Published in final edited form as:

Mol Biosyst. 2010 April ; 6(4): 721–728. doi:10.1039/b908047c.

Interplay of heritage and habitat in the distribution of bacterial signal transduction systems

Michael Y. Galperin^a, Roger Higdon^{b,c}, and Eugene Kolker^{b,c,d}

^a National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, 8600 Rockville Pike, Bethesda, Maryland 20894, USA

^b Bioinformatics & High-throughput Analysis Laboratory, Seattle Children's Research Institute, 1900 9th Ave., Seattle, Washington 98101, USA

^c Predictive Analytics, Seattle Children's Hospital, 4800 Sand Point Way NE, Seattle, Washington 98105, USA

^d Biomedical and Health Informatics Division, Medical Education and Biomedical Health Department, School of Medicine, University of Washington, Seattle, Washington 98105, USA

Abstract

Comparative analysis of the complete genome sequences from a variety of poorly studied organisms aims at predicting ecological and behavioral properties of these organisms and help in characterizing their habitats. This task requires finding appropriate descriptors that could be correlated with the core traits of each system and would allow meaningful comparisons. Using the relatively simple bacterial models, first attempts have been made to introduce suitable metrics to describe the complexity of organism's signaling machinery, which included introducing the "bacterial IQ" score. Here, we use an updated census of prokaryotic signal transduction systems to improve this parameter and evaluate its consistency within selected bacterial phyla. We also introduce a more elaborate descriptor, a set of profiles of relative abundance of members of each family of signal transduction proteins encoded in each genome. We show that these family profiles are well conserved within each genus and are often consistent within families of bacteria. Thus, they reflect evolutionary relationships between organisms as well as individual adaptations of each organism to its specific ecological niche.

Keywords

comparative genomics; evolution; protein phosphorylation; receptor; *Mycobacterium*; *Shewanella*

Introduction

All organisms adapt to their environment by perceiving environmental signals and accordingly modifying their behavior and/or metabolism. The recent progress in genome and metagenome sequencing revealed that organismal (primarily microbial) diversity in natural habitats far exceeds most previous estimates. The influx of a large number of complete genomes from environmental microorganisms inhabiting a variety of ecological niches requires new approaches to deal with this previously hidden world¹. Several years ago we have conducted a census of signal transduction systems encoded in genomes of 167

prokaryotic species that provided sufficient background data to compare microorganisms in terms of their ability to adjust to environmental changes². It showed that most proteins involved in signal transduction – sensor histidine kinases, methyl-accepting chemotaxis proteins, adenylate cyclases, diguanylate cyclases, c-di-GMP-specific phosphodiesterases, and Ser/Thr protein kinases – are unevenly distributed even in closely related microorganisms. While the total number of signaling proteins encoded in any given genome correlated well with the genome size, the contribution of each individual class of signaling proteins varied greatly between organisms belonging to different phylogenetic lineages. These data were consistent with the observation that different microorganisms use different signal transduction pathways to respond to the same environmental challenges³ and supported the idea that these pathways were somewhat interchangeable^{4,5}. As an example, environmental modulation of the cyclic diguanylate (c-di-GMP) synthesis can occur through a dedicated environmental sensor with a GGDEF output domain (e.g. AdrA⁶), through a sensor histidine kinase with a GGDEF-containing response regulator (e.g. PleD⁷), or through a methyl-accepting chemotaxis sensor protein, from which the signal is transmitted to the chemotaxis-specific histidine kinase CheA and then to a GGDEF-containing response regulator (e.g. WspR⁸). This, in turn, meant that the sheer complexity of the signal transduction system, manifested in the total number of the interacting components, was an important parameter reflecting the ability of the host organism to adapt to the environmental challenges. To account for the complexity of signal transduction even in the most primitive bacterial cells, two new parameters have been introduced. The signal adaptation index, commonly referred to as “bacterial IQ”, reflects the abundance of signal transduction components encoded in a given organism as compared to others of a similar genome size². The “degree of extrovertness” reflects the fraction of the outward-looking environmental sensors among all signal transduction proteins encoded in a given genome². While these parameters offered a new way to describing environmental organisms, they were defined in a largely *ad hoc* manner and were not supported by any statistical analysis. As part of our attempts to introduce statistically sound metrics to define bacterial genomes and proteomes¹, we present here an improved way to calculate IQ scores and introduce family profiles of signal transduction proteins, a new metric that reflects the number of members of each particular family of signal transduction protein encoded in the given genome.

Results

Improving the “bacterial IQ” score

Previous analyses of bacterial signal transduction^{2,9,10} revealed a strong correlation between the total number of signaling proteins encoded in any given genome and the genome size. Both values correlated well with the total number of proteins encoded in the corresponding genome². Therefore, the number of signaling proteins encoded in a given genome was judged to be a poor a measure of its signaling capabilities, as it would introduce a bias towards the organisms with larger genome sizes. The fraction of the total proteome dedicated to signal transduction also grew with genome size, indicating that this value is also biased towards the organisms with larger genome sizes. To account for these observations, the bacterial IQ metric was devised to be proportional to the square root of the total number of genes encoding signal transduction proteins per 1 Mbp of a given genome². In this work, we made several adjustments aimed at improving the robustness of the IQ parameter while preserving its independence from the influence of the genome size.

During the four years that elapsed since the completion of the previous signal transduction census², there was a dramatic increase in the number of completely sequenced prokaryotic genomes. Although a fair share of this increase was due to the sequencing of multiple isolates of selected pathogenic bacteria (*Bacillus anthracis*, *Escherichia coli*, *Salmonella enterica*, *Staphylococcus aureus*, *Streptococcus pneumoniae*, *Streptococcus pyogenes*), joint

efforts of several genome sequencing centers resulted in a rapidly improving coverage of microbial diversity¹¹. By the end of 2008, the size of the non-redundant genome set, covering all individual bacterial and archaeal species with completely sequenced genomes, increased to 555 genomes. This set included first genome sequences from several poorly studied bacterial phyla and of a number of relatively large genomes that demonstrate great variety in their protein content (Table 1). This prompted an update of the census of prokaryotic signal transduction proteins, http://www.ncbi.nlm.nih.gov/Complete_Genomes/SignalCensus.html, and a re-evaluation of the best way to calculate bacterial IQ scores.

In addition, the set of signal transduction proteins included in the total count was further expanded to include two-component response regulators^{9,12} and Ser/Thr protein phosphatases of the PP2C superfamily^{13,14}. These additions were not expected to affect the general trends, as, according to the previous observations, the number of response regulators encoded in any prokaryotic genome are close to the number of their cognate histidine kinases⁹ (see Supplementary Figure S1), whereas the number of Ser/Thr protein phosphatases is typically comparable to the number of Ser/Thr kinases encoded in the same genome^{14,15}. Indeed, despite these changes, the total number of signal transduction proteins continued to show strong correlation with the genome size (Supplementary Figure S2). In accordance with the earlier observations, the relationship was close to quadratic (the slope of the regression line, 1.91). This correlation held also for the histidine kinases and response regulators and, albeit less strongly, for the total sum of other signal transduction proteins (Supplementary Figure S3). A strong correlation with a nearly-quadratic relationship was also observed between the total number of signal transduction proteins and the total number of proteins encoded in that genome (Figure 1). Based on these data, we have modeled the relationship between the number of signaling proteins and the genome size aiming to put the IQ parameter on a solid statistical footing (see Supplementary Methods). The IQ scores were normalized to make their distribution was roughly symmetric and scaled so that the minimum IQ score was equal to zero and the mean was close to 100. This resulted in the following approximate formula for calculating the IQ score:

$$IQ = 50 * (\ln N_{SP} - 2 \ln N_T) + 700$$

where N_{SP} is the number of signaling proteins and N_T is the total number of proteins in an organism. This formula was then used to compare the signaling capacities of diverse microorganisms (Table 2).

IQ variation across different phyla

The availability of complete genomes of at least 7 different species from 11 bacterial and two archaeal phyla (Table 2) provided a sufficient data set to calculate the average IQ scores and evaluate the consistency of this parameter across various prokaryotic taxa. Despite the significant variance in the mean IQ scores within most bacterial and archaeal phyla (Table 2), there were statistically significant differences between certain phyla and between different classes within *Proteobacteria* and *Firmicutes*. A key reason for this variance, identified in Table 2, was the presence in many phyla of both free-living and obligately pathogenic bacteria. In the latter ones, adaptation to the parasitic lifestyle typically involves genome compaction, coupled with the loss of metabolic and signaling genes. As a result, closely related organisms often differ in genome sizes and encode dramatically different gene sets. Another factor contributing to the variance in signal protein content is the ability of organisms to use oxygen and other terminal electron acceptors. Strictly anaerobic fermentative lifestyle often results in a decrease in the signaling capacity, whereas the ability to utilize distinct electron acceptors always correlates with a high number of sensory and

signal transduction proteins. Among anaerobes, the highest versatility in utilizing various terminal electron acceptors is seen among members of *Thermotogae*, certain clostridia, and delta-proteobacteria, which is reflected in the relatively high average IQ values in these groups of organisms (Table 2).

The groups showing consistently low IQ values include *Tenericutes* (mycoplasmas), which have undergone extreme genome compaction in the course of their evolution, and *Crenarchaeota*, which are represented mostly by extremophiles that occupy stable and unique ecological niches. It should be noted, however, that two recently sequenced genomes provided the first examples of several types of signal transduction proteins in both these lineages – and significantly increased data variance within these phyla. The genome of the mollicute *Acholeplasma laidlawii* showed the presence of several classes of proteins – histidine kinases, response regulators, diguanylate cyclases and c-di-GMP phosphodiesterases and even an adenylate cyclase – that have not been seen in any previously sequenced mollicute genome (Supplementary Figure S4). Likewise, the recently sequenced genome of the first non-extremophilic crenarchaeon, an ammonia-oxidizing, carbon-fixing marine archaeal group 1 member *Nitrosopumilus maritimus*¹⁶, encodes 11 histidine kinases and 10 response regulators, which were absent in all previously sequenced crenarchaeal genomes (Supplementary Figure S5). These examples show that we are still dealing with only a minor slice of microbial diversity and our genome analyses are still subject to sample bias.

Signal transduction protein family profiles

The above examples show that an integral parameter like the IQ score needs to be supplemented with more detailed metrics that would allow more targeted comparisons of the signaling capabilities within closely related or diverse organisms. Following the example of the Clusters of Orthologous Groups of proteins (COG) functional profiles¹⁷, as discussed in Ref. ¹, we introduce here family profiles of signal transduction proteins, which reflect the number of members of each particular family of signal transduction protein encoded in the given genome. For example, Figure 2 shows family profiles of 10 species from the genus *Mycobacterium* (phylum *Actinobacteria*). This genus unifies a wide variety of bacteria, including obligate parasites *Mycobacterium leprae* and *M. tuberculosis* (named after the diseases they cause), facultative pathogens, such as *M. marinum* and *M. smegmatis*, and free-living environmental isolates such as *M. gilvum* and *M. vanbaalenii*^{18,19}. With the exception of the highly degraded genome of *M. leprae*²⁰, all other members of genus have similar IQ values in the range of 80–100. Active pathogens, such as *M. leprae*, *M. tuberculosis*, *M. bovis*, and *M. avium*, have smaller genome sizes and encode fewer proteins than their free-living relatives. The family profiles, plotted in the order of increasing number of total proteins per genome, show that increased genome size is accompanied with an increase in the number of encoded two-component systems (histidine kinases and response regulators), Ser/Thr protein kinases and protein phosphatases (Figure 2). The complete absence of chemotaxis sensors (MCPs) is consistent with the observation that all mycobacteria are non-motile. Other systems show more complex distribution patterns with a remarkable increase in the number of adenylate cyclases and Ser/Thr protein kinases in the genome of *M. marinum*, a fast-growing aquatic organism that infects marine and freshwater fish. Protein kinases have been shown to regulate metabolism in mycobacteria²¹ and could be important for the environmental adaptations of this versatile pathogen, which is capable of infecting fish, frogs, and snakes and occasionally causing skin lesions in humans. At the same time *M. marinum* lacks any enzymes for synthesis or hydrolysis of c-di-GMP, which has been reported to be required for long-term survival of *M. smegmatis* under nutritional stress²². Indeed, by far the largest number of c-di-GMP-metabolizing enzymes is seen in *M. gilvum* and *M. vanbaalenii*, two free-living bacteria that can degrade a wide range of

environmentally toxic chemicals, including high-molecular-weight polycyclic aromatic hydrocarbons, such as benz[a]pyrene, and are of great interest for use in bioremediation^{18,23}.

In most bacterial genera, the numbers of signal transduction proteins do not vary as much as in *Mycobacterium* and their family profiles offer a convenient tool to visualize and compare organisms isolated from diverse ecological niches. Figure 3 shows the distribution of signal transduction proteins in 12 species of *Shewanella*, a very interesting genus of predominantly marine bacteria that are found in diverse environments and are capable of reducing a variety of terminal electron acceptors^{24,25}. This genus, and its best studied representative *Shewanella oneidensis* strain MR-1 are being systematically studied using a variety of high-throughput methods²⁴. Figure 3 shows that these versatile organisms encode members of every kind of signal transduction proteins, indicating their ability to respond to environmental challenges on several different levels – transcriptional (two-component systems, cAMP-dependent signaling), posttranslational (protein phosphorylation on Ser and Thr residues), whole-cell level (chemotaxis) and multicellular level (biofilm formation). Despite their different habitats and significant differences in respiratory metabolism, all *Shewanella* species have similar distributions of key families of signal transduction proteins.

Adaptation to extreme and simple environments

As part of the flexible genomic “shell” (as opposed to the core genome²⁶) signal transduction proteins are commonly acquired and lost in the course of evolution²⁷. As noted previously, adaptation to extreme (or stable) conditions often results in a loss of metabolic flexibility, which leads to a diminished demand for cellular regulation and therefore affects environmental sensing^{2,4}. A recent genome analysis of the moderately thermophilic bacterium *Anoxybacillus flavithermus* WK1, isolated from a super-saturated silicate solution, concluded that evolution of the family *Bacillaceae* from a common ancestor of all *Firmicutes* involved acquisition of numerous genes, which was followed by a substantial gene loss in the *Anoxybacillus/Geobacillus* branch in the course of the specific niche adaptation²⁸. A signal transduction protein family profile of various *Bacillaceae* (Figure 4A) is consistent with that conclusion (which was achieved without considering these proteins). This profile shows that *A. flavithermus* and two *Geobacillus* species encode all the same families of signal transduction proteins as *Bacillus* spp., albeit in a smaller number. In an interesting twist, *Bacillus anthracis*, a pathogen, encodes a far richer complement of environmental sensors than its free-living relatives (Figure 4A). Remarkably, an even smaller set of signal transduction proteins is seen in an extremely halotolerant and alkaliphilic species *Oceanobacillus iheyensis*, which is otherwise closely related to *Bacillus subtilis* and *B. halodurans*²⁹. The moderately halotolerant alkaliphile *B. halodurans* and *O. iheyensis* occupy similar habitats, which cannot account for the differences in signal protein content. It seems likely therefore that this difference is due to the distinct evolutionary histories of these organisms. The phylogenetic position of *O. iheyensis* with respect to other bacilli remains unresolved^{28,30}, but it seems likely that it had diverged from the main bacillar branch before that branch has experienced the massive gene influx. Thus, while *A. flavithermus* and *Geobacillus* appear to have lost some of their signal transduction genes, *O. iheyensis* probably never had them in the first place.

Expanding the phylogenetic coverage of Figure 4A to include other members of the order *Bacillales* adds family profiles both on the lower [genus *Listeria* (family *Listeriaceae*), genus *Staphylococcus* (family *Staphylococcaceae*)] and on the upper [genus *Lysinibacillus* (family *Planococcaceae*)] part of the spectrum (Figure 4B). This is consistent with the IQ scores of these organisms and their lifestyle as facultative pathogens, capable of saprophytic growth outside the host. Unlike *Bacillaceae* members, listeriae and staphylococci lack the ability to sporulate and the complex signal transduction machinery that regulates sporulation and germination. Moving further down the evolutionary tree to the class *Bacilli* adds the

order *Lactobacillales* with extensively sequenced genera *Lactobacillus* (family *Lactobacillaceae*) and *Streptococcus* (family *Streptococcaceae*). Representatives of these genera demonstrate further loss of signaling capacity (Figure 4C), most likely as a result of adaptation to the nutrient-rich media, milk for the former group and host tissues for the latter one. Further integration of signal transduction protein family profiles is possible through the use of heat maps (Figure 4D, Supplementary Figure S6). In such maps, grouping of organisms will not necessarily be defined by their phylogenetic proximity. Instead, it will reflect similarities in the quantity and distribution of their signal transduction systems which are more likely to be determined by their respective habitats.

Discussion

The availability of numerous genome sequences presents a challenge of using the sequence data to characterize metabolic and signaling pathways and ecological properties of the respective organisms and to extract useful information about their evolutionary origin, role in the biosphere, potential use in biotechnology, and other traits. So far, genomic data has been a mixed blessing: having dramatically expanded our knowledge of the molecular biology and biochemistry, it also illuminated major deficiencies in our understanding of the functioning of the live cell. The problem goes beyond the famous ‘70% hurdle’, the fact that about a third of all proteins encoded in any bacterial cell have unknown functions³¹; for most eukaryotic genome sequences, including the human genome, the fraction of characterized genes is far lower than that. Even for proteins with known biochemical functions, the exact substrate specificity often remains obscure^{24,32}. In signal transduction, recognition of a given protein as a sensor histidine kinase or a response regulator rarely provides any clues as to what stimuli are being sensed and which cellular responses are being triggered by them^{4,5}. Major unresolved questions about the functioning of the signal transduction machinery remain even for such model organisms as *E. coli* and *B. subtilis*³³, which are commonly used as reference points for describing metabolic and signaling pathways of lesser studied organisms.

We have recently pointed out that comparative genomics needs a new language that would combine (i) a standard set of parameters that describe statistical properties of each sequenced genome, as well as physiology and the natural environment of the host organism, and (ii) a set of new (still to be devised) metrics that would allow placing the newly sequenced genome in a proper evolutionary and ecological context¹. The first part of this proposal got instant acceptance in the community and contributed to the formulation of the minimum information about a genome sequence (MIGS) specification by the Genomic Standards Consortium³⁴. The second part proved to be more difficult, as development of new descriptors is not a straightforward task. Even utilization of the promising examples that we have cited – incorporation of a genome in the three-dimensional Tree of Life³⁵ and genome profiling by COG functional categories³⁶ require a certain effort. The ‘bacterial IQ’ score, which is not difficult to calculate, did not fare any better. Here, we sought to develop a better formula for the IQ score, putting it on a solid statistical footing, and to use this new formula for comparing different microbial phyla. However, the fluidity of the signal transduction systems often shows up at smaller phylogenetic distances, at the family and genus levels. The signal transduction protein family profiles, introduced in this work (Figures 2–4 and S4–S6), aim at capturing the differences in signaling capacity between closely related organisms. We hope that these new tools will help in comprehending the information stored in the genomic data and will eventually lead to a better understanding of how every genome is shaped by its phylogeny and its own adaptations to the environment, i.e. by interplay of its heritage and habitat.

Future developments

The family profiles of signal transduction proteins were designed to capture both phylogenetic and ecological properties of diverse bacteria. Figures 2–4 and S4–S6 show that these profiles largely accomplish these goals and provide an easy way to compare any given organism with its relatives and/or neighbors. Nevertheless, we envisage several ways in which these profiles could be improved.

Firstly, the list of the signal transduction systems included in these profiles could be expanded by incorporating (i) additional families of Ser/Thr and Tyr protein phosphatases¹⁵ (ii) the components of the phosphoenolpyruvate-dependent sugar:phosphotransferase system (PTS) that are involved in chemotaxis and regulation of the activity of the enterobacterial adenylate cyclase and various membrane permeases³⁷ (iii) the anti-sigma and anti-anti-sigma components that modulate the activity of RNA polymerase sigma subunits³⁸; (iv) systems that produce and respond to alternative nucleotide messengers such as ppGpp and the recently described cyclic diadenylate³⁹, and (v) additional signaling components that still remain to be characterized.

Secondly, the specificity of the profiles could be increased if, instead of treating all response regulators as a single group, they were separated into families, e.g. based on their domain architectures⁹. Preliminary data show that family profiles of various response regulators are consistent within most taxonomic groups of bacteria (MYG, manuscript in preparation).

Thirdly, it might make sense to include so-called “one-component” transcriptional regulators¹⁰, although their variety might make these profiles too complex and complicate their comparison. Finally, these profiles could be further modified to reflect the fraction of environmental (membrane-bound) and intracellular (cytoplasmic) receptors in each sensor protein family, e.g. by shading or hatching. This way, family profiles could be also used to compare the “extrovertness” of different genomes.

Conclusions

There is little doubt that almost all features of an organism, at least a simple unicellular one, could be deduced from its complete genome sequence. A key goal of genome analysis is using the sequence data to uncover salient features of the host organism and put them into phylogenetic and ecological context. That is why a significant part of comparative-genomics studies is devoted to finding the genomics correlates of such features as thermophily, psychrophily, halophily, alkalophily, resistance to radiation and the ability to survive stress, and using these parameters as metrics for genome description. This work continues the previous study that introduced the “bacterial IQ” score for comparing the sets of various receptor proteins encoded in bacterial and archaeal genomes. Using a larger set of genomes and an expanded list of signal transduction proteins, we compare the IQ scores within various groups of bacteria. We also present a new way of comparing signal transduction machineries of various bacteria by plotting the abundance of members of each signal transduction protein family in each analyzed genome, either as a 3D histogram or as a heat map. We show that these family profiles are well conserved within each genus and are often consistent within more distant groups of bacteria. At the same time, these profiles are affected by the lifestyle of each particular organism and the set of environmental signals it responds to. Thus, these family profiles reflect evolutionary relationships between organisms (their heritage) as well as individual adaptations of each organism to its specific ecological niche (its habitat).

Experimental

Data sets

The genome sequence data, genome sizes and the total numbers of encoded proteins were extracted from the Entrez Genome Project database at the US National Center for Biotechnology Information (NCBI), <http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi>. The analyzed set contained only one representative genome per species, typically the first one to be released to the public, as listed on the above web site. Exceptions included two strains of *E. coli*, K-12 substrain MG1655 and O157:H7 strain Sakai, three strains of *Salmonella enterica*, and two strains of *Orientia tsutsugamushi*, Boryong and Ikeda. The numbers of signaling proteins of each type encoded in every given genome were calculated and manually curated essentially as described previously^{2,9} and reconciled with the data in the Pfam (<http://pfam.sanger.ac.uk/>)⁴⁰ and MiST (<http://genomics.ornl.gov/mist/>)⁴¹ databases, where available. The updated census of signal transduction proteins is available at the web site http://www.ncbi.nlm.nih.gov/Complete_Genomes/SignalCensus.html.

Bacterial IQ scores

The previously observed trend that the number of signal transduction proteins (N_{SP}) grows approximately as square of the genome size² was verified using the expanded genome set with two additional classes of proteins and replacing the genome size with the total number of proteins (N_T) encoded in every given genome. The overall relationship between N_{SP} and N_T was estimated by fitting $\ln N_{SP}$ as linear regression of $\ln N_T$. The genomes with $N_{SP} < 2$ were excluded from the analysis and assigned IQ = 0 (alternatively, arbitrary constant of 0.5 proteins/genome could be added to accommodate cases of $N_{SP} = 0$; this had only a minor effect on the overall results). On a log-log scale, N_{SP} and N_T showed a linear relationship with the slope approximately equal to 2 (Figure 1). The bacterial IQ score was then estimated as a measure of the deviation of N_{SP} for the given genome from the regression line. To make the distribution of IQ scores more symmetric and bell shaped than in the previous work², the IQ scores were normalized by taking natural logarithms of the protein counts. Also, the IQ scores were shifted and scaled by constants so that the minimum value was equal to zero and the mean was approximately equal to 100 (see the Supplemental Methods for details).". Additionally, inclusion in the set of genomes with N_{SP} from 2 to 5 resulted in a much better coverage of small genomes.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by the NIH Intramural Research Program at the National Library of Medicine (MYG) and by grants to EK from the NIGMS (R01 GM076680-02), NIDDK (U01 DK072473), and the NSF (DBI-0544757 and NSF-07140).

References

1. Galperin MY, Kolker E. Curr Opin Biotechnol. 2006; 17:440–447. [PubMed: 16978854]
2. Galperin MY. BMC Microbiol. 2005; 5:35. [PubMed: 15955239]
3. Zhulin IB, Nikolskaya AN, Galperin MY. J Bacteriol. 2003; 185:285–294. [PubMed: 12486065]
4. Galperin MY. Environ Microbiol. 2004; 6:552–567. [PubMed: 15142243]
5. Galperin MY, Gomelsky M. ASM News. 2005; 71:326–333.
6. Römmling U, Gomelsky M, Galperin MY. Mol Microbiol. 2005; 57:629–639. [PubMed: 16045609]
7. Jenal U, Malone J. Annu Rev Genet. 2006; 40:385–407. [PubMed: 16895465]

8. Hickman JW, Tifrea DF, Harwood CS. *Proc Natl Acad Sci USA*. 2005; 102:14422–14427. [PubMed: 16186483]
9. Galperin MY. *J Bacteriol*. 2006; 188:4169–4182. [PubMed: 16740923]
10. Ulrich LE, Koonin EV, Zhulin IB. *Trends Microbiol*. 2005; 13:52–56. [PubMed: 15680762]
11. Galperin MY. *Environ Microbiol*. 2006; 8:1313–1317. [PubMed: 16872395] Galperin MY. *Environ Microbiol*. 2008; 10:3187–3192. [PubMed: 19025554]
12. Gao R, Mack TR, Stock AM. *Trends Biochem Sci*. 2007; 32:225–234. [PubMed: 17433693]
13. Kennelly PJ, Potts M. *Front Biosci*. 1999; 4:D372–385. [PubMed: 10077542]
14. Shi L, Potts M, Kennelly PJ. *FEMS Microbiol Rev*. 1998; 22:229–253. [PubMed: 9862122]
15. Kennelly PJ. *FEMS Microbiol Lett*. 2002; 206:1–8. [PubMed: 11786249]
16. Könneke M, Bernhard AE, de la Torre JR, Walker CB, Waterbury JB, Stahl DA. *Nature*. 2005; 437:543–546. [PubMed: 16177789]
17. Tatusov RL, Galperin MY, Natale DA, Koonin EV. *Nucleic Acids Res*. 2000; 28:33–36. [PubMed: 10592175]
18. Galperin MY. *Environ Microbiol*. 2007; 9:839–845. [PubMed: 17359256]
19. Stinear TP, Seemann T, Harrison PF, Jenkin GA, Davies JK, Johnson PD, Abdellah Z, Arrowsmith C, Chillingworth T, Churcher C, Clarke K, Cronin A, Davis P, Goodhead I, Holroyd N, Jagels K, Lord A, Moule S, Mungall K, Norbertczak H, Quail MA, Rabinowitsch E, Walker D, White B, Whitehead S, Small PL, Brosch R, Ramakrishnan L, Fischbach MA, Parkhill J, Cole ST. *Genome Res*. 2008; 18:729–741. [PubMed: 18403782] Demangel C, Stinear TP, Cole ST. *Nat Rev Microbiol*. 2009; 7:50–60. [PubMed: 19079352]
20. Cole ST, Eiglmeier K, Parkhill J, James KD, Thomson NR, Wheeler PR, Honore N, Garnier T, Churcher C, Harris D, Mungall K, Basham D, Brown D, Chillingworth T, Connor R, Davies RM, Devlin K, Duthoy S, Feltwell T, Fraser A, Hamlin N, Holroyd S, Hornsby T, Jagels K, Lacroix C, Maclean J, Moule S, Murphy L, Oliver K, Quail MA, Rajandream MA, Rutherford KM, Rutter S, Seeger K, Simon S, Simmonds M, Skelton J, Squares R, Squares S, Stevens K, Taylor K, Whitehead S, Woodward JR, Barrell BG. *Nature*. 2001; 409:1007–1011. [PubMed: 11234002] Madan Babu M. *Trends Microbiol*. 2003; 11:59–61. [PubMed: 12598124]
21. O'Hare HM, Duran R, Cervenansky C, Bellinzoni M, Wehenkel AM, Pritsch O, Obal G, Baumgartner J, Vialaret J, Johnsson K, Alzari PM. *Mol Microbiol*. 2008; 70:1408–1423. [PubMed: 19019160]
22. Kumar M, Chatterji D. *Microbiology*. 2008; 154:2942–2955. [PubMed: 18832301]
23. Kim SJ, Kweon O, Jones RC, Freeman JP, Edmondson RD, Cerniglia CE. *J Bacteriol*. 2007; 189:464–472. [PubMed: 17085566] Kim SJ, Kweon O, Jones RC, Edmondson RD, Cerniglia CE. *Biodegradation*. 2008; 19:859–881. [PubMed: 18421421]
24. Kolker E, Picone AF, Galperin MY, Romine MF, Higdon R, Makarova KS, Kolker N, Anderson GA, Qiu X, Auberry KJ, Babnigg G, Beliaev AS, Edlefsen P, Elias DA, Gorby YA, Holzman T, Klappenbach JA, Konstantinidis KT, Land ML, Lipton MS, McCue LA, Monroe M, Pasa-Tolic L, Pinchuk G, Purvine S, Serres MH, Tsapin S, Zakrajsek BA, Zhu W, Zhou J, Larimer FW, Lawrence CE, Riley M, Collart FR, Yates JR 3rd, Smith RD, Giometti CS, Nealson KH, Fredrickson JK, Tiedje JM. *Proc Natl Acad Sci USA*. 2005; 102:2099–2104. [PubMed: 15684069]
25. Fredrickson JK, Romine MF, Beliaev AS, Auchtung JM, Driscoll ME, Gardner TS, Nealson KH, Osterman AL, Pinchuk G, Reed JL, Rodionov DA, Rodrigues JL, Saffarini DA, Serres MH, Spormann AM, Zhulin IB, Tiedje JM. *Nat Rev Microbiol*. 2008; 6:592–603. [PubMed: 18604222]
26. Makarova KS, Aravind L, Galperin MY, Grishin NV, Tatusov RL, Wolf YI, Koonin EV. *Genome Res*. 1999; 9:608–628. [PubMed: 10413400] Shi T, Falkowski PG. *Proc Natl Acad Sci USA*. 2008; 105:2510–2515. [PubMed: 18268351] Lapierre P, Gogarten JP. *Trends Genet*. 2009; 25:107–110. [PubMed: 19168257]
27. Koretke KK, Lupas AN, Warren PV, Rosenberg M, Brown JR. *Mol Biol Evol*. 2000; 17:1956–1970. [PubMed: 11110912] Alm E, Huang K, Arkin A. *PLoS Comput Biol*. 2006; 2:e143. [PubMed: 17083272]
28. Saw JH, Mountain BW, Feng L, Omelchenko MV, Hou S, Saito JA, Stott MB, Li D, Zhao G, Wu J, Galperin MY, Koonin EV, Makarova KS, Wolf YI, Rigden DJ, Dunfield PF, Wang L, Alam M. *Genome Biol*. 2008; 9:R161. [PubMed: 19014707]

29. Takami H, Takaki Y, Uchiyama I. *Nucleic Acids Res.* 2002; 30:3927–3935. [PubMed: 12235376]
30. Hao W, Golding GB. *Genome Res.* 2006; 16:636–643. [PubMed: 16651664]
31. Bork P. *Genome Res.* 2000; 10:398–400. [PubMed: 10779480]
32. Galperin MY. *Comp Funct Genomics.* 2001; 2:14–18. [PubMed: 18628897] Galperin MY, Koonin EV. *Nucleic Acids Res.* 2004; 32:5452–5463. [PubMed: 15479782] Kolker E, Makarova KS, Shabalina S, Picone AF, Purvine S, Holzman T, Cherny T, Armbruster D, Munson RS Jr, Kolesov G, Frishman D, Galperin MY. *Nucleic Acids Res.* 2004; 32:2353–2361. [PubMed: 15121896]
33. Fabret C, Feher VA, Hoch JA. *J Bacteriol.* 1999; 181:1975–1983. [PubMed: 10094672] Galperin, MY. *Systems Biology and Biotechnology of Escherichia coli.* Lee, SY., editor. Springer; Berlin: 2009. p. 133–148.
34. Field D, Garrity G, Gray T, Morrison N, Selengut J, Sterk P, Tatusova T, Thomson N, Allen MJ, Angiuoli SV, Ashburner M, Axelrod N, Baldauf S, Ballard S, Boore J, Cochrane G, Cole J, Dawyndt P, De Vos P, DePamphilis C, Edwards R, Faruque N, Feldman R, Gilbert J, Gilna P, Glockner FO, Goldstein P, Guralnick R, Haft D, Hancock D, Hermjakob H, Hertz-Fowler C, Hugenholtz P, Joint I, Kagan L, Kane M, Kennedy J, Kowalchuk G, Kottmann R, Kolker E, Kravitz S, Kyrpides N, Leebens-Mack J, Lewis SE, Li K, Lister AL, Lord P, Maltsev N, Markowitz V, Martiny J, Methe B, Mizrachi I, Moxon R, Nelson K, Parkhill J, Proctor L, White O, Sansone SA, Spiers A, Stevens R, Swift P, Taylor C, Tateno Y, Tett A, Turner S, Ussery D, Vaughan B, Ward N, Whetzel T, Gil I San, Wilson G, Wipat A. *Nat Biotechnol.* 2008; 26:541–547. [PubMed: 18464787]
35. Kunin V, Goldovsky L, Darzentas N, Ouzounis CA. *Genome Res.* 2005; 15:954–959. [PubMed: 15965028]
36. Norton JM, Klotz MG, Stein LY, Arp DJ, Bottomley PJ, Chain PS, Hauser LJ, Land ML, Larimer FW, Shin MW, Starkenburg SR. *Appl Environ Microbiol.* 2008; 74:3559–3572. [PubMed: 18390676] Oda Y, Larimer FW, Chain PS, Malfatti S, Shin MV, Vergez LM, Hauser L, Land ML, Braatsch S, Beatty JT, Pelletier DA, Schaefer AL, Harwood CS. *Proc Natl Acad Sci USA.* 2008; 105:18543–18548. [PubMed: 19020098] Chain PS, Denef VJ, Konstantinidis KT, Vergez LM, Agullo L, Reyes VL, Hauser L, Cordova M, Gomez L, Gonzalez M, Land M, Lao V, Larimer F, LiPuma JJ, Mahenthalingam E, Malfatti SA, Marx CJ, Parnell JJ, Ramette A, Richardson P, Seeger M, Smith D, Spilker T, Sul WJ, Tsoi TV, Ulrich LE, Zhulin IB, Tiedje JM. *Proc Natl Acad Sci USA.* 2006; 103:15280–15287. [PubMed: 17030797]
37. Deutscher J, Francke C, Postma PW. *Microbiol Mol Biol Rev.* 2006; 70:939–1031. [PubMed: 17158705]
38. Helmann JD. *Adv Microb Physiol.* 2002:4647–110. Campbell EA, Westblade LF, Darst SA. *Curr Opin Microbiol.* 2008; 11:121–127. [PubMed: 18375176]
39. Potrykus K, Cashel M. *Annu Rev Microbiol.* 2008; 62:35–51. [PubMed: 18454629] Witte G, Hartung S, Buttner K, Hopfner KP. *Mol Cell.* 2008; 30:167–178. [PubMed: 18439896]
40. Finn RD, Tate J, Mistry J, Coghill PC, Sammut SJ, Hotz HR, Ceric G, Forslund K, Eddy SR, Sonnhammer EL, Bateman A. *Nucleic Acids Res.* 2008; 36:D281–D288. [PubMed: 18039703]
41. Ulrich LE, Zhulin IB. *Nucleic Acids Res.* 2007; 35:D386–D390. [PubMed: 17135192]
42. Johnston S, Lin S, Lee P, Caffrey SM, Wildschut J, Voordouw JK, da Silva SM, Pereira IA, Voordouw G. *Environ Microbiol.* 2009; 11:981–991. [PubMed: 19077010]

Figure 1. Relationship between the number of signal transduction proteins and the total number of proteins encoded in that genome

The phylogenetic affiliation of the source genomes is indicated as follows: *Actinobacteria*, yellow circles; *Cyanobacteria*, cyan circles; *Proteobacteria*, empty squares; *Firmicutes*, blue triangles; members of the CFB group (phyla *Bacteroidetes* and *Chlorobi*), green triangles; members of other bacterial phyla, pink triangles. Archaeal genomes are not shown but have been included in the calculation of the best fit.

Figure 2. Signal transduction protein family profiles for 10 mycobacterial genomes

The source organisms are indicated by the following colors: *Mycobacterium leprae* TN, purple; *M. bovis* AF2122/97, pink; *M. tuberculosis* H37Rv, red; *M. ulcerans* Agy99, buff; *M. avium* subsp. *paratuberculosis* K-10, blue; *M. abscessus*, salmon; *M. marinum* M, green; *M. gilvum* PYR-GCK, lavender; *M. vanbaalenii* PYR-1, pale lavender; *M. smegmatis* str. MC2 155, light blue.

Figure 3. Signal transduction protein family profiles for *Shewanella* species

The source organisms are shown in the following order: *Shewanella denitrificans* OS217, *Shewanella halifaxensis* HAW-EB4, *Shewanella pealeana* ATCC 700345, *Shewanella frigidimarina* NCIMB 4, *Shewanella putrefaciens* CN-32, *Shewanella piezotolerans* WP3, *Shewanella loihica* PV-4, *Shewanella amazonensis* SB2B, *Shewanella oneidensis* MR-1, *Shewanella baltica* OS155, *Shewanella woodyi* ATCC 51908, *Shewanella sediminis* HAW-EB3.

Figure 4. Signal transduction protein family profiles for various bacilli

Family profiles of the representatives of family *Bacillaceae* (A), order *Bacillales* (B), class *Bacilli* (C) and the phylum *Firmicutes* (D).

A. The source organisms are shown in the following order: *Oceanobacillus iheyensis* HTE831, *Anoxybacillus flavithermus* WK1, *Geobacillus kaustophilus* HTA426, *Geobacillus thermodenitrificans* NG80-2, *Bacillus pumilus* SAFR-032, *B. subtilis* subsp. *subtilis* str. 168, *B. amyloliquefaciens* FZB42, *B. licheniformis* ATCC 14580, *B. clausii* KSM-K16, *B. halodurans* C-125, *B. anthracis* str. Ames, *B. cereus* ATCC 14579, *B. thuringiensis* serovar *konkukian* str. 97-27, *B. weihenstephanensis* KBAB4.

B. The source organisms are as in panel A, with the addition of *Staphylococcus epidermidis* ATCC 12228, *Staph. haemolyticus* JCSC1435, *Staph. aureus* subsp. *aureus* N315, *Listeria monocytogenes* EGD-e, *Listeria innocua* Clip11262, and *Lysinibacillus sphaericus* C3-41.

C. The source organisms are as in panel B, with the addition of *Lactobacillus helveticus* DPC 4571, *Lactococcus lactis* subsp. *lactis* I11403, *Streptococcus pyogenes* M1 GAS, *Strep. mutans* UA159, and *Strep. pneumoniae* TIGR4.

D. A heat map of signal transduction protein family profiles of 78 *Firmicute* species, clustered by the number of proteins of different types normalized by the proteome size. For each organism, the number of signal transduction proteins was divided by the square of the total number of encoded proteins (a log of this value is roughly equivalent to the IQ score) and then transformed to ranks to remove the influence of outliers and asymmetry in the distributions. The neighbor-joining tree of these ranks clearly separates members of the *Bacilli* (upper half) and the *Clostridia* (bottom half). The organisms with low IQs (shades of

blue) tend to cluster toward the top of heat map while the higher IQs (shades of red) cluster toward the bottom.

Table 1Improved coverage of microbial diversity by complete genomes^a

	2005 ^b	2009 ^b
Represented phyla		
Bacterial ^c	13 ^c	20 ^c
Archaeal	3	4
Total number of genomes	167	555
Bacterial	147	504
Archaeal	20	51
Larger than 7.5 Mbp	3	23
The number of encoded proteins		
Total	497,733	1,808,414
Histidine kinases	3,789	16,060
Chemotaxis sensors (MCPs)	1,128	4,742
Diguanylate cyclases ^d	1823	6,669
c-di-GMP phosphodiesterases ^d	1512	4,824
Adenylate cyclases	285	1,087
Ser/Thr protein kinases	765	3,471

^a All data refer to a non-redundant set of microbial genomes that includes a single representative genome from every bacterial and archaeal species. See http://www.ncbi.nlm.nih.gov/Complete_Genomes/SignalCensus.html for details.

^b The data as of January 1st of the respective year.

^c In both cases, *Tenericutes* (mycoplasmas) are counted as a separate phylum.

^d Including proteins with fused GGDEF and EAL domains.

Table 2

Mean IQ scores within key groups of prokaryotes

Phylum, class	No. genomes	Obligate parasites ^a	Obligate anaerobes ^b	Signal transduction proteins (% total)	Mean IQ score ± SD
BACTERIA					
Actinobacteria	44	10	–	5,515 (2.84%)	89.1 ± 25.0
Bacteroidetes	13	6	5	1,090 (2.78%)	77.3 ± 37.9
Chlamydiae	7	7	–	79 (1.01%)	113.4 ± 25.1
Chlorobi	10	–	–	431 (1.88%)	92.9 ± 29.2
Chloroflexi	7	–	–	1,280 (5.72%)	140.1 ± 21.5
Cyanobacteria	22	–	–	3,428 (3.82%)	97.0 ± 25.0
Firmicutes	78	14	28	7,520 (3.30%)	108.4 ± 18.7
Bacilli	50	12		3,696 (2.60%)	97.9 ± 14.2
Clostridia	28	2	28	3,824 (4.47%)	127.1 ± 8.5
Proteobacteria	264	70	15	41,540 (4.22%)	105.3 ± 30.6
Alpha	79	31	–	9,901 (3.62%)	102.9 ± 20.8
Beta	44	8	–	8,477 (4.01%)	97.6 ± 27.5
Gamma	107	22	–	15,930 (4.10%)	100.4 ± 34.5
Delta	21	1	13 ^a	6,166 (7.06%)	135.4 ± 14.2
Epsilon	13	8	2	1,066 (4.31%)	138.8 ± 30.6
Spirochaetes	12	10	–	822 (3.61%)	141.0 ± 25.5
Tenericutes	20	19	–	54 (0.36%)	60.2 ± 44.5
Thermotogae	8	–	8	449 (3.00%)	138.6 ± 11.6
ARCHAEA					
Crenarchaeota	18	–	7	155 (0.42%)	40.1 ± 23.9
Euryarchaeota	31	–	23	1,339 (1.87%)	79.6 ± 34.8

^aIncluding obligate symbionts of eukaryotes.^bThese data are incomplete and may be biased, see Johnston *et al.*⁴² and references therein for the discussion of oxygen tolerance in *Desulfovibrio vulgaris*.