using the ML gene-tree topologies. Sites in the protein or DNA sequence were rank-ordered (between 0 and 1, with 0 being the slowest- and 1 the fastest-evolving) by the number of substitutions inferred to have occurred in the site in all branches of the phylogeny (Figure 3.1A). A substitution-rate cutoff (*k*, also between 0 and 1) was then chosen to delineate slow (few substitutions in the phylogeny) and fast (many substitutions) sites. Invariant amino acid sites (with no observed substitutions) were excluded, but invariant nucleotide sites were retained in the DNA analyses for consistency in comparison with dN/dS. S:F was computed as follows, after excluding branches with F=0:

Equation 1.

S:F ratio = number of substitutions per slow-evolving site / number of substitutions per fast-evolving site

The number of substitutions per site was corrected for multiple hits using a Poisson correction for protein sequence (Equation 2) or a Jukes-Cantor correction for DNA (Equation 3).

Equation 2.

$$d = a\left[\left(1 - \frac{c}{s}\right)^{-1/a} - 1\right]$$

Equation 3.

$$d = -\frac{3}{4}\ln\left(1 - \left(\frac{4}{3}\cdot\frac{c}{s}\right)\right)$$

where d is the corrected number of substitutions per site, c is the number of observed substitutions, s is the number of sites (fast or slow), and the parameter a is set to 2.4, as suggested for the JTT substitution model (Nei & Kumar, 2000).

The deviation of each branch from the expected S:F ratio was evaluated using a Fisher Exact test. For each branch i in a gene tree of N branches, we define S_i as the number of slow-site substitutions, and F_i as the number of fast-site substitutions in branch i. We define the total numbers of slow- and fast-site substitutions in all other branches ($x \neq i$) of the gene tree, $S_{tot} = \sum_{x \neq i}^{N} S_x$ and $F_{tot} = \sum_{x \neq i}^{N} F_x$, respectively. We compute Fisher's Odds Ratio statistic, O.R. = $(S_i/F_i) / (S_{tot}/F_{tot})$, and associated p-value to assess confidence in a branch i having S:F significantly greater (O.R. > 1) or less than (O.R. < 1) than the rest of the gene tree.

3.13 Setting the cutoff (*k*) between Slow and Fast-sites

We describe two methods to choose *k*. The *minFDR* method aims to maximize the sensitivity while controlling for the false-discovery rate (FDR) in the dataset of ~1000 genes. The *minSD* method aims to allocate sites into distinct 'slow' and 'fast' distributions, making the distributions as non-overlapping and as 'tight' as possible about their respective means.

In the *minFDR* method, we choose *k* to minimize the false discovery rate (FDR), a measure of the ratio of signal to noise in the data (Storey & Tibshirani, 2003). To do so, we range *k* from 0.05 to 0.95 in increments of 0.05 and re-compute S:F ratios and associated *p*-values for all branches and gene families for each value of *k*. The result is a distribution of *p*-values (over all genes and all branches) associated with each value of *k*. We choose the *k* yielding the distribution with the highest ratio of true:false positives, or the minimum FDR (Figure 3.1C; Supplementary Methods). We note that *minFDR* is only valid if the data actually contain a detectable signal of selection at the protein level, and the estimated *p*-values are unbiased.

In contrast, *minSD* chooses *k* to minimize instances of slow-sites being miscategorized as fast (or *vice versa*) within a single gene (thus, a separate *k* can be chosen for each gene). Briefly, for each choice of *k*, slow- and fast-sites of the alignment are considered as two separate distributions, each used to infer the likeliest gene-tree and branch lengths. If an excessively high *k* is chosen, some 'true' fast-sites will be miscategorized as slow (and 'true' slow-sites miscategorized as fast), thereby introducing greater variance into both distributions and both branch-length estimates. Our approach is thus to estimate standard deviations for both slow- and fast-site distributions, and choose the *k* that minimizes the standard deviation of both distributions (Figure 3.1B; Supplementary Methods).

Supplementary methods.

3.14. Construction of gene- and species-trees

Protein sequences were aligned using MUSCLE (Edgar, 2004), and all gaps were removed, including one flanking residue on either side. Nucleotide alignments were also performed using MUSCLE, ensuring that codons were consistent with the protein alignment. Gene tree topologies and branch lengths were estimated using PhyML 2.4.5 (Guindon & Gascuel, 2003) with a JTT substitution model and 4 γ-distributed rate categories. As we previously noted, use of the WAG substitution model instead of JTT did not significantly influence the branch lengths or topologies of these proteins (Shapiro & Alm, 2008). The consensus species tree topology was estimated as previously described, using a concatenation of 80

COGs present in single copy in each of the 30 genomes (Shapiro & Alm, 2008). Divergence times in the species tree were estimated using a relaxed molecular clock model implemented in multidivtime (Thorne et al, 1998) with parameters rttm = 1.75 By, rttmsd = 1, rtrate = 3.9, rtratesd = 3.9, brownmean = 0.6, brownsd = 0.6, bigtime = 5.0, and the constraint that *E. coli* and *Salmonella* diverged between 0.057 and 0.176 billion years ago (Battistuzzi et al, 2004, Ochman & Wilson, 1987).

3.15. Detailed procedures for choosing *k*.

For DNA data, *k* is the (approximate) proportion of sites in the multiple sequence alignment classified as 'slow'. *k* is 'approximate' because multiple sites might be tied at the same rate (substitutions/site), so that the cutoff is drawn to minimize the deviation from the desired *k* (*i.e.* if the desired *k* were 0.67, falling between sites with rank 0.60 and 0.70, the cutoff would be chosen at 0.70). For AA data, *k* is the rank-order of the site where the cutoff is drawn. For example, in an alignment of 10 sites with 1, 2, 3, and 4 substitutions in 4, 3, 1, and 2 sites, respectively, *k* would be 0.25 (1/4) if the cutoff were drawn between 1 and 2 substitutions/site, even though this would result in a proportion of 0.40 (4/10) of all sites being classed as 'slow'.

i. *minFDR* method

In this method, we compute S:F ratios for all genes in our dataset to yield a distrubutiuon of *p*-values for each possible choice of *k*. We then want to choose the value of *k* producing the distribution with the lowest FDR. This is done by performing a 1-sided Kolmogorov-Smirnov (KS) test of the observed *p*-value distribution against a uniform distribution of *p*-values. The higher the value of the KS test *D* statistic, the more favorable the ratio of true:false positives. As verification that the *D* statistic is indeed a good indicator of FDR, we plotted the observed *p*-value distribution (ranging from 0 to 1, binned in increments of 0.05) for each value of *k*, and calculated FDR by dividing the average number of counts in bins with $p>0.5$ by the number of counts in the $p<0.05$ bin. We then chose the value of *k* that minimized FDR. These two methods always yielded the same choice of *k*.

ii. *minSD* method

As in the *minFDR* method, we vary *k* from 0.05 to 0.95 in increments of 0.05. For each *k*, the sequence is split into slow and fast-sites, according to their rank-order. Branch lengths are then estimated twice independently with PhyML, once using only the slow-site distribution and once using only the fast-site distribution. The result is two gene trees: one estimated using slow-sites, the other using fast-sites. If, for example, some fast-sites are miscategorized as slow, most of the slow-sites will support short branch lengths, but a few (miscateogrized) sites will support long branch lengths. So, a branch in the gene-tree

could be 'stretched' (inferred to have undergone more substitutions per site) without significantly reducing the likelihood. Such 'stretching' of each branch, can be used to estimate the standard deviation of each branch in the gene tree, which is our goal. Starting from the maximum-likelihood branch length, the branch length is stretched to 300% of its ML length, in increments of 5%, and then compressed to zero, also in increments of 5%. The likelihood of the gene tree is re-estimated with PhyML for each incremental stretch of the branch to produce a distribution of N branch lengths and their associated likelihoods, for each branch. For branches with non-zero ML lengths, the likelihood of a particular branch length could be well-approximated by a lognormal distribution, transformed for simplicity such that the mean (ML branch length) is zero (Equation 4). The log-likelihood distribution can therefore be approximated by a parabola with *L*-intercept equal to ln $(\sigma(2\pi)^{-1/2})$ (Equation 5).

Equation 4.

$$L = p_x(x) \sim N(0,\sigma^2)$$

Equation 5.

$$\ln L = \ln \frac{1}{\sigma\sqrt{2\pi}} - \frac{x^2}{2\sigma^2}$$

where the likelihood (L) is a function of log branch length (x), given by a lognormal probability density function with mean set to zero (by subtracting the PhyML-estimated ML branch length) and variance σ^2. Letting $y = ln\ (\sigma(2\pi)^{-1/2}) - ln\ L$ and $m = 1/2\sigma^2$, and substituting the estimated maximum likelihood for the term $1/\sigma(2\pi)^{1/2}$, yields the parabola $y = mx^2$. We then estimated the variance and standard deviation by linearizing the parabola and solving for the best-fit slope (m). For each branch, we combined slow and fast-sites to calculate a pooled standard deviation. We repeated this procedure for a range of values of k, and chose the k that minimized the mean pooled standard deviation over all branches in the gene tree. Note that with this method, each gene family can take a different value of k.

3.16. Simulated models of evolution

We used the *evolver* program from the PAML package (Yang, 2000) to generate sequences of 300 codons for 30 species, using the γ-proteobacteria species tree. All simulations had 3 site-classes, as described in the main text. S:F was computed for each branch as described for the real data. The slow:fast cutoff of k = 0.33 was chosen using the *minSD* method (Supplementary Methods).

3.17 Estimation of synonymous and nonsynonymous substitution rates (dS and dN)

We used two methods to estimate dN and dS for each branch in each gene tree: (i) ML estimation, using the codeml program from the PAML 4.0 package (Yang, 2000), allowing dN/dS to vary freely among branches ('free-ratio' model), and (ii) using the NG86 'counting' method of Nei and Gojobori (Nei & Gojobori, 1986) to estimate dN and dS between each branch and its ancestor sequence (from the joint reconstruction). Because ancestral reconstructions were performed using PAML's likelihood model, transition/transversion biases, nonuniform codon usage, and variable rate categories are accounted for in the NG86 estimates of dN and dS.

Supplementary Note 1. Validation of S:F by comparison to dN/dS.

We applied the S:F metric to codon sequences of 917 gene families present in at most one copy per genome in the majority of 30 species of γ-proteobacteria (Methods). With the cutoff between slow and fast-sites set such that 2/3 of nucleotide sites are classed as 'slow' ($k = 0.67$, based on the rough approximation that ~2/3 of sites within a codon are nonsynonymous and thus slow-evolving), the S:F ratio is expected to converge on the dN/dS ratio. Indeed, there is very good agreement between S:F and dN/dS (Figure 3.S2; Pearson's correlation = 0.91, $p < 2.2e\text{-}16$). This suggests that, despite being naive as to whether a site is synonymous or nonsynonymous, the S:F method still tends to categorize sites correctly into these categories by virtue of synonymous sites generally evolving faster than nonsynonymous sites. S:F correlates best with the NG86 method for estimating dN/dS (Nei & Gojobori, 1986), probably due to the similar substitutions models used (Table 3.S1; Methods). Estimates of dN/dS from PAML are made using a more complex substitution model permitting much higher values of dS (Yang, 2000). This results in a poorer correlation with S:F, which is improved if branches saturated with substitutions at synonymous sites (dS > 2) are removed (Table 3.S1).

The agreement between S:F and dN/dS is not exact, probably due to imperfect mapping from 'slow' to nonsynonymous sites, and from 'fast' to synonymous sites. One reason for this is that 'slow' and 'fast' labels apply to an entire column (site) of a multiple sequence alignment, whereas a column may contain some sites that are nonsynonymous and some that are synonymous, depending on the pair of codons being compared. Therefore, a column may be classified as 'slow' and still contain a mixture of synonymous and nonsynonymous sites. It is likely that many of the synonymous sites in this 'slow' column will undergo more substitutions than nonsynonymous sites in the column, resulting in a systematic inflation of S:F relative to dN/dS. Consistent with this explanation, we observe a slope greater than 1 in the log plot of S:F versus dN/dS (Figure 3.S2). Nonetheless, this effect is minor, and the strong correlation leads us to conclude that S:F behaves similarly to dN/dS when applied to codon sequences. This comparison to dN/dS is intended only to illustrate proof of concept, and we stress that dN/dS is to be preferred over S:F when (unsaturated) codon sequences are available.

Supplementary Note 2. Additional Fisher *p*-value filter on Enrichment/Depletion of Cellular Functions among high-S:F subset of genes.

When we further restrict the 'high-S:F' subset to branches with Odds Ratio > 1 ($p < 0.05$), the same major results hold, but fewer categories show significant enrichment or depletion (Figure 3.S3). Because the Odds Ratio cutoff is fairly stringent, the lack of significance may simply be due to an insufficient amount of data from which to draw conclusions. Nevertheless, even with the additional *p*-value cutoff, the main

findings hold: categories N (motility/secretion) and M (cell envelope) are over-represented in the high-S:F set of genes, while category J (ribosome & translation) is under-represented. The specific differences between the *p*-value filtered vs. unfiltered datasets should be interpreted with caution, since lack of significance in the filtered dataset may often be due to low counts. For example, motility/secretion enrichment is more significant in the amino acid-based methods, while cell envelope enrichment is more significant in the DNA-based methods. Yet is this because species-specific positive selection on the cell envelope is more readily detectable at the DNA level than the protein-level (*e.g.* selection tends to have been more recent)? Or is this simply an artifact of low counts in the hypergeometric test? We leave resolutions to these questions for more detailed future studies.

Supplementary Note 3. Relationship of S:F to dN/dS and other methods.

The theoretical underpinnings of S:F are analogous to dN/dS: namely, that deviations from a nearly neutral model of evolution can be detected when an excess of substitutions are observed in the most constrained set of protein or nucleotide sites except that fast-sites may contain a mixture of nearly neutral and positively selected sites.

The expectation is that substitutions in slow-sites will usually occur as the result of a shift in the strength or direction of natural selection. By categorizing sites into relatively slow- and fast-evolving classes, S:F effectively normalizes-out the baseline level of purifying (or diversifying) selection acting on a protein family in order to identify branches that deviate from the baseline – for example, high S:F values would result from either excess positive or relaxed purifying selection on slow-sites. Meanwhile, low values of S:F are expected when purifying selection is consistently strong on slow-sites, as observed in ribosomal proteins (main text).

The S:F substitution ratio is more general than dN/dS because it is applicable not only to codons, but to amino acids or noncoding DNA sequences (although it is likely less powerful than codon-based methods for protein coding sequences with unsaturated dS). In this work, we focus on amino acid sequence divergence over time scales where dS is often saturated with multiple substitutions, reducing the power of dN/dS to detect departures from neutral evolution (Gojobori, 1983, Smith & Smith, 1996). In effect, by applying S:F to protein sequence, we separate dN into its slow- and fast-evolving components. Other methods have been developed to separate dN into 'conservative' and 'radical' amino acid substitutions (Hanada et al, 2007), but these require learning genome-wide substitution matrices that apply to all proteins. We do not assume that all protein families follow the same substitution rules, and instead define 'slow' and 'fast' sites empirically for each protein family separately. Most protein families contain some

amino acid sites that are unconstrained, evolving freely over time, and other sites that are heavily constrained and conserved (Felsenstein & Churchill, 1996). By identifying these sites, we can use the S:F ratio to detect positive or relaxed negative selection on slow-sites in a protein. Conceptually, this type of empirical, amino acid-based (as opposed to codon-based) approach to detect selection resembles Coin and Durbin's technique to distinguish pseudogenes from functional genes - also based entirely on patterns of amino acid substitutions (Coin & Durbin, 2004). In this earlier work, the authors found pseudogenes to have distinct substitution patterns from functional proteins, and recognized that a similar distinction might exist between proteins under purifying selection.

We note that the S:F ratio may be affected if the identities of 'slow' and 'fast' sites differ between subclades of the tree, as predicted under the 'covarion' model of evolution (Miyamoto & Fitch, 1995). Provided that subclades contain roughly equal numbers of 'slow' and 'fast' sites, no systematic bias in S:F ratios is expected, although their variance could be greatly increased and *p*-values affected. Covarion effects are thus worth considering when interpreting S:F ratios, and future work could measure both simultaneously.

Figure legends

Figure 3.1. Overview of S:F methodology.

(A) Hypothetical 5-species phylogeny and multiple sequence alignment for a protein of 6 amino acids. Substituted amino acids are highlighted in black, and substitutions are ranked by the number of substitutions per site. Excluding the invariant site, there are 2 sites in the slowest category (1 sub/site), 2 sites in the fastest category (3 subs/site), and 1 site in an intermediate category (2 subs/site). If the cutoff (k) were drawn such that the intermediate category is classed as 'fast', species 1 would have 1 substitution in 2 slow-sites and 3 substitutions in 3 fast-sites, yielding S:F = (1/2) / (3/3) = 0.5.

(B) The *minSD* method to choose *k*. The sites in a protein are binned in a histogram (top panel) according to their evolutionary rate (relative number of subs/site, normalized to range from 0 to 1). Three possible choices of *k* are considered. For each *k*, slow- and fast-sites are considered separately to estimate branch-lengths and likelihoods for the phylogeny. Branch-length distributions are shown for a representative branch (bottom panel). In practice, variances of all branch-length distributions in the phylogeny are computed and pooled. In this example, the intermediate choice (k_2) yields the lowest pooled variance and is thus the best choice.

(C) The *minFDR* method to choose *k*. For each choice of *k*, S:F ratios and *p*-values are computed for all branches of all gene trees to produce a distribution of thousands of p-values, which are plotted in a histogram. The false-discovery rate (FDR) for the $p<0.05$ bin is estimated as the average number of branches in bins with $p>0.5$ (dashed lines) divided by the number of branches in the $p<0.05$ bin. The value of *k* producing the lowest FDR (in this case, FDR=0.05, meaning that of the 100 branches with significantly unusual S:F at the $p<0.05$ confidence level, 5 are expected to be false positives) is chosen as the optimal cutoff.

Figure 3.2. Response of S:F to different selection scenarios

S:F ratios for a single branch (*V. cholerae*) under selection at slow-sites (light-gray bars). 8 branches (*V. cholerae*, *V. vulnificus*, *X. oryzae*, *P. syringae*, *S. oneidensis*, *P. multocida*, *E. coli*, *B. aphidicola APS*) under selection (black bars). Clade of 4 species (*P. syringae*, *P. aeruginosa*, *P. fluorescens*, *P. putida*) under selection (dark-grey bars). The x-axis shows dN/dS in 'slow-evolving' sites (class 3), comprising 10% of each sequence, and set to 0.1 in all non-target branches in the tree. Each bar shows the mean S:F in the target branch(es) for 100 replicate simulations, with error bars showing +/- the standard error of the mean. When multiple branches are targeted, a single representative branch is displayed, chosen at random.

Figure 3.3. Enrichment/depletion of cellular functions in the high-S:F subset of genes

(A) Schematic of Hypergeometric test results for enrichment or depletion of COG functional categories of genes in the top 10% highest values of S:F within each genome, pooled over all 57 branches in the species tree. Functional categories over-represented in the high-S:F set of genes are colored in maroon, and those under-represented in blue, with color saturation proportional to the significance of the Hypergeometric test for enrichment/depletion. The results are repeated using 5 different metrics: (1) S:F applied to amino acid sequences (AA), estimating *k* with *minFDR* (k=0.55), or (2) estimating *k* with *minSD*, (3) S:F applied to nucleotide (DNA) sequences of the same set of genes, estimating *k* with *minFDR*, or (4) setting k=0.67, such that 2/3 of sites are considered slow and 1/3 fast, and (5) dN/dS estimated with the NG86 method.

(B) Functional enrichment/depletion is mapped onto each branch of the γ-proteobacterial species tree, with branch lengths (time±standard deviation) estimated using a relaxed molecular clock model (Supp. Methods). The number of genes in each branch for which S:F was calculated (N) is shown below each branch. Genes were only included in an internal branch if that internal branch was present in the gene tree, otherwise it was excluded. Enrichment/depletion of each functional category among the high-S:F gene set (top 10% S:F values in the branch; *minSD* method) is shown in maroon/blue colored boxes to the left of each species (terminal branch), or above each internal branch. Branches with positively shifted S:F distributions are highlighted in maroon (genome-wide distribution of S:F is shifted to higher values than all-genome pooled distribution; assessed by K-S test D statistic, $p < 0.05$ after Bonferroni correction for 57 branches).

Figure 3.4. Genes involved in energy production have elevated S:F in pseudomonads

Gene-by-branch heatmap for genes in category C (energy production) in top 10% of S:F in one or more branches of the *Pseudomonas* clade (blue box). Columns represent either terminal branches, or internal nodes (highlighted in grey on the tree). Data is presented for three different methods: S:F applied to amino acid data using *minSD* to choose *k* (top), using *minFDR* to choose k (middle), or dN/dS (NG86 method) applied to codons (bottom). Red: Gene in top 10% of S:F (or dN/dS) values in the branch, with saturation proportional to the magnitude of the S:F (or dN/dS) ratio. Black: gene is present in the branch but not among top 10%. Grey: gene is not present in the branch, or the branch in the gene tree does not correspond to a monophyetic clade in the species tree. Blue: Ratio not estimated because the denominator (F or dS) is saturated with substitutions. Genes are listed by the short name of their *E. coli* ortholog, with COG number in parentheses. Branch lengths in the species tree are not to scale. Genes on the same operon in *E. coli* (Price et al, 2005) are grouped together with curly brackets.

Figure 3.5. Alignment and structure of proteins with high S:F.

(A) Multiple sequence alignment (MSA) of SdhC transmembrane helix I (left), with positions numbered according to the *E. coli* structure (Yankovskaya et al, 2003) (right). Columns of the MSA are colored by conservation, with site categories (slow or fast), as determined by both *minSD* and *minFDR* methods, as well as the majority-rule consensus, shown below each column. Perfectly conserved columns were not assigned a slow/fast category. Slow-site substitutions in *P. fluorescens*, assigned by the *minSD* method, are boxed in red, *P. syringae* in yellow. The structure shows subunits SdhC (green), SdhD (blue), and part of SdhB (grey). Slow-sites with substitutions in *P. fluorescens* are colored in red, *P. syringae* in yellow, and *P. aeruginosa* in orange. The *E. coli* side chains, not the substituted *Pseudomonas* residues, are depicted. Other molecules in the structure are: ubiquinone (purple), heme b (cyan), cardiolipin (magenta), and the 3Fe-4S iron-sulfur cluster (yellow/orange spheres). Branch lengths in the species tree are not to scale. Structure image generated using PyMOL (DeLano, 2002); MSA image using Jalview (Clamp et al, 2004).
(B) MSA of OmpW (left), with positions numbered according to the *E. coli* structure (Hong et al, 2006) (right). *V. cholerae* substitutions in slow-sites are boxed in orange if identified by the *minFDR* method, or in red if by the *minSD* method. The same color scheme is used on the structure, with *E. coli* side chains depicted. Species added to the MSA but not among the 30 species used in other analyses are shown in grey text.

Table legends

Table 3.1. Substitutions in slow and fast-sites in *P. fluorescens* SdhC (COG 2009)
Data are shown for S:F applied to amino acid data (AA) and nucleotide data (DNA), and for dN/dS, each estimated by two different methods. For each method, sites were percentile-ranked based on the number of substitutions/site, and divided into slow and fast at the rank cutoff (*k*) indicated (Methods). The 'Rank" column indicates the percent-rank of the ratio (S:F or dN/dS) among genes in the *P. fluorescens* genome, with 1% indicating very high S:F. In the Fisher Test column, O.R. >1 indicate that the S:F ratio is greater in *P. fluorescens* than other branches of the gene tree.

Table 3.2. Substitutions in slow and fast-sites in *V. cholerae* OmpW (COG 3047)
Data as described in Table 3.1 legend. The 'Rank" column indicates the percent-rank of the ratio (S:F or dN/dS) among genes in the *V. cholerae* genome, with 1% indicating very high S:F.

Supplementary figure legends

Figure 3.S1. Finding an optimal cutoff (*k*) between fast and slow-sites.
Choice of slow/fast percentile cutoff (*k*) vs: (i) correlation between a given fixed *k* and *minSD* methods for calculating S:F ratios over all branches and gene families (dashed red line); (ii) true-discovery rate (1 - FDR), as estimated by the Kolgomorov-Smirnov *D* statistic comparing the observed distribution of Fisher *p*-values for a given *k* versus a uniform *p*-value distribution (black line, corresponding to left y-axis scale); and (iii) correlation of S:F at the given *k* with dN/dS (dotted green line; DNA data only). Pearson's correlations (right y-axis) and KS test *D* statistics (left y-axis) are normalized so that the maximum value ($D = 0.18$) is set to 1. (A) For consistency in comparing to dN/dS, invariant sites (those with no substitutions) are included. (B) Protein data exclude invariant sites. In both graphs, *k* is the proportion of sites classified as 'slow', treating ties as described in Methods.

Figure 3.S2. Agreement between dN/dS and S:F applied to codons
Values of S:F for all branches in 917 γ-proteobacterial gene trees are plotted on a $\log_{10}$ scale against dN/dS (calculated using the NG86 method) for all branches in all gene families for which both ratios are defined (N = 17,142). Pearson's correlation = 0.91, $p < 2.2e\text{-}16$. The gray line represents y = x.

Figure 3.S3. Additional Fisher *p*-value filter on Enrichment/Depletion of Cellular Functions among high-S:F subset of genes.
Results of Hypergeometric tests for enrichment or depletion of COG functional categories of genes in the top 10% highest values of S:F within each genome, pooled over all genomes. Genes in the top 10% high-S:F set are only counted provided their Odds Ratio statistic is significantly greater than 1 (Fisher exact test, $p<0.05$). Because Odds Ratios were not computed for dN/dS, the dN/dS values were instead filtered in the following way: genes in the top 10% of dN/dS were only counted provided that their value of dN/dS was above the mean dN/dS for all branches in the gene tree. Functional categories over-represented in the high-S:F set of genes are colored in red, and those under-represented in blue, with color saturation proportional to the significance of the Hypergeometric test. Other aspects of the figure are as described for Figure 3.3A in the main text.

Figure 3.S4. Relationships between internal and terminal branch lengths, S:F and dN/dS
Branch lengths are plotted against dN/dS (A), S:F applied to amino acid sequences using the minFDR method (B), and the minSD method (C). Branch length is defined as substitutions per synonymous site (A) or per fast-site (B,C). Points on the left represent mean values (average over all genes) for each of the 57 branches (30 terminal, 27 internal). The plot on the left shows each individual branch from each gene tree as a separate point, on a $\log_{10}$ scale. Best-fit regression lines and Pearson's correlations (R) are only shown when a significant (p<0.1) correlation is observed. Error bars represent the standard error of the mean within each branch.

Supplementary tables legends

Table 3.S1. Pearson's correlations between dN/dS and S:F
Pearson's correlations are reported between 3 S:F methods (applied to DNA with k=0.67, or applied to amino acids (AA) with k set using *minSD* or using *minFDR*, k=0.55) and methods for computing dN/dS (full maximum likelihood using PAML, or maximum likelihood ancestral reconstruction using PAML, followed by counting substitutions using the NG86 method). Data are only included if both ratios being compared are defined (i.e. denominators are not saturated or equal to zero), resulting in some variation in the sample sizes (N). p < 2.2e-16 for all correlations (r).

Table 3.S2. Comparison of Internal and Terminal branch lengths, S:F, and dN/dS ratios
Values of the ratio (either dN/dS or S:F, as appropriate) or branch length (in units of substitutions per fast-site (F) for S:F methods, dS for dN/dS) are compared between internal and terminal branches. Values represent the mean over all branches and all genes. Shown in parentheses are the branch-mean values, first averaged within each branch, then averaged over the total number of branches (30 and 27, for terminal and internal branches, respectively). In all cases, internal and terminal branches are significantly different by both a two-sample t-test and a Wilcoxon rank sum test (p < 2.2e-16 for pooled data; p < 0.01 for branch-means). Branches with saturated values of dS or F were not included in the analysis.

Figure 3.1

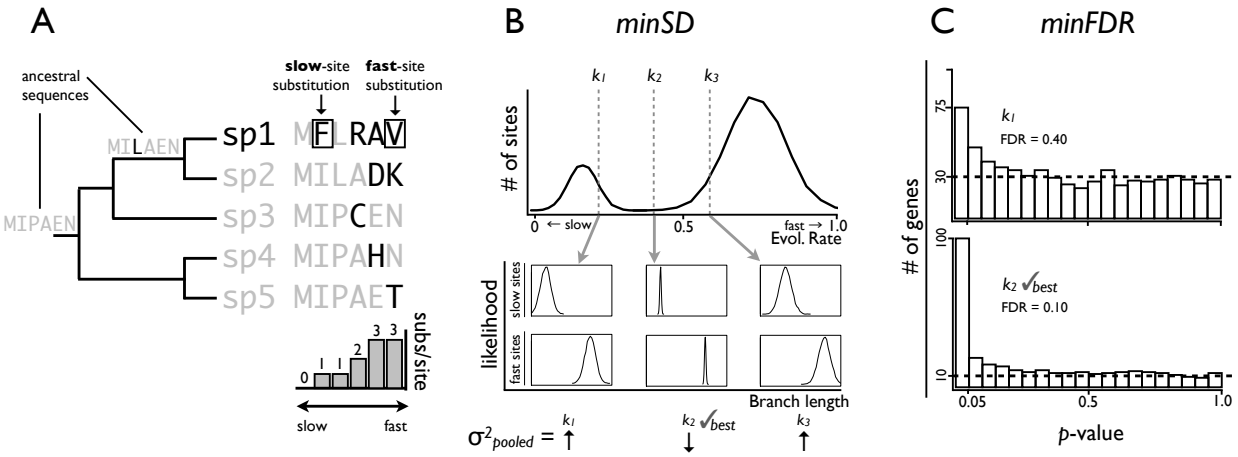


Figure 3.2

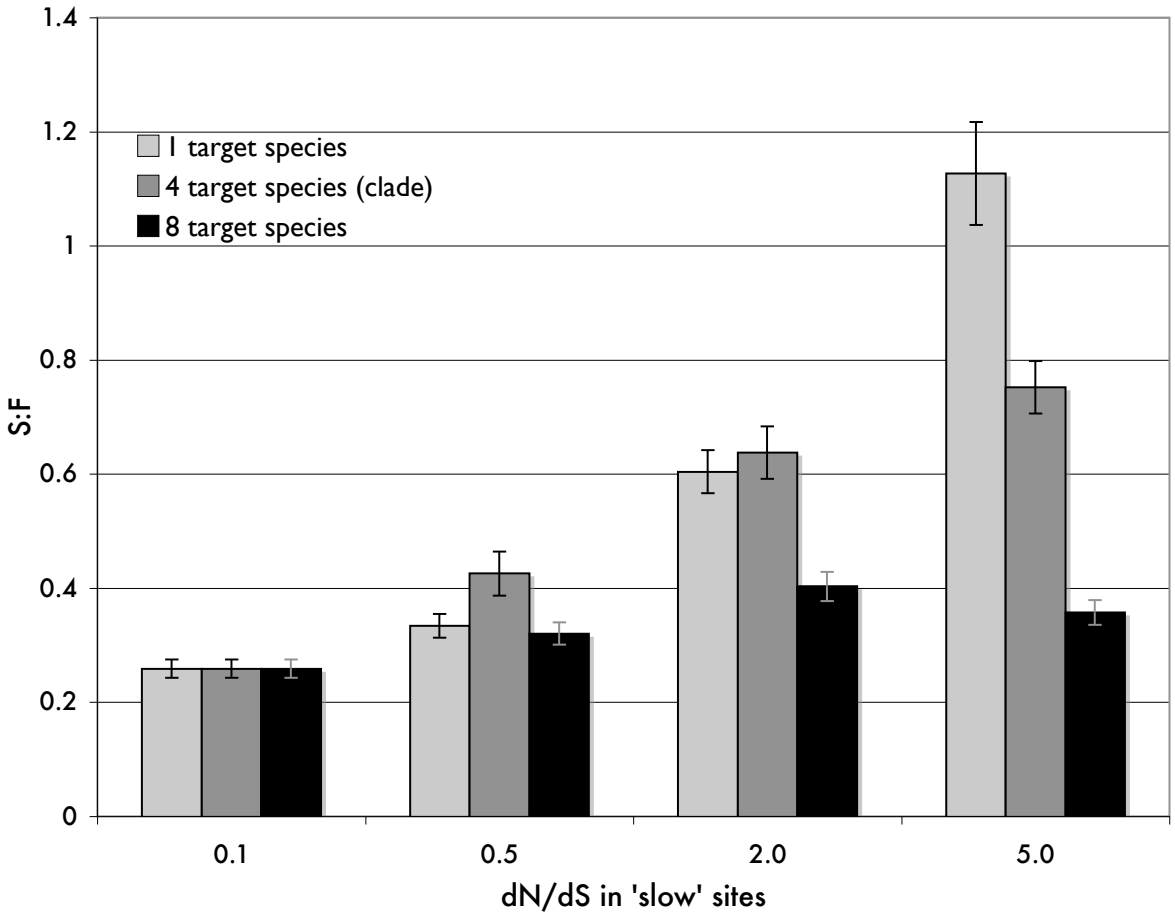


Figure 3.3

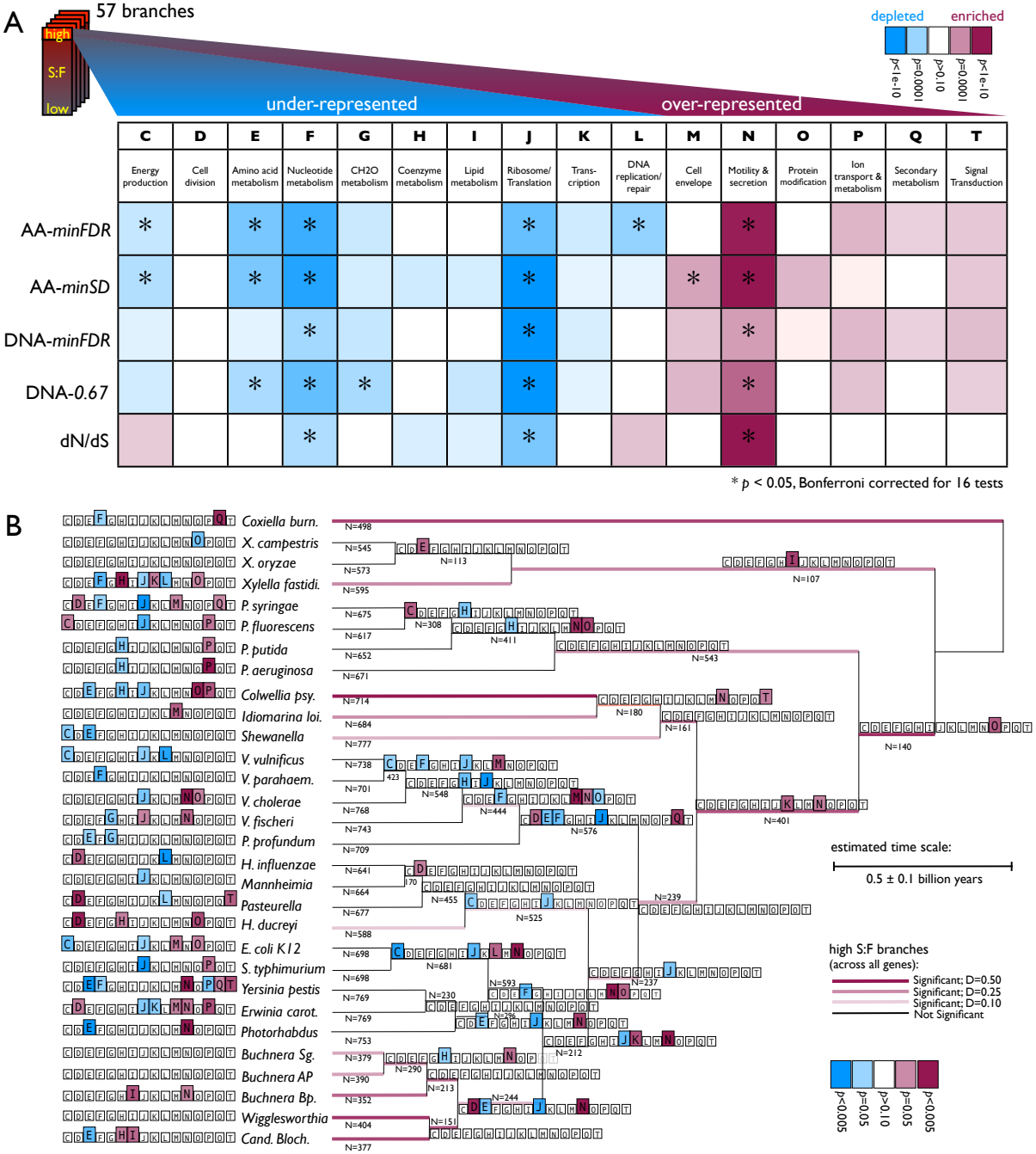

A
57 branches
high
S:F
low
under-represented
over-represented
depleted
enriched
p<1e-10
p=0.0001
p>0.10
p=0.0001
p<1e-10
C Energy production
D Cell division
E Amino acid metabolism
F Nucleotide metabolism
G CH2O metabolism
H Coenzyme metabolism
I Lipid metabolism
J Ribosome/ Translation
K Trans- cription
L DNA replication/ repair
M Cell envelope
N Motility & secretion
O Protein modification
P Ion transport & metabolism
Q Secondary metabolism
T Signal Transduction
AA-minFDR
AA-minSD
DNA-minFDR
DNA-0.67
dN/dS
* p < 0.05, Bonferroni corrected for 16 tests
B
Coxiella burn.
X. campestris
X. oryzae
Xylella fastidi.
P. syringae
P. fluorescens
P. putida
P. aeruginosa
Colwellia psy.
Idiomarina loi.
Shewanella
V. vulnificus
V. parahaem.
V. cholerae
V. fischeri
P. profundum
H. influenzae
Mannheimia
Pasteurella
H. ducreyi
E. coli K12
S. typhimurium
Yersinia pestis
Erwinia carot.
Photorhabdus
Buchnera Sg.
Buchnera AP
Buchnera Bp.
Wigglesworthia
Cand. Bloch.
estimated time scale:
0.5 ± 0.1 billion years
high S:F branches
(across all genes):
Significant; D=0.50
Significant; D=0.25
Significant; D=0.10
Not Significant
p<0.005
p=0.05
p>0.10
p=0.05
p<0.005

Figure 3.4

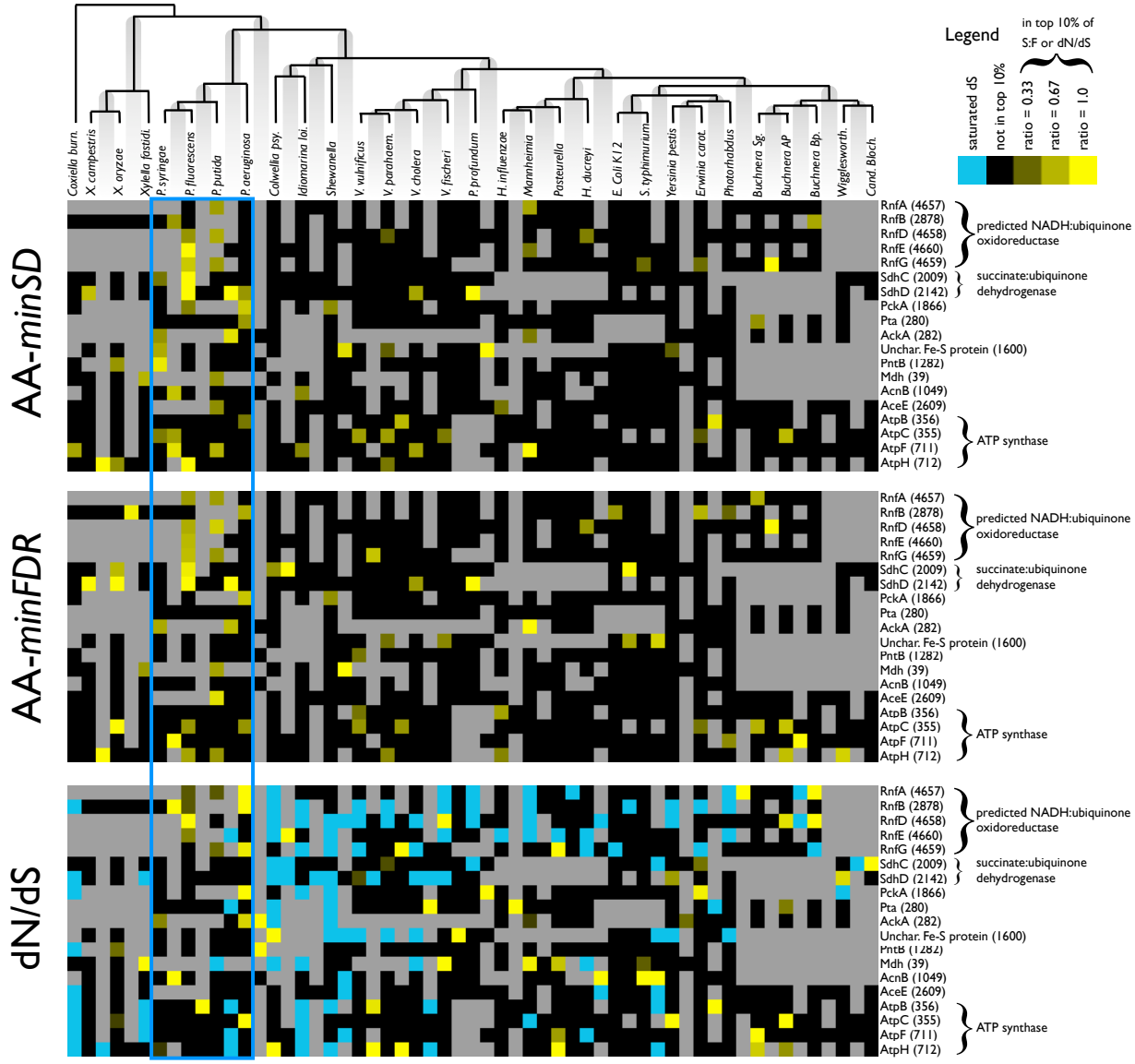

Coxiella burn.
X. campestris
X. oryzae
Xylella fastidi.
P. syringae
P. fluorescens
P. putida
P. aeruginosa
Colwellia psy.
Idiomarina loi.
Shewanella
V. vulnificus
V. parahaem.
V. cholera
V. fischeri
P. profundum
H. influenzae
Mannheimia
Pasteurella
H. ducreyi
E. Coli K12
S. typhimurium
Yersinia pestis
Erwinia carot.
Photorhabdus
Buchnera Sg
Buchnera AP
Buchnera Bp.
Wigglesworth.
Cand. Bloch.
Legend
in top 10% of S:F or dN/dS
saturated dS
not in top 10%
ratio = 0.33
ratio = 0.67
ratio = 1.0
AA-minSD
AA-minFDR
dN/dS
RnfA (4657)
RnfB (2878)
RnfD (4658)
RnfE (4660)
RnfG (4659)
predicted NADH:ubiquinone oxidoreductase
SdhC (2009)
SdhD (2142)
succinate:ubiquinone dehydrogenase
PckA (1866)
Pta (280)
AckA (282)
Unchar. Fe-S protein (1600)
PntB (1282)
Mdh (39)
AcnB (1049)
AceE (2609)
AtpB (356)
AtpC (355)
AtpF (711)
AtpH (712)
ATP synthase

Figure 3.5

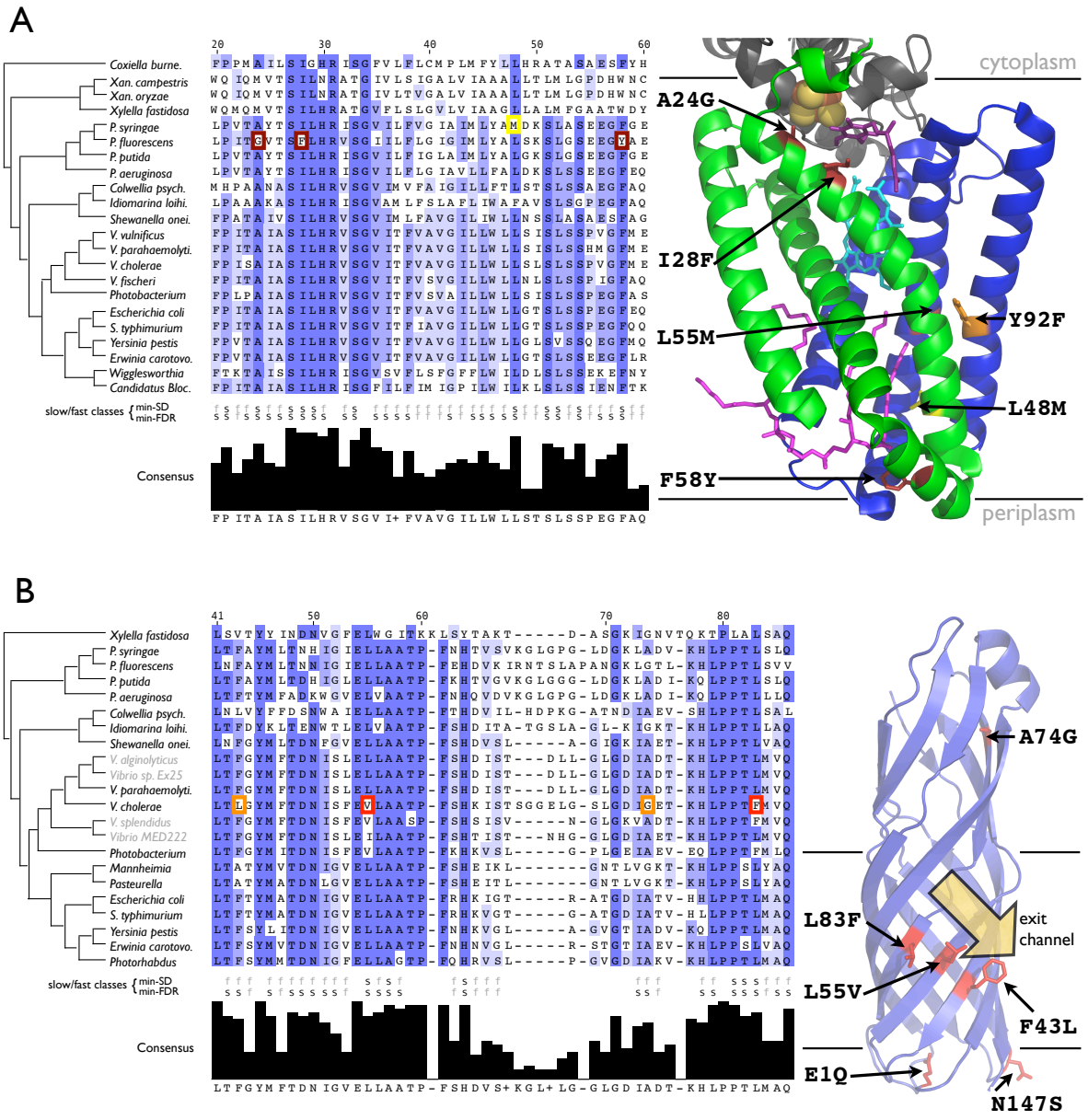

A
Coxiella burne.
Xan. campestris
Xan. oryzae
Xylella fastidosa
P. syringae
P. fluorescens
P. putida
P. aeruginosa
Colwellia psych.
Idiomarina loihi.
Shewanella onei.
V. vulnificus
V. parahaemolyti.
V. cholerae
V. fischeri
Photobacterium
Escherichia coli
S. typhimurium
Yersinia pestis
Erwinia carotovo.
Wigglesworthia
Candidatus Bloc.
slow/fast classes
min-SD
min-FDR
Consensus
cytoplasm
periplasm
A24G
I28F
L55M
F58Y
Y92F
L48M
B
Xylella fastidosa
P. syringae
P. fluorescens
P. putida
P. aeruginosa
Colwellia psych.
Idiomarina loihi.
Shewanella onei.
V. alginolyticus
Vibrio sp. Ex25
V. parahaemolyti.
V. cholerae
V. splendidus
Vibrio MED222
Photobacterium
Mannheimia
Pasteurella
Escherichia coli
S. typhimurium
Yersinia pestis
Erwinia carotovo.
Photorhabdus
slow/fast classes
min-SD
min-FDR
Consensus
A74G
L83F
exit
channel
L55V
F43L
E1Q
N147S

Table 3.1

		slow (S)		fast (F)		S:F		Fisher Test	
Method	***k***	**# subs.**	**# sites**	**# subs.**	**# sites**	**Ratio^**	**Rank**	**O.R.**	***p***
AA *minFDR*	0.55	5	66	3	37	0.93	2%*	2.42	0.28
AA *minSD*	0.35	3	30	5	73	1.50	1%*	5.26	0.043
DNA *minFDR*	0.75	17	275	23	88	0.20	61%	0.82	0.63
DNA *k*=0.67	0.67	8	228	32	135	0.13	55%	0.59	0.22
dN/dS NG86	n/a	dN = 0.042		dS = 0.39		0.11	37%	n/a	
dN/dS PAML	n/a	dN = 0.042		dS = 0.29		0.14	27%	n/a	

^ Corrected for multiple substitutions
* S:F ratio is among the top 10% highest in the genome

Table 3.2

		slow (S)		fast (F)		S:F		Fisher Test	
Method	***k***	**# subs.**	**# sites**	**# subs.**	**# sites**	**Ratio^**	**Rank**	**O.R.**	***p***
AA *minFDR*	0.55	6	75	6	49	0.63	1.5%*	1.71	0.38
AA *minSD*	0.35	2	16	10	108	1.38	0.1%*	8.25	0.037
DNA *minFDR*	0.75	34	360	86	135	0.07	25%	0.46	0.00015
DNA *k*=0.67	0.67	25	325	95	170	0.08	49%	0.46	0.00056
dN/dS NG86	n/a	dN = 0.07		dS > 0.75**		< 0.093	n/a**	n/a	
dN/dS PAML	n/a	dN = 0.08		dS = 3.70		< 0.022	42%	n/a	

^ Corrected for multiple substitutions
* S:F ratio is among the top 10% highest in the genome
** saturated in NG86, rank & ratio are only approximate

Figure 3.S1

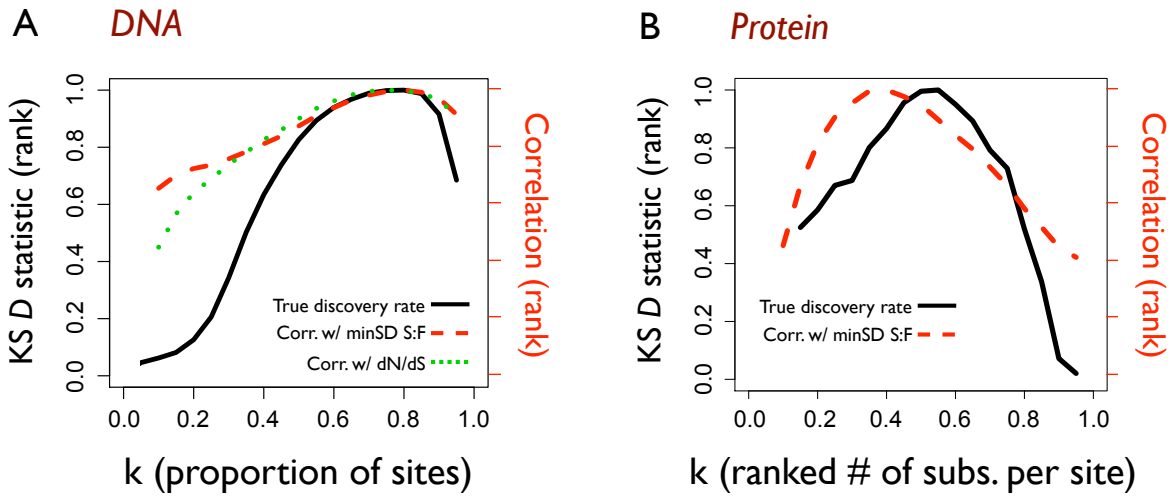


Figure 3.S2

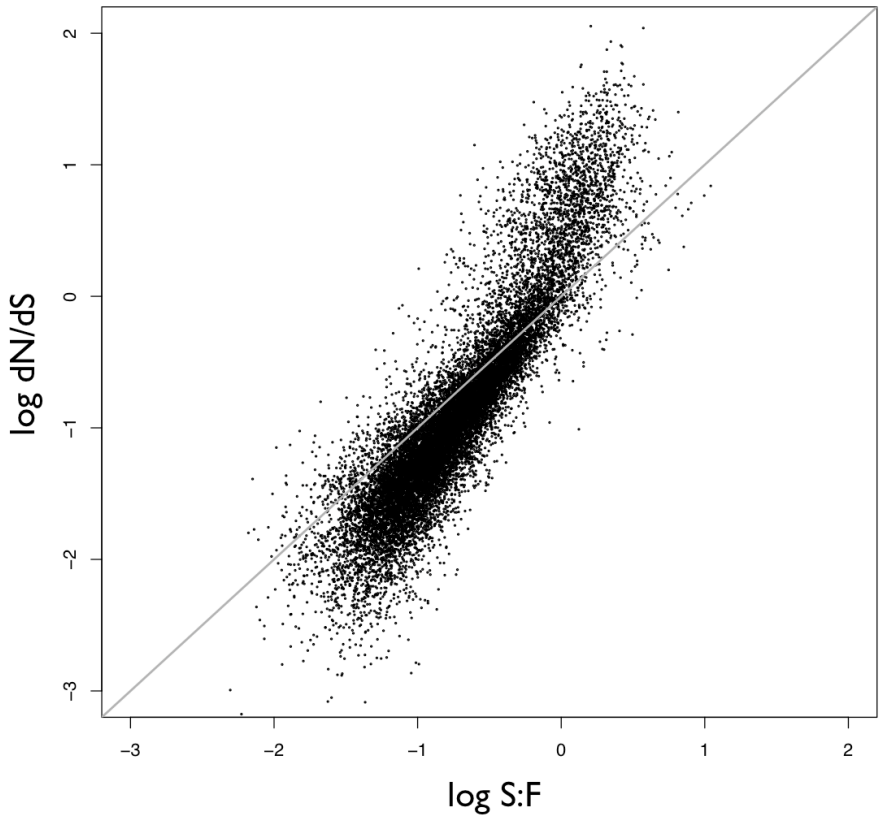


Figure 3.S3

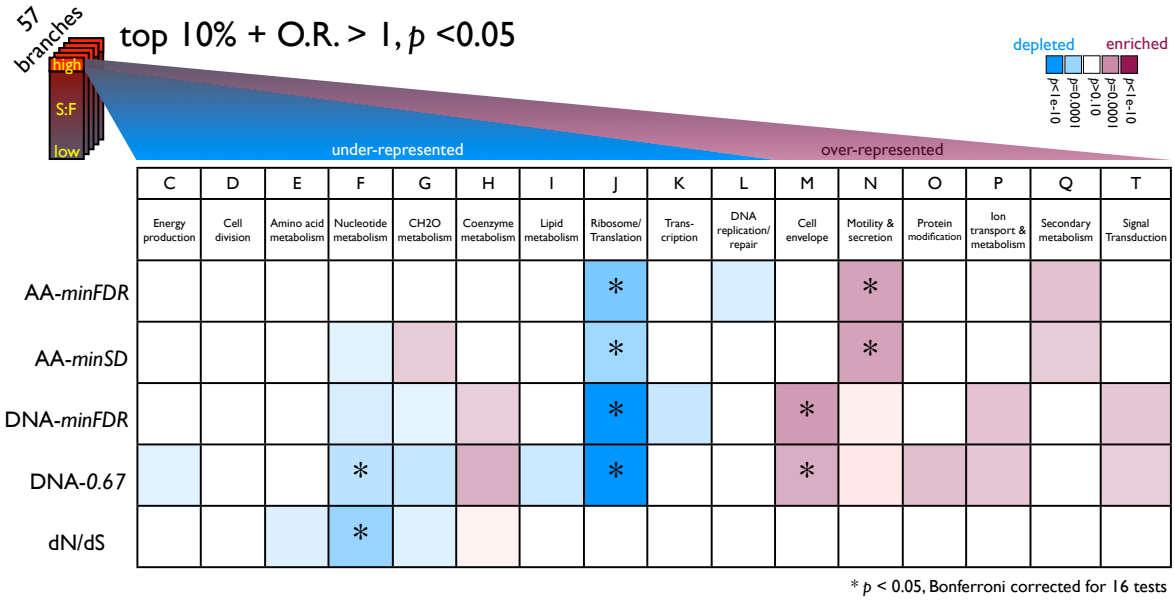

57 branches
high
S:F
low
top 10% + O.R. > 1, p <0.05
depleted
enriched
p<1e-10
p=0.0001
p>0.10
p=0.0001
p<1e-10
under-represented
over-represented
C Energy production
D Cell division
E Amino acid metabolism
F Nucleotide metabolism
G CH2O metabolism
H Coenzyme metabolism
I Lipid metabolism
J Ribosome/ Translation
K Trans- cription
L DNA replication/ repair
M Cell envelope
N Motility & secretion
O Protein modification
P Ion transport & metabolism
Q Secondary metabolism
T Signal Transduction
AA-minFDR
AA-minSD
DNA-minFDR
DNA-0.67
dN/dS
* p < 0.05, Bonferroni corrected for 16 tests

Figure 3.S4

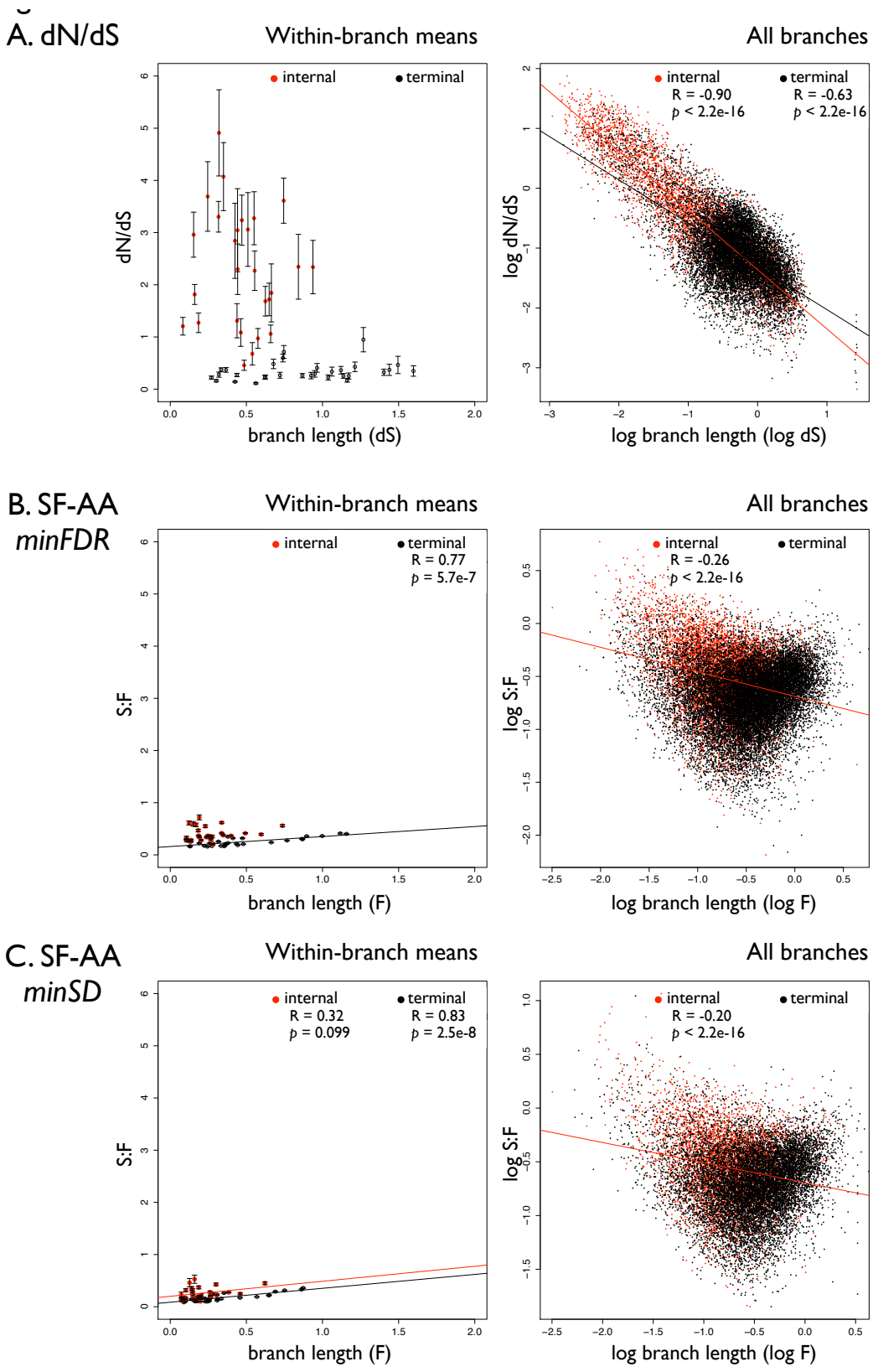

A. dN/dS
Within-branch means
All branches
internal
terminal
dN/dS
branch length (dS)
R = -0.90
p < 2.2e-16
R = -0.63
p < 2.2e-16
log dN/dS
log branch length (log dS)
B. SF-AA
minFDR
R = 0.77
p = 5.7e-7
S:F
branch length (F)
R = -0.26
p < 2.2e-16
log S:F
log branch length (log F)
C. SF-AA
minSD
R = 0.32
p = 0.099
R = 0.83
p = 2.5e-8
R = -0.20
p < 2.2e-16

Table 3.S1

		S:F method					
		DNA, *k*=0.67		AA, *minSD*		AA, *minFDR*	
dN/dS method	**set of genes**	**r**	**N**	**r**	**N**	**r**	**N**
PAML	all	0.70	29298	0.22	17936	0.20	27279
	dS < 2	0.78	19651	0.33	11227	0.36	17768
NG86	dS < 0.75	0.91	17142	0.26	10163	0.31	15495

Table 3.S2

	Ratio		Branch length	
Method	**internal**	**terminal**	**internal**	**terminal**
dN/dS	2.25 (2.31)	0.30 (0.34)	0.48 (0.48)	0.77 (0.87)
S:F *minFDR*	0.38 (0.41)	0.24 (0.25)	0.26 (0.20)	0.44 (0.47)
S:F *minSD*	0.23 (0.26)	0.17 (0.18)	0.20 (0.20)	0.32 (0.34)

Conclusions and future directions

In this thesis, I have described three new methods for comparing microbial genomes and inferring recombination events or changes in the regime of natural selection. I showed how these methods can be applied to better understand mechanisms of ecological differentiation, both over very recent or ancient evolutionary time scales.

While these methods collectively add substantially to the toolkit of microbial population genomics, the toolkit is still quite sparse. A recurring theme throughout this thesis has been the interplay of recombination and selection: on the one hand, recombination is essential to avoid genomewide purges of diversity every time a new adaptive mutation occurs and rises in frequency; on the other hand, some recombination events are themselves adaptive events and become fixed in a population by positive selection. We are tantalizingly close to being able to identify these adaptive recombination events, but still lack robust quantitative tests for their assessment. For example, the habitat-specific alleles discussed in Chapter 1, which are highly divergent between zooplankton- and small particle-associated populations, and often have clear origins outside either of these populations, are very good candidate positively selected recombination events. Yet traditional tests (*e.g.* M-K test) tend to lack power in detecting positive selection on such alleles because they have high levels of both nonsynonymous *and* synonymous substitutions. More work is needed in developing formal tests for the selective value of recombination events that we are currently only able to qualitatively label as 'non-neutral'. Such a test may take the form of an adapted version of the long-range haplotype test – where exceptionally long recombinant tracts (or regions with an exceptionally high density of SNPs) that are also at high frequency in a population are identified as outliers from the genomewide distribution. Assuming that microbial genomes can be sampled at sufficiently fine time scales, as was shown to be possible for the *Vibrio splendidus* populations, such a test might be very powerful to detect recent adaptations, and would be relatively straightforward to build in to the STARRInIGHTS framework.

The concept of neutral (within-species) versus adaptive (cross-species) recombination might also be harnessed to address the problem of how to define microbial species. One can imagine a situation in which housekeeping genes (*e.g.* DNA replication machinery, ribosomal proteins) are recombined neutrally within a species, and may even occasionally be transferred between species but will rarely if ever be driven to high frequency in the acceptor population due to lack of adaptive value. Conversely, membrane proteins or 'accessory' metabolic genes are more likely to become fixed in a new population, given the right selective pressures. Thus, it might be possible to design a test for unevenness of recombination frequency across different functional categories of genes. When strains within the same

species are compared, there should be no unevenness, with all categories being recombined equally. However once strains from two or more distinct species, subject to different selective pressures, are included, certain gene functions should be recombined more frequently than others. This type of species definition assumes that species boundaries are inherently fuzzy, so that differential selection, not recombination, defines species. Such a definition may be somewhat unsatisfactory, or at least potentially conservative. For example, there was no significant increase in unevenness across COG functional categories when the boundary between zooplankton- and small particle-associated vibrios was crossed (Figure 1.S3). These two populations would therefore, perhaps correctly, not be classed as distinct species, although they have clear ecological differences. Clearly, more work is needed on this front. Both empirical work (for example, comparing genomes of more strains from the Polz lab isolate collection, over a wider range of genetic distances than what was encompassed in the closely-related 8 strains described in Chapter 1), and modeling will be useful in understanding the process of sympatric speciation in microbes.

As more and more microbial genomes are sequenced, the main challenge will be in organizing them into ecologically (and clinically) relevant units. While this can to some extent be achieved by simply clustering aligned genome sequences, the resulting clusters are still of questionable biological relevance: Could there be functional diversity, or more fine-scale niche-partitioning within clusters? What is the role of recombination within and between clusters? While the STARRInIGHTS pipeline represents a step forward in answering these questions, we still need more refined models of microbial demographics and recombination. For example, we still do not fully understand the factors that limit homologous recombination in bacteria. In some bacteria, homologous recombination efficiency drops off proportionally with genetic distance (Thomas & Nielsen, 2005), but more complex barriers involving polymorphic competence peptides may come into play in some lineages (Cornejo et al, 2010). Incorporating such complexity, perhaps in a lineage-specific fashion, will be an important aspect of future population genetic models.

The methods described here are currently being applied to gain insight into the evolution of several different organisms. The selective signatures approach is being used to study patterns of lineage-specific variation of natural selection in mammals (Mar Alba; Evolutionary Genomics Group, Barcelona, Spain). I have also begun collaborations to apply STARRInIGHTS analysis to populations of *Streptococcus pneumoniae* genomes, some highly virulent and others not (with Tiffany Williams and George Weinstock, Washington University, St. Louis), and *Enterococcus faecalis* strains which vary in their levels of antibiotic resistance (with Kelli Palmer and Michael Gilmore, Schepens Eye Institute and

Harvard Medical School). These projects will hopefully yield biological insight into the evolution and diversification of different populations, ultimately demonstrating the utility and generality of the new tools I have developed.

Bibliography

Alm EJ, Huang KH, Price MN, Koche RP, Keller K, Dubchak IL*, et al* (2005) The MicrobesOnline Web site for comparative genomics. *Genome Res.* **15**:1015-1022

Alm E, Huang K and Arkin A (2006) The evolution of two-component systems in bacteria reveals different strategies for niche adaptation. *PLoS Comput.Biol.* **2**:e143

Anisimova M and Liberles DA (2007) The quest for natural selection in the age of comparative genomics. *Heredity* **99**:567-579

Anisimova M, Nielsen R and Yang Z (2003) Effect of recombination on the accuracy of the likelihood method for detecting positive selection at amino acid sites. *Genetics* **164**:1229-1236

Arjunan P, Nemeria N, Brunskill A, Chandrasekhar K, Sax M, Yan Y*, et al* (2002) Structure of the pyruvate dehydrogenase multienzyme complex E1 component from Escherichia coli at 1.85 A resolution. *Biochemistry* **41**:5213-5221

Astuti D, Latif F, Dallol A, Dahia PL, Douglas F, George E*, et al* (2001) Gene mutations in the succinate dehydrogenase subunit SDHB cause susceptibility to familial pheochromocytoma and to familial paraganglioma. *Am.J.Hum.Genet.* **69**:49-54

Atwood KC, Schneider LK and Ryan FJ (1951) Selective mechanisms in bacteria. *Cold Spring Harb.Symp.Quant.Biol.* **16**:345-355

Bantinaki E, Kassen R, Knight CG, Robinson Z, Spiers AJ and Rainey PB (2007) Adaptive divergence in experimental populations of Pseudomonas fluorescens. III. Mutational origins of wrinkly spreader diversity. *Genetics* **176**:441-453

Barrett RD, MacLean RC and Bell G (2006) Mutations of intermediate effect are responsible for adaptation in evolving Pseudomonas fluorescens populations. *Biology letters* **2**:236-238

Barrick JE, Yu DS, Yoon SH, Jeong H, Oh TK, Schneider D*, et al* (2009) Genome evolution and adaptation in a long-term experiment with Escherichia coli. *Nature* **461**:1243-1247

Battistuzzi F, Feijao A and Hedges SB (2004) A genomic timescale of prokaryote evolution: insights into the origin of methanogenesis, phototrophy, and the colonization of land. *BMC Evolutionary Biology* **4**:44

Beja O, Aravind L, Koonin EV, Suzuki MT, Hadd A, Nguyen LP*, et al* (2000) Bacterial rhodopsin: evidence for a new type of phototrophy in the sea. *Science* **289**:1902-1906

Bielawski JP, Dunn KA, Sabehi G and Beja O (2004) Darwinian adaptation of proteorhodopsin to different light intensities in the marine environment. *Proc.Natl.Acad.Sci.U.S.A.* **101**:14824-14829

Black WC,4th, Baer CF, Antolin MF and DuTeau NM (2001) Population genomics: genome-wide sampling of insect populations. *Annu.Rev.Entomol.* **46**:441-469

Blount ZD, Borland CZ and Lenski RE (2008) Historical contingency and the evolution of a key innovation in an experimental population of Escherichia coli. *Proc.Natl.Acad.Sci.U.S.A.* **105**:7899-7906

Bustamante CD, Fledel-Alon A, Williamson S, Nielsen R, Hubisz MT, Glanowski S*, et al* (2005) Natural selection on protein-coding genes in the human genome. *Nature* **437**:1153-1157

Charlesworth J and Eyre-Walker A (2006) The rate of adaptive evolution in enteric bacteria. *Mol.Biol.Evol.* **23**:1348-1356

Chen SL, Hung CS, Xu J, Reigstad CS, Magrini V, Sabo A, *et al* (2006) Identification of genes subject to positive selection in uropathogenic strains of Escherichia coli: a comparative genomics approach. *Proc.Natl.Acad.Sci.U.S.A.* **103**:5977-5982

Chimpanzee Sequencing and Analysis Consortium (2005) Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437**:69-87

Clamp M, Cuff J, Searle SM and Barton GJ (2004) The Jalview Java alignment editor. *Bioinformatics* **20**:426-427
Cohan FM (2001) Bacterial species and speciation. *Syst.Biol.* **50**:513-524

Cohan FM and Koeppel AF (2008) The Origins of Ecological Diversity in Prokaryotes. *Current Biology* **18**:1024-1034

Coin L and Durbin R (2004) Improved techniques for the identification of pseudogenes. *Bioinformatics* **20 Suppl 1**:I94-I100

Cornejo OE, McGee L and Rozen DE (2010) Polymorphic competence peptides do not restrict recombination in Streptococcus pneumoniae. *Mol.Biol.Evol.* **27**:694-702

Darling AC, Mau B, Blattner FR and Perna NT (2004) Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Res.* **14**:1394-1403

Darwin C (1859) The Origin of Species, 1st edition. (Burrow JW, ed.). Penguin Classics, London, UK, p. 226.

Das M, Chopra AK, Cantu JM and Peterson JW (1998) Antisera to selected outer membrane proteins of Vibrio cholerae protect against challenge with homologous and heterologous strains of V. cholerae. *FEMS Immunol.Med.Microbiol.* **22**:303-308

de Queiroz K (2005) Different species problems and their resolution. *Bioessays* **27**:1263-1269

de Visser JA (2002) The fate of microbial mutators. *Microbiology* **148**:1247-1252

DeLano WL (2002) The PyMOL Molecular Graphics System. DeLano Scientific, Palo Alto, CA.

Didelot X and Falush D (2007) Inference of bacterial microevolution using multilocus sequence data. *Genetics* **175**:1251-1266

Doolittle WF and Papke RT (2006) Genomics and the bacterial species problem. *Genome Biol.* **7**:116

Drake JW (1991) A Constant Rate of Spontaneous Mutation in DNA-Based Microbes. *Proceedings of the National Academy of Sciences* **88**:7160-7164

Drummond DA, Bloom JD, Adami C, Wilke CO and Arnold FH (2005) Why highly expressed proteins evolve slowly. *Proc.Natl.Acad.Sci.U.S.A.* **102**:14338-14343

Durbin R, Eddy SR, Krogh A and Mitchison G (1998) Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids, 10th edition. Cambridge University Press, Cambridge, UK.

Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**:1792-1797

Falush D and Bowden R (2006) Genome-wide association mapping in bacteria? *Trends Microbiol.* **14**:353-355
Falush D, Kraft C, Taylor NS, Correa P, Fox JG, Achtman M, *et al* (2001) Recombination and mutation during long-term gastric colonization by Helicobacter pylori: estimates of clock rates, recombination size, and minimal age. *Proc.Natl.Acad.Sci.U.S.A.* **98**:15056-15061

Fay JC and Wu CI (2000) Hitchhiking under positive Darwinian selection. *Genetics* **155**:1405-1413

Fearnhead P, Smith NG, Barrigas M, Fox A and French N (2005) Analysis of recombination in Campylobacter jejuni from MLST population data. *J.Mol.Evol.* **61**:333-340

Feil EJ, Maiden MC, Achtman M and Spratt BG (1999) The relative contributions of recombination and mutation to the divergence of clones of Neisseria meningitidis. *Mol.Biol.Evol.* **16**:1496-1502

Feldgarden M, Byrd N and Cohan FM (2003) Gradual evolution in bacteria: evidence from Bacillus systematics. *Microbiology* **149**:3565-3573

Felsenstein J (1993) PHYLIP (Phylogeny Inference Package) v.3.5. Department of Genetics, University of Washington, Seattle, USA.

Felsenstein J and Churchill GA (1996) A Hidden Markov Model approach to variation among sites in rate of evolution. *Mol.Biol.Evol.* **13**:93-104

Fraser C, Hanage WP and Spratt BG (2007) Recombination and the nature of bacterial speciation. *Science* **315**:476-480

Fraser C, Hanage WP and Spratt BG (2005) Neutral microepidemic evolution of bacterial pathogens. *Proc.Natl.Acad.Sci.U.S.A.* **102**:1968-1973

Fraser C, Alm EJ, Polz MF, Spratt BG and Hanage WP (2009) The bacterial species challenge: making sense of genetic and ecological diversity. *Science* **323**:741-746

Fraser HB, Hirsh AE, Steinmetz LM, Scharfe C and Feldman MW (2002) Evolutionary rate in the protein interaction network. *Science* **296**:750-752

Fry AJ and Wernegreen JJ (2005) The roles of positive and negative selection in the molecular evolution of insect endosymbionts. *Gene* **355**:1-10

Gevers D, Cohan FM, Lawrence JG, Spratt BG, Coenye T, Feil EJ*, et al* (2005) Opinion: Re-evaluating prokaryotic species. *Nature reviews.Microbiology* **3**:733-739

Gogarten JP, Doolittle WF and Lawrence JG (2002) Prokaryotic evolution in light of gene transfer. *Mol.Biol.Evol.* **19**:2226-2238

Gojobori T (1983) Codon substitution in evolution and the "saturation" of synonymous changes. *Genetics* **105**:1011-1027

Guindon S and Gascuel O (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst.Biol.* **52**:696-704

Gutierrez-Rios RM, Rosenblueth DA, Loza JA, Huerta AM, Glasner JD, Blattner FR*, et al* (2003) Regulatory network of Escherichia coli: consistency between literature knowledge and microarray profiles. *Genome Res.* **13**:2435-2443

Guttman DS, Gropp SJ, Morgan RL and Wang PW (2006) Diversifying selection drives the evolution of the type III secretion system pilus of Pseudomonas syringae. *Mol.Biol.Evol.* **23**:2342-2354

Guttman DS and Dykhuizen DE (1994) Detecting selective sweeps in naturally occurring Escherichia coli. *Genetics* **138**:993-1003

Hanada K, Shiu SH and Li WH (2007) The nonsynonymous/synonymous substitution rate ratio versus the radical/conservative replacement rate ratio in the evolution of mammalian genes. *Mol.Biol.Evol.* **24**:2235-2241

Hanage WP, Spratt BG, Turner KM and Fraser C (2006) Modelling bacterial speciation. *Philosophical transactions of the Royal Society of London.Series B, Biological sciences* **361**:2039-2044

Hazen T, Kennedy K, Chen S, Yi S and Sobecky P (2009) Inactivation of mismatch repair increases the diversity of Vibrio parahaemolyticus. *Environ.Microbiol.* **11**:1254-1266

Hegreness M, Shoresh N, Hartl D and Kishony R (2006) An equivalence principle for the incorporation of favorable mutations in asexual populations. *Science* **311**:1615-1617

Herbeck JT, Funk DJ, Degnan PH and Wernegreen JJ (2003) A conservative test of genetic drift in the endosymbiotic bacterium Buchnera: slightly deleterious mutations in the chaperonin groEL. *Genetics* **165**:1651-1660

Hey J (2006) Recent advances in assessing gene flow between diverging populations and species. *Curr.Opin.Genet.Dev.* **16**:592-596

Holt KE, Parkhill J, Mazzoni CJ, Roumagnac P, Weill FX, Goodhead I*, et al* (2008) High-throughput sequencing provides insights into genome variation and evolution in Salmonella Typhi. *Nat.Genet.* **40**:987-993

Hong H, Patel DR, Tamm LK and Berg Bvd (2006) The outer membrane protein OmpW forms an eight-stranded beta-barrel with a hydrophobic channel. *The Journal of biological chemistry* **281**:7568-7577

Hou S, Saw JH, Lee KS, Freitas TA, Belisle C, Kawarabayasi Y*, et al* (2004) Genome sequence of the deep-sea gamma-proteobacterium Idiomarina loihiensis reveals amino acid fermentation as a source of carbon and energy. *Proc.Natl.Acad.Sci.U.S.A.* **101**:18036-18041

Hughes AL, Friedman R, Rivailler P and French JO (2008) Synonymous and nonsynonymous polymorphisms versus divergences in bacterial genomes. *Mol.Biol.Evol.* **25**:2199-2209

Hughes AL (2007) Looking for Darwin in all the wrong places: the misguided quest for positive selection at the nucleotide sequence level. *Heredity* **99**:364-373

Hughes AL (2005) Evidence for abundant slightly deleterious polymorphisms in bacterial populations. *Genetics* **169**:533-538

Hughes AL and Nei M (1988) Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. *Nature* **335**:167-170

Hunt DE, David LA, Gevers D, Preheim SP, Alm EJ and Polz MF (2008a) Resource partitioning and sympatric differentiation among closely related bacterioplankton. *Science* **320**:1081-1085

Hunt DE, Gevers D, Vahora NM and Polz MF (2008b) Conservation of the chitin utilization pathway in the Vibrionaceae. *Appl.Environ.Microbiol.* **74**:44-51

International HapMap Consortium (2005) A haplotype map of the human genome. *Nature* **437**:1299-1320

Ishii N, Fujii M, Hartman PS, Tsuda M, Yasuda K, Senoo-Matsuda N*, et al* (1998) A mutation in succinate dehydrogenase cytochrome b causes oxidative stress and ageing in nematodes. *Nature* **394**:694-697

Itoh T, Martin W and Nei M (2002) Acceleration of genomic evolution caused by enhanced mutation rate in endocellular symbionts. *Proceedings of the National Academy of Sciences* **99**:12944-12948

Jolley KA, Wilson DJ, Kriz P, McVean G and Maiden MC (2005) The influence of mutation, recombination, population history, and selection on patterns of genetic diversity in Neisseria meningitidis. *Mol.Biol.Evol.* **22**:562-569

Jones DT, Taylor WR and Thornton JM (1992) The rapid generation of mutation data matrices from protein sequences. *Computer applications in the biosciences : CABIOS* **8**:275-282

Jordan IK, Kondrashov FA, Rogozin IB, Tatusov RL, Wolf YI and Koonin EV (2001) Constant relative rate of protein evolution and detection of functional diversification among bacterial, archaeal and eukaryotic proteins. *Genome Biol.* **2**:RESEARH0053

Kalinowski ST and Hedrick PW (2001) Estimation of linkage disequilibrium for loci with multiple alleles: basic approach and an application using data from bighorn sheep. *Heredity* **87**:698-708

Kelley JL, Madeoy J, Calhoun JC, Swanson W and Akey JM (2006) Genomic signatures of positive selection in humans and the limits of outlier approaches. *Genome Res.* **16**:980-989

Kettler GC, Martiny AC, Huang K, Zucker J, Coleman ML, Rodrigue S*, et al* (2007) Patterns and implications of gene gain and loss in the evolution of Prochlorococcus. *PLoS Genet* **3**:2515-2528

Kimura M (1977) Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution. *Nature* **267**:275-276

Kishino H and Hasegawa M (1989) Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in hominoidea. *J.Mol.Evol.* **29**:170-179

Kishony R and Leibler S (2003) Environmental stresses can alleviate the average deleterious effect of mutations. *J.Biol.* **2**:14

Koeppel A, Perry EB, Sikorski J, Krizanc D, Warner A, Ward DM*, et al* (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**:2504-2509

Kondrashov AS (1986) Multilocus Model of Sympatric Speciation III. Computer Simulations. *Theor.Popul.Biol.* **29**:1-15

Kondrashov AS and Mina MV (1986) Sympatric speciation: when is it possible? *Biol.J.Linn.Soc.* **27**:201-223
Kondrashov FA and Kondrashov AS (2001) Multidimensional epistasis and the disadvantage of sex. *Proc.Natl.Acad.Sci.U.S.A.* **98**:12089-12092

Kosiol C, Vinar T, Fonseca RRd, Hubisz MJ, Bustamante CD, Nielsen R*, et al* (2008) Patterns of positive selection in six Mammalian genomes. *PLoS Genet* **4**:e1000144

Kryazhimskiy S and Plotkin JB (2008) The Population Genetics of dN/dS. *PLoS Genet* **4**:e1000304

Lee SJ and Gralla JD (2002) Promoter use by sigma 38 (rpoS) RNA polymerase. Amino acid clusters for DNA binding and isomerization. *The Journal of biological chemistry* **277**:47420-47427

Lefebure T and Stanhope MJ (2007) Evolution of the core and pan-genome of Streptococcus: positive selection, recombination, and genome composition. *Genome Biol.* **8**:R71

Lerat E, Daubin V, Ochman H and Moran NA (2005) Evolutionary origins of genomic repertoires in bacteria. *Plos Biol* **3**:e130

Letunic I and Bork P (2007) Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics* **23**:127-128

Ley RE, Turnbaugh PJ, Klein S and Gordon JI (2006) Microbial ecology: human gut microbes associated with obesity. *Nature* **444**:1022-1023

Li L, Jr CJS and Roos DS (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res*. **13**:2178-2189

Li YF, Costello JC, Holloway AK and Hahn MW (2008) Reverse ecology and the power of population genomics. *Evolution* **62**:2984-2994

Lin W, Fullner KJ, Clayton R, Sexton JA, Rogers MB, Calia KE*, et al* (1999) Identification of a vibrio cholerae RTX toxin gene cluster that is tightly linked to the cholera toxin prophage. *Proc.Natl.Acad.Sci.U.S.A.* **96**:1071-1076

Luikart G, England PR, Tallmon D, Jordan S and Taberlet P (2003) The power and promise of population genomics: from genotyping to genome typing. *Nat.Rev.Genet.* **4**:981-994

Ma W, Dong FF, Stavrinides J and Guttman DS (2006) Type III effector diversification via both pathoadaptation and horizontal transfer in response to a coevolutionary arms race. *PLoS Genet* **2**:e209

Maiden MC (2008) Population genomics: diversity and virulence in the Neisseria. *Curr.Opin.Microbiol.* **11**:467-471

Majewski J and Cohan FM (1999) Adapt globally, act locally: the effect of selective sweeps on bacterial sequence diversity. *Genetics* **152**:1459-1474

Mandel MJ, Wollenberg MS, Stabb EV, Visick KL and Ruby EG (2009) A single regulatory gene is sufficient to alter bacterial host range. *Nature* **457**:215-218

Martin DP, Walt Evd, Posada D and Rybicki EP (2005) The evolutionary value of recombination is constrained by genome modularity. *PLoS Genet* **1**:e51

Martiny AC, Coleman ML and Chisholm SW (2006) Phosphate acquisition genes in Prochlorococcus ecotypes: evidence for genome-wide adaptation. *Proc.Natl.Acad.Sci.U.S.A.* **103**:12552-12557

Massingham T and Goldman N (2005) Detecting amino acid sites under positive selection and purifying selection. *Genetics* **169**:1753-1762

Mau B, Glasner JD, Darling AE and Perna NT (2006) Genome-wide detection and analysis of homologous recombination among sequenced strains of Escherichia coli. *Genome Biol.* **7**:R44

Mavrodi DV, Blankenfeldt W and Thomashow LS (2006) Phenazine compounds in fluorescent Pseudomonas spp. biosynthesis and regulation. *Annu.Rev.Phytopathol.* **44**:417-445

McDonald JH and Kreitman M (1991) Adaptive protein evolution at the Adh locus in Drosophila. *Nature* **351**:652-654

McInerney JO (2006) The causes of protein evolutionary rate variation. *Trends in ecology & evolution (Personal edition)* **21**:230-232

McKane M and Milkman R (1995) Transduction, restriction and recombination patterns in Escherichia coli. *Genetics* **139**:35-43

McVean G, Awadalla P and Fearnhead P (2002) A coalescent-based method for detecting and estimating recombination from gene sequences. *Genetics* **160**:1231-1241

Meibom KL, Li XB, Nielsen AT, Wu C, Roseman S and Schoolnik GK (2004) The Vibrio cholerae chitin utilization program. *Proc.Natl.Acad.Sci.U.S.A.* **101**:2524-2529

Mering Cv, Hugenholtz P, Raes J, Tringe SG, Doerks T, Jensen LJ*, et al* (2007) Quantitative Phylogenetic Assessment of Microbial Communities in Diverse Environments. *Science* **315**:1126-1130

Mes TH, Doeleman M, Lodders N, Nubel U and Stal LJ (2006) Selection on protein-coding genes of natural cyanobacterial populations. *Environ.Microbiol.* **8**:1534-1543

Michel AP, Sim S, Powell THQ, Taylor MS, Nosil P and Feder JL (2010) Widespread genomic divergence during sympatric speciation. *Proceedings of the National Academy of Sciences*: Epub May 10.

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

Milkman R and Bridges MM (1990) Molecular evolution of the Escherichia coli chromosome. III. Clonal frames. *Genetics* **126**:505-517

Miyamoto MM and Fitch WM (1995) Testing the covarion hypothesis of molecular evolution. *Mol.Biol.Evol.* **12**:503-513

Moran NA (1996) Accelerated evolution and Muller's rachet in endosymbiotic bacteria. *Proceedings of the National Academy of Sciences* **93**:2873-2878

Muse SV and Weir BS (1992) Testing for equality of evolutionary rates. *Genetics* **132**:269-276

Mwangi MM, Wu SW, Zhou Y, Sieradzki K, Lencastre Hd, Richardson P*, et al* (2007) Tracking the in vivo evolution of multidrug resistance in Staphylococcus aureus by whole-genome sequencing. *Proc.Natl.Acad.Sci.U.S.A.* **104**:9451-9456

Myers LC, Terranova MP, Ferentz AE, Wagner G and Verdine GL (1993) Repair of DNA methylphosphotriesters through a metalloactivated cysteine nucleophile. *Science* **261**:1164-1167

Nakamura Y, Itoh T, Matsuda H and Gojobori T (2004) Biased biological functions of horizontally transferred genes in prokaryotic genomes. *Nat Genet* **36**:760-6

Nei M and Gojobori T (1986) Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol.Biol.Evol.* **3**:418-426

Nei M and Kumar S (2000) Molecular Evolution and Phylogenetics. Oxford University Press, New York, NY.

Nozawa M, Suzuki Y and Nei M (2009) Reliabilities of identifying positive selection by the branch-site and the site-prediction methods. *Proc.Natl.Acad.Sci.U.S.A.* **109**:6700-6705

Ochman H, Elwyn S and Moran NA (1999) Calibrating bacterial evolution. *Proc.Natl.Acad.Sci.U.S.A.* **96**:12638-12643

Ochman H and Wilson AC (1987) Evolution in bacteria: evidence for a universal substitution rate in cellular genomes. *J.Mol.Evol.* **26**:74-86

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

Pal C, Papp B and Lercher MJ (2005) Adaptive evolution of bacterial metabolic networks by horizontal gene transfer. *Nat.Genet.* **37**:1372-1375

Papa R, Martin A and Reed RD (2008) Genomic hotspots of adaptation in butterfly wing pattern evolution. *Current opinion in genetics & development* **18**:559-564

Paulsen IT, Press CM, Ravel J, Kobayashi DY, Myers GS, Mavrodi DV*, et al* (2005) Complete genome sequence of the plant commensal Pseudomonas fluorescens Pf-5. *Nat.Biotechnol.* **23**:873-878

Pepperell C, Hoeppner VH, Lipatov M, Wobeser W, Schoolnik GK and Feldman MW (2010) Bacterial genetic signatures of human social phenomena among M. tuberculosis from an Aboriginal Canadian population. *Mol.Biol.Evol.* **27**:427-440

Polz MF, Hunt DE, Preheim SP and Weinreich DM (2006) Patterns and mechanisms of genetic and phenotypic differentiation in marine microbes. *Philos T R Soc B* **361**:2009-2021

Price MN, Huang KH, Alm EJ and Arkin AP (2005) A novel method for accurate operon predictions in all sequenced prokaryotes. *Nucleic Acids Res.* **33**:880-892

Price MN, Dehal PS and Arkin AP (2010) FastTree 2--approximately maximum-likelihood trees for large alignments. *PLoS ONE* **5**:e9490

Price-Whelan A, Dietrich LE and Newman DK (2007) Pyocyanin alters redox homeostasis and carbon flux through central metabolic pathways in Pseudomonas aeruginosa PA14. *J.Bacteriol.* **189**:6372-6381

Price-Whelan A, Dietrich LE and Newman DK (2006) Rethinking 'secondary' metabolism: physiological roles for phenazine antibiotics. *Nature chemical biology* **2**:71-78

Pupko T, Pe'er I, Shamir R and Graur D (2000) A fast algorithm for joint reconstruction of ancestral amino acid sequences. *Mol.Biol.Evol.* **17**:890-896

Rambaut A and Grassly NC (1997) Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Computer applications in the biosciences : CABIOS* **13**:235-238

Rasmussen MD and Kellis M (2007) Accurate gene-tree reconstruction by learning gene- and species-specific substitution rates across multiple complete genomes. *Genome Res.* **17**:1932-1942

Rice P, Longden I and Bleasby A (2000) EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet.* **16**:276-277

Rocha EP (2006) The quest for the universals of protein evolution. *Trends Genet.* **8**:412-416

Rocha EP, Smith JM, Hurst LD, Holden MT, Cooper JE, Smith NH*, et al* (2006) Comparisons of dN/dS are time dependent for closely related bacterial genomes. *J.Theor.Biol.* **239**:226-235

Rocha EP and Danchin A (2004) An analysis of determinants of amino acids substitution rates in bacterial proteins. *Mol.Biol.Evol.* **21**:108-116

Rozen DE, Schneider D and Lenski RE (2005) Long-term experimental evolution in Escherichia coli. XIII. Phylogenetic history of a balanced polymorphism. *J.Mol.Evol.* **61**:171-180

Sabehi G, Beja O, Suzuki MT, Preston CM and DeLong EF (2004) Different SAR86 subgroups harbour divergent proteorhodopsins. *Environ.Microbiol.* **6**:903-910

Sabeti PC, Varilly P, Fry B, Lohmueller J, Hostetter E and Cotsapas C (2007) Genome-wide detection and characterization of positive selection in human populations. *Nature* **449**:913-918

Sabeti PC, Schaffner SF, Fry B, Lohmueller J, Varilly P, Shamovsky O*, et al* (2006) Positive natural selection in the human lineage. *Science* **312**:1614-1620

Sabeti PC, Reich DE, Higgins JM, Levine HZ, Richter DJ, Schaffner SF*, et al* (2002) Detecting recent positive selection in the human genome from haplotype structure. *Nature* **419**:832-837

Saeed R and Deane CM (2006) Protein protein interactions, evolutionary rate, abundance and age. *BMC Bioinformatics* **7**:128

Sarich VM and Wilson AC (1967) Rates of albumin evolution in primates. *Proc.Natl.Acad.Sci.U.S.A.* **58**:142-148

Schluter D and Conte GL (2009) Genetics and ecological speciation. *Proc.Natl.Acad.Sci.U.S.A.* **106**:9955-9962

Schmidt B, Sinha R, Beresford-Smith B and Puglisi SJ (2009) A fast hybrid short read fragment assembly algorithm. *Bioinformatics* **25**:2279-2280

Schmidt HA, Strimmer K, Vingron M and Haeseler Av (2002) TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics* **18**:502-504

Segre D, Deluna A, Church GM and Kishony R (2005) Modular epistasis in yeast metabolism. *Nat.Genet.* **37**:77-83

Shapiro BJ and Alm EJ (2008) Comparing Patterns of Natural Selection across Species Using Selective Signatures. *PLoS Genet* **4**:e23

Shapiro BJ, David LA, Friedman J and Alm EJ (2009) Looking for Darwin's footprints in the microbial world. *Trends Microbiol.* **17**:196-204

Shapiro JA, Huang W, Zhang C, Hubisz MJ, Lu J, Turissini DA*, et al* (2007) Adaptive genic evolution in the Drosophila genomes. *Proc.Natl.Acad.Sci.U.S.A.* **104**:2271-2276

Sheppard SK, McCarthy ND, Falush D and Maiden MC (2008) Convergence of Campylobacter species: implications for bacterial evolution. *Science* **320**:237-239

Simmons SL, DiBartolo G, Denef VJ, Goltsman DSA, Thelen MP and Banfield JF (2008) Population genomic analysis of strain variation in Leptospirillum group II bacteria involved in acid mine drainage formation. *Plos Biol* **6**:1427-1442

Smith JM and Smith NH (1996) Synonymous nucleotide divergence: what is saturation? *Genetics* **142**:1033-1036

Smith JM, Smith NH, O'Rourke M and Spratt BG (1993) How clonal are bacteria? *Proc.Natl.Acad.Sci.U.S.A.* **90**:4384-4388

Sokurenko EV, Feldgarden M, Trintchina E, Weissman SJ, Avagyan S, Chattopadhyay S*, et al* (2004) Selection footprint in the FimH adhesin shows pathoadaptive niche differentiation in Escherichia coli. *Mol.Biol.Evol.* **21**:1373-1383

Storey JD and Tibshirani R (2003) Statistical significance for genomewide studies. *Proc.Natl.Acad.Sci.U.S.A.* **100**:9440-9445

Susko E, Leigh J, Doolittle WF and Bapteste E (2006) Visualizing and assessing phylogenetic congruence of core gene sets: a case study of the gamma-proteobacteria. *Molecular biology and evolution* **23**:1019-1030

Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**:585-595

Tamas I, Klasson L, Canback B, Naslund AK, Eriksson AS, Wernegreen JJ*, et al* (2002) 50 million years of genomic stasis in endosymbiotic bacteria. *Science* **296**:2376-2379

Tatusov RL, Koonin EV and Lipman DJ (1997) A genomic perspective on protein families. *Science* **278**:631-637

Thomas CM and Nielsen KM (2005) Mechanisms of, and barriers to, horizontal gene transfer between bacteria. *Nat.Rev.Microbiol.* **3**:711-721

Thompson JR, Pacocha S, Pharino C, Klepac-Ceraj V, Hunt DE, Benoit J*, et al* (2005) Genotypic diversity within a natural coastal bacterioplankton population. *Science* **307**:1311-1313

Thorne JL, Kishino H and Painter IS (1998) Estimating the rate of evolution of the rate of molecular evolution. *Molecular biology and evolution* **15**:1647-1657

Thornton KR, Jensen JD, Becquet C and Andolfatto P (2007) Progress and prospects in mapping recent selection in the genome. *Heredity* **98**:340-348

Tishkoff SA, Reed FA, Ranciaro A, Voight BF, Babbitt CC, Silverman JS*, et al* (2007) Convergent adaptation of human lactase persistence in Africa and Europe. *Nat.Genet.* **39**:31-40

Turnbaugh PJ, Ley RE, Mahowald MA, Magrini V, Mardis ER and Gordon JI (2006) An obesity-associated gut microbiome with increased capacity for energy harvest. *Nature* **444**:1027-1031

Turner T, Hahn M and Nuzhdin S (2005) Genomic islands of speciation in Anopheles gambiae. *Plos Biol* **3**:1572-1578

Velde KVd and Kiekens P (2004) Structure analysis and degree of substitution of chitin, chitosan and dibutyrylchitin by FT-IR spectroscopy and solid state 13C NMR. *Carbohydr.Polym.* **58**:409-416

Volkman SK, Sabeti PC, DeCaprio D, Neafsey DE, Schaffner SF, Jr DAM*, et al* (2007) A genome-wide map of diversity in Plasmodium falciparum. *Nat.Genet.* **39**:113-119

Vos M and Didelot X (2009) A comparison of homologous recombination rates in bacteria and archaea. *ISME J* **3**:199-208

Wall DP, Hirsh AE, Fraser HB, Kumm J, Giaever G, Eisen MB*, et al* (2005) Functional genomic analysis of the rates of protein evolution. *Proc.Natl.Acad.Sci.U.S.A.* **102**:5483-5488

Weber E and Koebnik R (2006) Positive selection of the Hrp pilin HrpE of the plant pathogen Xanthomonas. *J.Bacteriol.* **188**:1405-1410

Weigel LM, Clewell DB, Gill SR, Clark NC, McDougal LK, Flannagan SE*, et al* (2003) Genetic analysis of a high-level vancomycin-resistant isolate of Staphylococcus aureus. *Science* **302**:1569-1571

Whelan S and Goldman N (2001) A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol.Biol.Evol.* **18**:691-699

Whittam TS, Ochman H and Selander RK (1983) Geographic components of linkage disequilibrium in natural populations of Escherichia coli. *Mol.Biol.Evol.* **1**:67-83

Wiehlmann L, Wagner G, Cramer N, Siebert B, Gudowius P, Morales G*, et al* (2007) Population structure of Pseudomonas aeruginosa. *Proceedings of the National Academy of Sciences* **104**:8101

Wilmes P, Simmons SL, Denef VJ and Banfield JF (2009) The dynamic genetic repertoire of microbial communities. *FEMS Microbiol.Rev.* **33**:109-132

Xu C, Wang S, Ren H, Lin X, Wu L and Peng X (2005) Proteomic analysis on the expression of outer membrane proteins of Vibrio alginolyticus at different sodium concentrations. *Proteomics* **5**:3142-3152

Xu L, Chen H, Hu X, Zhang R, Zhang Z and Luo ZW (2006) Average gene length is highly conserved in prokaryotes and eukaryotes and diverges only between the two kingdoms. *Mol.Biol.Evol.* **23**:1107-1108

Yang Z and Nielsen R (2002) Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Mol.Biol.Evol.* **19**:908-917

Yang Z (2000) PAML: Phylogenetic Analysis by Maximum Likelihood. University College, London

Yang Z (1998) Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Mol Biol Evol* **15**:568-73

Yankovskaya V, Horsefield R, Tornroth S, Luna-Chavez C, Miyoshi H, Leger C*, et al* (2003) Architecture of succinate dehydrogenase and reactive oxygen species generation. *Science* **299**:700-704

Zeng K, Fu YX, Shi S and Wu CI (2006) Statistical tests for detecting positive selection by utilizing high-frequency variants. *Genetics* **174**:1431-1439