# Transfer of noncoding DNA drives regulatory rewiring in bacteria

Yaara Oren[a], Mark B. Smith[b], Nathan I. Johns[c], Millie Kaplan Zeevi[a], Dvora Biran[d], Eliora Z. Ron[d,e], Jukka Corander[f], Harris H. Wang[c], Eric J. Alm[g,1], and Tal Pupko[a,1]

Departments of [a]Cell Research and Immunology and [d]Microbiology and Biotechnology, George S. Wise Faculty of Life Sciences, Tel Aviv University, Tel Aviv 69978, Israel; [b]Microbiology Graduate Program and [g]Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139; [c]Department of Systems Biology, Columbia University Medical Center, New York, NY 10032; [e]MIGAL, Galilee Research Institute, Kiryat Shmona 11016, Israel; and [f]Department of Mathematics and Statistics, University of Helsinki, Helsinki, FIN-00014, Finland

Edited by Richard E. Lenski, Michigan State University, East Lansing, MI, and approved September 15, 2014 (received for review July 15, 2014)

**Understanding the mechanisms that generate variation is a common pursuit unifying the life sciences. Bacteria represent an especially striking puzzle, because closely related strains possess radically different metabolic and ecological capabilities. Differences in protein repertoire arising from gene transfer are currently considered the primary mechanism underlying phenotypic plasticity in bacteria. Although bacterial coding plasticity has been extensively studied in previous decades, little is known about the role that regulatory plasticity plays in bacterial evolution. Here, we show that bacterial genes can rapidly shift between multiple regulatory modes by acquiring functionally divergent nonhomologous promoter regions. Through analysis of 270,000 regulatory regions across 247 genomes, we demonstrate that regulatory "switching" to nonhomologous alternatives is ubiquitous, occurring across the bacterial domain. Using comparative transcriptomics, we show that at least 16% of the expression divergence between *Escherichia coli* strains can be explained by this regulatory switching. Further, using an oligonucleotide regulatory library, we establish that switching affects bacterial promoter architecture. We provide evidence that regulatory switching can occur through horizontal regulatory transfer, which allows regulatory regions to move across strains, and even genera, independently from the genes they regulate. Finally, by experimentally characterizing the fitness effect of a regulatory transfer on a pathogenic *E. coli* strain, we demonstrate that regulatory switching elicits important phenotypic consequences. Taken together, our findings expose previously unappreciated regulatory plasticity in bacteria and provide a gateway for understanding bacterial phenotypic variation and adaptation.**

bacterial evolution | regulatory evolution | HRT | core genes

The acquisition of genes from nonparental lineages through horizontal gene transfer (HGT) has been shown to transform bacterial capabilities radically, influencing key processes, including pathogenicity, antibiotic resistance, and utilization of novel energy substrates (1–4). These striking findings have led many to believe that changes in gene content underlie the rapid pace of bacterial evolution (5, 6). However, an overlooked corollary of this ubiquitous exchange of DNA (7, 8) is that noncoding regions can be similarly subject to transfer and recombination, enabling rapid rewiring of regulatory networks (9, 10). Consistent with this hypothesis, recent studies have uncovered several cases of regulatory rearrangements, whereby regulatory regions have "switched" to nonhomologous alternatives with remarkable phenotypic consequences (11–13). For example, the inversion of a single promoter was shown to convert a commensal to a pathogen (12). Similarly, in *Escherichia coli*, citrate utilization was shown to evolve through promoter capture, enabling expression of an otherwise silent transporter (13). These discoveries demonstrate that regulatory "switching" to divergent alternative sequences is possible and can produce functional transformations. Nonetheless, it remains unclear whether these intriguing observations reflect exceptional anecdotes restricted to highly mobile genes in unusual strains or early representatives of a broader paradigm.

## Results

**Regulatory Switching in *E. coli* Core Genes.** To assess the significance of regulatory switching on bacterial evolution, we first considered core genes, which are present in all members of a clade and typically encode basal cellular "housekeeping" functions. Core genes are subject to strong purifying selection and are viewed as islands of stability within the dynamic bacterial genome [although exceptions exist (14, 15)]. Accordingly, regulatory switching in core genes is particularly unexpected, and is also easily detectable against the background of sequence conservation in coding regions.

We compared multiple sequence alignments of the 1,479 core genes present in all 46 publicly available *E. coli* genomes and up to 300 base pairs of the upstream regulatory region for each gene. As expected, the regulatory regions of most core genes are highly conserved (median nucleotide identity of 94%); however, a significant minority (13%) appear to be nonhomologous, sharing less than 50% nucleotide identity (Fig. S1). Because such poor conservation is inconsistent with the traditional view that core genes are slow-evolving (5), we investigated this divergent subpopulation further.

We first focused on *hemH*, as a representative of the nonhomologous upstream regulatory regions (Fig. 1*A*). *hemH* is a single gene operon that encodes ferrochelatase, the terminal enzyme in heme biosynthesis. *hemH* and its upstream gene, *adk*, display

### Significance

The rapid pace of evolution in bacteria is widely attributed to the promiscuous horizontal transfer and recombination of protein-coding genes. However, it has not been investigated if the same forces also drive the evolution of noncoding regulatory regions. Here, we establish that regulatory regions can "switch" between nonhomologous alternatives and that switching is ubiquitous, occurring across the bacterial domain. We show that regulatory switching has a strong impact on promoter architecture and expression divergence. Further, we demonstrate that regulatory transfer facilitates rapid phenotypic diversification of a human pathogen. This regulatory mobility enables bacterial genes to access a vast pool of potential regulatory elements, facilitating efficient exploration of the regulatory landscape.

Author contributions: Y.O., D.B., E.Z.R., H.H.W., E.J.A., and T.P. designed research; Y.O., M.B.S., N.I.J., M.K.Z., D.B., and J.C. performed research; Y.O., M.B.S., J.C., E.J.A., and T.P. analyzed data; and Y.O., M.B.S., E.J.A., and T.P. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Data deposition: The sequences reported in this paper have been deposited in the Gene Expression Omnibus (GEO) database, www.ncbi.nlm.nih.gov/geo (accession no. GSE59468).

[1]To whom correspondence may be addressed. Email: ejalm@mit.edu or talp@post.tau.ac.il.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1413272111/-/DCSupplemental.

www.pnas.org/cgi/doi/10.1073/pnas.1413272111

near-perfect conservation (>98% amino acid identity) across all 46 *E. coli* strains. However, the regulatory region between these genes comprises a 155-bp region that can be classified into two distinct, nonhomologous sequence types (less than 42% average pairwise nucleotide identity between clusters). In contrast, within clusters, there is almost perfect homology (>96% nucleotide identity). Thus, *hemH* represents a canonical example of regulatory switching between two alternative, nonhomologous regulatory sequences.

To determine the overall prevalence of such switching among *E. coli* core genes, we devised an algorithm that could systematically identify core genes with at least two distinct types of regulatory sequences (*SI Text* and Fig. S2). Remarkably, we found 166 unambiguous cases of regulatory switching (11% of all core genes in *E. coli*). The vast majority (83%) of these divergent regions contain bona fide promoters (16), as opposed to interoperonic regions, which is significantly more than expected by chance (Fisher's exact test, $P < 0.005$), indicating that switching is enriched among promoters, where it can facilitate regulatory rewiring.

Moreover, we found that regulatory switching often creates new transcription factor binding sites. In 41% of the 44 diverged core genes for which high-quality transcription factor binding site annotations exist (17), alternative regulatory types were associated with divergent binding patterns (Table S1). For example, in *hemH* (Fig. 1*A*), all type 1 sequences contain an experimentally validated OxyR binding site (18) that is missing from all type 2 sequences. Type 2 sequences, instead, harbor canonical binding sites for both ArgP and DnaA (Fig. 1).

**Horizontal Regulatory Transfer as a Switching Mechanism.** To elucidate the evolutionary mechanisms that lead to regulatory switching, we returned to our representative example of *hemH* and mapped its regulatory regions onto the *E. coli* species tree (Fig. S3; generated by concatenation of all core genes). We found that the distribution of the alternative promoter types is incongruent with the *E. coli* species phylogeny, consistent with evolution by horizontal regulatory transfer (HRT) (Fig. 1*B*). For the observed distribution to be explained by vertical transmission, multiple independent genomic rearrangement events with identical boundaries would have to be posited, with independent acquisition of the identical SNPs shared within each regulatory type; clearly, this alternative interpretation is implausible.

To determine if horizontal transfer has an impact on other regulatory regions in *E. coli*, we used the approximately unbiased (AU) test, a maximum-likelihood–based methodology (19). Specifically, we statistically tested for incongruence between the topology of the promoter sequences against the species tree. The null hypothesis of this test is vertical inheritance (as defined by the species tree); therefore, rejection of the null hypothesis is a strong indication of HRT. We found that 51% of all core gene promoters are incongruent with the species phylogeny, indicating that regulatory regions, similar to coding genes, are frequently transferred. However, in many of these cases, the promoter and its upstream gene might have been cotransferred. To tease out the cases in which the promoters were transferred independent of their genes, we compared the topology of each core gene with the topology of

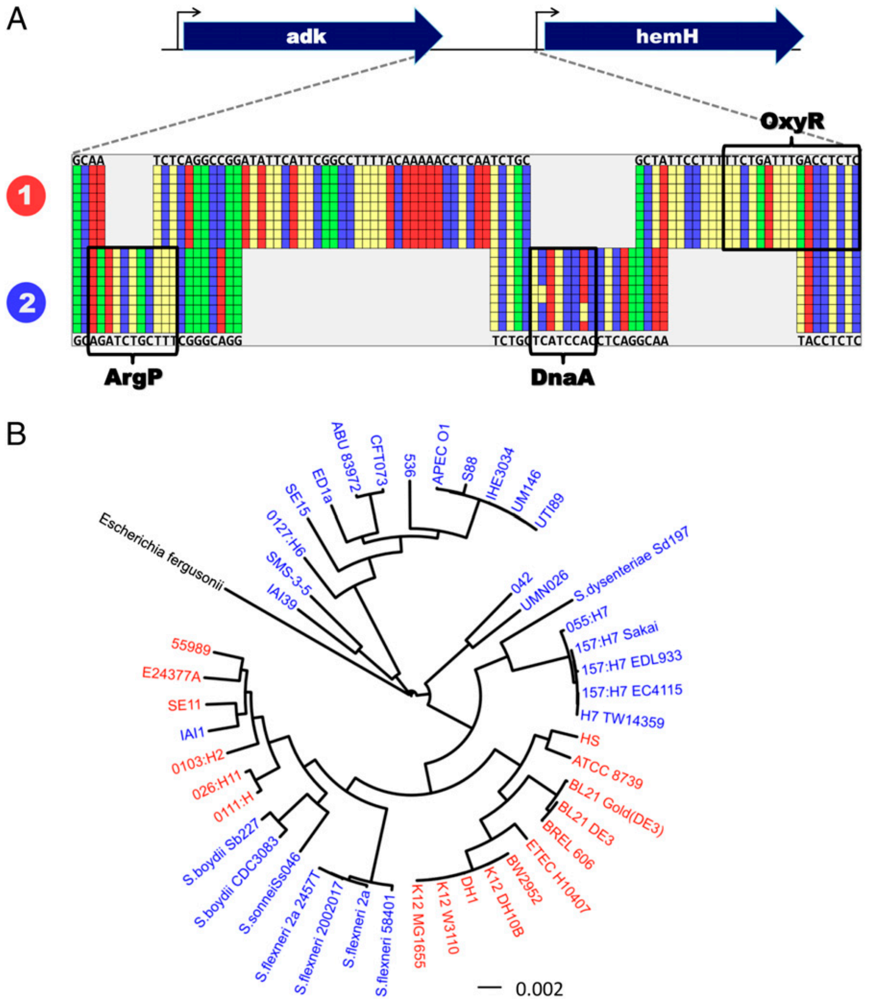

**Fig. 1.** Regulatory switching and horizontal transfer of the *hemH* promoter. (*A*) Operonic structure and a representative multiple sequence alignment of the regulatory region of *hemH* (20 of 46 sequences are shown). The first line (*E. coli* K-12 MG1655) and the last line (*E. coli* O157:H7 Sakai) depict the nucleotide sequence of representative strains from each sequence type. Gray boxes represent gaps in the alignment. The nucleotides are colored red (Ade), yellow (Thy), green (Gua), and blue (Cys). Binding sites of ArgP, DnaA, and OxyR are boxed. The numbered labels in the left margin indicate the two alternative regulatory types that are found in *hemH*. (*B*) Two types mapped onto the *E. coli* species tree rooted with *Escherichia fergusonii*. Promoter type 1 is shown in red, and type 2 is shown in blue. (Scale bar: nucleotide substitutions per site.) The patchy distribution of these alternative sequence types is inconsistent with vertical transmission.

its associated upstream regulatory region (using the same methodology described above, with more details provided in *SI Text*). In the case of *hemH*, both the promoter history and the gene history significantly differed from the species tree, yet their topologies were not statistically different from each other. Therefore, we cannot exclude the possibility that these two regions were cotransferred. Nevertheless, for 32% of all promoters, we detected a clear signal that they were transferred independent of the gene they regulate.

**Intergenera HRT Between *E. coli* and *Enterobacter*.** Given the frequency of HRT among *E. coli* strains, we expanded our analysis to investigate if HRT can cross species boundaries and discovered intergenera HRT between *E. coli* and *Enterobacter*. As shown in Fig. 2, we found that among 22 *E. coli* strains, the leader sequence of the biosynthesis gene *metE* exhibits a greater sequence similarity to the leader sequence found in *Enterobacter* than to its homologs in more closely related *E. coli* strains. Although most *E. coli* have a long leader sequence (169 bp), a subset of *E. coli* (most of which are uropathogenic *E. coli*) has, instead, a short (49 bp) AT-rich leader sequence that is shared with *Enterobacter*. In contrast to this incongruent regulatory region, phylogenies of the surrounding core genes match the species phylogeny, suggesting that the incongruence of the intervening regulatory sequence is best explained by horizontal transfer of the regulatory region alone (Fig. 2*A*). The direction of this regulatory transfer is most likely from *Enterobacter* to *E. coli*, because other Enterobacteriaceae species close to *E. coli* all harbor the long allele (Fig. 2*B*). Furthermore, all of the short *E. coli* regulatory alleles are nearly identical, suggesting a recent regulatory transfer.

**Regulatory Switching Is Also Prevalent in the Accessory Genome.** Thus far, our analysis focused on core genes, for which regulatory switching was especially unexpected. Next, we examined the prevalence of regulatory switching among all gene classes. Among 2,286 noncore accessory genes in *E. coli* strain MG1655, we detected a similar level of switching (11.8%) to that observed across core genes in *E. coli* (11.2%). Moreover, we found that switching occurs across all functional categories, including global regulators (Fig. S4 and Table S2). The finding that global regulators exhibit regulatory switching is especially significant, because *cis* rewiring of a single regulatory protein could create large-scale downstream effects *in trans*. We also found that regulatory switching occurs more frequently in signal transduction pathways (Fisher's exact test, $P < 0.05$). Regulatory switching in signal transduction pathways could help these vital environmental interfaces more rapidly align their response to environmental conditions upon shifts in ecological niches.

**Regulatory Switching Affects *E. coli* Promoter Architecture.** To assess the impact of regulatory switching, we first examined if promoter switching is associated with changes in the positioning of the gene transcription start site (TSS). To this end, we synthesized an *E. coli* promoter library, which allows detection of TSS from multiple bacterial strains in parallel. A similar approach was successfully applied to study TSS composition in *E. coli* (20). After filtering core genes for which the TSS could not be reliably determined due to annotation biases, we were left with 822 core gene clusters (*SI Text*). These core gene clusters were classified as either switched (166 core gene clusters) or unswitched (656 core gene clusters). From each core gene cluster, we selected at least two promoter regions, leading to a total of 1,693 promoters. The selected promoters were synthesized by Agilent Technologies using the oligo library synthesis method (21). This library was transformed into *E. coli* K-12 MG1655, and expression on LB was measured using RNA-sequencing (RNA-Seq). Expression data were used to accurately determine TSS positions of 485 promoter sequences from 40 different *E. coli*. Orthologous TSS positions were used to compute TSS divergence: average distance in base pairs between TSSs of orthologous genes. The mean divergence between switched orthologs was fivefold higher than that between unswitched orthologs ($P < 0.01$; Fig. 3*A*). Switched orthologs also exhibited significantly more TSS divergence than unswitched genes when multiple TSSs in a single gene were taken into account ($P < 0.03$; *SI Text*). Based on our results, we conclude that regulatory switching drives promoter architecture divergence.

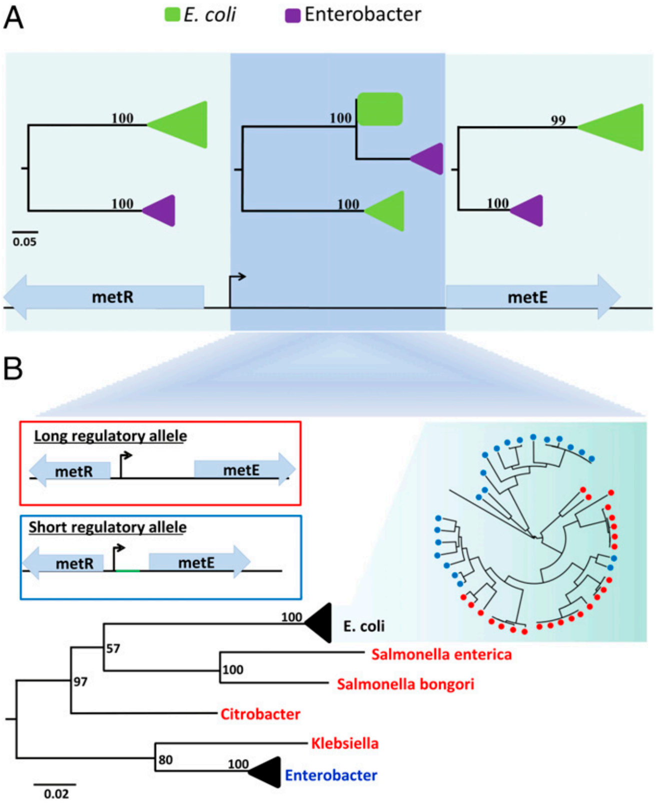

**Fig. 2.** HRT between *E. coli* and *Enterobacter*. (*A*) Phylogenetic tree for *metR*, *metE*, and the leader sequence of *metE*. Clades are collapsed into triangles or marked by a square (which represents that all sequences are 100% identical). *Enterobacter* is shown in purple, and *E. coli* is shown in green. For both protein-coding genes, *E. coli* and *Enterobacter* each form a monophyletic group. In contrast, the phylogeny of the intergenic region is incongruent with the phylogeny for the surrounding genes, suggesting horizontal transfer of the intergenic region independent of the surrounding genes. (*B*) Long and short regulatory alleles and their mapping onto the Enterobacteriaceae species tree. The long regulatory allele is shown in red, and the short regulatory allele is shown in blue. The phylogenetic pattern of the long and short alleles is consistent with HRT from *Enterobacter* to *E. coli*. The high AT content stretch of the short regulatory allele is marked in green. (Scale bar: nucleotide substitutions per site.) Statistical support for the internal branches was computed using 100 bootstrap repetitions.

**Regulatory Switching Drives Expression Diversification of *E. coli* Strains.** To test if regulatory switching alters the transcriptional response, we performed high-throughput RNA-Seq to compare the expression patterns of two *E. coli* strains that occupy distinct ecological niches: a gastrointestinal commensal (MG) and a urinary tract pathogen (CFT). We measured gene expression levels for all 3,293 orthologous genes present in both strains when grown on either defined minimal potassium morpholinopropane sulfonate (MOPS) media or pooled, sterile human urine (Fig. S5). Despite their ecological differences and more than 5 My of evolutionary divergence, most genes exhibited similar expression between strains exposed to the same conditions (Fig. 3; MOPS: $R^2 = 0.95$, urine: $R^2 = 0.98$). Nonetheless, as shown in Fig. 3, 266

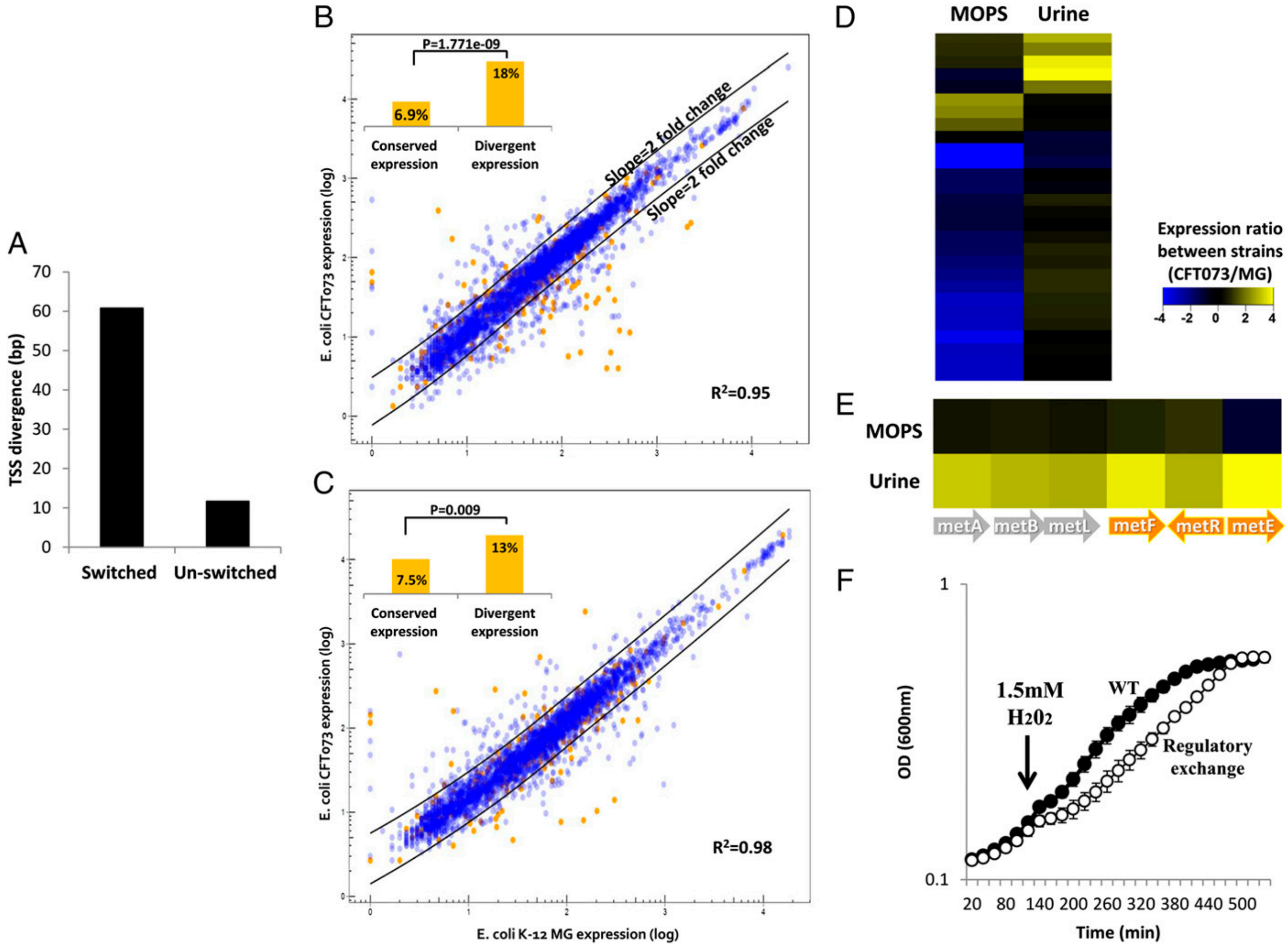

**Fig. 3.** Regulatory switching drives expression diversification and adaptation of *E. coli*. (*A*) Switched orthologs are more diverged with respect to their TSS position compared with unswitched genes. Expression diversification of *E. coli* strains grown on MOPS (*B*) and on pooled human urine (*C*). Each circle represents the average transcript level of an orthologous gene across three independent experiments. Genes that underwent regulatory switching are shown in orange, and those genes not affected by switching are shown in blue. The black lines, estimated by locally weighted scatterplot smoothing regression, indicate twofold change in expression between strains. (*Insets*) Level of regulatory switching in genes showing divergent expression vs. conserved expression. (*D*) Expression of switched genes exhibiting condition-specific expression divergence. (*E*) Condition-dependent divergence of the methionine biosynthesis pathways. Genes affected by regulatory switching are marked by orange arrows. Although both strains express the pathway in a similar manner when grown on MOPS, the pathogenic strain shows up to 16-fold higher expression of the pathway when strains are grown on urine. (*F*) Replacement of the short *metE* regulatory allele with the long ancestral allele renders the pathogenic bacteria more sensitive to oxidative stress. The strains were grown on MOPS media without methionine. After 2 h, oxidative stress was induced by adding $H_20_2$ to a final concentration of 1.5 mM (marked by an arrow). Filled circles (●) denote CFT073 WT, and empty circles (○) denote CFT073 with a K-12 regulatory region. The data represent three independent experiments.

genes in MOPS and 219 genes in urine exhibited statistically significant and substantial (over twofold change) expression divergence. The frequency of switched genes within this divergent expression group was found to be threefold higher than in the conserved expression group (Fig. 3*B*, *Inset*). The tendency of switched genes to exhibit higher expression divergence was also indicated by ~1.4-fold higher median expression divergence compared with unswitched genes (MOPS: $P = 9.65 \times 10^{-9}$, urine: $P = 6.75 \times 10^{-5}$; Wilcoxon rank-sum test).

Notably, 45% of the genes exhibiting switching-associated expression divergence are condition-specific (i.e., their expression diverges in one condition only) (Fig. 3*D*). Thus, switching may alter the response of bacteria only in a subset of environmental conditions. For example, condition-dependent expression divergence was observed in genes belonging to the methionine biosynthesis pathway (Fig. 3*E*). These genes exhibited similar expression levels in both strains when grown on MOPS but displayed higher expression in the uropathogenic *E. coli* when grown on urine. Three of these genes underwent switching, including the regulator of this pathway (*metR*), *metF*, and the last enzyme in the pathway (*metE*), which exhibited the highest expression divergence (up to 16-fold) (Fig. 3*E*).

**HRT Affects the Fitness of Pathogenic *E. coli*.** The gene which exhibits the greatest urine specific expression divergence, *metE*, is known for its high sensitivity to oxidation (22). Consequently, cells exposed to oxidative stress develop methionine auxotropy (23). This sensitivity poses a challenge to uropathogenic *E. coli*, which is often exposed to oxidative stress generated by host immune cells (24). We reasoned that the switching observed in the regulatory region of *metE* (common to all uropathogenic *E. coli* isolates) might confer a fitness advantage under oxidizing conditions. To test this hypothesis, we constructed an isogenic pathogenic strain that was identical to its parent strain except that the short *metE* regulatory allele was replaced with the longer ancestral allele found in commensal *E. coli*. The resulting strain exhibited a similar growth rate on MOPS media lacking methionine. In contrast, under oxidative stress, this replacement strain exhibited a marked growth defect relative to the WT strain harboring the shorter *metE* allele (Fig. 3*F*). These results demonstrate that a single regulatory switching event, in which the coding region remains unmodified, can confer a significant fitness advantage.

**Regulatory Switching Is Ubiquitous Across the Bacterial Domain.** To determine if regulatory switching affects other clades beyond *E. coli*, we extended our analysis to nine additional taxa from across the bacterial domain with diverse physiological characteristics (Table S3). We found that all clades experienced switching, highlighting the phylogenetic breadth of this phenomenon (Fig. 4). Remarkably, the frequency of regulatory switching in core genes varies by more than an order of magnitude, from 0.5% in *Chlamydia trachomatis*, an obligate intracellular human pathogen, to more than 15% in *Neisseria meningitidis*, a highly recombinogenic pathogen that causes meningitis and septicemia. The variation in switching level among these bacterial clades could not be explained by sampling bias (Fig. S6) or phylogeny (Fig. 4).

These findings raise the question as to what is driving variation in switching levels. Donor accessibility, ecology, and recombination efficiency were all found to affect gene transfer (25), and therefore are expected to affect regulatory transfer. Indeed, the level of switching is associated with the overall recombination-to-mutation (r/m) ratio (Table S4). Specifically, species with low r/m ratios are characterized by a low level of switching (e.g., *C. trachomatis* and *Mycobacterium tuberculosis*), whereas species with high r/m ratios are characterized by a high level of switching (e.g., *Helicobacter pylori* and *N. meningitidis*). However, this factor alone cannot explain the full extent of variation in the levels of regulatory switching. For instance, although *Salmonella enterica* and *E. coli* exhibit similar r/m ratios (0.14 and 0.38, respectively), *E. coli* exhibits more than a 10-fold higher level of regulatory switching. This difference might stem from the different lifestyle of the two species. Whereas *S. enterica* is an intracellular pathogen, *E. coli* is largely extracellular, and thus might be exposed to more foreign DNA during the course of its infection. Another factor that can affect the overall level of switching is the ability of bacteria to acquire DNA from the environment. Indeed, the highest levels of regulatory switching were found in the naturally competent bacteria *H. pylori* and *N. meningitidis*. Future work is needed to elucidate how mechanistic constraints and ecological barriers affect regulatory switching.

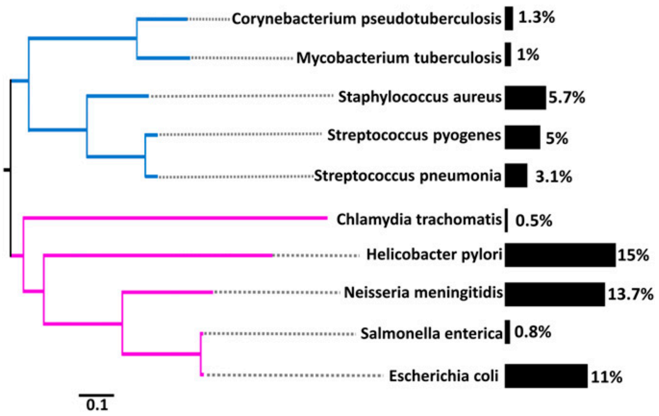

**Fig. 4.** Regulatory switching is ubiquitous across the bacterial domain. The bacterial species phylogeny based on 29 concatenated ribosomal proteins obtained from a study by Williams et al. (37) is shown. Bars indicate the level of regulatory switching observed across all genomes within each clade. Numbers at the end of each bar correspond to the percentage of core genes exhibiting regulatory switching. Gram-negative and Gram-positive taxa are shown in purple and blue, respectively. (Scale bar: substitutions per site.)

## Discussion

Our observation that core genes exhibit ubiquitous regulatory switching contradicts the common assumption that core genes do not play a role in diversification (5). Previous studies have focused on protein-level conservation and overlooked regulatory switching as an orthogonal source of phenotypic variation in core genes. Switching enables a cell to bypass deleterious intermediates generated through the accumulation of point mutations, allowing even essential genes, such as *hemH*, to undergo regulatory modification. By enabling a “quantum leap” between the fitness peaks of functional regulatory elements, switching could facilitate efficient exploration of alternative promoter architectures.

The molecular mechanism most likely underlying the bacterial ability to switch from one regulatory sequence to another is homologous recombination. A short region of sequence identity is required to initiate this mechanism, and its efficiency decreases with increased sequence divergence between genomes (26, 27). Because core genes are highly conserved both between strains and often across distant species, they may enable regulatory switching between otherwise diverged bacteria. In line with this view, we find that 13.8% of the switched regulatory regions reside within a conserved region in which both the upstream and downstream genes are orthologous. Further support for the association between conservation and in situ replacement is the observation that xenologous recombination, the replacement of a gene by a distant homolog, was previously found to be prevalent within conserved operons (28). Of note, we expect regulatory switching to be even more frequent than in situ gene replacement, because regulatory regions are shorter than genes and can fit on a single *E. coli* recombination segment, which is, on average, 242 bp (*SI Text*).

We have shown that regulatory regions, similar to coding regions of bacteria, can be subjected to recombination and exchange. Several theories have been suggested to explain the differential frequencies with which genes undergo HGT. For example, the complexity hypothesis posits that HGT is rare in genes coding for proteins with many interactions compared with those genes coding for proteins with only a few interactions (29, 30). Other studies have detected functional and ecological barriers to horizontal transfer of protein-coding genes (25, 31). The barriers to HRT remain to be discovered, leaving many unanswered questions. Is it restricted by the number of regulatory interactions? Is it promoted by the availability of transcription factors that are shared between the donor and the acceptor? The sheer increase in the availability of fully sequenced bacterial genomes, together with the development of more specific tools for HRT analysis, should shed light on the evolutionary forces shaping the regulatory genome.

The ability of bacteria to tap a broad pool of regulatory sequences suggests that in addition to an environment-specific metagenome, there is an unexplored parallel pool of sequences, the metaregulome. In response to environmental changes, bacteria not only acquire new proteins; they may also acquire novel regulatory sequences to enable more appropriate control of their existing protein repertoire. Our results demonstrate the importance of mobile DNA in regulatory evolution, opening a new window for exploring the mechanisms that bacteria use to respond to environmental changes.

## Materials and Methods

Additional details are available in *SI Text*.

**Regulatory Switching Pipeline.** We detected orthologous genes using reciprocal Translated BLAST (tblastx) (32) best hits with at least 95% amino acid identity (for the core gene analysis; only genes that were shared among all strains of a given species were considered). Next, we detected orthologous gene clusters, requesting 90% identity among all members of a cluster. The regulatory region of each gene cluster, defined as 300 bp upstream of the

TSS, was extracted. Last, the orthologous regulatory regions of each gene were clustered. Genes were considered switched if their regulatory regions formed more than one cluster.

**HRT Detection.** HRT was detected by searching for statistical significant incongruence between the species tree and the regulatory region tree. Specifically, maximum-likelihood trees were reconstructed using PhyML (33) with the general time reversible model (34), and incongruence was tested using the AU (19) test as implemented in CONSEL software (35). To test whether a core gene and its regulatory region were independently transferred, we repeated this procedure comparing the core gene tree and the regulatory region tree.

**Promoter Library TSS Determination.** We synthesized a library of 1,693 promoters from 40 *E. coli* strains and used the RNA-Seq–based approach described by Kosuri et al. (20) to determine the TSS of orthologous genes. For each of the 485 genes expressed under the experimental condition, we computed a distance score reflecting shifts in TSS positioning across strains. A bootstrap-based approach was used to test whether TSS shifts were significantly enriched among switched genes.

**RNA-Seq.** *E. coli* CFT073 and *E. coli* K-12 MG1655 were grown with shaking at 37 °C in 12 mL of MOPS media supplemented with 0.2% tryptone and 0.2% glucose until the $OD_{600}$ reached 0.2. Five milliliters of the bacterial media was then passed through a 0.2-mm pore-sized filter and resuspended in either urine (pooled from six healthy volunteers) or MOPS. The resuspended bacteria were grown for an additional 15 min with shaking at 37 °C and then harvested. Detailed information and sequences are available in the Gene Expression Omnibus (GEO) database (accession no. GSE59468).

**Allelic Exchange and Exposure to Oxidative Stress.** MetE allelic exchange was achieved by using the λ-red recombination system (36). For the oxidative stress experiments, bacteria were grown for 2 h on minimal MOPS media with 0.2% glucose. After 2 h, $H_2O_2$ at a final concentration of 1.5 mM was added to the culture and growth was monitored.

**ACKNOWLEDGMENTS.** We thank Sriram Kosuri for help in obtaining the Agilent oligo library. Y.O. is an Edmond J. Safra Fellow. T.P. is supported by Israeli Foundation Grant 1092/13. J.C. is supported by European Research Council Grant 239784 and by Academy of Finland Grant 251170. D.B. and E.Z.R. are supported by Deutsch-Israelische Projektkooperation Grant RO 2612/1-1. H.H.W. is supported by NIH Director's Early Independence Award 1DP5OD009172-01. N.I.J. is supported by a National Science Foundation (NSF) graduate research fellowship. E.J.A. is supported by the NSF Grant DEB-0936234.

1. Lester CH, Frimodt-Møller N, Sørensen TL, Monnet DL, Hammerum AM (2006) In vivo transfer of the vanA resistance gene from an Enterococcus faecium isolate of animal origin to an E. faecium isolate of human origin in the intestines of human volunteers. *Antimicrob Agents Chemother* 50(2):596–599.
2. Chen J, Novick RP (2009) Phage-mediated intergeneric transfer of toxin genes. *Science* 323(5910):139–141.
3. Hehemann JH, et al. (2010) Transfer of carbohydrate-active enzymes from marine bacteria to Japanese gut microbiota. *Nature* 464(7290):908–912.
4. Ochman H, Lawrence JG, Groisman EA (2000) Lateral gene transfer and the nature of bacterial innovation. *Nature* 405(6784):299–304.
5. Medini D, et al. (2008) Microbiology in the post-genomic era. *Nat Rev Microbiol* 6(6): 419–430.
6. Treangen TJ, Rocha EP (2011) Horizontal transfer, not duplication, drives the expansion of protein families in prokaryotes. *PLoS Genet* 7(1):e1001284.
7. Dagan T, Artzy-Randrup Y, Martin W (2008) Modular networks and cumulative impact of lateral transfer in prokaryote genome evolution. *Proc Natl Acad Sci USA* 105(29):10039–10044.
8. Koonin EV, Makarova KS, Aravind L (2001) Horizontal gene transfer in prokaryotes: Quantification and classification. *Annu Rev Microbiol* 55:709–742.
9. Ragan MA, Beiko RG (2009) Lateral genetic transfer: Open issues. *Philos Trans R Soc Lond B Biol Sci* 364(1527):2241–2251.
10. Matus-Garcia M, Nijveen H, van Passel MW (2012) Promoter propagation in prokaryotes. *Nucleic Acids Res* 40(20):10032–10040.
11. Lee DH, Palsson BO (2010) Adaptive evolution of Escherichia coli K-12 MG1655 during growth on a Nonnative carbon source, L-1,2-propanediol. *Appl Environ Microbiol* 76(13):4158–4168.
12. Somvanshi VS, et al. (2012) A single promoter inversion switches Photorhabdus between pathogenic and mutualistic states. *Science* 337(6090):88–93.
13. Blount ZD, Barrick JE, Davidson CJ, Lenski RE (2012) Genomic analysis of a key innovation in an experimental Escherichia coli population. *Nature* 489(7417):513–518.
14. Chen SL, et al. (2006) Identification of genes subject to positive selection in uropathogenic strains of Escherichia coli: A comparative genomics approach. *Proc Natl Acad Sci USA* 103(15):5977–5982.
15. Retchless AC, Lawrence JG (2012) Ecological adaptation in bacteria: Speciation driven by codon selection. *Mol Biol Evol* 29(12):3669–3683.
16. Mendoza-Vargas A, et al. (2009) Genome-wide identification of transcription start sites, promoters and transcription factor binding sites in E. coli. *PLoS ONE* 4(10):e7526.
17. Gama-Castro S, et al. (2011) RegulonDB version 7.0: Transcriptional regulation of Escherichia coli K-12 integrated within genetic sensory response units (Gensor Units). *Nucleic Acids Res* 39(Database issue):D98–D105.
18. Zheng M, et al. (2001) DNA microarray-mediated transcriptional profiling of the Escherichia coli response to hydrogen peroxide. *J Bacteriol* 183(15):4562–4570.
19. Shimodaira H (2002) An approximately unbiased test of phylogenetic tree selection. *Syst Biol* 51(3):492–508.
20. Kosuri S, et al. (2013) Composability of regulatory sequences controlling transcription and translation in Escherichia coli. *Proc Natl Acad Sci USA* 110(34):14024–14029.
21. LeProust EM, et al. (2010) Synthesis of high-quality libraries of long (150mer) oligonucleotides by a novel depurination controlled process. *Nucleic Acids Res* 38(8):2522–2540.
22. Leichert LI, Jakob U (2004) Protein thiol modifications visualized in vivo. *PLoS Biol* 2(11):e333.
23. Hondorp ER, Matthews RG (2004) Oxidative stress inactivates cobalamin-independent methionine synthase (MetE) in Escherichia coli. *PLoS Biol* 2(11):e336.
24. Rama G, Chhina DK, Chhina RS, Sharma S (2005) Urinary tract infections-microbial virulence determinants and reactive oxygen species. *Comp Immunol Microbiol Infect Dis* 28(5-6):339–349.
25. Popa O, Dagan T (2011) Trends and barriers to lateral gene transfer in prokaryotes. *Curr Opin Microbiol* 14(5):615–623.
26. Shen P, Huang HV (1986) Homologous recombination in Escherichia coli: Dependence on substrate length and homology. *Genetics* 112(3):441–457.
27. Vulić M, Dionisio F, Taddei F, Radman M (1997) Molecular keys to speciation: DNA polymorphism and the control of genetic exchange in enterobacteria. *Proc Natl Acad Sci USA* 94(18):9763–9767.
28. Omelchenko MV, Makarova KS, Wolf YI, Rogozin IB, Koonin EV (2003) Evolution of mosaic operons by horizontal gene transfer and gene displacement in situ. *Genome Biol* 4(9):R55.
29. Cohen O, Gophna U, Pupko T (2011) The complexity hypothesis revisited: Connectivity rather than function constitutes a barrier to horizontal gene transfer. *Mol Biol Evol* 28(4):1481–1489.
30. Doolittle WF (1999) Lateral genomics. *Trends Cell Biol* 9(12):M5–M8.
31. Sorek R, et al. (2007) Genome-wide experimental determination of barriers to horizontal gene transfer. *Science* 318(5855):1449–1452.
32. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol* 215(3):403–410.
33. Guindon S, Gascuel O (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol* 52(5):696–704.
34. Simon T (1986) Some probabilistic and statistical problems in the analysis of DNA sequences. *Some Mathematical Questions in Biology: DNA Sequence Analysis*, Lectures on Mathematics in the Life Sciences, ed Minura RM (American Mathematical Society, Providence, RI), Vol 17, pp 57–86.
35. Shimodaira H, Hasegawa M (2001) CONSEL: For assessing the confidence of phylogenetic tree selection. *Bioinformatics* 17(12):1246–1247.
36. Datsenko KA, Wanner BL (2000) One-step inactivation of chromosomal genes in Escherichia coli K-12 using PCR products. *Proc Natl Acad Sci USA* 97(12):6640–6645.
37. Williams D, et al. (2011) A rooted net of life. *Biol Direct* 6:45.

# Supporting Information

Oren et al. 10.1073/pnas.1413272111

## SI Text

## 1. *Escherichia coli* Analysis

**1.1. Detecting the Genes Comprising the Core Genome of *E. coli*.** To infer the core genome of *E. coli*, we analyzed 46 *E. coli* strains, including eight *Shigella* strains that are obligate intraintestinal pathogens belonging to the *E. coli* species (the full list is provided in *SI Text*, section V). *E. coli* orthologs were identified using pairwise reciprocal Translated BLAST (tblastx) best hits (1) against the K-12 MG1655 reference strain. Based on these homology search results, we identified the *E. coli*'s core genome (i.e., the set of genes present in all strains of a given species, with high sequence conservation). We demanded at least 95% amino acid sequence identity for the region of homology identified by tblastx as high-scoring segment pairs (hsps). Tblastx, rather than Protein BLAST, was used to improve the sensitivity of orthology detection, which is especially important for low-quality genomes. Because the hsp may be only a fraction of the total length of a protein, we also required that the length of the hsp, excluding gaps, should be longer than 50% of the total query length. We additionally demanded that the length of the putative ortholog would not differ by more than 20% from the length of the query sequence. To ensure high conservation among all orthologs within each orthologous group, we used Cluster Database at High Identity with Tolerance (CD-HIT) (2) to filter out all core clusters in which some members show less than 90% nucleotide identity. Such an approach was previously shown to maximize specificity and ensure proper ortholog detection (3). Finally, we filtered out all core clusters that potentially include paralogous genes. Potential paralogous genes were defined as cases in which two different genes in K-12 were mapped, using tblastx, to the same protein.

**1.2. Reconstructing *E. coli*'s Core Regulatory Regions Database.** Because there might be changes in core gene boundaries and length that are the result of using different annotation programs for different strains rather than bona fide sequence changes (4), we performed several steps of preprocessing of the data before aligning them. These steps ensure that the variation in the regulatory regions we detect is not biased by the annotation method. First, a regulatory region is defined as the 300 base pairs immediately upstream of the transcription start site (TSS) of a gene. This definition is frequently used in regulatory studies of bacteria (5, 6).

For each core gene cluster of orthologs, we performed the following steps:

*i*) Extract the gene and the 300 base pairs upstream to its TSS.

*ii*) Remove the length of the longest gene from the 3′ prime of all of the sequences in the cluster, thus removing any sequence that is suspected of being part of the coding core gene. Differences in gene lengths commonly reflect differences in the annotation of TSSs. This procedure ensures that the regulatory region we analyze is upstream of all potential TSSs.

*iii*) Align the regulatory regions with MAFFT (7), trim off indels that are at the 3′ prime and 5′ prime edges of the sequences, and unalign the resulting sequences (by removing all indels). This step is a data cleaning step, which ensures that all of the regulatory regions in each cluster span homologous regions within their genomes.

*iv*) The resulting sequences are realigned with PRANK (3), an indel-sensitive alignment program. The region obtained can include the promoter of the gene, its 5′ UTR, and also parts of the upstream coding gene. Note that by using this unbiased approach to regulatory region definition, we can also detect changes in operonic structure that might influence gene expression.

**1.3. Regulatory Regions Clustering.** For each orthologous group, the unaligned regulatory sequences were clustered at the identity level of 80% using CD-HIT (2). Because clusters that contain only a single sequence may reflect sequencing errors, when counting the number of clusters for each core gene, we only counted clusters with at least two members. We also required that the divergence between clusters would be at least 1.5-fold higher than the divergence within clusters. This criterion ensures that the clusters are fundamentally different, and therefore are likely to possess different regulatory properties. Cluster divergence is calculated based on the PRANK alignment of the regulatory regions. Of the 1,479 orthologous groups tested, 166 were characterized by two or more clusters. We term these groups diverged regulatory core genes or "switched genes." We tested how many of these diverged regulatory core genes are the first gene in their harboring operon in K-12. Operonic structures were taken from RegulonDB (8). Of the 166 switched genes, 138 were first in their operon. This rate is significantly higher than the background rate of 865 (of 1,479) total core genes that are first in their operon (Fisher exact test, $P < 0.005$).

**1.4. Orthology Surrounding Switched Regulatory Regions.** To determine if the genes adjacent to core genes are also orthologous, we extracted the genes upstream of each core gene cluster and clustered them using CD-HIT. We considered an upstream cluster to be a cluster of orthologs if at least 45 of the 46 genes exhibited at least 90% nucleotide identity.

**1.5. Detection of Transcription Factor Binding Sites.** Known transcription factor binding sites (TFBSs) of *E. coli* model strain MG1655 K-12 were downloaded from RegulonDB (8). Predicted TFBSs were computed based on position weight matrices (PWMs) of all available bacteria downloaded from PRODORIC, version 8.9 (9). We scanned the regulatory regions of all core genes using a sliding window with the size of the PWM using Perl scripts. A "hit" was considered significant when the score of the hit was 100-fold higher than the score of the background (the log-odds ratio of the PWM vs. a background model, in which the nucleotide frequencies were determined based on the concatenation of all *E. coli*'s regulatory sequences).

**1.6. Reconstructing the *E. coli* Species Phylogeny.** We performed multiple sequence alignments for the 1,479 *E. coli* core genes using PRANK (3). Gap columns at the 3′ and 5′ ends of genes (if present) were removed before concatenation. The concatenated alignment of all core genes was used as input to PhyML (10) to reconstruct the species tree. The maximum-likelihood tree was computed with the general time reversible model accounting for among-site rate variation (the GTR + $\Gamma$ + I model). To assess the robustness of the topology, 100 bootstrap repeats were conducted. All bootstrap values were higher than 95%, except for four internal branches (shown in Fig. S3). The phylogrouping obtained (Fig. 1) is consistent with the phylogrouping presented in previous studies (11, 12).

**1.7. Horizontal Regulatory Transfer Detection in *E. coli*.** To test for horizontal regulatory transfer (HRT), we tested each orthologous group for incongruence between the tree reconstructed from its regulatory regions and the species tree (inferred based on the concatenation of all core alignments). This testing was done using the approximately unbiased (AU) test (13) as implemented in CONSEL software (14). The input to CONSEL is the log-likelihood of each site under the two possible tree topologies. These log-likelihoods were computed using PhyML (10), assuming the GTR + Γ + I model. The *P* values of this AU test (13) were corrected for multiple testing at an alpha value of 0.05 with a Bonferroni correction. For each orthologous group, we also tested whether the tree obtained for the regulatory region and the tree obtained for the corresponding coding region were congruent. Such congruency suggests a single horizontal transfer event that affected both the gene and its coding region. Incongruences suggest different evolutionary histories for the gene and its regulatory region. This test was conducted using CONSEL as described above, except that we conducted two reciprocal tests here: comparing the gene tree with the regulatory sequences-based tree and comparing the regulatory sequences-based tree with the gene tree. Only if both tests were statistically significant did we consider the two trees as incongruent. The *P* values of the AU test (13) were assessed at an alpha value of 0.05 after Bonferroni correction.

**1.8. HRT Between *E. coli* and *Enterobacter*.** To decipher the evolutionary history of the regulatory region of *metE*, we reconstructed three separate phylogenetic trees: a tree for each coding gene flanking the regulatory region and a tree for the intergenic region (Fig. 2*A*). The trees were reconstructed based on 45 taxa, 43 *E. coli* and two *Enterobacter aerogenes*, that were found to be homologous to the *E. coli* sequences using Nucleotide BLAST (blastn) with default settings. The tree for each section was computed using PhyML (10), assuming the GTR + Γ + I model.

To infer the evolutionary history of the two *metE* alleles, we first inferred the species tree among representatives of the Enterobacteriaceae (Fig. 2*B*). The following strains were used for the analysis:

| Strain | Gene identifier (GI) |
|---|---|
| *E. aerogenes* KCTC 2190 | 336246508 |
| *E. aerogenes* EA1509E | 444350194 |
| *Citrobacter koseri* ATCC BAA-895 | 157081501 |
| *Salmonella bongori* N268-08 | 526125113 |
| *Salmonella enterica* serovar Thompson str. RM6836 | 548713695 |
| *Klebsiella pneumoniae* CG43 | 549815675 |

Phylogenetic relationships among these strains were established based on coding regions of seven housekeeping genes used for multilocus sequence typing (MLST) analysis: *adk*, *fumC*, *gyrB*, *icd*, *mdh*, *purA*, and *recA*. The genes were aligned with MAFFT. The concatenated MLST tree was computed using PhyML, assuming the GTR + Γ + I model. *Pseudomonas aeruginosa* LESB58 (GI: 218888746) was used as an outgroup to root the tree.

**1.9. Accessory Genome Analysis.** We followed the same general approach described above for core genes to identify regulatory switching in *E. coli* accessory genes; however, we did not demand that the gene be found in all *E. coli* strains this time. For the purposes of this analysis, we define accessory genes as those genes found in the *E. coli* MG1655 reference strain and in at least one other *E. coli* strain.

## 2. Experimental Validation

**2.1. Promoter Library Design.** There are 1,479 core gene clusters in *E. coli* (*SI Text*, section 1.1).

To avoid errors arising from using different annotation programs (more details are provided in *SI Text*, section 1.2), we first filtered all promoter clusters that, after alignment, had indels directly upstream of their TSS. The remaining 822 core gene clusters were divided into two groups: 166 switched core gene clusters and 656 unswitched core gene clusters, based on the pipeline for detecting switched core genes described above. From each unswitched core gene cluster, we selected the two most diverged promoter regions (sequence divergence was computed as 1 − frequency of matched base pairs). For each switched core gene, we randomly selected a single representative from each promoter cluster (determination of these promoter clusters is described in *SI Text*, section 1.3). In total, 1,693 promoters were selected from 40 *E. coli* strains. The 165 base pairs upstream of the core gene TSS were synthesized by Agilent Technologies using the oligo library synthesis method.

**2.2. Promoter Library Construction and Growth.** Synthesized promoters contained BamHI and PstI restriction sites, unique 12-bp DNA barcodes, and common end sequences for amplification. To reduce barcode bias, each promoter was synthesized with three different barcodes, thus leading to a library of 5,229 unique oligonucleotides. Promoters were amplified from an ∼1 pmol pool and cloned into a custom p15A orivector, upstream of superfolder GFP (sfGFP) (15), using BamHI-HF (high fidelity) and PstI-HF restriction enzymes and T4 Ligase (New England Biolabs). Ligated constructs were transformed into *E. coli* MegaX DH10B T1R Electrocomp cells (Invitrogen) and subsequently retransformed into *E. coli* MG1655, both at greater than 100-fold coverage and under carbenicillin selection. Library culture was grown to an $OD_{600}$ of ∼0.4 from a 1:300 dilution. Cultures were immediately cooled to 4 °C, pelleted, and frozen for future RNA analysis.

**2.3. Promoter Library TSS Determination.** We sequenced whole 5′ UTRs to determine TSSs using a method similar to a previously described study (16). Total RNA was isolated from pellets using a Qiagen RNeasy Midi Kit, and rRNA was removed using an Epicentre Ribo-Zero rRNA Magnetic Removal Kit for Gram-negative bacteria and purified using a Qiagen RNeasy MinElute Kit. Remaining mRNA was dephosphorylated using 5′ RNA Polyphosphatase (Epicentre) as follows: 12 μL of RNA from the previous step, 2 μL of 10× RNA 5′ polyphosphatase reaction buffer, 0.5 μL of RiboGuard RNase Inhibitor (Epicentre), 1 μL of RNA 5′ Polyphosphatase (60 units), and 4.5 μL of RNase-free water at 37 °C for 30 min.

The reaction was cleaned up using an RNeasy MinElute Kit. We then ligated an RNA adaptor to mRNA 5′ ends with the sequence GAGUUCAGACGUGUGCUCUUCCGAUCUNN to dephosphorylated mRNA as follows: 14 μL of RNA from the previous step, 2 μL of 250 μM RNA adaptor, 2.5 μL of 10× ligase buffer, 2 μL of Epicentre T4 RNA Ligase (10 units), 2 μL of 10 mM ATP, 1 μL of RiboGuard RNase Inhibitor, and 1 μL of DMSO at 22.5 °C for 3 h, followed by 65 °C for 10 min for deactivation.

Ligated products were purified using an RNeasy MinElute Kit. RT was performed on mRNA using an internal sfGFP primer. We combined the following in an RNase-free PCR tube: 0.2 μL of 10 μM RT primer (ACCGTTGACATCACCATCCAGTTCC), 12 μL of RNA from ligation, and 1 μL of 10 mM dNTP mix at 65 °C for 5 min and on ice for 1 min.

The following components were then added to the PCR tube from the last step: 4 μL of 5× First-Strand Buffer (Invitrogen), 1 μL of 0.1 M DTT, 1 μL of RNaseOUT (Invitrogen), and 1 μL of SuperScript III Reverse Transcriptase (200 units; Invitrogen).

The reaction was mixed by gentle pipetting and then incubated for 60 min at 55 °C and inactivated by heating to 70 °C for 15 min. The resulting cDNA was prepared for next-generation sequencing by adding Illumina adapters and indexes through two sequential quantitative PCR reactions (Kapa SYBR Fast; Kapa Biosystems) while minimizing cycles to prevent overamplification. Samples were sequenced using Illumina 250-bp paired-end sequencing. Reads were mapped to promoters by identifying the N-terminal barcode in read 1. TSSs were determined by mapping read 2 (5′ end of transcripts) to promoter sequences using custom Python scripts.

**2.4. TSS Analysis.** For each promoter, we obtained an average of 382 reads. Often, these reads point to different TSSs. This inconsistency stems from either methodological noise or from the existence of two or more alternative TSSs. We have written a bioinformatics pipeline: (*i*) to determine whether there is a single TSS or multiple TSSs and (*ii*) to determine the most probable location for each TSS after removing noise. To this end, we first filtered out all those promoters for which the coverage was less than 30 reads.

The number of reads supporting each possible TSS position (between 1 and 165) was summarized. Next, the promoter positions that correspond to the 25th percentile and the 75th percentile were computed. In cases where the difference between these two positions was smaller than 5 bp, we concluded that the data support the existence of a single TSS. In all other cases, the existence of multiple TSSs was considered.

***Analyzing single TSS cases.*** The promoter position supported by the highest number of TSSs was identified. Only promoter positions supported by at least 30 reads were furthered considered (all other cases were discarded as cases for which not enough data exist to determine the TSS accurately). Next, we demanded that this position be supported by at least two of the three barcodes. Support was defined as the existence of a supporting read in the vicinity of the TSS position (from position −5 bp to +5 bp of the suspected TSS). Furthermore, to validate that the inferred TSS position is not sensitive to random noise, we applied the following bootstrapping approach. We repeated the above analysis for 100 bootstrap samples. A TSS was considered reliable if at least 90 bootstrap replicates yielded the exact same TSS.

***Analyzing multiple TSS cases.*** The obtained TSS distribution was used as input to an expectation maximization method to determine the number of modes and their locations (17). The number of modes was restricted by the constraint that each mode should be supported by at least 10% of the reads. We further requested that the primary TSS be supported by at least 30 reads. Finally, the primary TSS should also be supported by at least 90 bootstrap replicates, as described above.

**2.5. Comparing TSSs Between Switched and Unswitched Promoter Clusters.** For each promoter cluster, we computed a DTP score: the difference between primary TSSs among promoter alleles. The DTP is simply the difference in base pairs between the two primary TSSs. When more than two alleles exist, the DTP score is defined as the average among the pairwise DTP scores.

For over 80% of the switched core gene clusters and over 81% of the unswitched core gene clusters, the DTP score is 0. This finding suggests that determination of the DTP (after all of the filtering steps described above) is reliable and that the TSS is relatively conserved among diverged *E. coli* strains. However, we further noted that among those core gene clusters with high DTP values, there were differences in the DTP values between the switched and unswitched categories (the switched core genes have higher DTPs). To test if this difference is statistically significant, we first computed the difference between average DTPs among the 10 top DTP values for the two groups (switched and unswitched gene clusters). To determine if the observed DTP difference is significantly higher than the random expectation, we repeated this procedure 1,000 times for random labeling of core genes (switched or unswitched), resulting in an empirical null distribution. The difference between switched and unswitched average top decile scores was found to be significant (60.8 for switched and 11.7 for unswitched; $P < 0.01$). These results suggest that switching alters promoter architecture.

The DTP score defined above only considers distance in TSS among primary TSSs. We repeated the above analyses with an alternative definition of TSS distance, the DATP score (difference between all TSSs, primary and nonprimary, among promoter alleles). The difference between the DTP and the DATP is only for cases in which multiple TSSs exist. Whereas the DTP only considers the primary TSS, the DATP considers alternative TSSs. If at least one allele has multiple TSSs, the DATP score is computed as the average between the best-matching TSSs between both alleles (where the mapping between the TSSs of alternative alleles is the one that minimizes the distance). A significant statistical difference ($P < 0.03$) between the average top decile score of switched (59 DATP) and unswitched (37 DATP) genes was also found when accounting for multiple TSSs.

**2.6. Comparative Transcriptomics of *E. coli* Strains.** *E. coli* CFT073 was isolated from the blood and urine of a woman with acute pyelonephritis (18). *E. coli* K-12 MG1655 is gastrointestinal commensal stool isolate (19). Whereas the pathogenic *E. coli* CFT073 has 5,338 genes, the commensal *E. coli* has only 4,321 genes. The CFT073-MG set of orthologs was inferred by identifying pairwise reciprocal tblastx best hits. We demanded at least 95% identity in amino acid sequence for the region of homology identified by tblastx as hsps. There were 3,293 orthologous genes that satisfied these criteria.

The regulatory sequences for each of these orthologs were clustered at the identity level of 80% using CD-HIT. We defined a gene cluster as directly switched if its regulatory region forms at least two clusters. Genes affected by switching are either directly switched genes or genes in an operon whose regulatory region was switched [operonic organization was taken from a study by Gama-Castro et al. (8)]. Using these criteria, 193 genes were directly switched and 64 genes reside in an operon with a switched promoter, bringing the total number of genes affected by switching to 257 (7.8% of the total number of shared genes).

**2.7. Bacterial Growth Conditions.** Urine was collected from three healthy male and female volunteers ($n = 6$) aged 20–40 y who had no history of a urinary tract infection or antibiotic use in the prior 3 mo. The urine samples were polled and immediately filter-sterilized (0.2-m pore size). The pooled urine was stored at 4 °C for use within 3 d. For RNA preparations, both strains were grown with shaking at 37 °C in 12 mL of potassium morpholinopropane sulfonate (MOPS) media (20) supplemented with 0.2% tryptone and 0.2% glucose until the OD$_{600}$ reached 0.2. Five milliliters of the bacterial media was then passed through a 0.2-m pore-sized filter and resuspended with either urine or MOPS. The resuspended bacteria were grown for an additional 15 min, with shaking at 37 °C. To stop bacterial growth, ice was added to the MOPS and urine-grown samples, and the cultures were harvested by centrifugation (10 min, 8,000 × *g*, 4 °C). The supernatant was discarded, and 500 µL of saline (0.9% NaCl) was added to each sample. Samples were then treated with RNA Protect Bacterial Reagent (Qiagen) to stabilize RNA according to the manufacturer's instructions. The bacterial pellet was frozen at −20 °C until RNA extraction. For the oxidative stress experiments, bacteria were grown for 2 h on minimal MOPS media with 0.2% glucose (not supplemented with tryptone). After 2 h, $H_2O_2$ at a final concentration of 1.5 mM was added to the culture and growth was monitored.

**2.8. RNA Isolation.** Total RNA was isolated from the bacterial pellet using the RNeasy Kit (Qiagen) and treated with DNase according to the manufacturer's instructions. The RNA was depleted of rRNA using a Ribo-Zero rRNA Removal Gram-Negative Kit (Epicentre) according to the manufacturer's protocol.

**2.9. RNA-Sequencing Analysis.** RNA-sequencing libraries were prepared using TruSeq RNA Sample Prep Kits (Illumina). High RNA quality was confirmed before library construction using Agilent's TapeStation 2200. The 50-bp single-end reads were then sequenced on an Illumina HiSeq 2500. Reads were aligned using Rockhopper software (21) with default parameters. Following alignment, reads from each experiment were normalized by upper quartile normalization (22).

**2.10. Identification of Orthologous Genes with Expression Diversification.** For each of the 3,293 CFT073-MG orthologs, we calculated the average expression level in MOPS and urine across three independent biological repeats. A gene was considered to have divergent expression only if the ratio in expression between the two strains [CFT073 reads per kilobase per million (RPKM)/MG1655 RPKM] was higher than 2. Formally, we requested that the expression of the CFT strain be either more than twice or less than half of the expression under a perfect correlation between the two strains, computed using a locally weighted scatterplot smoothing function (Fig. 3). To filter out cases in which high ratio reflects random noise, we further requested that the difference in expression between the two strains be statistically significant. To this end, we used a negative-binomial test, correcting for multiple testing using a false discovery rate at the 1% level, implemented by edger software (23). If the gene exhibited expression divergence only at one of the conditions tested (MOPS or urine) and not at the other and there was a fold change higher than twofold between conditions, the gene was considered to display a condition-specific expression divergence.

**2.11. Overall Contribution of Regulatory Switching to Expression Divergence.** For each condition (MOPS or urine), the percentage of divergence that can be explained by regulatory switching was calculated by dividing the number of genes that are both switched and differentially expressed in this condition by the total number of genes that are differentially expressed in this condition. The overall contribution of regulatory switching to expression divergence was calculated by dividing the number of genes that are both switched and exhibit diverged expression in at least one of the conditions by the number of genes exhibiting diverged expression in at least one of the conditions.

**2.12. Allelic Exchange of metE Regulatory Region.** A two-step protocol was used to obtain a "clean" allelic exchange of the metE regulatory region. Both steps were done using the λ-red recombination system (24). At first, competent WT O73 bacteria were transformed with a pKD46 plasmid. The transformants were grown in ampicillin-containing LB, induced by arabinose. The bacteria were then transformed with the PCR fragment encoding for kanamycin with the flanking region of the O73 *metE* regulatory region. The resulting recombinants were screened for kanamycin-resistant and methionine auxotrophy. Next, kanamycin-resistant recombinants containing pKD46 plasmid were transformed with linear PCR products containing the regulatory region between MG1655 *metE* and *metR*. Recombinants were screened for methionine autotrophy and kanamycin sensitivity. The pKD46 plasmid was cured by growth on LB at 42 °C. The final exchanged strain was verified by PCR assay.

## 3. Regulatory Switching Pipeline

The same methodological steps that were used for identifying regulatory switching in *E. coli* were applied for nine additional bacterial species: *Chlamydia trachomatis*, *Salmonella enterica*, *Mycobacterium tuberculosis*, *Corynebacterium pseudotuberculosis*, *Streptococcus pneumoniae*, *Streptococcus pyogenes*, *Staphylococcus aureus*, *Helicobacter pylori*, and *Neisseria meningitidis*. The full list of genomes analyzed for each bacterial species is provided in *SI Text*, section 6.

## 4. Recombination Inference with BRAT NextGen

For each species, BRAT NextGen (25) was used to determine recombinant segments in the core genome alignment (reconstructed from all concatenated core genes). Inference was performed using the same fixed value of the hyperparameter alpha = 1 for all datasets to ensure maximal comparability of the results and to avoid potential estimation sensitivity due to the smaller number of genomes available for some species. Twenty iterations of the estimation algorithm were used in all cases; this number of iterations was assessed to be sufficient for convergence because changes in the hidden Markov model parameters were already negligible over approximately the latter half of iterations for all species. Significance of a recombinant segment was determined as in the study by Marttinen et al. (25) using a permutation test with 100 permutations executed in parallel on a cluster computer. A threshold of 5% was used to conclude significance for each putative recombination.

## 5. *E. coli* Strains Used in This Study

| Strain | UID |
|---|---|
| *E. coli* K 12 substr. MG1655 | 57779 |
| *E. coli* ETEC H10407 | 161993 |
| *E. coli* O111 H 11128 | 41023 |
| *E. coli* O103 H2 12009 | 41013 |
| *E. coli* O26 H11 11368 | 41021 |
| *E. coli* BW2952 | 59391 |
| *E. coli* IAI1 | 59377 |
| *E. coli* 55989 | 59383 |
| *E. coli* K 12 substr. DH10B | 58979 |
| *E. coli* HS | 58393 |
| *E. coli* K-12 substr. W3110 | 58567 |
| *Shigella boydii* CDC 3083 94 | 58415 |
| *S. boydii* Sb227 | 58215 |
| *E. coli* E24377A | 58395 |
| *E. coli* SE11 | 59425 |
| *Shigella flexneri* 2002017 | 159233 |
| *S. flexneri* 5 8401 | 58583 |
| *S. flexneri* 2a 2457T | 57991 |
| *S. flexneri* 2a | 62907 |
| *E. coli* DH1 | 161951 |
| *Shigella sonnei* Ss046 | 58217 |
| *E. coli* BL21 DE3 | 59245 |
| *E. coli* B REL606 | 58803 |
| *E. coli* BL21 Gold DE3 pLysS AG | 59245 |
| *E. coli* ATCC 8739 | 58783 |
| *E. coli* O55 H7 CB9615 | 46655 |
| *E. coli* O157 H7 Sakai | 57781 |
| *E. coli* O157 H7 EDL933 | 57831 |
| *E. coli* IAI39 | 59381 |
| *E. coli* SMS 3 5 | 58919 |
| *E. coli* UM146 | 162043 |
| *E. coli* ABU 83972 | 161975 |
| *E. coli* IHE3034 | 162007 |
| *E. coli* ED1a | 59379 |
| *E. coli* S88 | 62979 |
| *E. coli* APEC O1 | 58623 |

| | |
|---|---|
| *E. coli* 536 | 58531 |
| *E. coli* UTI89 | 58541 |
| *E. coli* CFT073 | 57915 |
| *E. coli* O127 H6 E2348 69 | 59343 |
| *E. coli* SE15 | 161939 |
| *E. coli* O157 H7 TW14359 | 59235 |
| *E. coli* O157 H7 EC4115 | 59091 |
| *E. coli* UMN026 | 62981 |
| *Shigella dysenteriae* Sd197 | 58213 |
| *E. coli* 042 | 161985 |
| *Escherichia fergusonii* ATCC 35469 | 59375 |

ATCC, American Type Culture Collection; UID, unique identifier.

## 6. Bacterial Strains Used in the Regulatory Divergence Pipeline

### 6.1. *H. pylori*

| Strain | UID |
|---|---|
| *H. pylori* 2017 | 161151 |
| *H. pylori* 2018 | 161159 |
| *H. pylori* 26695 | 57787 |
| *H. pylori* 35A | 49903 |
| *H. pylori* 51 | 161925 |
| *H. pylori* 83 | 161153 |
| *H. pylori* 908 | 159985 |
| *H. pylori* B38 | 59415 |
| *H. pylori* B8 | 49873 |
| *H. pylori* Cuz20 | 159987 |
| *H. pylori* ELS37 | 158157 |
| *H. pylori* F16 | 161145 |
| *H. pylori* F30 | 159991 |
| *H. pylori* F32 | 161139 |
| *H. pylori* F57 | 161143 |
| *H. pylori* G27 | 59305 |
| *H. pylori* Gambia94 24 | 159493 |
| *H. pylori* HPAG1 | 58517 |
| *H. pylori* HUP B14 | 162213 |
| *H. pylori* India7 | 161149 |
| *H. pylori* J99 | 57789 |
| *H. pylori* Lithuania75 | 159491 |
| *H. pylori* P12 | 59327 |
| *H. pylori* PeCan18 | 162211 |
| *H. pylori* PeCan4 | 53539 |
| *H. pylori* Puno120 | 159611 |
| *H. pylori* Puno135 | 161157 |
| *H. pylori* SJM180 | 53541 |
| *H. pylori* SNT49 | 159615 |
| *H. pylori* Sat464 | 159467 |
| *H. pylori* Shi112 | 162207 |
| *H. pylori* Shi169 | 162209 |
| *H. pylori* Shi417 | 162205 |
| *H. pylori* Shi470 | 59165 |
| *H. pylori* SouthAfrica7 | 159989 |
| *H. pylori* XZ274 | 165869 |
| *H. pylori* 52 | 159983 |
| *H. pylori* v225d | 159639 |

### 6.2. *N. meningitidis*

| Strain | UID |
|---|---|
| *N. meningitidis* 053442 | 58587 |
| *N. meningitidis* 8013 | 161967 |
| *N. meningitidis* FAM18 | 57825 |
| *N. meningitidis* G2136 | 162085 |
| *N. meningitidis* H44 76 | 162083 |
| *N. meningitidis* M01 240149 | 162079 |
| *N. meningitidis* M01 240355 | 162075 |
| *N. meningitidis* M04 240196 | 162081 |
| *N. meningitidis* MC58 | 57817 |
| *N. meningitidis* NZ 05 33 | 162077 |
| *N. meningitidis* WUE 2594 | 162093 |
| *N. meningitidis* Z2491 | 57819 |
| *N. meningitidis* alpha14 | 61649 |
| *N. meningitidis* alpha710 | 161971 |

### 6.3. *S. aureus*

| Strain | UID |
|---|---|
| *S. aureus* 04 02981 | 161969 |
| *S. aureus* 11819 97 | 159981 |
| *S. aureus* 71193 | 162141 |
| *S. aureus* COL | 57797 |
| *S. aureus* ECT R 2 | 159389 |
| *S. aureus* ED133 | 159689 |
| *S. aureus* ED98 | 41455 |
| *S. aureus* HO 5096 0412 | 162163 |
| *S. aureus* JH1 | 58457 |
| *S. aureus* JH9 | 58455 |
| *S. aureus* JKD6008 | 159855 |
| *S. aureus* JKD6159 | 159691 |
| *S. aureus* LGA251 | 159391 |
| *S. aureus* M013 | 88065 |
| *S. aureus* MRSA252 | 57839 |
| *S. aureus* MSHR1132 | 89393 |
| *S. aureus* MSSA476 | 57841 |
| *S. aureus* MW2 | 57903 |
| *S. aureus* Mu3 | 58817 |
| *S. aureus* Mu50 | 57835 |
| *S. aureus* N315 | 57837 |
| *S. aureus* NCTC 8325 | 57795 |
| *S. aureus* Newman | 58839 |
| *S. aureus* RF122 | 57661 |
| *S. aureus* S0385 | 159247 |
| *S. aureus* T0131 | 159861 |
| *S. aureus* TCH60 | 159859 |
| *S. aureus* TW20 | 159241 |
| *S. aureus* USA300 FPR3757 | 58555 |
| *S. aureus* USA300 TCH1516 | 58925 |
| *S. aureus* VC40 | 88071 |

### 6.4. *S. pyogenes*

| Strain | UID |
|---|---|
| *S. pyogenes* Alab49 | 162171 |
| *S. pyogenes* M1 GAS | 57845 |
| *S. pyogenes* MGAS10270 | 58571 |
| *S. pyogenes* MGAS10394 | 58105 |
| *S. pyogenes* MGAS10750 | 58575 |
| *S. pyogenes* MGAS15252 | 158037 |
| *S. pyogenes* MGAS1882 | 158061 |
| *S. pyogenes* MGAS2096 | 58573 |
| *S. pyogenes* MGAS315 | 57911 |
| *S. pyogenes* MGAS5005 | 58337 |
| *S. pyogenes* MGAS6180 | 58335 |
| *S. pyogenes* MGAS8232 | 57871 |
| *S. pyogenes* MGAS9429 | 58569 |
| *S. pyogenes* Manfredo | 57847 |
| *S. pyogenes* NZ131 | 59035 |
| *S. pyogenes* SSI 1 | 57895 |

### 6.5. *S. pneumoniae*

| Strain | UID |
|---|---|
| *S. pneumoniae* 670 6B | 52533 |
| *S. pneumoniae* 70585 | 59125 |
| *S. pneumoniae* AP200 | 52453 |
| *S. pneumoniae* ATCC 700669 | 59287 |
| *S. pneumoniae* CGSP14 | 59181 |
| *S. pneumoniae* D39 | 58581 |
| *S. pneumoniae* G54 | 59167 |
| *S. pneumoniae* Hungary19A 6 | 59117 |
| *S. pneumoniae* INV104 | 162039 |
| *S. pneumoniae* INV200 | 162035 |
| *S. pneumoniae* JJA | 59121 |
| *S. pneumoniae* OXC141 | 162037 |
| *S. pneumoniae* P1031 | 59123 |
| *S. pneumoniae* R6 | 57859 |
| *S. pneumoniae* ST556 | 162191 |
| *S. pneumoniae* TCH8431 19A | 49735 |
| *S. pneumoniae* TIGR4 | 57857 |
| *S. pneumoniae* Taiwan19F 14 | 59119 |

### 6.6. *C. pseudotuberculosis*

| Strain | UID |
|---|---|
| *C. pseudotuberculosis* 1002 | 159677 |
| *C. pseudotuberculosis* 1 06 A | 159665 |
| *C. pseudotuberculosis* 258 | 167260 |
| *C. pseudotuberculosis* 267 | 162175 |
| *C. pseudotuberculosis* 316 | 89381 |
| *C. pseudotuberculosis* 31 | 162167 |
| *C. pseudotuberculosis* 3 99 5 | 83609 |
| *C. pseudotuberculosis* 42 02 A | 159669 |
| *C. pseudotuberculosis* C231 | 159675 |
| *C. pseudotuberculosis* CIP 52 97 | 159667 |
| *C. pseudotuberculosis* Cp162 | 168258 |
| *C. pseudotuberculosis* FRC41 | 50585 |
| *C. pseudotuberculosis* I19 | 159673 |
| *C. pseudotuberculosis* P54B96 | 157909 |
| *C. pseudotuberculosis* PAT10 | 159671 |

### 6.7. *M. tuberculosis*

| Strain | UID |
|---|---|
| *M. tuberculosis* CCDC5079 | 161943 |
| *M. tuberculosis* CCDC5180 | 161941 |
| *M. tuberculosis* CDC1551 | 57775 |
| *M. tuberculosis* CTRI 2 | 161997 |
| *M. tuberculosis* F11 | 58417 |
| *M. tuberculosis* H37Ra | 58853 |
| *M. tuberculosis* H37Rv | 170532 |
| *M. tuberculosis* H37Rv | 57777 |
| *M. tuberculosis* KZN 1435 | 59069 |
| *M. tuberculosis* KZN 4207 | 83619 |
| *M. tuberculosis* KZN 605 | 54947 |
| *M. tuberculosis* RGTB327 | 157907 |
| *M. tuberculosis* RGTB423 | 162179 |
| *M. tuberculosis* UT205 | 162183 |

### 6.8. *S. enterica*

| Strain | UID |
|---|---|
| *S. enterica* arizonae serovar 62 z4 z23 RSK2980 | 58191 |
| *S. enterica* serovar Agona SL483 | 59431 |
| *S. enterica* serovar Choleraesuis SC B67 | 58017 |
| *S. enterica* serovar Dublin CT 02021853 | 58917 |
| *S. enterica* serovar Enteritidis P125109 | 59247 |
| *S. enterica* serovar Gallinarum 287 91 | 59249 |
| *S. enterica* serovar Gallinarum pullorum RKS5078 | 87035 |
| *S. enterica* serovar Heidelberg B182 | 162195 |
| *S. enterica* serovar Heidelberg SL476 | 58973 |
| *S. enterica* serovar Newport SL254 | 58831 |
| *S. enterica* serovar Paratyphi A AKU 12601 | 59269 |
| *S. enterica* serovar Paratyphi A ATCC 9150 | 58201 |
| *S. enterica* serovar Paratyphi B SPB7 | 59097 |
| *S. enterica* serovar Paratyphi C RKS4594 | 59063 |
| *S. enterica* serovar Schwarzengrund CVM19633 | 58915 |
| *S. enterica* serovar Typhi CT18 | 57793 |
| *S. enterica* serovar Typhi P stx 12 | 87001 |
| *S. enterica* serovar Typhi Ty2 | 57973 |
| *S. enterica* serovar Typhimurium 14028S | 86059 |
| *S. enterica* serovar Typhimurium 798 | 158047 |
| *S. enterica* serovar Typhimurium LT2 | 57799 |
| *S. enterica* serovar Typhimurium SL1344 | 86645 |
| *S. enterica* serovar Typhimurium ST4 74 | 84393 |
| *S. enterica* serovar Typhimurium T000240 | 84397 |
| *S. enterica* serovar Typhimurium U.K. 1 | 87049 |
| *S. enterica* serovar Typhimurium | 86061 |

### 6.9. *C. trachomatis*

| Strain | UID |
|---|---|
| *C. trachomatis* 434 Bu | 61633 |
| *C. trachomatis* A2497 | 159863 |
| *C. trachomatis* A2497 | 159993 |
| *C. trachomatis* A HAR 13 | 58333 |
| *C. trachomatis* B Jali20 OT | 59351 |
| *C. trachomatis* B TZ1A828 OT | 59349 |
| *C. trachomatis* D EC | 159881 |
| *C. trachomatis* D LC | 159879 |
| *C. trachomatis* D UW 3 CX | 57637 |
| *C. trachomatis* E 11023 | 161369 |
| *C. trachomatis* E 150 | 161403 |
| *C. trachomatis* E SW3 | 167483 |
| *C. trachomatis* F SW4 | 167484 |
| *C. trachomatis* F SW5 | 167485 |
| *C. trachomatis* G 11074 | 161409 |
| *C. trachomatis* G 11222 | 161361 |
| *C. trachomatis* G 9301 | 161377 |
| *C. trachomatis* G 9768 | 161353 |
| *C. trachomatis* L2b UCH 1 proctitis | 61635 |
| *C. trachomatis* L2c | 68843 |
| *C. trachomatis* Sweden2 | 161995 |

1. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol* 215(3):403–410.
2. Li W, Godzik A (2006) Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22(13):1658–1659.
3. Löytynoja A, Goldman N (2005) An algorithm for progressive multiple alignment of sequences with insertions. *Proc Natl Acad Sci USA* 102(30):10557–10562.
4. Nijveen H, Matus-Garcia M, van Passel MW (2012) Promoter reuse in prokaryotes. *Mob Genet Elements* 2(6):279–281.
5. Vanet A, Marsan L, Sagot MF (1999) Promoter sequences and algorithmical methods for identifying them. *Res Microbiol* 150(9-10):779–799.
6. Dekhtyar M, Morin A, Sakanyan V (2008) Triad pattern algorithm for predicting strong promoter candidates in bacterial genomes. *BMC Bioinformatics* 9:233.

7. Katoh K, Misawa K, Kuma K, Miyata T (2002) MAFFT: A novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res* 30(14):3059–3066.
8. Gama-Castro S, et al. (2011) RegulonDB version 7.0: Transcriptional regulation of Escherichia coli K-12 integrated within genetic sensory response units (Gensor Units). *Nucleic Acids Res* 39(Database issue):D98–D105.
9. Grote A, et al. (2009) PRODORIC (release 2009): A database and tool platform for the analysis of gene regulation in prokaryotes. *Nucleic Acids Res* 37(Database issue): D61–D65.
10. Guindon S, Gascuel O (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol* 52(5):696–704.
11. Touchon M, et al. (2009) Organised genome dynamics in the Escherichia coli species results in highly diverse adaptive paths. *PLoS Genet* 5(1):e1000344.
12. Sims GE, Kim SH (2011) Whole-genome phylogeny of Escherichia coli/Shigella group by feature frequency profiles (FFPs). *Proc Natl Acad Sci USA* 108(20):8329–8334.
13. Shimodaira H (2002) An approximately unbiased test of phylogenetic tree selection. *Syst Biol* 51(3):492–508.
14. Shimodaira H, Hasegawa M (2001) CONSEL: For assessing the confidence of phylogenetic tree selection. *Bioinformatics* 17(12):1246–1247.
15. Pédelacq JD, Cabantous S, Tran T, Terwilliger TC, Waldo GS (2006) Engineering and characterization of a superfolder green fluorescent protein. *Nat Biotechnol* 24(1):79–88.
16. Kosuri S, et al. (2013) Composability of regulatory sequences controlling transcription and translation in Escherichia coli. *Proc Natl Acad Sci USA* 110(34):14024–14029.
17. Benaglia T, Chauveau D, Hunter DR, Young DS (2009) Mixtools: An R Package for Analyzing Finite Mixture Models. *J Stat Softw* 32(6):1–29.
18. Mobley HL, et al. (1990) Pyelonephritogenic Escherichia coli and killing of cultured human renal proximal tubular epithelial cells: Role of hemolysin in some strains. *Infect Immun* 58(5):1281–1289.
19. Bachmann BJ (1996) *Derivations and Genotypes of Some Mutant Derivatives of Escherichia coli* (ASM Press, Washington, DC), 2nd Ed.
20. Neidhardt FC, Bloch PL, Smith DF (1974) Culture medium for enterobacteria. *J Bacteriol* 119(3):736–747.
21. McClure R, et al. (2013) Computational analysis of bacterial RNA-Seq data. *Nucleic Acids Res* 41(14):e140.
22. Bullard JH, Purdom E, Hansen KD, Dudoit S (2010) Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics* 11:94.
23. Robinson MD, McCarthy DJ, Smyth GK (2010) edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26(1): 139–140.
24. Datsenko KA, Wanner BL (2000) One-step inactivation of chromosomal genes in Escherichia coli K-12 using PCR products. *Proc Natl Acad Sci USA* 97(12): 6640–6645.
25. Marttinen P, et al. (2012) Detection of recombination events in bacterial genomes from large population samples. *Nucleic Acids Res* 40(1):e6.

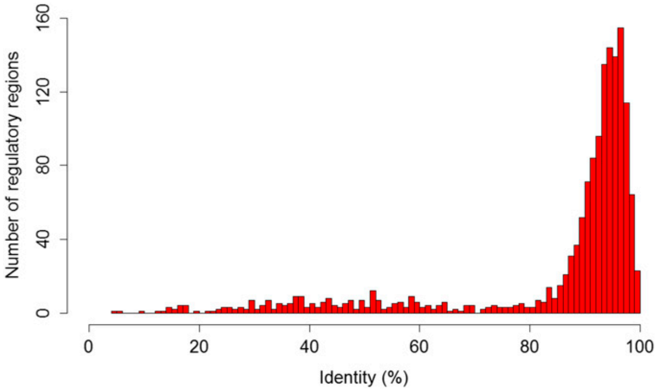

**Fig. S1.** Sequence divergence among core regulatory regions. Percent identity is defined as the fraction of positions that are exactly the same across all analyzed *E. coli* strains. Although most core regulatory regions are conserved, there is a subpopulation of core genes that exhibit high divergence in their regulatory regions.

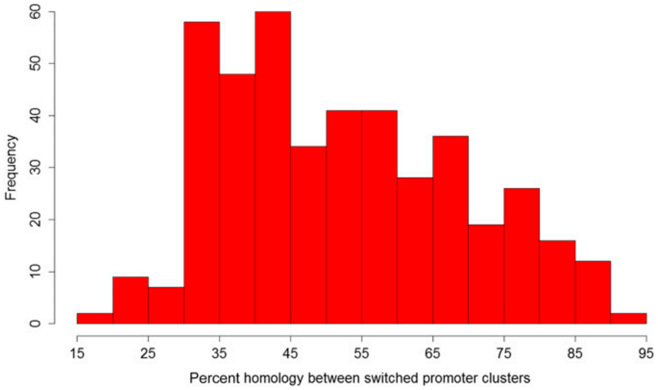

**Fig. S2.** Average percent identity between switched *E. coli* promoter clusters. Average percent identity is defined as the average pairwise identity (fraction of positions that are identical) between all of the promoters that do not belong to the same cluster. Promoter pairwise alignment was performed by MAFFT. The average sequence identity between promoter clusters of switched genes is 52%. Cases in which the percent identity between switched promoter clusters is high are those cases in which switching is restricted to a small fragment of the entire regulatory region.

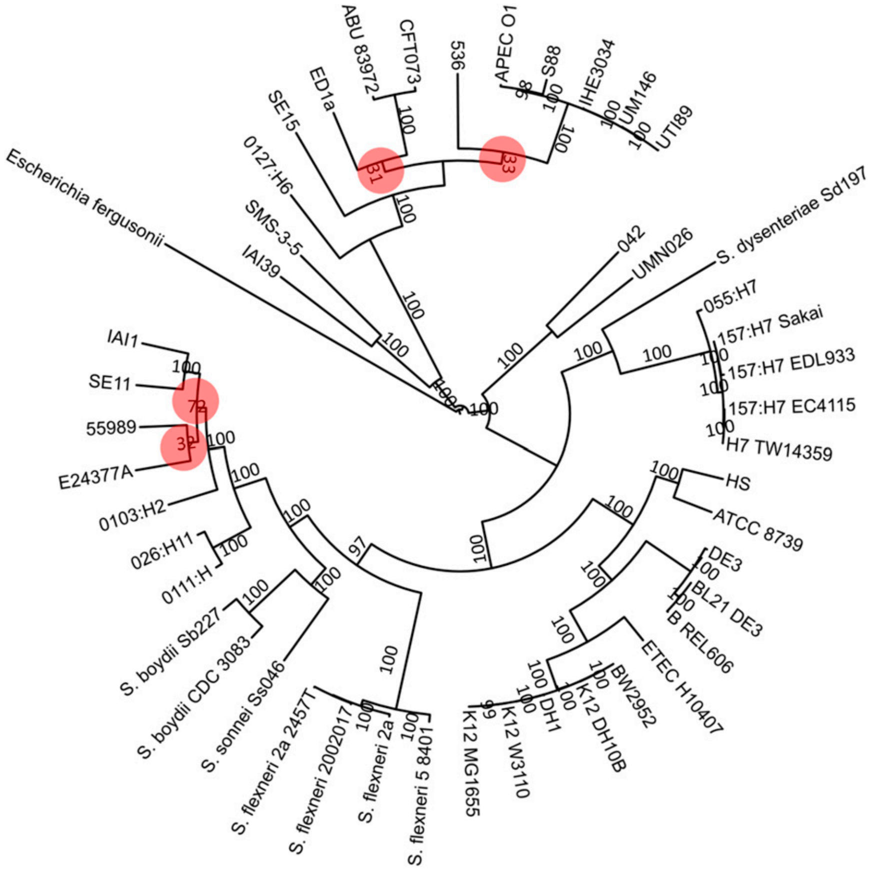

**Fig. S3.** *E. coli* species tree. The maximum-likelihood tree is based on 1,479 concatenated core genes. Statistical support for the internal branches was computed using bootstrapping with 100 repetitions. Branches with low support (below 95%) are marked by a red circle.

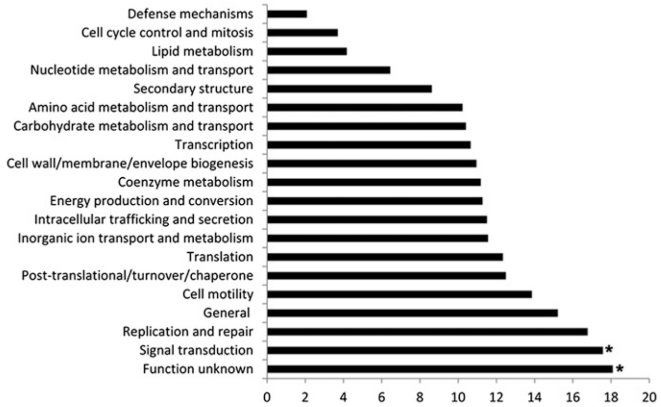

**Fig. S4.** Regulatory switching occurs in all gene functional groups. Bars represent the percentage of genes in each cluster of orthologous group category that exhibits regulatory switching in *E. coli* out of the total number of genes in each category. Asterisks indicate categories that are significantly enriched in genes that underwent regulatory switching ($P < 0.05$ after false discovery rate correction).

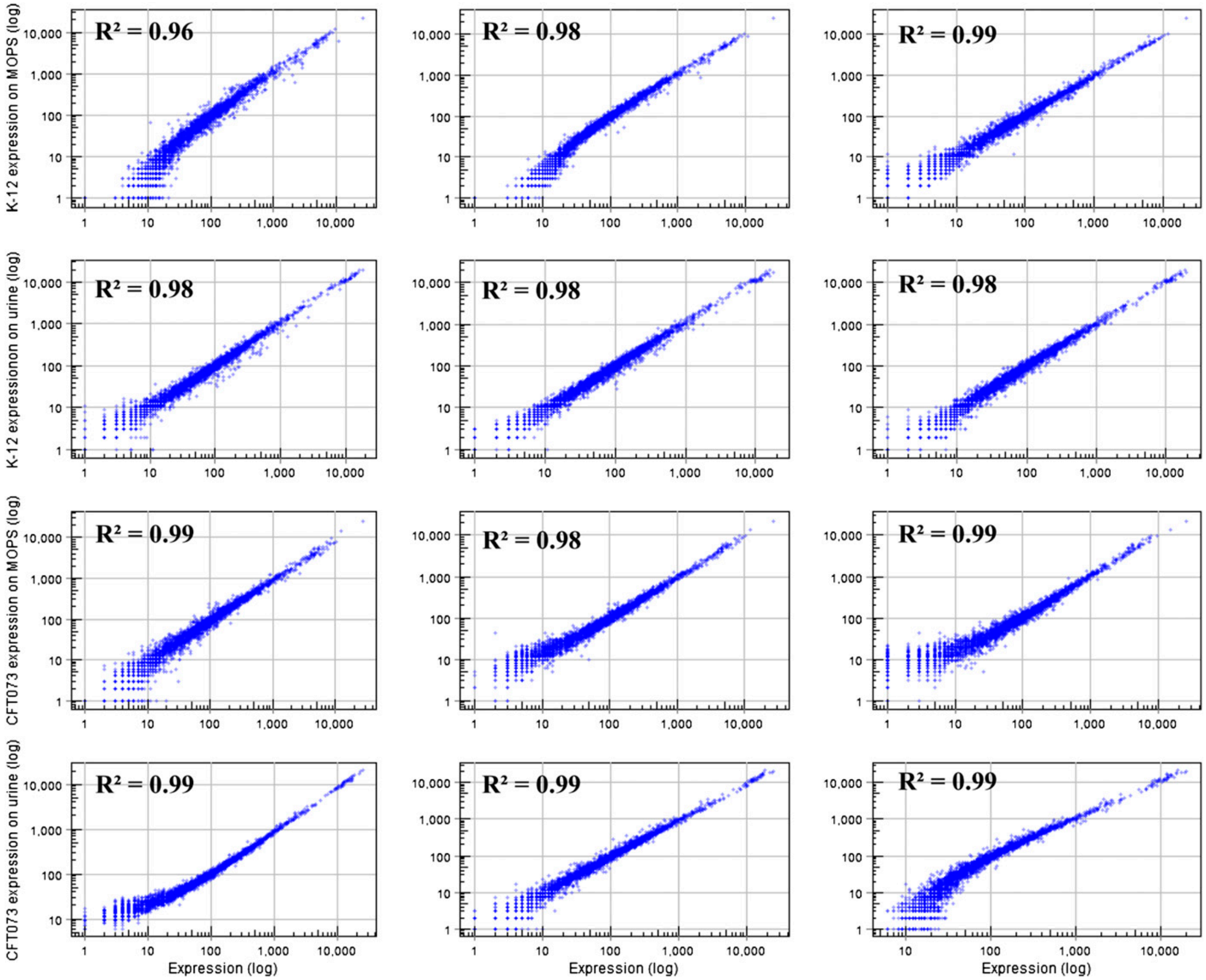

**Fig. S5.** Pairwise correlation between biological replicates of the RNA sequencing. The gene-wise correlation of transcriptomes for each growth condition (*E. coli* MG grown on MOPS and urine and *E. coli* CFT073 grown on MOPS and urine) across all biological replicates ($n = 3$) and their coefficient of determination are shown.

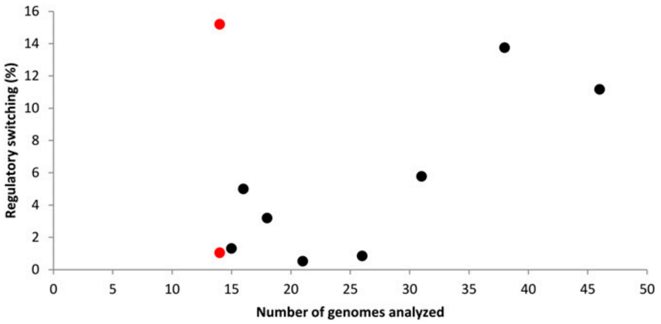

**Fig. S6.** Level of regulatory switching does not correlate with the number of genomes sampled. The relative percentage of regulatory switching is plotted against the number of genomes analyzed in each species. Marked in red are *N. meningitidis* (upper point) and *M. tuberculosis* (lower point). Although both *Neisseria* and *Mycobacterium* had 14 genomes sampled, the percentage of switching in *Neisseria* is 15-fold higher than in *Mycobacterium*. The Pearson correlation is not statistically different from 0 (asymptotic $P = 0.19$).

**Table S1. Regulatory clusters with altered transcription factor binding patterns**

| Protein GI | Known transcription factor binding sites that are only found in one of the regulatory types | Predicted transcription factor binding sites |
|---|---|---|
| 16130166 | CspA, Fis | Fnr |
| 16128459 | OxyR | dnaA, argP |
| 16130817 | ArgP | OxyR |
| 90111537 | exuR | OxyR |
| 16128575 | Fur*, RutR* | |
| 16129213 | Fur | GcvA |
| 16131596 | PhoB | NhaR |
| 16128788 | CRP, AscG | Fnr |
| 16130639 | NsrR | IHF |
| 16130585 | | OxyR[†] |
| 16128912 | | CspA[†] |
| 16131536 | | AraC[†], GntR[†] |
| 16130824 | ArgP, Lrp | OxyR, GcvA |
| 16128390 | Fis | lexA |
| 16131678 | MetJ* | |
| 49176358 | Lrp | OxyR |
| 16131856 | | AlgU[†,‡] |
| 16128961 | TorR | TorR[§] |

*Second type has a deletion in the specific transcription factor binding site compared with the *E. coli* K-12 strain.
[†] Second type has an insertion of the specified transcription factor binding site compared with the *E. coli* K-12 strain.
[‡] Transcription factor binding site is predicted based on the position-specific scoring matrix built for AlgU of *P. aeruginosa*.
[§] Regulatory types differ in the number of TorR binding sites (between two and four).

**Table S2. Transcription factors that underwent regulatory switching**

| Protein GI | Name | No. of target genes* |
|---|---|---|
| 49176356 | gntR | 12 |
| 16129236 | cysB | 24 |
| 16130817 | argP | 14 |
| 16128106 | pdhR | 42 |
| 16128961 | torr | 13 |
| 16130456 | iscR | 31 |
| 16130130 | narP | 64 |
| 16131539 | uhpA | 1 |
| 49176329 | nanR | 8 |
| 90111289 | marR | 3 |
| 90111537 | exuR | 8 |
| 16129295 | Fnr | 304 |
| 16130621 | ascG | 5 |
| 16131677 | metR | 6 |
| 49176012 | lacI | 3 |
| 90111079 | caiF | 10 |
| 90111679 | zur | 6 |
| 94541116 | yoeB | NA[†] |
| 145698338 | fabrR | 2 |

NA, not available.
*Genes directly regulated by the transcription factor. Data are taken from RegulonDB.
[†] No known target genes.

**Table S3. Characteristics of the bacterial species used in the regulatory divergence pipeline**

| Species | No. of genomes | No. of core genes | Nucleotide diversity* | Relative percent of core[†] | Gram | Lifestyle[‡] |
|---|---|---|---|---|---|---|
| *C. trachomatis* | 21 | 764 | 0.003 | 85 | − | Obligate intercellular parasite |
| *S. enterica* | 26 | 1,798 | 0.01 | 40 | − | Pathogen |
| *M. tuberculosis* | 14 | 1,720 | 0.0002 | 44 | + | Human pathogen |
| *C. pseudotuberculosis* | 15 | 1,372 | 0.005 | 66 | + | Ruminant pathogen |
| *S. pneumoniae* | 18 | 1,193 | 0.008 | 57 | + | Pathogen |
| *S. pyogenes* | 16 | 1,100 | 0.008 | 60 | + | Pathogen |
| *S. aureus* | 31 | 520 | 0.01 | 20 | + | Opportunistic pathogen |
| *H. pylori* | 38 | 365 | 0.03 | 24 | − | Opportunistic pathogen |
| *N. meningitidis* | 14 | 1,116 | 0.02 | 57 | − | Opportunistic pathogen |
| *E. coli* | 46 | 1,479 | 0.01 | 32 | − | Diverse |

−, Gram-negative bacteria; +, Gram-positive bacteria.

*Nucleotide diversity, π, measures the average of nucleotide differences per site between all pairwise sequences. The core genes of each bacterial species were concatenated and used as input for pipeline for diversity analyses.

[†] Number of core genes divided by the average genome size.

[‡] Data are taken from Integrated Microbial Genomes.

**Table S4. Recombination analysis**

| Species name | r/m ratio | No. of recombination events | Mean recombination segment length, bp |
|---|---|---|---|
| *C. trachomatis* | 0.66 | 95 | 10,005 |
| *S. enterica* | 0.14 | 554 | 660 |
| *M. tuberculosis* | 0 | 0 | 0 |
| *C. pseudotuberculosis* | 0 | 0 | 0 |
| *S. pneumoniae* | 5.17 | 960 | 1,970 |
| *S. pyogenes* | 3.07 | 546 | 3,031 |
| *S. aureus* | 0.58 | 380 | 1,131 |
| *E. coli* | 0.38 | 5,239 | 242 |
| *H. pylori* | 21.11 | 830 | 921 |
| *N. meningitidis* | 14.62 | 1,280 | 1,504 |

r/m ratio, recombination-to-mutation ratio.