Transfer of noncoding DNA drives regulatory rewiring in bacteria

The MIT Faculty has made this article openly available. Please share how this access benefits you. Your story matters.
Transfer of noncoding DNA drives regulatory rewiring in bacteria

Yaara Oren, Mark B. Smith, Nathan I. Johns, Millie Kaplan Zeevi, Dvora Biran, Eliora Z. Ron, Jukka Corander, Harris H. Wang, Eric J. Alm, and Tal Pupko

Departments of *Cell Research and Immunology and *Microbiology and Biotechnology, George S. Wise Faculty of Life Sciences, Tel Aviv University, Tel Aviv 69978, Israel; †Microbiology Graduate Program and ‡Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139; §Department of Systems Biology, Columbia University Medical Center, New York, NY 10032; ¶MIGAL, Galilee Research Institute, Kiryat Shmona 11016, Israel; and †Department of Mathematics and Statistics, University of Helsinki, Helsinki, Finland, 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 transcriptranscripts, 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.

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.
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. L1), 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. 1B). For the observed distribution to be explained by vertical transmission, multiple independent genomic rearrangement events with identical boundaries would have to be postulated, 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 phylogeny. 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...
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. 2A). 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. 2B). 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 changes 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. 3A). 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.

#### 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
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. 3B, 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. 3D). 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. 3E). 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. 3E).

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 auxotrophy (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. 3F). 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 recombining pathogen that causes meningitis and septicaemia. 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.

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 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 j-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$_2$O$_2$ at a final concentration of 1.5 mM was added to the culture and growth was monitored.

ACKNOWLEDGMENTS. We thank Siriram 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 20 823171-1, H.J.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.