Molecular Biology and Evolution

Research article

# Function-Specific Accelerations in Rates of Sequence Evolution Suggest Predictable Epistatic Responses to Reduced Effective Population Size

Tobias Warnecke*†,[1] and Eduardo P. C. Rocha[2,3]

[1]Department of Biology and Biochemistry, University of Bath, Bath, United Kingdom

[2]Département Génomes et Génétique, Microbial Evolutionary Genomics, Institut Pasteur, Paris, France

[3]Centre National de la Recherche Scientifique, Unité de Recherche Associée 2171, Paris, France

†Present address: Bioinformatics and Genomics Program, Centre for Genomic Regulation, Universitat Pompeu Fabra, Barcelona, Spain.

***Corresponding author:** E-mail: Tobias.Warnecke@crg.es.

**Associate editor:** Jennifer Wernegreen

## Abstract

Changes in effective population size impinge on patterns of molecular evolution. Notably, slightly deleterious mutations are more likely to drift to fixation in smaller populations, which should typically also lead to an overall acceleration in the rates of evolution. This prediction has been validated empirically for several endosymbiont and island taxa. Here, we first show that rate accelerations are also evident in bacterial pathogens whose recent shifts in virulence make them prime candidates for reduced effective population size: *Bacillus anthracis*, *Bordetella parapertussis*, *Mycobacterium leprae*, *Salmonella enterica* typhi, *Shigella* spp., and *Yersinia pestis*. Using closely related genomes to analyze substitution rate dynamics across six phylogenetically independent bacterial clades, we demonstrate that relative rates of coding sequence evolution are biased according to gene functional category. Notably, genes that buffer against slightly deleterious mutations, such as chaperones, experience stronger rate accelerations than other functional classes at both nonsynonymous and synonymous sites. Although theory predicts altered evolutionary dynamics for buffer loci in the face of accumulating deleterious mutations, to observe even stronger rate accelerations is surprising. We suggest that buffer loci experience elevated substitution rates because the accumulation of deleterious mutations in the remainder of the genome favors compensatory substitutions in *trans*. Critically, the hyper-acceleration is evident across phylogenetically independent clades, supporting the hypothesis that reductions in effective population size predictably induce epistatic responses in genes that buffer against slightly deleterious mutations.

**Key words:** effective population size, quality control, rate acceleration, gene function, epistasis.

## Introduction

In finite populations, the likely fate of a sequence variant is determined by the leverage of selection versus genetic drift (Wright 1931). An increase in the relative power of drift—conceptualized as a reduction in the effective population size ($N_e$)—can therefore have strong repercussions for subsequent patterns of molecular evolution. Notably, depending on the distribution of fitness effects and the severity of reduction in $N_e$, a greater or lesser fraction of slightly deleterious mutations that would have previously been purged by selection now behave as effectively neutral variants and are more likely to rise to fixation (see Charlesworth 2009 for a recent review). In populations where slightly deleterious mutations outnumber slightly beneficial mutations, these altered fixation dynamics should also manifest themselves as an overall acceleration in the rate of nucleotide substitutions (Woolfit 2009).

These theoretical predictions have been confirmed empirically in populations where $N_e$ is likely reduced because of low census population sizes or peculiar lifestyles. Arguably, the strongest and most detailed evidence in this regard comes from obligate endosymbiotic bacteria (Moran 1996; Wernegreen and Moran 1999; Woolfit and Bromham 2003; Canback et al. 2004; Kuo et al. 2009) where transmission dynamics dictate repeated bottlenecks, no recombination, and strong population subdivision, resulting in dramatically reduced $N_e$ compared with free-living relatives. Consistent with predictions, 16S ribosomal RNA genes show accelerated rates of sequence evolution along several endosymbiont lineages (Woolfit and Bromham 2003). Similarly, genome-wide rates of nonsynonymous to synonymous substitutions (dN/dS)—interpreted as a proxy for the power of purifying selection (Ohta 1992)—are elevated in obligate endosymbionts (Kuo et al. 2009). Comparisons between island and mainland taxa have yielded comparable results (Johnson and Seger 2001; Woolfit and Bromham 2005; Legrand et al. 2009), with rates of evolution typically faster in island species, where $N_e$ is presumably smaller. This supports the notion that the observed rate accelerations are prompted, at least in part, by reductions in $N_e$ rather than by a specific ecological scenario (endosymbiosis).

To date, the majority of studies concerned with rate accelerations under reduced $N_e$ have either covered few loci or





http://mbe.oxfordjournals.org/

considered genome-wide averages with little regard to functional differences between genes. If the fitness effects of individual mutations are independent, then the population genetic processes described above should indeed be agnostic to gene function. However, the effect of any one mutation frequently depends on its genetic background and that background changes over time as the population persists at reduced $N_e$ and slightly deleterious variants continue to fix (Silander et al. 2007). In other words, mutations interact epistatically and earlier mutations condition the fitness landscape for subsequent mutations.

One particularly poignant case concerns epistasis between mutations in quality control (QC) machinery and mutations in the remainder of the expressed genome. The deleterious effect of a mutation may be reduced by the intervention of QC mechanisms that identify faulty products and mitigate their effects. Famously, chaperones can buffer against mutations that would otherwise destabilize protein structure (Rutherford and Lindquist 1998; Tokuriki and Tawfik 2009a). Although less extensively explored, other molecular systems can in principle also function as buffers. For example, in a number of eukaryotic taxa, mis-spliced mRNAs are targeted by the nonsense-mediated decay pathway (Maquat and Carmichael 2001), which degrades erroneous messages and thereby prevents them from being translated into nonsensical—and potentially harmful—proteins.

When effective population size shrinks, slightly deleterious mutations are predicted to accumulate in the genome, thus increasing the burden imposed on QC pathways. In turn, patterns of sequence evolution of QC genes may reflect this elevated burden. Indeed, theoretical work suggests that "chaperone-like" genes—genes that, by some mechanism, confer robustness to slightly deleterious mutations—should experience an altered selection regime under reduced $N_e$ (Krakauer and Plotkin 2002; Gros and Tenaillon 2009).

Here, we contrast the effect of reduced $N_e$ on the evolutionary regime of QC genes with substitution patterns in genes of different functional classes. For this, we shift from standard works on highly evolved endosymbiotic systems to recently emerged lineages of bacterial pathogens that exhibit more stringent host specificity, reduced recombination, and more frequent bottlenecks than their sister strains and are therefore prime candidates for reduced $N_e$. The genomes we analyze (table 1) are closely related allowing high-confidence reconstruction of synonymous and nonsynonymous substitutions. At the same time, although the taxa described below share the label "pathogenic," they occupy a diverse set of niches and exhibit very different virulence strategies and growth rates. For example, *M. leprae* is an exceedingly slow-growing gram-positive obligate intracellular pathogen that infects humans via the respiratory route and accumulates in body extremities, notably in macrophages and the peripheral nervous system (Britton and Lockwood 2004). *Shigella* spp. on the other hand are fast-growing gram-negative facultative pathogens transmitted through a fecal–oral route that provoke acute bloody diarrhea and, while producing intracellular infections, thrive in the absence of a host under favorable conditions (Kaper et al. 2004). Hence, although one may find broad commonalities between these emerging lineages of pathogens, for example, in terms of heightened stress induced by the host immune system, commonalities in sequence evolution between them cannot be readily attributed to a shared response to a shared novel environment.

## Materials and Methods

### General Strategy

We screened multiple clusters of closely related bacterial genomes to identify candidate clades where, based on comparative disease ecology, clonality, and genomic features (genome size, pseudogene content, and density of transposable elements; see table 1), we suspected at least one taxon to have strongly reduced $N_e$.

Based on a preliminary assessment of these criteria, we initially identified eight candidate clusters (*Bacillus*, *Bordetella*, *Escherichia*, *Mycobacterium*, *Neisseria*, *Salmonella*, *Xanthomonas*, and *Yersinia*) for which we reconstructed phylogenetic relationships from concatenated alignments of all orthologous protein-coding genes (see below for details). Within this larger phylogeny, we identified, if possible, four taxa that mapped onto the topology given in figure 1. The four taxa were chosen to include one genome with likely reduced $N_e$ in a derived position on the topology (taxon A in fig. 1) and three taxa (B–D) of likely larger effective population sizes. The only exception here is *Bordetella* where $N_e$ of *Bo. pertussis* (taxon C, table 1) is probably even smaller than $N_e$ of *Bo. parapertussis* (taxon A). However, the latter should still be smaller than $N_e$ of *Bo. bronchiseptica* (taxon B) justifying the inclusion of this clade in the analysis.

We discarded *Neisseria* and *Xanthomonas* from the analyses because the topology of available genomes did not match our requirements, and we considered indicators for differences in effective population sizes to be weak, respectively. For *Escherichia* spp., we analyzed two subclades, one including *Sh. boydii* and the other *Sh. flexneri*. In all analyses that involve combining data across clades, we only consider one of the two *Escherichia* sets of taxa because not all branches are phylogenetically independent. We present results for *Sh. boydii* but all results hold regardless of which *Shigella* taxon is dropped from the analysis.

### Ortholog Identification and Alignment

Coding sequences for all genomes were extracted from GenBank (RefSeq, ftp://ftp.ncbi.nih.gov/genomes/Bacteria). We excluded genes with internal in-frame stop codons and genes whose length was not a multiple of three nucleotides. One genome from each cluster (typically all available genomes with the same genus name) was randomly chosen as a pivot and compared with each other genome in a pairwise fashion. For every gene in the pivot genome, we conducted a FASTA search in the other genome and vice versa.

**Table 1.** Taxa Used in the Analyses and Traits Differentiating the Taxon With Reduced $N_e$ Relative to the Average of the Clade ($\Delta$).

| Clade | Taxon[a] | Core (kb)[c] | IS in Taxon A[d] | Clonal[e] | Δ (kb)[f] | Δ Coding Density[g] | Δ Niche[h] | References |
|---|---|---|---|---|---|---|---|---|
| *Bacillus* | *Ba. anthracis* Ames (A)<br>*Ba. cereus* E33L (B)<br>*Ba. cereus* Q1 (C)<br>*Ba. cereus* ATCC 14579 (D) | 2,949 | 14 | + | −82 | −2% | — | (Mock and Fouet 2001; Read et al. 2003; Helgason et al. 2004; Touchon and Rocha 2007) |
| *Bordetella* | *Bo. parapertussis* 12822 (A)[b]<br>*Bo. bronchiseptica* RB50 (B)<br>*Bo. pertussis* Tohama I (C)<br>*Bo. petrii* DSM 12804 (D) | 1,804 | 112 | + | −131 | −2% | — | (Yuk et al. 1998; Parkhill et al. 2003) |
| *Escherichia/Shigella* | *Sh. boydii* CDC 3083-94 (A)<br>*Sh. flexneri* 2a str. 301(A)<br>*E. coli* 55989 (B)<br>*E. coli* MG1655 (C)<br>*E. coli* ED1a (D) | 2,608<br>2,716 | 314<br>403 | +?<br>+? | −385<br>−394 | −6%<br>−10% | —<br>— | (Jin et al. 2002; Yang et al. 2005; Hershberg et al. 2007; Balbi et al. 2009) |
| *Mycobacterium* | *M. leprae* TN (A)<br>*M. tuberculosis* H37Rv (B)<br>*M. marinum* M (C)<br>*M. avium* 104 (D) | 782 | 0 | + | −2240 | −45% | — | (Cole et al. 2001; Stinear et al. 2007) |
| *Salmonella enterica* | *str Typhi* CT18 (A)<br>*str typhimurium* LT2 (B)<br>*str Schwarzengrund* CVM19633(C)<br>*str arizonae* (D) | 2,836 | 19 | + | +86 | −3% | — | (Beltran et al. 1988; Roumagnac et al. 2006; Touchon and Rocha 2007) |
| *Yersinia* | *Y. pestis* Angola (A)<br>*Y. pseudotuberculosis* IP 32953 (B)<br>*Y. pseudotuberculosis* YPIII (C)<br>*Y. pseudotuberculosis* IP 31758 (D) | 2,938 | 140 | + | −198 | −7% | — | (Achtman et al. 1999; Chain et al. 2004) |

NOTE.—IS, insertion sequence.

[a]Capitals in parentheses indicate the position of the strain in figure 1. "A" corresponds to the taxon with reduced $N_e$.

[b]The comparisons are for *Bo. parapertussis* and its sister group *Bo. bronchiseptica*. Yet, the outgroup *Bo. pertussis* has independently acquired even more IS and has an even shorter genome.

[c]Size of the concatenate of genes of the core genome.

[d]Number of insertion sequences in the genome.

[e]Lineage regarded as essentially clonal (+).

[f]Difference in genome size between A and the average of the other genomes of the clade.

[g]Difference in coding density between A and the average of the other genomes of the clade.

[h]Niche breadth in A relative to the other genomes of the clade (− narrower/+ broader).

We retained the ten best hits for every gene. For these ten hits, we constructed exact pairwise alignments using the Needleman–Wunsch algorithm where we did not penalize gaps at the edge of smaller sequences (end-gap free alignment following Erickson and Sellers (1983). At this stage, we considered as putative orthologs reciprocal best hits with >40% similarity and <20% difference in the length. In the next step, each putatively orthologous pair was tested for gene order conservation. Gene A in genome A was considered orthologous to a gene B in genome B if at least four genes in a neighborhood of ten genes around gene A (five genes upstream and five genes downstream of the focal gene) have an ortholog in a neighborhood of ten genes around gene B. Genes that do not satisfy this constraint were removed. We then removed all orthologous pairs that exhibit <85% amino acid conservation to the pivot. If for a given genome we eliminated more than 5% of the putative positional orthologs, we discarded the entire genome. We then defined the list of orthologs as the intersection of all pairwise lists. The use of exact alignment and gene order conservation allows minimizing false positive ortholog assignments in the face of horizontal gene transfer. This is important in the context of this work because falsely inferred orthologs tend to have very high substitution rates relative to true orthologs and might produce spurious results. Finally, we aligned these orthologs at the protein level using Muscle (Edgar 2004) and back-translated them to DNA.

## Phylogenetic Reconstruction

Following alignment of all genes, we concatenated all DNA alignments into one large alignment to reconstruct phylogenetic relationships of each clade. Given the very large size of core genome alignments and the low density of polymorphisms, we built a distance matrix with Tree-Puzzle (Schmidt et al. 2002) using maximum likelihood reconstruction under the Hasegawa–Kishino–Yano model. We then built the phylogenetic tree from the distance matrix using BIONJ (Gascuel 1997), which we consider to be the species tree. The robustness of the tree was evaluated by bootstrapping the concatenated alignment of the core genome, using BOOT and CONSENSE from the PHYLIP package (http://evolution.genetics.washington.edu/phylip.html). This tree then served as the guide tree for rate analyses in PAML (see below).

## Clusters of Orthologous Groups

We concatenated all orthologous genes by Clusters of Orthologous Group (COG) functional category

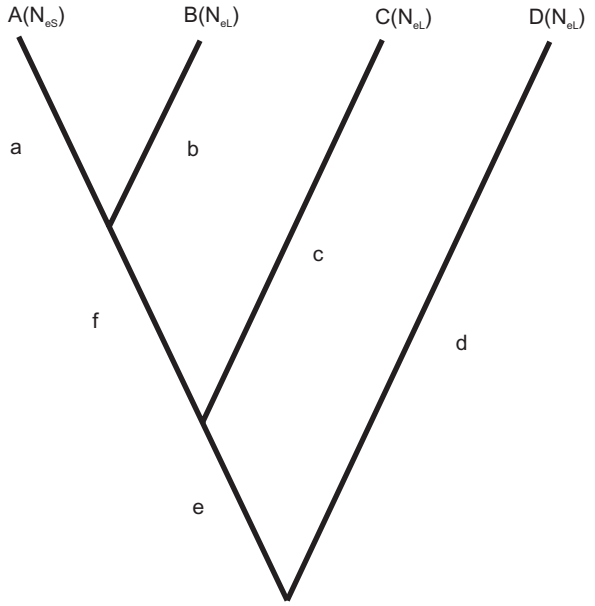


**Fig. 1.** Guide topology for four-taxon analysis of rate disparities (see table 1 for how taxa map onto topology by clade).

(http://www.ncbi.nlm.nih.gov/COG/grace/fiew.cgi) according to their assignment in the COG database, downloaded from NCBI (ftp://ftp.ncbi.nih.gov/genomes/Bacteria). If a gene was assigned to more than one COG functional category, it was allocated to all concatenates concerned. Note that this should exacerbate similarities between concatenates, whereas our aim here is to reveal differences. Note also, that all results presented below are qualitatively identical when we allocate a gene to only a single COG chosen randomly from the COG classes assigned to that gene (data not shown).

COG functional categories A, B, W, and Z (see supplementary table S1, Supplementary Material online for a list of all COG functional categories) were excluded from the analysis because very few, if any, genes belong to these categories so that we could not obtain robust rate estimates for these genes.

We also generated two custom functional categories. First, all orthologs without a COG assignment were defined to belong to a custom category "X" and concatenated accordingly. Genes that have not been identified as belonging to any COG domain are genes for which the COG algorithm cannot detect at least three mutually consistent genome-specific best hits to other genomes in the COG database (see Tatusov et al. 2000). These genes might therefore represent phylogenetically isolated lineage-specific genes. Second, we defined a functional category "QC" to represent QC genes. Although COG functional category O ("Posttranslational modification, protein turnover, chaperones") contains a large number of genes that are implicated in QC pathways, it also harbors a number of genes that we considered unlikely to be directly involved in QC/buffering (e.g., ATP-binding cassette-type transporter components). "QC" is therefore a subset of COG functional category O with non-QC genes removed (supplementary table S2, Supplementary Material online).

## Analysis of Substitution Rates

We computed branch-specific synonymous and nonsynonymous substitution rates under a free ratio model using the codeml algorithm implemented in PAML (Yang 2007). The species trees identified above were used as guide trees. Some phylogenetic models may yield misleading substitution rate estimates when the genomes concerned differ markedly in nucleotide composition. All our clades differ by $<1\%$ in genomic GC content with the exception of the mycobacterial clade where compositional differences across taxa are much more pronounced (58% GC for *M. leprae* and 66% for *M. tuberculosis*). To corroborate our conclusions with an evolutionary model accounting for the evolution of nucleotide composition, we also implemented the Galtier and Gouy substitution model (Galtier and Gouy 1995) in baseml (model T92 + GC) and repeated all relative rate analyses as described below. Although, as expected, the codeml model tends to overestimate ancestral branch lengths and underestimate the length of branch *a* (not shown), the conclusion of cross-clade hyperacceleration of QC genes (see below) is strongly supported by the T92 + GC model (supplementary fig. S1, Supplementary Material online).

## Protein Structural Analysis

Compensation at the level of protein structure (in *cis*) might involve residues that are, on average, located more closely together in protein structural space than random pairs of substitutions. In order to test whether substitution patterns are consistent with a greater role for *cis*-compensatory evolution in QC genes, we downloaded 181,944 protein chains from the Protein Data Bank (www.pdb.org) and conducted a protein Blast search against all proteins (from taxa A) in our final data set. Following the approach of Bloom et al. (2006), we aligned all hits with an *E* value $<10^{-5}$ using Muscle (Edgar 2004) and retained alignments with sequence identity $>80\%$, selecting the alignment with maximum sequence identity in case several protein chains matched the same query protein. After downloading the corresponding protein structures from the Protein Data Bank, we determined pairwise distances between residues where an amino acid substitution had occurred along branch *a*. In addition, we computed pairwise distances between all residues in the structure for normalization purposes (see below). For a total of 503 proteins that have the necessary minimum of two mapped substitutions along branch *a* to compute intersubstitution distances, we found a corresponding structure with sufficiently high sequence identity. Only eight of these proteins belonged to COG QC. As COGs vary in average protein size, and therefore in typical interresidue distance, we calculated a *Z*-score for each protein as

$$Z = \frac{D_{\text{sub}} - D_{\text{all}}}{\sigma_{\text{all}} \Big/ \sqrt{n}},$$

where $D_{\text{sub}}$ and $D_{\text{all}}$ are the mean log-transformed distances between residues where substitutions occurred and between

all residues, respectively, and $n$ is the number of residues in the structure. The distributions of $Z$ scores from QC and non-QC genes (pooled across clades) were then compared by means of a two-sided $t$-test.

## Results

### Accelerated Rates of Sequence Evolution in Pathogenic Bacteria

We identified six phylogenetically independent clades of closely related bacterial genomes including one genome that—based on comparative disease ecology, clonality, and genomic features—we expect to have a relatively reduced effective population size (see Materials and Methods, table 1). For each clade, we defined the core genome, inferred its phylogeny (see Materials and Methods), and focused on a four-taxon phylogeny of the type depicted in figure 1, where the taxon with presumed smaller $N_e$ [taxon A($N_{eS}$)] has derived status. In order to unravel function-specific signatures of sequence evolution in response to reduced $N_e$ while obtaining robust rate estimates from the given phylogenies, we concatenated the alignments of protein-coding core genes by their COG functional category (see Materials and Methods). Alongside the predefined COG classes, we included a QC category, which contained chaperones, proteasome components, transfer-messenger RNA machinery, and other genes we considered to function in QC pathways (see Materials and Methods). For each of these COG concatenates, we determined branch-specific dN and dS as well as the dN/dS ratio (see Materials and Methods).

We compared the rates of evolution on the terminal branch leading up to taxon A (branch $a$ in fig. 1) with the rate of the branch leading to its sister taxon (branch $b$) as a rate disparity (RD):

$$\mathrm{RD} = \frac{r_a}{r_a + r_b},$$

where $r_a$ and $r_b$ are the rates on branches $a$ and $b$, respectively. RD provides a measure of relative rate acceleration that is effectively decoupled from the absolute rates of evolution along the two branches as well as from the ancestral rate (supplementary table S3, Supplementary Material online). In other words, genes that evolve more slowly (low absolute rate) show no significant tendency to experience faster accelerations (high RD) under reduced $N_e$. Further, being decoupled from absolute rates, we do not expect RDs to covary with gene-specific features that are often highly predictive of absolute rates of evolution, such as, most notably, expression level (reviewed in Pal et al. 2006). Indeed, although there is a strong correlation between the frequency of optimal codons (Fop), a proxy of gene expression levels, and absolute rates of evolution, RD and Fop do not correlate (supplementary fig. S2, Supplementary Material online). Finally, we found no tendency for COG concatenates with less sequence information to have more extreme values of RD (supplementary fig. S3, Supplementary Material online).

If taxon A has indeed experienced a reduction in $N_e$, RD should be >0.5 (i.e., more than 50% of substitutions after the split of A and B have occurred along branch $a$). This prediction is met in all candidate clades (fig. 2, supplementary fig. S4 and table S4, Supplementary Material online) both for synonymous and nonsynonymous substitutions. We also observe elevated dN/dS ratios. This is consistent with reduced purifying selection in the emerging strain with reduced $N_e$ and suggests that rate accelerations are not due to increased mutation rates (Ohta 1992; Kuo et al. 2009). Accelerations are also evident relative to the ancestral branch $f$ for most rate comparisons (supplementary fig. S5, Supplementary Material online). Rate accelerations along all the lineages with reduced $N_e$ for practically all functional classes suggest a strong global effect of lower effective population sizes on these emerging clonal pathogens.

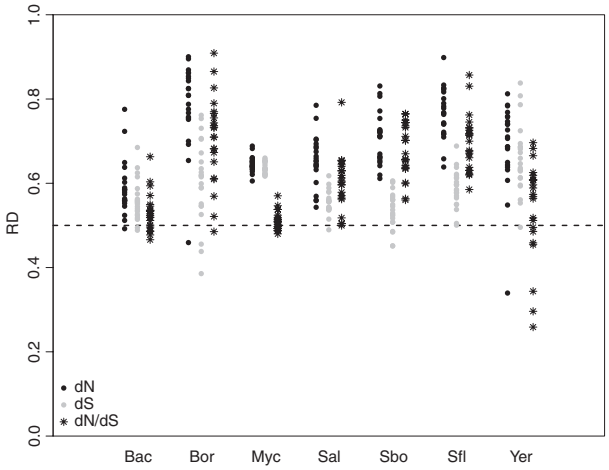


**Fig. 2.** Rate disparities across clades. The taxon with presumed reduced $N_e$ [taxon A($N_{eS}$)] is compared with its sister taxon [taxon B($N_{eL}$)] to determine a RD as defined in the main text for nonsynonymous substitutions (dN, black circles), synonymous substitutions (dS, gray circles), and their ratio (dN/dS, stars). Each data point represents the RD for one COG functional category. The dashed line represents rate equality (RD = 0.5). Bac: *Bacillus*; Bor: *Bordetella*; Myc: *Mycobacterium*; Sal: *Salmonella*; Sbo: *Escherichia* (taxon A: *Shigella boydii*); Sfl: *Escherichia* (taxon A: *Sh. flexneri*); and Yer: *Yersinia*.

Average values of RD vary considerably across clades (fig. 2, supplementary fig. S4, Supplementary Material online). Although we discern no clear-cut relationship to other genomic indicators of reduced $N_e$ (table 1), variability in average RD is likely a function, at least in part, of the timing and intensity of $N_e$ reduction in the focal lineage. Within each clade, RD values for individual COGs (supplementary table S4, Supplementary Material online) will further vary because different functions experience different intensities of purifying and/or positive selection following shifts in virulence (Canback et al. 2004). Finally, although our core genomes are relatively large, these emerging lineages are largely monomorphic, and the density of polymorphisms is therefore very low. This may induce some uncertainty in the inference of individual values of RD for each individual COG in each individual clade, even though collectively they show a strong coherent signal indicating acceleration of evolutionary rates.

**Table 2.** Quality control genes are hyper-accelerated relative to the remainder of the coding genome.

| | RD (dN) | | RD (dS) | |
|---|---|---|---|---|
| Clade | QC | Non-QC | QC | Non-QC |
| *Bacillus* | **0.72** | **0.58** | **0.69** | **0.56** |
| *Bordetella* | **0.87** | **0.80** | **0.73** | **0.60** |
| *Escherichia (Sh. boydii)* | **0.81** | **0.69** | **0.57** | **0.53** |
| *Escherichia (Sh. flexneri)* | **0.90** | **0.75** | **0.60** | **0.58** |
| *Mycobacterium* | **0.66** | **0.64** | **0.63** | **0.64** |
| Salmonella | **0.69** | **0.65** | **0.60** | **0.56** |
| *Yersinia* | **0.64** | **0.66** | **0.84** | **0.65** |

### A Shared Trend for Hyper-Acceleration of QC Genes in Genomes of Lineages with Reduced Effective Population Size

Although variability in relative rates across COGs per se is not unexpected, we were primarily interested to ascertain whether there are COGs that evolve systematically faster or slower than other COGs across clades as that might indicate predictable epistatic responses to reduced $N_e$. In particular, we wished to test if the evolutionary behavior of QC genes is in any way distinct from the other functional classes. When we compare RDs of QC genes with RDs in the remainder of the protein-coding genome, we find that QC genes appear hyper-accelerated along reduced-$N_e$ lineages in the majority of clades studied (table 2), a pattern we do not expect to see under a random model ($P = 0.033$, see-supplementary text, Supplementary Material online). This is in agreement with the hypothesis that QC proteins endure a different selective regime than the average protein when taxa undergo a depression in $N_e$.

But do QC genes really stand out from the crowd? To address this question in a way that allows dissection of the relations between different functions and is alert to multiple testing, we developed a rank-based framework that draws inspiration from Condorcet electoral statistics. The approach is illustrated in figure 3. Briefly, we treat each clade as an individual ballot where COGs are compared in a pairwise fashion. Combination of these ballots across clades reveals shared propensities of one COG to evolve faster/slower than a second COG across multiple phylogenetically independent instances. Figure 4 provides a graphical summary of these shared propensities for different RD comparisons (RD for dN, dS, and dN/dS). On average, both synonymous and nonsynonymous RDs of QC genes are larger relative to other COGs. Note that, interestingly, this is not the case for RD for dN/dS. Typically, inconspicuous dN/dS might have been interpreted as a lack of acceleration in dN relative to dS and therefore maintenance of purifying selection. This is clearly not the case here: both dN and dS are relatively accelerated.

Is this hyper-acceleration relative to other functional categories different from what we would have expected by chance? To determine statistical significance, we computed the marginal total of a particular COG in the cross-clade ballot matrix as an index of its propensity to show greater RD relative to all other COGs in the genome. We then compared the observed marginal totals with randomized marginal totals (see fig. 3). Figure 5 illustrates that—uniquely across COG classes—observed marginal totals for QC are significantly higher than expected for both synonymous and nonsynonymous RD comparisons. In other words, the sequences of QC genes have a significant cross-clade tendency to be hyper-accelerated relative to other functional categories.

To understand the biological significance of these results, we conducted a series of tests. First, hyper-acceleration at synonymous sites might trivially reflect single mutational events that affect both nonsynonymous and neighboring synonymous sites. However, we found neighboring substitutions to be rare (<1% of all substitutions) and not more common along branch *a* compared with branch *b*. Further, the frequency of these events is below average for QC genes. More generally, RD (dN) and RD (dS) are uncorrelated in all but one clade (see supplementary text,Supplementary Material online), further suggesting that these two measures reflect largely independent substitution events. Second, as the chaperonin GroEL was previously shown to experience accelerated rates of nonsynonymous evolution along several pathogenic lineages (Williams et al. 2010), we excluded GroEL from the analysis but found that above results remain virtually unchanged. Finally, we verified that hyper-acceleration of QC genes remains strongly supported when we compute evolutionary rates under a substitution model that controls for base composition (supplementary fig. S1, Supplementary Material online, Materials and Methods).

Note that all other ostensibly significant trends (fig. 5) lose support under this model, with the exception of hypo-acceleration of COG Q genes. We currently lack any convincing explanation as to why COG Q genes ("Secondary metabolites biosynthesis, transport, and catabolism"), typically amongst the fastest-evolving genes in absolute terms, should experience low synonymous RD across clades. In contrast, theory certainly suggests that QC loci might exhibit altered evolutionary dynamics under reduced $N_e$ (Gros and Tenaillon 2009). Intuitively, however, one might suspect that these loci interact with accumulating slightly deleterious mutations via negative epistasis—that is, deleterious mutation arising in a QC locus would exacerbate effects of deleterious mutation in the substrate. One might therefore expect purifying selection at QC loci to intensify. Clearly, this simple prediction is not supported by the data. Instead, our data suggests that QC genes typically experience accelerated evolution.

## Discussion

Population genetic theory predicts accelerated rates of nucleotide substitutions following a reduction in effective population size. This prediction has previously been upheld in a number of endosymbiont and island taxa. Here, we demonstrate on a genome-wide basis that rate accelerations can also be observed, over much shorter evolutionary distances, in several clades of pathogenic bacteria. This

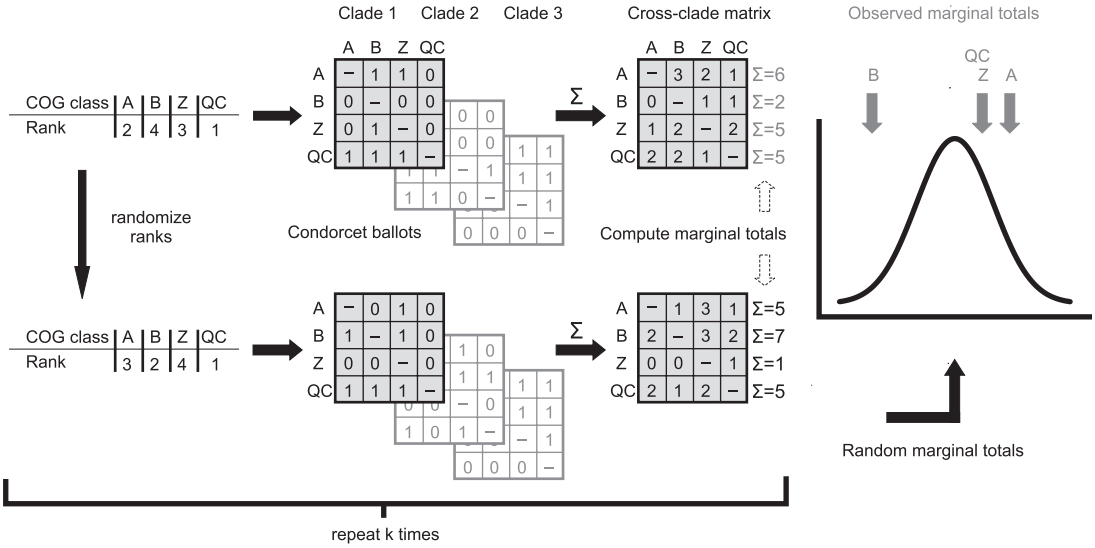


**Fig. 3.** Condorcet analysis to detect function-specific responses to reduced $N_e$ shared across clades. Step 1: Concatenated COG alignments are ranked within each clade according to their rate disparity (RD, highest RD ranks first). Step 2: Rankings are translated into a Condorcet ballot matrix where COGs are subjected to pairwise comparisons. For each pair *x*, *y*, where *x* is the COG by row and *y* is the COG by column, if rank(*x*) $<$ rank(*y*), that cell is assigned 1, if rank(*x*) $>$ rank(*y*), the cell is assigned 0. The COG with the largest RD (QC in this example) therefore scores exclusively 1s. Step 3: Condorcet ballots are added via matrix addition to yield a cross-clade matrix. Step 4: As an overall measure of the cross-clade tendency of a COG to have higher/lower RD than all other COGs, compute the marginal row total for each COG. Step 5: Randomize ranks across COGs and repeat steps 2–4 $k = 10{,}000$ times for each COG to obtain a distribution of marginal totals. Step 6: Compare the observed marginal totals with the distribution obtained through randomization.

follows previous single-clade analyses pointing to lower purifying selection in *Shigella* (Balbi et al. 2009), ongoing genome degradation in *Yersinia* (Achtman et al. 1999) and *Bordetella* (Parkhill et al. 2003), extensive gene decay in *M. leprae* (Cole et al. 1998; Gomez-Valero et al. 2007), and sexual isolation in *Sa. typhi* (Holt et al. 2008).

Reduced $N_e$ is arguably the most parsimonious explanation for the global trend toward elevated RD, but other factors might contribute as well. First, average generation times might be systematically shorter for taxon A so that rate accelerations may be caused by a larger number of generations on branch *a*. This seems unlikely. As far as we know, there is no evidence that increased pathogenicity and/or host specificity is commonly associated with faster growth rates. In fact, there is some evidence that enhanced pathogenicity is associated with increased minimal generation times, for example, in the reduced-$N_e$ taxa *M. leprae* (240 h) and *Y. pestis* (1.25 h) relative to their sister taxa *M. tuberculosis* (19 h) and *Y. pseudotuberculosis* (30 min), respectively (Vieira-Silva and Rocha 2010).

Second, selection to evade immune responses might lead to positive selection in surface-exposed proteins. Yet, these proteins are a small fraction of the proteome and tend to be excluded from the core genomes. For example, in *E. coli*, these genes correspond to less than 1% of the core genome (Nogueira et al. 2009). Furthermore, most

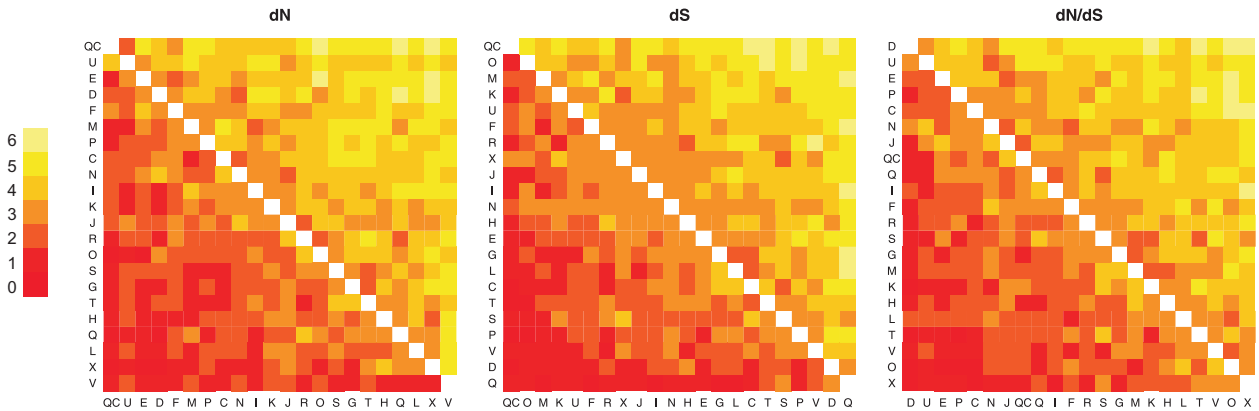


**Fig. 4.** Visualization of cross-clade Condorcet matrices for different rate classes (RD for dN, dS, and dN/dS). Any COG *x* (row) can score a maximum of 6 in each pairwise comparison with any other COG *y* (column) if it outranks (i.e., has higher RD than) COG *y* in all 6 clades. For example, looking at RD for synonymous rates (dS) COG QC scores 6 in comparison with COG T, which means that QC has a higher synonymous RD than T in all clades considered. Matrices are ordered by the marginal row totals.

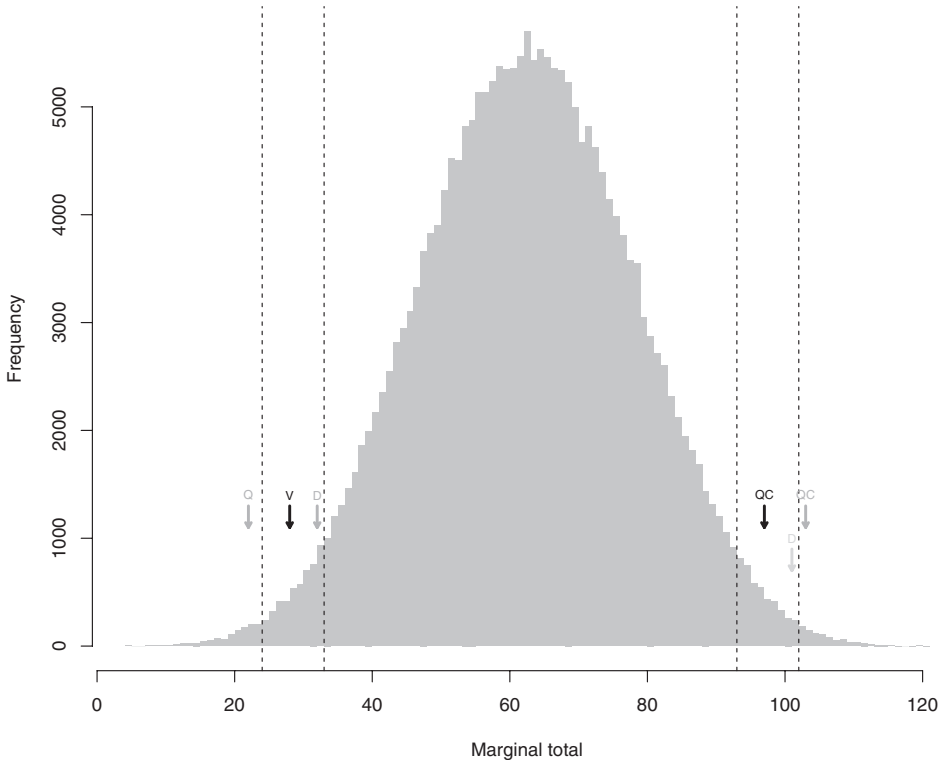


**Fig. 5.** Cross-clade tendency for high or low COG-specific RDs. The random distribution of marginal totals (see fig. 3) is compared with observed marginal totals. Dashed lines represent the two-tailed 5% and 1% significance thresholds Bonferroni-corrected for multiple testing. Arrows indicate COGs with significant marginal totals regardless of which Shigella is dropped from the analysis. Arrows are shaded by rate class (dN—black; dS—grey; and dN/dS—light grey, lower plane). QC has a significant cross-clade tendency to have an exaggerated RD relative to other COGs at both the nonsynonymous and synonymous level.

sister clades of the focal species with low $N_e$ (i.e., taxa B) in this analysis are themselves pathogens and thus subject to selection imposed by the immune system. Finally, we find that COGs with exclusively intracellular protein localizations (e.g., the translation-associated COG J) also endure accelerated sequence evolution.

Third, our candidate genomes may have lost some capacity for DNA damage repair, either because repair enzymes have been lost altogether or because they have accumulated mutations that affect adequate functioning or expression. There is certainly evidence that repair capacity is frequently lost or diminished in genomes undergoing reductive evolution (Dale et al. 2003; Silva et al. 2003). We acknowledge that this might be a contributory factor to accelerated rates, particularly in the mycobacterial clade where *M. leprae* differs markedly in nucleotide composition and is known to have reduced repair functions (Dawes and Mizrahi 2001). However, such differences in repair repertoires were found neither in *Shigella* (Balbi et al. 2009) nor in *Y. pestis* or *Ba. anthracis* (Hershberg and Petrov 2010). Further, repair biases should affect genes similarly regardless of gene function and are therefore unlikely to explain function-specific evolutionary trends. Finally, loss of repair capacity is thought to be a secondary effect of reduced $N_e$ rather than its cause (if $N_e$ is high, the deleterious inactivation of the repair machinery is efficiently purged). Overall, lower $N_e$ seems the most parsimonious explanation for the commonalities found in this work in terms of the global accelerations of evolutionary rates.

## Hyper-Acceleration of Quality Control Genes

Can reduced $N_e$ also explain why QC genes show greater rate disparities than other COGs across clades? There are several radically different models that might explain these observations. First, accelerations in substitution rates may be owing to relaxed selection. This deceptively simple explanation poses more problems than it solves. The QC class contains elements, such as chaperones, that are nearly ubiquitous among prokaryotes, highly expressed, and highly conserved in sequence. Classes with similar characteristics, such as COG J (translation), do not exhibit RD values different from the average genome. Moreover, there is no obvious reason why selection on QC genes should relax above and beyond what happens in other functional categories. On the contrary, theory suggests that QC machinery should become more critical to viability in populations with reduced $N_e$ (Krakauer and Plotkin 2002; Gros and Tenaillon 2009). This prediction seems to be born out in the upregulated expression of chaperones in a number of endosymbiotic taxa (e.g., Baumann et al. 1996) and in the effective buffering of deleterious mutations in cells engineered to overproduce GroEL (Fares et al. 2002).

Alternatively, differences in relative rates between functional categories might reflect adaptations to a new host/

niche (Canback et al. 2004; Toft and Fares 2009; Toft et al. 2009). As we observe shared responses along a number of independent branches, however, we would have to evoke a remarkable case of parallel evolution for bacteria that occupy very different niches; in addition, although our focal taxa are pathogenic, the same is often already true for their respective ancestors. In short, we do not think that either relaxed purifying or pervasive positive selection provide an adequate explanation for the function-specific variation in rate disparities.

We suggest instead that the patterns of molecular evolution we observe for QC genes are most consistent with a third alternative: compensatory evolution. A number of studies, notably on antibiotic resistance (Reynolds 2000; Maisnier-Patin et al. 2002; Trindade et al. 2009) in bacterial model systems, have demonstrated that the fitness costs of deleterious mutation are frequently ameliorated by subsequent mutations. Such compensatory changes can occur in *cis* (e.g., restabilizing a previously destabilized protein structure) or in *trans*, that is, in genes other than the one with the original deleterious mutation (Reynolds 2000). Intriguingly, compensatory evolution is expected to assume greater prominence under reduced $N_e$. Although compensatory mutations are a priori more likely to arise than true revertants because the mutational target is larger, under small $N_e$ chances are higher that a compensatory mutation of intermediate effect fixes before a large-effect revertant arises (Levin et al. 2000). In principle, this qualitative change in evolutionary dynamics should apply to all genes regardless of their functional role in the cell. However, QC genes can mitigate fitness effects of a large number of substrates in *trans* and may therefore constitute a natural playground for compensatory change. We therefore suggest that the excess acceleration we observe might primarily reflect compensatory changes *in trans* accompanying the accumulation of slightly deleterious mutations in genomes of lineages with reduced $N_e$.

Validating this hypothesis, especially as applied to genome-wide trends, is difficult in practice. Critically, diagnosing compensatory evolution requires knowledge of the fitness effects of the substitutions involved. Establishing, from sequence data alone, that an initial mutation is deleterious, whereas the subsequent mutation alleviates the deleterious effect has therefore only been possible where deleterious effect and amelioration are directly observable, for example, for human disease mutations or experimental microbial assays (see above), or where they can be strongly inferred as in the case of de- and restabilizing substitutions in transfer RNA stems (Meer et al. 2010).

As we lack residue-specific fitness data as well as strong expectations about the likely fitness effects of individual substitutions in our data, we cannot test our hypothesis of compensation in *trans* directly. However, we can rule out alternative explanations. For example, compensatory changes may occur in *cis* from selection for protein stability (Tokuriki and Tawfik 2009b). In this case, de- and restabilizing substitution pairs should cluster more closely together in protein structural space than random pairs of substitutions. We therefore compared the relative proximity of substitutions in QC versus non-QC genes in the corresponding protein structures. We did not detect a significant difference between the two protein classes ($t$-test $P = 0.49$; see Materials and Methods). Although this test is rather indirect, it nonetheless suggests that compensation in *cis* for protein stability does not explain our results. May compensation occur in *cis* but simply not in relation to protein stability?

We showed above that hyper-acceleration in QC genes takes place both at synonymous and nonsynonymous sites. We have ruled out that this might occur because of multiple contiguous substitutions, and we found no effect of absolute expression levels on relative acceleration rates that might justify this correlation. Recent evidence shows that changes in synonymous sites can strongly impact fitness (Kudla et al. 2009; Lind et al. 2010) so that it is certainly conceivable that compensation takes place, for example, at the level of mRNA structure or in the context of protein-DNA/RNA interactions (Kenigsberg et al. 2010). We currently see no convincing test to directly implicate these alternative terrains of compensatory evolution in our data, principally because our understanding of how nonsynonymous substitutions affect fitness in *trans* and how synonymous substitutions affect fitness in general remains extremely fragmentary. Therefore, although compensatory evolution remains a strong candidate to explain hyper-acceleration of QC genes, we acknowledge that further testing will be required to corroborate or dispel this hypothesis. Finally, we note that, if QC hyper-acceleration is, as we suggest, an epistatic effect independent of specific ecological scenarios, patterns of molecular evolution in QC genes might serve—given sufficient temporal and genomic resolution—as indicators of changes in effective population size.

## Supplementary Material

Supplementary tables S1–S4 and figures S1–S5 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

## Acknowledgments

We would like to thank Peter Vlasov for assistance with protein structural analysis, Laurence Hurst and Fyodor Kondrashov for valuable discussions, and Sara Vieira-Silva for comments on the manuscript. This work was supported by a Medical Research Council Capacity Building Studentship and a European Molecular Biology Organisation Long-term Fellowship to T.W. and funding from the Centre National de la Recherche Scientifique and the Institut Pasteur to E.P.C.R.

## References

Achtman M, Zurth K, Morelli G, Torrea G, Guiyoule A, Carniel E. 1999. Yersinia pestis, the cause of plague, is a recently emerged clone of Yersinia pseudotuberculosis. *Proc Natl Acad Sci U S A.* 96:14043–14048.

Balbi KJ, Rocha EP, Feil EJ. 2009. The temporal dynamics of slightly deleterious mutations in Escherichia coli and Shigella spp. *Mol Biol Evol*. 26:345–355.

Baumann P, Baumann L, Clark MA. 1996. Levels of *Buchnera aphidicola* Chaperonin GroEL during growth of the aphid *Schizaphis graminum*. *Curr Microbiol*. 32:279–285.

Beltran P, Musser JM, Helmuth R, et al. (11 co-authors). 1988. Toward a population genetic analysis of Salmonella: genetic diversity and relationships among strains of serotypes S. choleraesuis, S. derby, S. dublin, S. enteritidis, S. heidelberg, S. infantis, S. newport, and S. typhimurium. *Proc Natl Acad Sci U S A*. 85:7753–7757.

Bloom JD, Drummond DA, Arnold FA, Wilke CO. 2006. Structural determinants of the rate of protein evolution in yeast. *Mol Biol Evol*. 23:1751–1761.

Britton WJ, Lockwood DN. 2004. Leprosy. *Lancet*. 363:1209–1219.

Canback B, Tamas I, Andersson SG. 2004. A phylogenomic study of endosymbiotic bacteria. *Mol Biol Evol*. 21:1110–1122.

Chain PS, Carniel E, Larimer FW, et al. (23 co-authors). 2004. Insights into the evolution of Yersinia pestis through whole-genome comparison with Yersinia pseudotuberculosis. *Proc Natl Acad Sci U S A*. 101:13826–13831.

Charlesworth B. 2009. Fundamental concepts in genetics: effective population size and patterns of molecular evolution and variation. *Nat Rev Genet*. 10:195–205.

Cole ST, Brosch R, Parkhill J, et al. (42 co-authors). 1998. Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence. *Nature* 393:537–544.

Cole ST, Eiglmeier K, Parkhill J, et al. (44 co-authors). 2001. Massive gene decay in the leprosy bacillus. *Nature* 409:1007–1011.

Dale C, Wang B, Moran N, Ochman H. 2003. Loss of DNA recombinational repair enzymes in the initial stages of genome degeneration. *Mol Biol Evol*. 20:1188–1194.

Dawes SS, Mizrahi V. (44 co-authors). 2001. DNA metabolism in Mycobacterium leprae. *Lepr Rev*. 72:408–414.

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

Erickson BW, Sellers PH. 1983. Recognition of patterns in genetic sequences. In: David Sankoff, Joseph B. Kruskal, editors. Time warps, string edits, and macromolecules: the theory and practice of sequence comparison. Reading (MA): Addison-Wesley. p. 55–91.

Fares MA, Ruiz-González MX, Moya A, Elena SF, Barrio E. 2002. Endosymbiotic bacteria: groEL buffers against deleterious mutations. *Nature* 417:398.

Galtier N, Gouy M. 1995. Inferring phylogenies from DNA sequences of unequal base compositions. *Proc Natl Acad Sci U S A*. 92:11317–11321.

Gascuel O. 1997. BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. *Mol Biol Evol*. 14:685–695.

Gomez-Valero L, Rocha EP, Latorre A, Silva FJ. 2007. Reconstructing the ancestor of Mycobacterium leprae: the dynamics of gene loss and genome reduction. *Genome Res*. 17:1178–1185.

Gros PA, Tenaillon O. 2009. Selection for chaperone-like mediated genetic robustness at low mutation rate: impact of drift, epistasis and complexity. *Genetics* 182:555–564.

Helgason E, Tourasse NJ, Meisal R, Caugant DA, Kolsto AB. 2004. Multilocus sequence typing scheme for bacteria of the Bacillus cereus group. *Appl Environ Microbiol*. 70:191–201.

Hershberg R, Petrov DA. 2010. Evidence that mutation is universally biased towards AT in bacteria. *PLoS Genet* 6(9):e1001115.

Hershberg R, Tang H, Petrov DA. 2007. Reduced selection leads to accelerated gene loss in Shigella. *Genome Biol*. 8:R164.

Holt KE, Parkhill J, Mazzoni CJ, et al. (13 co-authors). 2008. High-throughput sequencing provides insights into genome variation and evolution in Salmonella Typhi. *Nat Genet*. 40:987–993.

Jin Q, Yuan Z, Xu J, et al. (33 co-authors). 2002. Genome sequence of Shigella flexneri 2a: insights into pathogenicity through comparison with genomes of Escherichia coli K12 and O157. *Nucleic Acids Res*. 30:4432–4441.

Johnson KP, Seger J. 2001. Elevated rates of nonsynonymous substitution in island birds. *Mol Biol Evol*. 18:874–881.

Kaper JB, Nataro JP, Mobley HL. 2004. Pathogenic Escherichia coli. *Nat Rev Microbiol*. 2:123–140.

Kenigsberg E, Bar A, Segal E, Tanay A. 2010. Widespread compensatory evolution conserves DNA-encoded nucleosome organization in yeast. *PLoS Comput Biol*. 6:e1001039.

Krakauer DC, Plotkin JB. 2002. Redundancy, antiredundancy, and the robustness of genomes. *Proc Natl Acad Sci U S A*. 99:1405–1409.

Kudla G, Murray AW, Tollervey D, Plotkin JB. 2009. Coding sequence determinants of gene expression in *Escherichia coli*. *Science* 324:255–258.

Kuo CH, Moran NA, Ochman H. 2009. The consequences of genetic drift for bacterial genome complexity. *Genome Res*. 19:1450–1454.

Legrand D, Tenaillon MI, Matyot P, Gerlach J, Lachaise D, Cariou ML. 2009. Species-wide genetic variation and demographic history of Drosophila sechellia, a species lacking population structure. *Genetics* 182:1197–1206.

Levin BR, Perrot V, Walker N. 2000. Compensatory mutations, antibiotic resistance and the population genetics of adaptive evolution in bacteria. *Genetics* 154:985–997.

Lind PA, Berg OG, Andersson DI. 2010. Mutational robustness of ribosomal protein genes. *Science* 330:825–827.

Maisnier-Patin S, Berg OG, Liljas L, Andersson DI. 2002. Compensatory adaptation to the deleterious effect of antibiotic resistance in Salmonella typhimurium. *Mol Microbiol*. 46:355–366.

Maquat LE, Carmichael GG. 2001. Quality control of mRNA function. *Cell* 104:173–176.

Meer MV, Kondrashov AS, Artzy-Randrup Y, Kondrashov FA. 2010. Compensatory evolution in mitochondrial tRNAs navigates valleys of low fitness. *Nature*. 464:279–282.

Mock M, Fouet A. 2001. Anthrax. *Annu Rev Microbiol*. 55:647–671.

Moran NA. 1996. Accelerated evolution and Muller's rachet in endosymbiotic bacteria. *Proc Natl Acad Sci U S A*. 93:2873–2878.

Nogueira T, Rankin DJ, Touchon M, Taddei F, Brown SP, Rocha EP. 2009. Horizontal gene transfer of the secretome drives the evolution of bacterial cooperation and virulence. *Curr Biol*. 19:1683–1691.

Ohta T. 1992. The nearly neutral theory of molecular evolution. *Annu Rev Ecol Syst*. 23:263–286.

Pal C, Papp B, Lercher MJ. 2006. An integrated view of protein evolution. *Nat Rev Genet*. 7:337–348.

Parkhill J, Sebaihia M, Preston A, et al. (53 co-authors). 2003. Comparative analysis of the genome sequences of Bordetella pertussis, Bordetella parapertussis and Bordetella bronchiseptica. *Nat Genet*. 35:32–40.

Read TD, Peterson SN, Tourasse N, et al. (52 co-authors). 2003. The genome sequence of Bacillus anthracis Ames and comparison to closely related bacteria. *Nature* 423:81–86.

Reynolds MG. 2000. Compensatory evolution in rifampin-resistant Escherichia coli. *Genetics* 156:1471–1481.

Roumagnac P, Weill FX, Dolecek C, et al. (11 co-authors). 2006. Evolutionary history of Salmonella typhi. *Science*. 314:1301–1304.

Rutherford SL, Lindquist S. 1998. Hsp90 as a capacitor for morphological evolution. *Nature* 396:336–342.

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

Silander OK, Tenaillon O, Chao L. 2007. Understanding the evolutionary fate of finite populations: the dynamics of mutational effects. *PLoS Biol*. 5:e94.

Silva FJ, Latorre A, Moya A. 2003. Why are the genomes of endosymbiotic bacteria so stable? *Trends Genet*. 19:176–180.

Stinear TP, Seemann T, Pidot S, et al. (23 co-authors). 2007. Reductive evolution and niche adaptation inferred from the genome of Mycobacterium ulcerans, the causative agent of Buruli ulcer. *Genome Res*. 17:192–200.

Tatusov RL, Galperin MY, Natale DA, Koonin EV. 2000. The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res*. 28:33–36.

Toft C, Fares MA. 2009. Selection for translational robustness in *Buchnera aphidicola*, endosymbiotic bacteria of aphids. *Mol Biol Evol*. 26:743–751.

Toft C, Williams TA, Fares MA. 2009. Genome-wide functional divergence after the symbiosis of proteobacteria with insects unraveled through a novel computational approach. *PLoS Comput Biol*. 5:e1000344.

Tokuriki N, Tawfik DS. 2009a. Chaperonin overexpression promotes genetic variation and enzyme evolution. *Nature* 459:668–673.

Tokuriki N, Tawfik DS. 2009b. Stability effects of mutations and protein evolvability. *Curr Opin Struct Biol*. 19:596–604.

Touchon M, Rocha EP. 2007. Causes of insertion sequences abundance in prokaryotic genomes. *Mol Biol Evol*. 24:969–981.

Trindade S, Sousa A, Xavier KB, Dionisio F, Ferreira MG, Gordo I. 2009. Positive epistasis drives the acquisition of multidrug resistance. *PLoS Genet*. 5:e1000578.

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

Wernegreen JJ, Moran NA. 1999. Evidence for genetic drift in endosymbionts (Buchnera): analyses of protein-coding genes. *Mol Biol Evol*. 16:83–97.

Williams TA, Codoner FM, Toft C, Fares MA. 2010. Two chaperonin systems in bacterial genomes with distinct ecological roles. *Trends Genet*. 26:47–51.

Woolfit M. 2009. Effective population size and the rate and pattern of nucleotide substitutions. *Biol Lett*. 5:417–420.

Woolfit M, Bromham L. 2003. Increased rates of sequence evolution in endosymbiotic bacteria and fungi with small effective population sizes. *Mol Biol Evol*. 20:1545–1555.

Woolfit M, Bromham L. 2005. Population size and molecular evolution on islands. *Proc R Soc B Biol Sci*. 272:2277–2282.

Wright S. 1931. Evolution in Mendelian Populations. *Genetics* 16:97–159.

Yang F, Yang J, Zhang X, et al. (27 co-authors). 2005. Genome dynamics and diversity of Shigella species, the etiologic agents of bacillary dysentery. *Nucleic Acids Res*. 33:6445–6458.

Yang Z. 2007. PAML 4: a program package for phylogenetic analysis by maximum likelihood. *Mol Biol Evol*. 24:1586–1591.

Yuk MH, Heininger U, Martinez de Tejada G, Miller JF. 1998. Human but not ovine isolates of Bordetella parapertussis are highly clonal as determined by PCR-based RAPD fingerprinting *Infection*. 26:270–273.