Selection for Unequal Densities of σ^{70} Promoter-Like Signals in Different Regions of Large Bacterial Genomes

Araceli M. Huerta[1*], **M. Pilar Francino**[2], **Enrique Morett**[3], **Julio Collado-Vides**[1]

1 Centro de Ciencias Genómicas, Universidad Nacional Autónoma de México, Cuernavaca, México, **2** Evolutionary Genomics Program, Department of Energy Joint Genome Institute, and Genomics Division, Lawrence Berkeley National Laboratory, Walnut Creek, California, United States of America, **3** Departamento de Ingeniería Celular y Biocatálisis, Instituto de Biotecnología, Universidad Nacional Autónoma de México, Cuernavaca, México

The evolutionary processes operating in the DNA regions that participate in the regulation of gene expression are poorly understood. In *Escherichia coli,* we have established a sequence pattern that distinguishes regulatory from nonregulatory regions. The density of promoter-like sequences, that could be recognizable by RNA polymerase and may function as potential promoters, is high within regulatory regions, in contrast to coding regions and regions located between convergently transcribed genes. Moreover, functional promoter sites identified experimentally are often found in the subregions of highest density of promoter-like signals, even when individual sites with higher binding affinity for RNA polymerase exist elsewhere within the regulatory region. In order to see the generality of this pattern, we have analyzed 43 additional genomes belonging to most established bacterial phyla. Differential densities between regulatory and nonregulatory regions are detectable in most of the analyzed genomes, with the exception of those that have evolved toward extreme genome reduction. Thus, presence of this pattern follows that of genes and other genomic features that require weak selection to be effective in order to persist. On this basis, we suggest that the loss of differential densities in the reduced genomes of host-restricted pathogens and symbionts is an outcome of the process of genome degradation resulting from the decreased efficiency of purifying selection in highly structured small populations. This implies that the differential distribution of promoter-like signals between regulatory and nonregulatory regions detected in large bacterial genomes confers a significant, although small, fitness advantage. This study paves the way for further identification of the specific types of selective constraints that affect the organization of regulatory regions and the overall distribution of promoter-like signals through more detailed comparative analyses among closely related bacterial genomes.

Citation: Huerta AM, Francino MP, Morett E, Collado-Vides J (2006) Selection for unequal densities of σ^{70} promoter-like signals in different regions of large bacterial genomes. PLoS Genet 2(11): e185. doi:10.1371/journal.pgen.0020185

Introduction

For both prokaryotes and eukaryotes, understanding of the organizational structure and mode of evolution of regulatory DNA sequences is incomplete. Although in the Bacteria gene regulation involves fewer proteins and *cis*-regulatory DNA sites, and less complex interactions among them, recent findings suggest that the classic description of bacterial promoter regions may have been significantly oversimplified [1]. In *Escherichia coli,* RNA polymerase (RNAP) is composed of a core complex of α, β, β′, and ω subunits and one of a variety of σ factors, the primary one being σ^{70}, which is essential for general transcription in exponentially growing cells. The canonical model of the σ^{70} promoter is defined as a simple pair of hexamers, positioned at −35 and −10 base pairs (bp) from the transcription start (+1), with respective consensus sequences TTGACA and TATAAT, and separated by a spacer of 15 to 21 bp [2]. RNAP-σ^{70} can recognize and bind −35 and −10 motifs that differ substantially from their consensus sequences, although mutations that bring these motifs closer to the consensi generally increase promoter strength [3]. On average, *E. coli* promoters preserve only eight of the 12 canonical bases of the −35 and −10 hexamers [4,5]. It has been recently shown that most of the regulatory regions in *E. coli* do not contain a single promoter sequence [1] but rather display high densities of potential RNAP-σ^{70} binding sites, forming clusters of overlapping promoter-like signals. In contrast, such signal densities are not detected in coding regions and regions located between convergently transcribed genes. Moreover, functional promoter sites identified experimentally are often found within the subregions of highest density of overlapping signals, even when individual sites with higher binding affinity for RNAP exist elsewhere within the region [1].

Even though the degeneracy of the σ^{70}-binding promoter motifs ensures that new sites can evolve (i.e., appear and become fixed in populations) via local point mutation on short timescales [6], random fixation of mutations by neutral

Editor: Ivan Matic, INSERM U571, France

Received August 12, 2005; **Accepted** September 12, 2006; **Published** November 10, 2006

A previous version of this article appeared as an Early Online Release on September 12, 2006 (doi:10.1371/journal.pgen.0020185.eor).

Abbreviations: bp, base pair(s); PWM, position weight matrix; RNAP, RNA polymerase

* To whom correspondence should be addressed. E-mail: amhuerta@ccg.unam.mx

Synopsis

The most important step in the regulation of genetic expression is the initiation of transcription. This process is accomplished by the association or specific binding of RNA polymerase to particular sequence segments present in the DNA, the promoters. Promoters are located in the upstream regions of the transcribed genes. The evolutionary processes operating in the DNA regions that participate in the regulation of gene expression are poorly understood. For a long time, the canonical picture of a σ^{70} promoter has been a 60 base pair region defined by the transcription start-point (+1) and two conserved hexanucleotide sequences centered 10 and 35 base pairs upstream from the +1. The authors have shown that in *Escherichia coli,* promoters exist in clusters, as a series of overlapping potentially competing RNAP interaction sites. The *E. coli* regulatory regions contain high densities of these promoter-like signals, in contrast to coding regions and regions located between convergently transcribed genes. They report that the differential densities between regulatory and nonregulatory regions are detectable in most eubacterial genomes, with the exception of those that have experienced severe genome degradation and size reduction. This suggests that the presence of this pattern in large bacterial genomes confers a significant, although small, fitness advantage.

drift could not explain the different promoter-like signal densities between regulatory and nonregulatory regions of *E. coli* reported in [1]. Moreover, it has been shown that natural selection acts to remove spurious occurrences of the two consensus words of the σ^{70} promoter (TTGACA and TATAAT) from both coding and noncoding regions in several eubacterial genomes, implying that it is disadvantageous to maintain misplaced sites which can strongly bind RNAP-σ^{70} and interfere with proper gene expression [7]. This suggests that the observed excess of promoter-like signals in regulatory regions is likely to be the result of natural selection for some past or present function.

Here we report that differential densities of promoter-like signals between regulatory and nonregulatory regions are detectable in most bacterial genomes, with the exception of those that have experienced severe size reduction. We argue that the phylogenetic distribution of this differential density pattern implies that this genomic feature is maintained by weak natural selection, and we discuss possible functional roles for the high redundancy of promoter-like signals in the regulatory regions of large bacterial genomes.

Results/Discussion

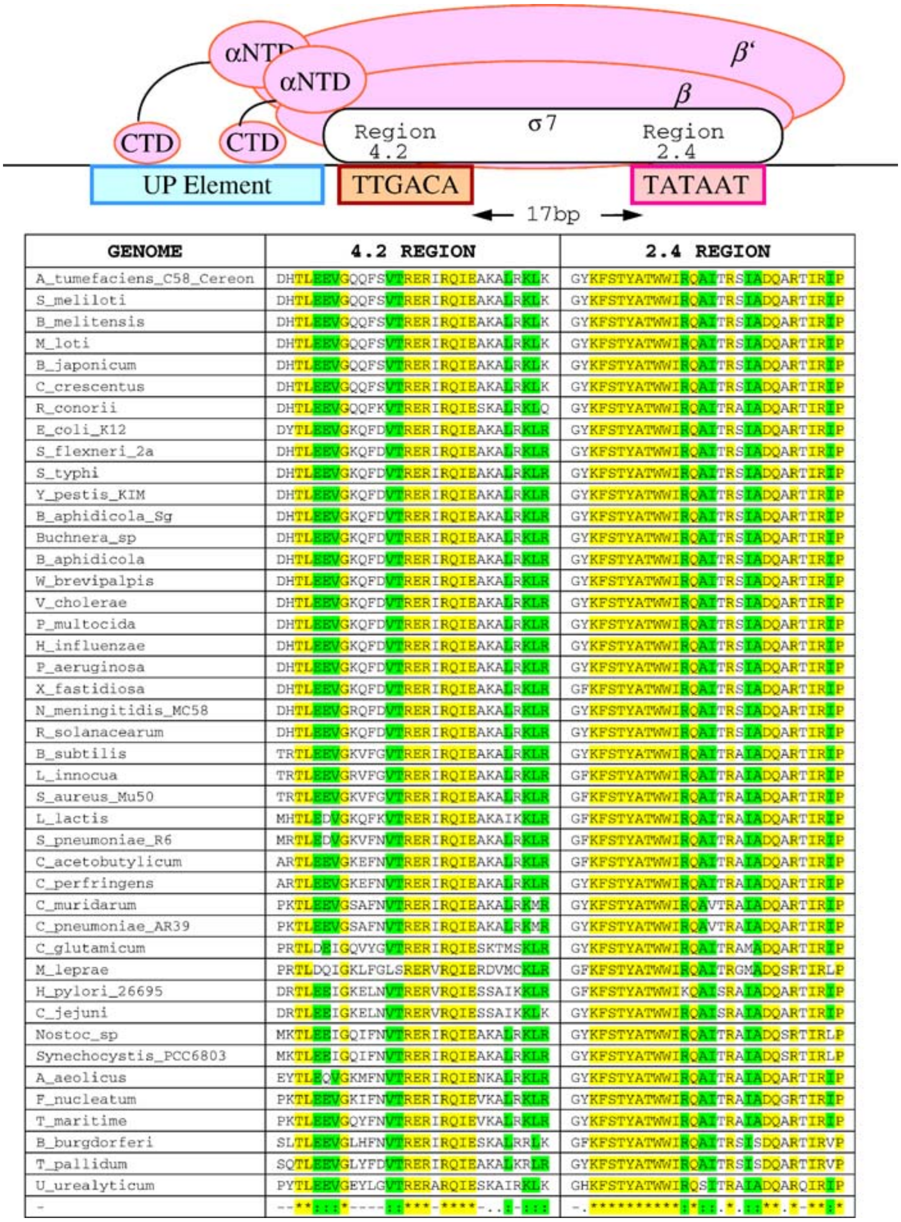
In order to explore whether the differential promoter signal density pattern discovered in *E. coli* is common to other bacterial species, we conducted similar analyses for a representative set containing 43 additional genomes belonging to different genera across all major bacterial phyla. This comparison is valid given that RNAP is evolutionarily conserved across Bacteria. Moreover, there seems to be only one housekeeping σ^{70} factor present in any given species [8,9], and all eubacterial σ^{70} protein sequences can be clearly aligned and contain highly similar motifs for the recognition and binding of −10 and −35 promoter sequences [10] (Figure 1). This implies that the DNA sequences of promoter motifs in these organisms must also be similar to those found in *E. coli.* Therefore, we used in our searches position weight matrices (PWMs) describing the −10 and −35 sequences determined in *E. coli* [1] and specifically calibrated to the base composition of strictly noncoding regions of each analyzed genome (Figure 2).

Table 1 reports for every genome the consensus and average score of the detected −10 and −35 motifs. The consensi were obtained after applying the COVER Function [1] on the set of promoter-like signals found by the PWM search in the functional regulatory regions of each genome (see Materials and Methods). From Table 1, it can be seen that all the additional 43 species analyzed have consensus sequences for both motifs highly similar to those of *E. coli,* as well as average scores of comparable magnitudes. In fact, large GC-rich genomes (*Pseudomonas, Ralstonia,* and most of the alpha-proteobacteria) display motif scores above those of *E. coli,* probably due to the greater compositional difference between the AT-rich motifs and the background genomic sequence. In contrast, small GC-poor genomes have lower motif scores, with the insect endosymbiont *Wigglesworthia* displaying the lowest.

Table 2 shows the promoter-like signal density patterns detected for all the genomes analyzed. Two main alternative profiles were obtained, as illustrated in Figure 3 (profiles for all species analyzed are available in Figure S1 and at http://www.ccg.unam.mx/Computational_Genomics/PromotorTools/PLoS06/Figure_S1.hmtl). Regulatory and nonregulatory regions contain differential densities of promoter-like signals in 24 genomes, including genera belonging to phyla distantly related to *E. coli,* such as the *Firmicutes, Actinobacteria, Cyanobacteria,* and *Thermotogae.* Clearly, the presence of the pattern is highly dependent on genome size (Pearson's correlation = 0.59, $p < 0.001$). All genomes above 4 Mb display marked differences in promoter-like signals between regulatory and nonregulatory regions, whereas none of the genomes under 1.5 Mb do so. Among the genomes of intermediate size, the pattern is detectable in 65% of the cases. There is also a notable effect of GC content (Pearson's correlation = 0.58, $p < 0.001$). Although the differential signal densities can be seen in genomes of very different GC contents (from 30% GC *Lactococcus lactis* to 62% GC *Ralstonia solanacearum* and *Caulobacter crescentus*), overall, 75% of GC-rich genomes display the differential pattern, in contrast to 53% of those under 50% GC. Multiple regression analysis reveals that up to 0.43 of the variance in pattern overrepresentation might be related to these two variables, genome size and GC content (for a correlation of about 0.66).

To test if the accumulation of promoter-like signals observed in regulatory regions is due to AT-richness of these sequences, we generated randomized regulatory sequences (see Materials and Methods) and we searched for promoter-like signals. As expected, the randomized regulatory sequences of large genomes show flattened distributions with significantly less promoter-like signals than real regulatory regions ($p < 0.001$), meaning that observed densities in regulatory regions are not due to an AT bias. Results for random regulatory regions are available in Figure S2 and at http://www.ccg.unam.mx/Computational_Genomics_PromotorTools/PLoS06/Figure_S2.html.

The observations here reported strongly suggest that the differential density of promoter-like signals in regulatory and nonregulatory regions of large bacterial genomes is main-



GENOME	4.2 REGION	2.4 REGION
A_tumefaciens_C58_Cereon	DHTLEEVGQQFSVTRERIRQIEAKALRKLK	GYKFSTYATWWIRQAITRSIADQARTIRIP
S_meliloti	DHTLEEVGQQFSVTRERIRQIEAKALRKLK	GYKFSTYATWWIRQAITRSIADQARTIRIP
B_melitensis	DHTLEEVGQQFSVTRERIRQIEAKALRKLK	GYKFSTYATWWIRQAITRSIADQARTIRIP
M_loti	DHTLEEVGQQFSVTRERIRQIEAKALRKLK	GYKFSTYATWWIRQAITRSIADQARTIRIP
B_japonicum	DHTLEEVGQQFSVTRERIRQIEAKALRKLK	GYKFSTYATWWIRQAITRSIADQARTIRIP
C_crescentus	DHTLEEVGQQFSVTRERIRQIEAKALRKLK	GYKFSTYATWWIRQAITRSIADQARTIRIP
R_conorii	DHTLEEVGQQFKVTRERIRQIESKALRKLQ	GYKFSTYATWWIRQAITRAIADQARTIRIP
E_coli_K12	DYTLEEVGKQFDVTRERIRQIEAKALRKLR	GYKFSTYATWWIRQAITRSIADQARTIRIP
S_flexneri_2a	DHTLEEVGKQFDVTRERIRQIEAKALRKLR	GYKFSTYATWWIRQAITRSIADQARTIRIP
S_typhi	DHTLEEVGKQFDVTRERIRQIEAKALRKLR	GYKFSTYATWWIRQAITRSIADQARTIRIP
Y_pestis_KIM	DHTLEEVGKQFDVTRERIRQIEAKALRKLR	GYKFSTYATWWIRQAITRSIADQARTIRIP
B_aphidicola_Sg	DHTLEEVGKQFDVTRERIRQIEAKALRKLR	GYKFSTYATWWIRQAITRSIADQARTIRIP
Buchnera_sp	DHTLEEVGKQFDVTRERIRQIEAKALRKLR	GYKFSTYATWWIRQAITRSIADQARTIRIP
B_aphidicola	DHTLEEVGKQFDVTRERIRQIEAKALRKLR	GYKFSTYATWWIRQAITRSIADQARTIRIP
W_brevipalpis	DHTLEEVGKQFDVTRERIRQIEAKALRKLR	GYKFSTYATWWIRQAITRSIADQARTIRIP
V_cholerae	DHTLEEVGKQFDVTRERIRQIEAKALRKLR	GYKFSTYATWWIRQAITRSIADQARTIRIP
P_multocida	DHTLEEVGKQFDVTRERIRQIEAKALRKLR	GYKFSTYATWWIRQAITRSIADQARTIRIP
H_influenzae	DHTLEEVGKQFDVTRERIRQIEAKALRKLR	GYKFSTYATWWIRQAITRSIADQARTIRIP
P_aeruginosa	DHTLEEVGKQFDVTRERIRQIEAKALRKLR	GYKFSTYATWWIRQAITRSIADQARTIRIP
X_fastidiosa	DHTLEEVGKQFDVTRERIRQIEAKALRKLR	GFKFSTYATWWIRQAITRSIADQARTIRIP
N_meningitidis_MC58	DHTLEEVGRQFDVTRERIRQIEAKALRKLR	GYKFSTYATWWIRQAITRSIADQARTIRIP
R_solanacearum	DHTLEEVGKQFDVTRERIRQIEAKALRKLR	GYKFSTYATWWIRQAITRSIADQARTIRIP
B_subtilis	TRTLEEVGKVFGVTRERIRQIEAKALRKLR	GYKFSTYATWWIRQAITRAIADQARTIRIP
L_innocua	TRTLEEVGRVFGVTRERIRQIEAKALRKLR	GFKFSTYATWWIRQAITRAIADQARTIRIP
S_aureus_Mu50	TRTLEEVGKVFGVTRERIRQIEAKALRKLR	GFKFSTYATWWIRQAITRAIADQARTIRIP
L_lactis	MHTLEDVGKQFKVTRERIRQIEAKAIKKLR	GFKFSTYATWWIRQAITRAIADQARTIRIP
S_pneumoniae_R6	MRTLEDVGKVFNVTRERIRQIEAKALRKLR	GFKFSTYATWWIRQAITRAIADQARTIRIP
C_acetobutylicum	ARTLEEVGKEFNVTRERIRQIEAKALRKLR	GFKFSTYATWWIRQAITRAIADQARTIRIP
C_perfringens	ARTLEEVGKEFNVTRERIRQIEAKALRKLR	GYKFSTYATWWIRQAITRAIADQARTIRIP
C_muridarum	PKTLEEVGSAFNVTRERIRQIEAKALRKMR	GYKFSTYATWWIRQAVTRAIADQARTIRIP
C_pneumoniae_AR39	PKTLEEVGSAFNVTRERIRQIEAKALRKMR	GYKFSTYATWWIRQAVTRAIADQARTIRIP
C_glutamicum	PRTLDEIGQVYGVTRERIRQIESKTMSKLR	GYKFSTYATWWIRQAITRAMADQARTIRIP
M_leprae	PRTLDQIGKLFGLSRERVRQIERDVMCKLR	GFKFSTYATWWIRQAITRGMADQSRTIRLP
H_pylori_26695	DRTLEEIGKELNVTRERVRQIESSAIKKLR	GFKFSTYATWWIKQAISRAIADQARTIRIP
C_jejuni	DRTLEEIGKELNVTRERVRQIESSAIKKLK	GYKFSTYATWWIRQAISRAIADQARTIRIP
Nostoc_sp	MKTLEEIGQIFNVTRERIRQIEAKALRKLR	GYKFSTYATWWIRQAITRAIADQSRTIRLP
Synechocystis_PCC6803	MKTLEEIGQIFNVTRERIRQIEAKALRKLR	GYKFSTYATWWIRQAITRAIADQSRTIRLP
A_aeolicus	EYTLEQVGKMFNVTRERIRQIENKALRKLR	GYKFSTYATWWIRQAITRAIADQARTIRIP
F_nucleatum	PKTLEEVGKIFNVTRERIRQIEVKALRKLR	GYKFSTYATWWIRQAITRAIADQGRTIRIP
T_maritime	PKTLEEVGQYFNVTRERIRQIEVKALRKLR	GYKFSTYATWWIRQAITRAIADQARTIRIP
B_burgdorferi	SLTLEEVGLHFNVTRERIRQIESKALRRLK	GFKFSTYATWWIRQAITRSISDQARTIRVP
T_pallidum	SQTLEEVGLYFDVTRERIRQIEAKALKRLR	GYKFSTYATWWIRQAITRSISDQARTIRVP
U_urealyticum	PYTLEEVGEYLGVTRERARQIESKAIRKLK	GHKFSTYATWWIRQSITRAIADQARQIRIP
-	--**:::*----::***-****-..:-:::	-.**********:*::.*.::**.*-**:*

Figure 1. Schematic Representation of the Main Interactions of RNAP with Promoter DNA and Alignment of the σ^{70} Motifs for Recognition and Binding of –10 (2.4 Region) and –35 Promoter Sequences (4.2 Region) for Representative Eubacteria

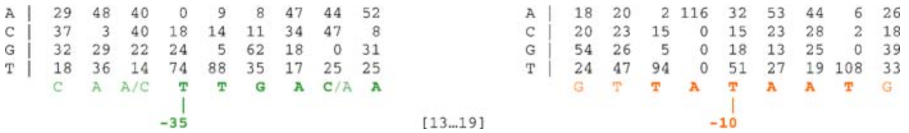
Sequence alignments of several σ^{70} factors from different bacteria reveal four conserved regions that can be further divided into subregions [39]. Only regions 2 and 4 are well conserved in all members of the σ^{70} family [40–43] and include subregions involved in binding to the core RNAP complex, promoter melting, and recognition of the –10 and –35 promoter sequences (regions 2.4 and 4.2, respectively) [10,40,44]. CLUSTALW was used to generate the alignment with default parameters (http://www.ebi.ac.uk/clustalw) [45].

doi:10.1371/journal.pgen.0020185.g001

tained by natural selection. First, the fact that this pattern is observed in phylogenetically distant genomes argues against mutational biases being its main source, since mutational biases are known to vary among genomes, particularly when there are large differences in GC content. Second, differential signal density is highly dependent on large genome size, and is completely absent from most of the small genomes of animal parasites and symbionts with an intracellular or predominantly host-restricted lifestyle (*Mycoplasma, Ureaplasma, Treponema, Borrelia, Campylobacter, Rickettsia, Buchnera,* and *Wigglesworthia*). And third, analysis on conservation of promoter-like signals in a group of 14 related species shows 79% signal conservation (see Materials and Methods).

Degradation of Regulatory Functions in the Small Genomes of Host-Restricted Bacteria

Acquisition of such obligate dependence on a host is known to have many deleterious consequences, collectively known as genome degradation or genome reduction [11]. Reduced genomes have independently evolved many times within several bacterial phyla, repeatedly undergoing rapid sequence evolution and acquiring extremely low GC content,



```
A |  29  48  40   0   9   8  47  44  52        A |  18  20   2 116  32  53  44   6  26
C |  37   3  40  18  14  11  34  47   8        C |  20  23  15   0  15  23  28   2  18
G |  32  29  22  24   5  62  18   0  31        G |  54  26   5   0  18  13  25   0  39
T |  18  36  14  74  88  35  17  25  25        T |  24  47  94   0  51  27  19 108  33
      C   A A/C   T   T   G   A C/A   A              G   T   T   A   T   A   A   T   G
                  |                                                  |
                -35                   [13…19]                       -10
```

Figure 2. Frequency Matrices for the −10 and −35 Motifs of σ^{70} Promoters in *E. coli*

This matrix pair (Matrix_18_15_13_2_1.5) was selected for searching across bacterial genomes from a collection of optimized matrices defined for *E. coli* in [1]. Note that in order to compare these matrices with the canonical patterns (TTGACA and TATAAT), the spacers of 13 bp to 19 bp between the two boxes correspond to the 15 bp to 21 bp reported in the literature, as the TGT triplet is considered as part of the −10 box. Before searching for promoter-like signals, these matrices were calibrated using the noncoding base composition of each target genome.
doi:10.1371/journal.pgen.0020185.g002

often with clearly maladaptive consequences, including accumulation of deleterious amino acid substitutions and loss of adaptive codon biases [12,13]. Typical changes also include a large increase in the frequency of mobile elements in the early phases of genome degradation, chromosomal rearrangements mediated by recombination among these elements, pseudogene formation, and deletions of varying size. There is recent evidence that genome degradation also

Table 1. −10 and −35 Consensi and Corresponding Average Scores for σ^{70} Promoter-Like Signals in Representative Bacterial Genomes

NCBI Accession Number	Organisms	−10 Box Consensus	−10 Average Score	−35 Box Consensus	−35 Average Score
NC_000853	*Thermotoga maritima*	GTTATAAT	2.26	AACTTGAAA	1.53
NC_000907	*Haemophilus influenzae*	GTTAAAAT	2.15	AAATTGAAA	1.42
NC_000908	*Mycoplasma genitalium*	GTTAAAAT	1.86	AAATTGAAA	1.43
NC_000911	*Synechocystis* PCC6803	GTTAAAAT	2.62	AAATTGAAA	1.55
NC_000912	*Mycoplasma pneumoniae*	ATTAAAAT	2.26	C|TTATTTAAA	1.58
NC_000913	*Escherichia coli* K12	GTTATAAT	2.79	AAATTGAAA	1.77
NC_000915	*Helicobacter pylori* 26695	TTTAAAAT	2.44	AAATTTAAA	1.39
NC_000918	*Aquifex aeolicus*	TTTAAAAT	2.52	ATCTTTAAA	1.47
NC_000919	*Treponema pallidum*	GGTATAAT	2.52	CTCTTGACG	1.54
NC_000964	*Bacillus subtilis*	GTTATAAT	2.77	ATATTGAAA	1.66
NC_001318	*Borrelia burgdorferi*	TTTATAAT	1.94	AAATTTAAA	1.33
NC_002162	*Ureaplasma urealyticum*	TTTAAAAT	1.68	AAATTTAAA	1.26
NC_002163	*Campylobacter jejuni*	TTTAAAAT	2.41	AAATTTAAA	1.34
NC_002179	*Chlamydophila pneumoniae* AR39	TTTATAAT	2.34	ATATTTAAA	1.48
NC_002488	*Xylella fastidiosa*	GTTATAAT	2.53	CAATTGAAA	1.70
NC_002505	*Vibrio cholerae*	GTTAAAAT	2.35	CAATTGAAA	1.57
NC_002516	*Pseudomonas aeruginosa*	GTTATAAT	3.75	CTCTTGAAA	2.51
NC_002620	*Chlamydia muridarum*	TTTATAAT	2.29	ATATTGAAA	1.44
NC_002662	*Lactococcus lactis*	GTTAAAAT	2.28	AAATTTAAA	1.44
NC_002663	*Pasteurella multocida*	GTTATAAT	2.40	AAATTGAAA	1.50
NC_002677	*Mycobacterium leprae*	GTTAAAAT	2.90	CAATTGACA	1.60
NC_002678	*Mesorhizobium loti*	ATCAATA|CT	2.98	AACTTGACA	1.95
NC_002696	*Caulobacter crescentus*	GTTATCAT	3.27	CGCTTGACG	2.00
NC_002758	*Staphylococcus aureus* Mu50	TTTATAAT	2.21	AAATTTAAA	1.33
NC_003030	*Clostridium acetobutylicum*	TTTATAAT	2.14	AAATTTAAA	1.24
NC_003047	*Sinorhizobium meliloti*	GTTATAAT	2.99	CGCTTGAAA	2.17
NC_003062	*Agrobacterium tumefaciens* C58	GTTATAAT	2.67	CTA|CTTGACA	1.96
NC_003098	*Streptococcus pneumoniae* R6	GTTATAAT	2.43	AAATTGAAA	1.55
NC_003103	*Rickettsia conorii*	GTTATAAT	2.06	AAATTTAAA	1.38
NC_003112	*Neisseria meningitidis* MC58	GTTAAAAT	2.79	AAATTGAAA	1.76
NC_003198	*Salmonella typhi*	GTTATAAT	2.99	AAATTGAAA	1.94
NC_003212	*Listeria innocua*	GTTATAAT	2.36	AAATTGAAA	1.41
NC_003272	*Nostoc* sp.	GTTAAAAT	2.34	AAATTGAAA	1.53
NC_003295	*Ralstonia solanacearum*	GTTATAAT	3.75	CAATTGAAG	2.25
NC_003317	*Brucella melitensis*	GTTATAAT	2.87	AAATTGAAA	1.97
NC_003366	*Clostridium perfringens*	TTTATAAT	1.94	AAATTTAAA	1.17
NC_003450	*Corynebacterium glutamicum*	GTTAAAAT	2.77	AAATTGAAA	1.80
NC_003454	*Fusobacterium nucleatum*	TTTATAAT	1.88	AAATTTAAA	1.21
NC_004088	*Yersinia pestis* KIM	GTTATAAT	2.73	CAATTGAAA	1.66
NC_004337	*Shigella flexneri* 2a	GTTATAAT	2.96	CTATTGAAA	1.82
NC_004344	*Wigglesworthia brevipalpis*	TTTATAAT	1.55	AAATTTAAA	0.67
NC_004463	*Bradyrhizobium japonicum*	GCTAAAAT	3.32	AACTTGACA	2.03
NC_004545	*Buchnera aphidicola*	TTTAAAAT	1.93	AAATTTAAA	0.74

Promoter-signal searches were performed with *E. coli* Matrix_18_15_13_2_1.5 (Figure 2), calibrated with the base composition of the strictly noncoding regions of the corresponding genome, as described in Materials and Methods. Consensi and average scores are reported for the subset of promoter-like signals most likely to contain functional promoters within each target genome.
doi:10.1371/journal.pgen.0020185.t001

Table 2. Density Patterns of σ^{70} Promoter-Like Signals in Eubacterial Genomes

Species	Genome Size (Mb)	% GC in Noncoding DNA	% Overall Similarity to *E. coli*	% Identity to *E. coli* rpoD	Overrepresentation of Signals	% Genes with Clusters	Signals by Cluster	% Signals in Clusters
M. genitalium	0.58	33	28	—	No	—	—	—
B. aphidicola	0.62	18	57	74	No	—	—	—
W. brevipalpis	0.70	15	55	73	No	—	—	—
U. urealyticum	0.75	23	28	49	No	—	—	—
M. pneumoniae	0.82	34	27	—	No	—	—	—
B. burgdorferi	0.91	23	29	31	No	—	—	—
C. muridarum	1.07	37	28	38	No	—	—	—
T. pallidum	1.14	55	27	36	No	—	—	—
C. pneumoniae AR39	1.23	35	28	40	No	—	—	—
R. conorii	1.27	31	32	45	No	—	—	—
A. aeolicus	1.55	41	30	38	No	—	—	—
C. jejuni	1.64	25	31	38	No	—	—	—
H. pylori 26695	1.67	32	30	37	Yes	100	3.66	88
H. influenzae	1.83	34	57	67	No	—	—	—
T. maritime	1.86	42	28	57	Yes	88	2.76	75
S. pneumoniae R6	2.04	34	30	65	Yes	98	3.53	83
B. melitensis	2.12	50	34	49	Yes	77	2.35	72
F. nucleatum	2.17	25	31	60	No	—	—	—
P. multocida	2.26	35	56	71	Yes	99	3.64	85
N. meningitidis MC58	2.27	46	42	52	Yes	90	2.91	78
L. lactis	2.37	30	31	61	Yes	99	3.59	85
X. fastidiosa	2.68	47	40	59	Yes	83	2.65	75
A. tumefaciens C58 Cereon	2.84	53	34	47	Yes	70	2.13	70
S. aureus Mu50	2.88	29	31	61	No	—	—	—
V. cholerae	2.96	43	54	76	Yes	94	3.21	84
L. innocua	3.01	34	32	61	Yes	98	3.52	83
C. perfringens	3.03	24	31	68	No	—	—	—
M. leprae	3.27	54	29	48	No	—	—	—
C. glutamicum	3.31	48	29	54	Yes	89	3.03	81
Synechocystis PCC6803	3.57	42	29	59	Yes	96	3.31	81
S. meliloti	3.65	58	34	48	Yes	62	1.82	65
R. solanacearum	3.72	62	40	57	Yes	63	1.82	70
C. acetobutylicum	3.94	27	31	63	No	—	—	—
C. crescentus	4.02	62	32	49	Yes	43	1.15	55
B. subtilis	4.21	38	32	68	Yes	90	2.84	76
Y. pestis KIM	4.60	42	70	91	Yes	97	3.76	84
S. flexneri 2a	4.61	44	97	100	Yes	95	3.38	82
E. coli K12	4.64	43	100	100	Yes	92	3.12	80
S. typhi	4.81	44	87	98	Yes	97	3.32	84
P. aeruginosa	6.26	61	45	66	Yes	60	1.79	69
Nostoc sp.	6.41	36	28	60	Yes	98	3.63	84
M. loti	7.04	58	33	48	Yes	55	1.53	60
B. japonicum	9.11	60	33	46	Yes	58	1.71	64

Species are ordered by genome size to highlight the impact of this character on the presence of the differential density pattern. The overall similarity to *E. coli* is a measure based on the similarities of all orthologs between two genomes [33,46]. The values in the last three columns are based on the most likely functional promoters selected by COVER and were obtained only for those genomes showing an excess of signals in regulatory versus convergent noncoding regions by a log-likelihood test ($p < 0.001$).

doi:10.1371/journal.pgen.0020185.t002

affects gene regulation, due to losses of certain promoter sequences, specialized σ factors, and regulatory proteins [14–16]. These common changes likely reflect a diminished capacity of reduced genomes to respond to natural selection, conducive to the accumulation of all types of moderately deleterious mutations that would normally be purged from the genomes of free-living bacteria. This lowered effectiveness of purifying selection is due to the subdivided population structure imposed by confinement within a host, which limits the effective size of the bacterial population by subjecting it to recurrent bottlenecks and by thwarting opportunities for recombination with close relatives [17]. In addition, different types of molecular evolutionary analyses indicate that the rate of generation of point mutations is accelerated in reduced genomes [18,19].

We argue that a decrease in the efficiency of purifying selection in host-restricted bacteria allows promoter-like signals to rapidly accumulate all along the genomic sequence of these organisms, causing the loss of differential signal density patterns. This interpretation is based on the following evidence: (i) simulation results demonstrating that transcription factor binding sites can rapidly appear as a consequence of local point mutations on short time scales without invoking selection [6]; (ii) the observation that natural selection can act to remove spurious transcription factor binding sites from nonregulatory regions in many bacterial genomes, with a weak strength similar to that of

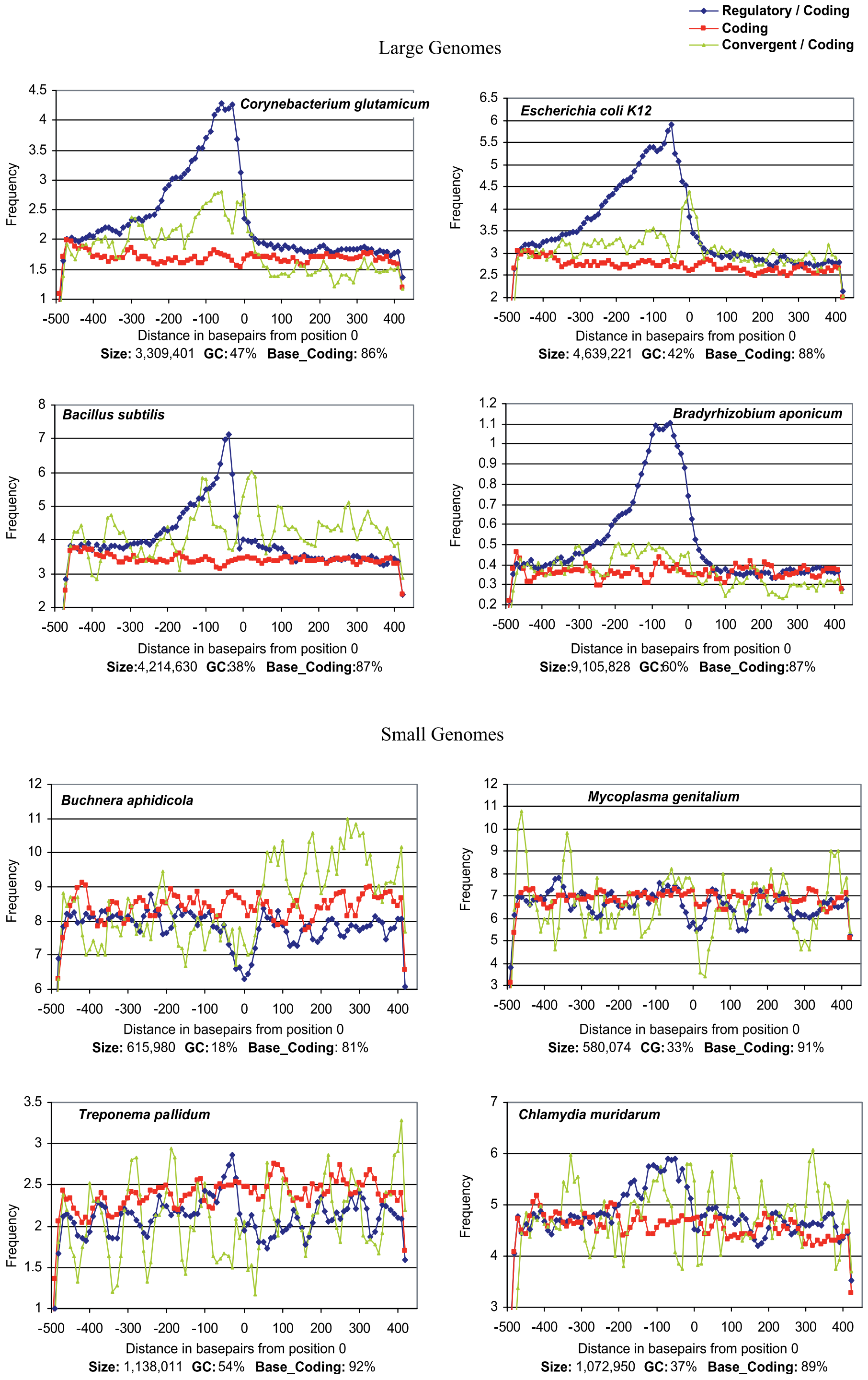


Figure 3. Signal Density in Regulatory versus Nonregulatory Regions of Large and Small Eubacterial Genomes

Regulatory regions correspond to the strictly noncoding regions located upstream of a gene start. For the set of genomes selected in this study, the average size of the strictly noncoding upstream regions is 182 bp. For the sake of the graph, the regions were extended to 500 bases upstream and 500 bases downstream of the start of the gene (position 0). Nonregulatory regions include the coding regions and the noncoding regions between convergently transcribed genes. For coding regions, the middle point of a gene was taken as the position 0 and 500 bases upstream and 500 bases downstream of this position were included. For the set of genomes analyzed here, the average size of the convergent regions is 194.5 bp. For the sake of the graph, the end of the 3′ gene was taken as position 0 and 500 bases upstream and 500 bases downstream of this position were included. The number of signals was averaged within intervals of 10 bp.

doi:10.1371/journal.pgen.0020185.g003

selection on adaptive codon bias [7]; (iii) empirical and theoretical results demonstrating that the effectiveness of selection is diminished in the small, highly structured populations of host-restricted bacteria [11–17]; (iv) the finding that promoter-like sequences in these organisms show a certain degree of degradation reflected in their decreased scores for the −10 and −35 motifs (Table 2), suggesting an overall reduction of the efficiency of selection

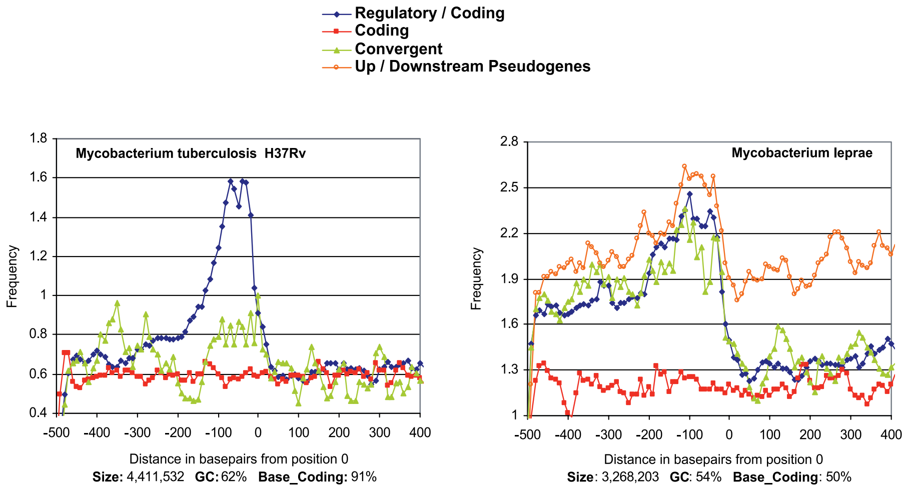


Figure 4. Signal Density in Regulatory versus Nonregulatory Regions of *M. tuberculosis* and *M. leprae*

M. leprae shows an increase of promoter-like signals in the upstream and downstream regions of pseudogenes relative to the regulatory regions. *M. leprae* contains over 1,115 recognizable pseudogenes [20]. Both 500 bp upstream and downstream of the start of pseudogenes (position 0) were analyzed for searches of promoter-like signals. All other region definitions and methodology are as in Figure 3.

doi:10.1371/journal.pgen.0020185.g004

on regulatory function; and (v) the fact that the nonregulatory regions of host-restricted bacteria present a frequency of promoter-like signals similar to the highest frequency peak found inside the regulatory regions of their commensal or free-living relatives (Figure 3). Figure 3 shows the frequency of promoter-like signals in *Buchnera aphidicola, Mycoplasma genitalium,* and their relatives *E. coli* and *Bacillus subtilis*. The increased frequency of promoter-like signals could have regulatory implications by making it more difficult for RNAP to correctly identify regulatory regions and the functional promoters within them. In that case, one could expect a slow response to changes in cellular conditions and/or a decreased level of gene expression.

To further test whether genome degradation produces loss of differential signal densities, we decided to compare signal density patterns between a genome that has recently entered a process of degradation and close relatives evolving under stronger purifying selection. Our initial analyses pinpointed *Mycobacterium leprae,* the obligate intracellular pathogen that causes leprosy, as one of the rare genomes above 3 Mb for which differential promoter signal densities are not detected. However, although the genome of *M. leprae* remains relatively large (3.27 Mb) and GC-rich (58% overall), it is clearly decreasing in size and GC content relative to its closest relative, *M. tuberculosis* (4.41 Mb and 65.6% overall GC). Genome comparisons between different mycobacterial species indicate that *M. leprae* has also undergone massive gene decay by pseudogenization and deletion, as well as numerous genomic rearrangements likely due to the proliferation of IS elements [20]. As we expected, analyses of promoter-like signals in two strains of *M. tuberculosis* do reveal the differential signal densities characteristic of large genomes (Figure 4 and Table 3). Both *M. tuberculosis* strains display higher average scores than *M. leprae* for the −35 and −10 promoter motifs (Table 3), supporting the idea that the leprosy bacillus is undergoing a general degradation of its regulatory sequences. Preliminary comparative analyses between the upstream regions of pseudogenes in *M. leprae* and the regulatory regions of *M. leprae* and *M. tuberculosis* reveal an evident increase of promoter-like signals in the upstream regions of pseudogenes (Figure 4). Studies about mechanisms that prevent the unnecessary expression of degenerated genes suggest that the accumulation of mutations in ribosome-binding sites and promoter regions provoke a translational and transcriptional silencing of pseudogenes [21]. Overall, the promoter-like signals upstream of *M. leprae* pseudogenes still retain the conserved −35 and −10 conseni of σ^{70} promoters ("TTGACA" and "TATAAT", respectively), indicating that promoter degeneration is not yet pervasive across the pseudogenes of this species. However, the average scores of the −35 and −10 motifs of promoters in pseudogenes are below the average scores of the promoters found in the regulatory regions, especially for the −35 motif (Table 3), implying a certain level of degradation of these signals. It is known that the −35 box is dispensable for transcription initiation as long as there is a regulatory mechanism to anchor the polymerase in the promoter, which can be achieved by regulatory proteins or extended −10 boxes. We suggest that excessive density of lightly degraded promoter-like signals may be silencing pseudogenes in two ways: (i) by attracting the polymerase to sites that could be competing with the "real" promoter and/or (ii) by presenting polymerase binding sites where the −35 degraded box is not sufficient to support open complex formation. Table S1 in Supporting Information shows selected comparisons between *M. leprae* pseudogenes and their corresponding functional equivalents in *M. tuberculosis*.

Potential Functions of the Promoter-Like Signals

In the course of our analyses, we identify two types of promoter-like signal densities: a global density and a local density (Figures 4 and 7 in [1]). The global density is obtained during the search for the −10 and −35 motifs using PWMs. In *E. coli,* this search produces an average of 38 promoter-like signals per 250-bp regulatory region, which may be distributed evenly across the sequence or present different levels of overlap. The second type of density, the local density, results from applying the COVER function [1], which establishes a competition on the

Table 3. −10 and −35 Consensi and Average Scores for the Degrading Genome of *M. leprae* and Its Close Relative *M. tuberculosis*

Species	Genome Size (Mb)	% GC	−10 Box Consensus	−10 Average Score	−35 Box Consensus	−35 Average Score
M. tuberculosis H37Rv	4.41	62	GTTATAAT	3.17	CACTTGACA	1.78
M. tuberculosis CDC1551	4.40	64	GTTATCAT	3.08	CACTTGACG	1.76
M. leprae	3.27	54	GTTAAAAT	2.90	CAATTGACA	1.60
M. leprae pseudogenes			GTTATAAT	2.52	CTATTGACA	0.82

The % GC is for noncoding DNA.
doi:10.1371/journal.pgen.0020185.t003

set of promoter-like signals in terms of −10 and −35 score and position of the signal regarding the gene start, selecting the more probable candidates to be functional promoters (see Material and Methods section for details). In *E. coli,* COVER produces 4.7 signals in average per 250-bp regulatory region. Most of these signals tend to exist as a series of overlapping potentially competing RNAP binding sites. The average size of these clusters of overlapped signals is 42 bp. Seventy-four percent of the *E. coli* experimentally mapped promoters are embedded in this kind of clusters [1]. We suggest that each of these types of densities may be maintained by natural selection for very different reasons.

Global density of promoter-like signals in regulatory regions. It is most likely that the global density is largely built by promoter-like signals that do not substitute the function of the primary promoter which has to respond to a given cellular condition; this would be the case if the majority of detected signals could bind RNAP and form a closed complex but were not able to proceed with the subsequent steps required to initiate transcription. However, a subset of these promoter-like signals could be just a single point mutation away from being able to operate as active transcription initiation sites. These could be called cryptic promoters, in analogy to cryptic genes that can be activated by single mutations. Some cryptic promoters could be relics of ancient promoters, indicating the high frequency of changes in regulatory regions. This high frequency correlates with the observed high flexibility in the evolution of transcriptional regulators in bacteria [22].

The permanence of cryptic promoters in the regulatory regions of bacteria could be facilitated by different kinds of evolutionary processes. First, they could constitute a collection of "back up" promoter sequences, maintained by selection for robustness. Bacteria having redundant signals capable of acquiring functionality with a single base change would increase in frequency in the population, because a fraction of their descendent cells would carry such activating mutations and would be able to survive in the advent of a deleterious mutation that destroyed the main promoter. In other words, the existence of multiple potential promoters would minimize the deleterious effects of genetic mutations on gene expression. In addition, for bacterial species prone to encounter a variety of environments, the existence of cryptic promoters of different strength could also allow for rapid evolution of changes in gene expression allowing adaptation to environmental changes. Taking into account that for σ^{70} transcription, the precise positioning of the regulatory proteins strongly determines their positive or negative role [23], this large availability of promoter-like sequences provides a fertile ground for quick changes in the role of regulatory proteins. Those changes have been proposed to depend on the demand of gene expression in a model where selection governs changes in gene regulation [24].

It is also possible that some of the signals detected in regulatory regions are not cryptic but rather fully functional promoters that are utilized only in special conditions. Although, in general, the site of transcription initiation is known to be rather precise, 25% of the transcription units in *E. coli* are known to be transcribed by more than one single promoter. The average number of mapped promoters coexisting in multiple-promoter regulatory regions is 3 [1,25]. The simultaneous availability of alternative promoters for a given gene would provide plasticity of gene expression in response to different conditions regularly encountered by the bacterial cell.

The regulatory region of the *E. coli lac* operon exhibits signals of both types. Six promoter sequences have been experimentally detected in close proximity to the primary *lac* promoter *(lacP1).* Four signals were created via single base pair mutations in the wild sequence, and two were detected in vivo as weak promoters that function when the primary promoter is impaired and its activator protein is absent [26,27].

Finally, the global density of promoter-like motifs in the regulatory regions could be bringing the polymerase near the promoter during the random DNA search prior to forming closed complexes, by attracting the enzyme to the general vicinity of the functional promoter. However, kinetic studies suggest that this might only result in a minor increase in the rate of formation of closed complexes (Jay D. Gralla, personal communication).

Local density of overlapping promoter-like signals. Overlapping promoter-like signals could play a regulatory role through functional interaction with the true promoter sequence, and their effect on regulation could be negative or positive.

Overlapping signals could negatively affect regulation by different means. The overlapping promoter-like signals might play a negative role if their interaction with RNAP were competitive. When two or more promoters are in close proximity, the potential exists for competition between them for the binding of RNAP [28]. Regulatory proteins play an important role in helping RNAP to choose the functional promoter sequence according to specific conditions. The regulation of the *gal* gene is an example of the effect of regulatory proteins on the positioning of RNAP between two competing promoters [29]. Also, the promoter-like signals could also induce pauses in the early steps of elongation when

σ^{70} is still bound to core RNAP. The strongest evidence indicating that σ^{70} can play a functional role during elongation comes from studies of the bacteriophage λ PR′ promoter and the lacUV5 promoter. Biochemical experiments have shown that σ^{70}-dependent pause occurs during early elongation in these promoters after RNAP has escaped from the promoter and synthesized a 16 or 17 nucleotide transcript. This pause is mediated by protein–DNA interaction between σ^{70} and a DNA sequence element in the initially transcribed region that resembles a promoter −10 element [30,31].

On the other hand, overlapping signals could also affect regulation positively. Some of these sites could be noncompetitive weak promoters that, in the absence of activation of the primary promoter, produce basal transcription of downstream genes. For example, transcription units that encode their own regulator would require constitutive levels of basal transcription. The overlapping promoter-like signals might also play a positive role by collecting RNAP molecules which could then be channeled to the primary promoter sequence [32].

Conclusion

Clearly, the distribution of the differential pattern of promoter-like signal densities among bacterial genomes mirrors that of genes and other genomic features that require weak selection to be effective in order to persist. This implies that the differential density of promoter-like signals between regulatory and nonregulatory regions confers some small but significant fitness advantage. Therefore, the outcome of gene regulation could be affected by factors much beyond the sequence of a single pair of RNAP binding sites in a given regulatory region, including the general abundance and organization of promoter-like signals in the region, as well as the presence of signals in the non-regulatory portions of the genome. Identification of the specific types of selective constraints that shape the number, position and arrangement of promoter-like signals across the different genomic regions of large bacterial genomes will require further comparative analyses among closely-related bacteria.

Materials and Methods

Sequence data. To evaluate the genomic frequency of promoter-like signals, three kinds of regions were defined in each genome according to National Center for Biotechnology Information (NCBI) annotations: coding, convergent, and strictly noncoding (excluding the convergent regions).

Coding regions contain genes with sizes above 1 kb. Convergent regions are noncoding regions located between the ends of two genes convergently transcribed. Convergent regions are analyzed separately because they are not expected to contain any functional promoters. In contrast, the strictly noncoding regions are located upstream of a gene start and are likely to contain regulatory regions.

To estimate the most likely functional promoters in each genome, we predicted the transcriptional units (single genes and operons) with the method of Moreno-Hagelsieb and Collado-Vides [33], which relies on the distribution of distances between genes in a given genome. The set of 250-bp sequences upstream of the first gene of every transcription unit is likely to constitute the smallest set of regulatory regions required for the expression of all genes in a genome and should contain the highest proportion of true functional promoters.

We used random sequences to evaluate if the overrepresentation of promoter-like signals in the regulatory regions is an artifact of AT-richness. We generated the random sequences according to a Markov chain model by using the RANDOM-SEQ program from RSAtools sofware [34]. For each genome, we obtained 1,000 random sequences, 1,000 bp sized, having the same nucleotide composition observed in regulatory regions using a Markov chain of order 0.

Promoter consensus matrices. The promoter model we adopted for our searches was Matrix__18__15__13__2__1.5 (Figure 2), defined through a thorough evaluation of more than 200 *E. coli* matrix pairs that optimized different criteria [1]. Matrices were obtained from a representative set of 116 promoters which have the description of the precise +1 nucleotide of transcription initiation, as well as the information on their regulation. The method used was the following: (i) The 116 sequences were aligned with respect to the position of the transcription initiation (+1). The first 18 bp upstream from the +1 were used as input for WCONSENSUS program [35] to identify the motif corresponding to the −10 conserved region. This size was selected considering that the distance from the −10 hexamer to the +1 varies from 4 to 12 bp. (ii) In order to identify the −35 box, we performed a realignment of the promoter sequences anchoring the −10 boxes identified by the program. New sequences of various lengths, initiating at 13 bp upstream of the −10, were selected. (iii) WCONSENSUS was run with this new set of sequences. Thus, for each one of the −10 alignments, several alternative sequence sizes were analyzed. The best pair, Matrix__18__15__13__2__1.5, was selected on the basis that it contains the canonical consensus sequences for both the −10 and −35 motifs and that it outperformed all others according to measures of sensitivity, specificity, precision, and accuracy.

Calibration of weight matrices. The elements of the frequency matrices in Figure 2 are simply the number of times that the indicated letter is observed at the indicated position in the alignment of 116 *E. coli* promoter sequences. Such elements must be calibrated before matrices can be used in other genomes by normalizing the frequencies with the *a priori* probability for the corresponding base in each genome. To estimate the a priori probabilities, we used the frequency observed for each nucleotide in the set of all noncoding regions for each analyzed genome. Calibrations were done by means of PATSER program [35,36]:

$$W(b,l) = f(b,l)\log_2\frac{f(b,l)}{p(b)} \tag{1}$$

where *f(b,l)* is the relative probability of the base *b* at the position *l* of the input *E. coli* matrix and *p(b)* is the a priori probability of base *b* in the noncoding regions of the analyzed genome.

Scoring a DNA promoter site. In order to define which sequences would be considered promoter-like signals in the different target genomes, we determined minimal cutoff scores that would retain 98% of the 116 original motifs from functional σ^{70} promoters that were used in generating the *E. coli* frequency matrices. With this aim, PATSER program was used to rescore the original *E. coli* motifs with the calibrated matrices corresponding to each genome. The mean and standard deviation of the new scores were obtained. In most of the analyzed genomes, 98% of the original motifs was retained using a cutoff of $\mu - 2.5\sigma$. PATSER calculates the score, I_{seq}, of a target motif of size L, as:

$$I_{seq} = \sum_{l=1}^{L} W(b_{seq}, l) \tag{2}$$

where b_{seq} is the base found in the l position of the target motif. The score of a promoter is the sum of scores corresponding to the −10 and −35 motifs.

Counting promoter-like signals. For each genome, the calibrated matrices were used to search for promoter-like signals in the four kinds of genomic regions (random, coding, convergent, and strictly noncoding). Only those signals scoring above the respective cutoff were retained and the number of promoter-like signals found in each genomic region was calculated. The size, in base pairs, of each genomic region was obtained. To conduct the analysis of under-representation/overrepresentation, we chose to compare the distribution of the promoter-like signals found in the strictly noncoding regions against that found in convergent regions. Before comparison, the number of detected signals was adjusted to the size of the smallest genomic region. A log-likelihood statistic was used for the calculated proportions to determine if the genome had an excess density of promoter-like signals in potentially regulatory regions (the strictly noncoding regions) when compared against noncoding regions between convergent genes.

For genomes with significant overrepresentation of promoter-like signals in regulatory versus convergent noncoding regions ($p < 0.001$), we estimated the most likely functional promoters in the genome. To this aim, we ran searches for promoter-like signals in the 250-bp upstream regions of the operons in each genome (see section on sequence data). We applied the COVER function (described below) on the collections of promoter-like signals found in the set of regulatory regions. The resulting COVER-predicted promoters in these regulatory regions were taken as the most likely functional promoters in that genome.

Cover function. The COVER program [1] is used to choose the most likely functional promoters from a conglomerate of promoter-like signals identified by weight matrices. First, COVER employs a "divide and conquer" strategy: Each promoter-like signal is assigned to a class based on the length of the spacer region between the −10 and −35 boxes. Then, the selection of the best signals within each class is based on a well-known partial order relation, the *inclusion relation* [37]. In our case, the relation uses both an "intrinsic score," the score resulting from adding the −10 and the −35 scores, and an "external score," based on the relative position of the predicted −10 in relation to the beginning of the gene. A promoter A is said to be included by another promoter B if, and only if, the scores of promoter B are both better than the corresponding scores of promoter A. This inclusion relation defines a "cover set", or collection of separate subsets, for each spacer class. Within each class, the predicted functional promoter is defined as the upper border of the subset with the highest cardinality, i.e., the one which includes the highest number of promoter signals. The signals detected by COVER are mostly found grouped into clusters.

Conservation of promoter-like signals in related bacteria. We have obtained 2,014 groups of orthologous genes from 14 enteric genomes that showed overrepresentation of promoter-like signals in the regulatory regions. The average size of the groups of orthologs was 10.4 genes. For all the promoter-like signals found by using weight matrices, we have measured the conservation of each promoter-like signal as the number of times that the signal is present in the same relative position in each upstream region for each set of orthologs. We report 79% signal conservation. Orthologs were kindly provided by G. Moreno-Hagelsieb and consisted of BLASTP reciprocal best hits obtained as explained previously in [33].

Graphs S1 and S2 in Supplementary Material were created using the XYGRAPH program from the RSAtools software [34]. Scripts were written in Perl [38].

Supporting Information

Figure S1. Signal Density in Regulatory versus Nonregulatory Regions for All Analyzed Eubacterial Genomes

Found at doi:10.1371/journal.pgen.0020185.sg001 (5.4 MB PDF).

Figure S2. Signal Density in Random Regulatory versus Regulatory Regions for All Analyzed Eubacterial Genomes

Found at doi:10.1371/journal.pgen.0020185.sg002 (493 KB PDF).

Table S1. Detailed Examples of *M. leprae* Pseudogenes with Orthologous Genes in *M. tuberculosis*

Found at doi:10.1371/journal.pgen.0020185.st001 (15 KB PDF).

Accession Numbers

The RegulonDB (http://regulondb.ccg.unam.mx/index.html) accession number for the regulatory region of the *E. coli lac* operon is ECK120014850 and that for *gal* gene is ECK120014842.

The National Center for Biotechnology Information (NCBI) (http://www.ncbi.nlm.nih.gov) accession numbers for representative bacterial genomes are given in Table 1.

Acknowledgments

The authors thank Victor del Moral and Cesar Bonavides for computational support and Gabriel Moreno-Hagelsieb, Jacques van Helden, and Jeff L. Froula for critical discussions and advice.

Author contributions. AMH, MPF, and JCV conceived and designed the experiments. AMH performed the experiments. AMH, MPF, EM, and JCV analyzed the data. AMH and JCV contributed reagents/materials/analysis tools. AMH and MPF wrote the paper.

Funding. AMH has been supported by a PhD grant from CONACyT (México) and postdoctoral support from the Program of Evolutionary Genomics, DOE JGI. EM acknowledges support from CONACYT, grant 37023-N. This work has been supported by grant 0028 from CONACyT and by grant GM62205 from the National Institutes of Health to JCV. The latest stages of this work were performed under the auspices of the US Department of Energy's Office of Science, Biological and Environmental Research Program and by the University of California, Lawrence Berkeley National Laboratory, under contract number DE-AC03-76SF00098.

Competing interests. The authors have declared that no competing interests exist.

References

1. Huerta AM, Collado-Vides J (2003) Sigma70 promoters in *Escherichia coli:* Specific transcription in dense regions of overlapping promoter-like signals. J Mol Biol 333: 261–278.
2. Gruber TM, Gross CA (2003) Multiple sigma subunits and the partitioning of bacterial transcription space. Annu Rev Microbiol 57: 441–466.
3. Hawley DK, McClure WR (1983) Compilation and analysis of *Escherichia coli* promoter DNA sequences. Nucleic Acids Res 11: 2237–2255.
4. Lisser S, Margalit H (1993) Compilation of *E. coli* mRNA promoter sequences. Nucleic Acids Res 21: 1507–1516.
5. Ozoline ON, Deev AA, Arkhipova MV (1997) Non-canonical sequence elements in the promoter structure. Cluster analysis of promoters recognized by *Escherichia coli* RNA polymerase. Nucleic Acids Res 25: 4703–4709.
6. Stone JR, Wray GA (2001) Rapid evolution of *cis*-regulatory sequences via local point mutations. Mol Biol Evol 18: 1764–1770.
7. Hahn MW, Stajich JE, Wray GA (2003) The effects of selection against spurious transcription factor binding sites. Mol Biol Evol 20: 901–906.
8. Wosten MM (1998) Eubacterial sigma-factors. FEMS Microbiol Rev 22: 127–150.
9. Mittenhuber G (2002) An inventory of genes encoding RNA polymerase sigma factors in 31 completely sequenced eubacterial genomes. J Mol Microbiol Biotechnol 4: 77–91.
10. Lonetto M, Gribskov M, Gross CA (1992) The sigma 70 family: Sequence conservation and evolutionary relationships. J Bacteriol 174: 3843–3849.
11. Moran NA, Plague GR (2004) Genomic changes following host restriction in bacteria. Curr Opin Genet Dev 14: 627–633.
12. Herbeck JT, Wall DP, Wernegreen JJ (2003) Gene expression level influences amino acid usage, but not codon usage, in the tsetse fly endosymbiont Wigglesworthia. Microbiology 149: 2585–2596.
13. Herbeck JT, Funk DJ, Degnan PH, Wernegreen JJ (2003) A conservative test of genetic drift in the endosymbiotic bacterium *Buchnera:* Slightly deleterious mutations in the chaperonin groEL. Genetics 165: 1651–1660.
14. Wilcox JL, Dunbar HE, Wolfinger RD, Moran NA (2003) Consequences of reductive evolution for gene expression in an obligate endosymbiont. Mol Microbiol 48: 1491–1500.
15. Madan Babu M (2003) Did the loss of sigma factors initiate pseudogene accumulation in *M. leprae*? Trends Microbiol 11: 59–61.
16. Moran NA, Dunbar HE, Wilcox JL (2005) Regulation of transcription in a reduced bacterial genome: Nutrient-provisioning genes of the obligate symbiont *Buchnera aphidicola*. J Bacteriol 187: 4229–4237.
17. Moran NA (1996) Accelerated evolution and Muller's rachet in endosymbiotic bacteria. Proc Natl Acad Sci U S A 93: 2873–2878.
18. Ochman H, Elwyn S, Moran NA (1999) Calibrating bacterial evolution. Proc Natl Acad Sci U S A 96: 12638–12643.
19. Itoh T, Martin W, Nei M (2002) Acceleration of genomic evolution caused by enhanced mutation rate in endocellular symbionts. Proc Natl Acad Sci U S A 99: 12944–12948.
20. Cole ST, Eiglmeier K, Parkhill J, James KD, Thomson NR, et al. (2001) Massive gene decay in the leprosy bacillus. Nature 409: 1007–1011.
21. Mira A, Pushker R (2005) The silencing of pseudogenes. Mol Biol Evol 22: 2135–2138.
22. Lozada-Chávez I, Janga SC, Collado-Vides J (2006) Bacterial regulatory networks are extremely flexible in evolution. Nucleic Acids Res 34: 3434–3445.
23. Collado-Vides J, Magasanik B, Gralla JD (1991) Control site location and transcriptional regulation in *Escherichia coli*. Microbiol Rev 55: 371–394.
24. Savageau MA (1998) Demand theory of gene regulation. I. Quantitative development of the theory. Genetics 149: 1665–1676.
25. Salgado H, Gama-Castro S, Peralta-Gil M, Diaz-Peredo E, Sanchez-Solano F, et al. (2006) RegulonDB version 5.0 RegulonDB (version 5.0): *Escherichia coli* K-12 transcriptional regulatory network, operon organization, and growth conditions. Nucleic Acids Res 34 (Database Issue): D394–D397.
26. Reznikoff W (1992). The lactose operon-controlling elements: A complex paradigm. Mol. Microbiol 6: 2419–2422.
27. Czarniecki D, Noel RJ Jr, Reznikoff WS (1997). The 245 region of the *Escherichia coli lac* promoter: CAP-dependent and CAP-independent transcription. J Bacteriol 179: 423–429.
28. Goodrich JA, McClure WR (1991) Competing promoters in prokaryotic transcription. Trends Biochem Sci 16: 394–397.
29. Goodrich JA, McClure WR (1992) Regulation of open complex formation at the *Escherichia coli* galactose operon promoters. Simultaneous interaction of RNA polymerase, gal repressor and CAP/cAMP. J Mol Biol 224: 15–29.
30. Nickels BE, Mukhopadhyay J, Garrity SJ, Ebright RH, Hochschild A (2004) The sigma 70 subunit of RNA polymerase mediates a promoter-proximal pause at the lac promoter. Nat Struct Mol Biol 11: 544–550.
31. Brodolin K, Zenkin N, Mustaev A, Mamaeva D, Heumann H (2004) The

sigma 70 subunit of RNA polymerase induces lacUV5 promoter-proximal pausing of transcription. Nat Struct Mol Biol 11: 551–557.

32. Reznikoff WS, Bertrand K, Donnelly C, Krebs M, Maquat LE, et al. (1987) Complex promoters. In: Reznikoff WS, Burgess RR, Dahlberg JE, Gross CA, Record MT Jr, et al., editors. RNA polymerase and the regulation of transcription. New York: Elsevier. pp. 105–113.
33. Moreno-Hagelsieb G, Collado-Vides J (2002) A powerful non-homology method for the prediction of operons in prokaryotes. Bioinformatics 18 (Suppl 1): S329–S336.
34. van Helden J (2003) Regulatory sequence analysis tools. Nucleic Acids Res 31: 3593–3596.
35. Stormo GD (1998) Information content and free energy in DNA–protein interactions. J Theor Biol 195: 135–137.
36. Hertz GZ, Stormo GD (1999) Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. Bioinformatics 15: 563–577.
37. Tremblay JP, Manohar R (1975) Discrete Mathematical Structures with Applications to Computer Science. New York: McGraw-Hill. 606 p.
38. Wall L, Christiansen T, Schwartz RL (1991) Programming Perl. Sebastopol (California): O'Reilly and Associates. 670 p.
39. Paget MS, Helmann JD (2003) The sigma70 family of sigma factors. Genome Biol 4: 203.1–203.6.
40. Isono K, Shimizu M, Yoshimoto K, Niwa Y, Satoh K, et al. (1997) Leaf-specifically expressed genes for polypeptides destined for chloroplasts with domains of sigma70 factors of bacterial RNA polymerases in *Arabidopsis thaliana*. Proc Natl Acad Sci U S A 94: 14948–14953.
41. Fenton MS, Lee SJ, Gralla JD (2000) *Escherichia coli* promoter opening and −10 recognition: Mutational analysis of sigma70. EMBO J 19: 1130–1137.
42. Gardella T, Moyle H, Susskind MM (1989) A mutant *Escherichia coli* sigma 70 subunit of RNA polymerase with altered promoter specificity. J Mol Biol 206: 579–590.
43. Waldburger C, Gardella T, Wong R, Susskind MM (1990) Changes in conserved region 2 of *Escherichia coli* sigma 70 affecting promoter recognition. J Mol Biol 215: 267–276.
44. Dombroski AJ (1997) Recognition of the −10 promoter sequence by a partial polypeptide of sigma70 in vitro. J Biol Chem 6: 3487–3494.
45. Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22: 4673–4680.
46. Janga SC, Moreno-Hagelsieb G (2004) Conservation of adjacency as evidence of paralogous operons. Nucleic Acids Res 32: 5392–5397.