JB Accepts, published online ahead of print on 10 October 2008
J. Bacteriol. doi:10.1128/JB.01084-08

Journal of Bacteriology submission (invited, ref. Igor Zhulin)

Title

Discovering functional novelty in metagenomes: examples from light-mediated processes

Running Title (54 chars allowed)

Novel function in light-mediated microbial processes

Authors

Amoolya H. Singh, Tobias Doerks, Ivica Letunic, Jeroen Raes, Peer Bork*

Affiliations

Structural and Computational Biology Unit

European Molecular Biology Laboratory

Meyerhofstrasse 1

69117 Heidelberg

***Corresponding Author**

Peer Bork, bork@embl.de

Manuscript Information

Number of figures: 6

Number of tables: 1

Number of pages: 34

Abstract

The emerging coverage of diverse habitats by metagenomic shotgun data opens new avenues of discovering functional novelty using computational tools. Here, we apply three different concepts for predicting novel functions within light-mediated microbial pathways in five diverse environments. Using phylogenetic approaches, we discovered two novel deep-branching subfamilies of photolyases (involved in light-mediated repair) distributed abundantly in high-UV environments. Using neighborhood approaches, we were able to assign to BLUF domain-containing proteins (involved in light-sensing) seven novel functional partners in luciferase synthesis, nitrogen metabolism, and quorum sensing. Finally, by domain analysis, we predict for RcaE proteins (involved in chromatic adaptation) sixteen novel domain architectures that indicate novel functionalities in habitats with little or no light. Quantification of protein abundance in the various environments supports our findings that bacteria utilize light for sensing, repair, and adaptation far more widely than previously thought. While the discoveries illustrate the opportunities in function discovery, we also discuss the immense conceptual and practical challenges that come along with this new type of data.

Introduction

One of the central questions in biology, starting from the time of Charles Darwin, has been the extent and distribution of biological diversity (68). The recent sequencing of several hundred bacterial and archaeal genomes and metagenomes, along with the discovery of large-scale lateral gene transfer (10) and recombination (25) in bacterial evolution, has not only renewed interest in the question of diversity, but also confounded it. The sequencing projects reveal that, contrary to

previous estimates, it is microbes that account for the vast majority of diversity in phenotype and genotype on earth (44, 47). Underlying this dazzling diversity in species and habitat is molecular diversity. Indeed, we are just beginning to scratch the surface of this molecular diversity (50). Even though our understanding of how the living world functions at the molecular level is far from complete, the discovery of novel molecules has important applications to medicine, agriculture, industry, and environmental conservation and remediation.

But how are we to discover functional novelty in the exponentially increasing amounts of sequenced genes and habitats (Fig. 1)? The naïve method, which is to search for homology to known molecules and mark everything else as novel, is prone to errors due to the existence of paralogous sequences, i.e. homologs with likely different functionalities, as well as paralogous domains within an otherwise homologous sequence that may lead to divergent function (17). To address these challenges, three major, non-exclusive concepts have been successfully used to establish functional similarity, and, conversely, to identify functional novelty: *(i)* operons and conserved gene neighborhoods, *(ii)* protein domain architectures, and *(iii)* protein subfamilies. Operon and gene neighborhood methods assume that if multiple genes are adjacent on a chromosome or contig, they are more likely to participate in the same cellular function (19, 48). The neighborhood approach is especially suitable when homology-based methods fail to detect sequences below the threshold for similarity (30). Domain-based methods infer function of similar segments within otherwise different sequences and are currently utilized by curated databases of known domains (e.g. (18)). This approach is useful in the analysis of multi-domain proteins that evolve in modular fashion, such that each domain may have high sequence similarity to a different gene and its evolution cannot be traced by homology alone (55). Finally,

47 subfamilies can be identified within a family of homologous sequences by abstracting the
48 information from the family's multiple sequence alignment into a generalized statistical profile
49 (e.g., using hidden Markov models or support vector machines) (11), and then searching for
50 shared properties (e.g. amino acids, hydrophobicity). This technique has been successful at
51 identifying novel biological function (2, 29, 36) and even novel species (13).
52
53 Although these methods are conceptually straightforward, identifying novelty from
54 environmental data remains difficult. The primary reason is the sheer volume of the data, for
55 which there is no centralized repository or standardized reporting of sampling conditions. The
56 size of the datasets is further exacerbated by problems with the data itself, such as the presence
57 of incomplete gene fragments, the uniformity of sequence coverage, and the use of shotgun
58 sequencing. Incomplete gene fragments, an artifact of current environmental sequencing
59 methodologies, limit the ability to correctly predict open reading frames and assign function.
60 Uniform sequence coverage implies that if a protein family is rare in a particular environment, or
61 belongs to a rare species, it might not be seen at all. Since functional novelty seems to be
62 primarily contributed by rare families (30, 51, 70) that mediate unusual niche-specific
63 adaptations, the inability to detect rare proteins fundamentally limits our ability to discover
64 novelty—although this may change as technical advances enable targeted deep sequencing in
65 high-diversity environments. Finally, the use of shotgun sequencing techniques (as opposed to,
66 for example, single-cell sequencing) makes it tricky to unite species and function identification.
67 The existing protocols to map taxonomy are either limited to querying against a small number
68 (e.g. 30–40) of marker genes (66), or falling back on error-prone annotation transfer from
69 homology (33).

In addition to problems with the size and nature of metagenomic data, computational tools must be adapted for reproducibly handling gigabytes or terabytes of data, leading to constraints in memory, CPU, and network bandwidth at every level of analysis. Tools must be adapted to process, filter, assemble, and align the sequence data; identify genes; annotate the genes with function; map genes or sequences to taxonomy; estimate species evenness and richness; construct phylogenetic trees; perform multivariate analyses against ecological metrics; build and validate population or metabolic models where time-series data are available; and visualize the results. Even as computational biologists adapt standard tools to complete these tasks, mathematicians and statisticians must rigorously re-assess the suitability of different methods to large datasets, identify sources of analytical and numerical errors, and revise estimates of sensitivity and specificity.

Even if novel techniques such as single cell sequencing reduce some of above problems in the future, certain challenges will remain unless our entire planet has been genetically explored in sufficient depth. This is because the remaining challenges are conceptual by nature. For example, identification of orthology is already extremely difficult with complete genomes in hand due to chromosomal inversions, gene fusions, alternative splicing, retrotranscription and a variety of genetic processes that dilute the necessary information .This genetic uncertainty is matched by a functional one: because the term ‘function’ remains in use with an operational rather than absolute definition (8, 9, 30), annotation processes will remain of insufficient depth for quite some time.

93 Despite these limitations, the benefits of "bioprospecting" for natural and naturally-derived
94 products are considerable, with potential to cure genetic and infectious diseases, arrest
95 environmental destruction, and offset global energy shortage. Here, we hope to raise awareness
96 of the potential and pitfalls of using environmental sequence data to discover novelty, and
97 illustrate the promise of our methods to discover novelty in light-mediated microbial pathways
98 functioning in sensing, repair, and adaptation.

Materials and Methods

Collecting genome, metagenome, and habitat data. We collected raw data on habitat and number of sequenced ORFs (Fig. 1) from the Genomes Online database (43) on 12 June 2008, consisting of 802 genomes (669 species of bacteria, 53 archaea, 80 eukaryotes) and 25 metagenomes. The date of publication was used as the sequencing date for a genome or metagenome. We classified the 86 distinct annotated habitats for these 827 genomes/metagenomes into 10 categories using the Habitat-Lite subset of terms (32) from the Environmental Ontology (www.environmentontology.org). When an organism was reported to have multiple habitats, the primary one was used as input to Fig. 1. Primary habitats were checked against Bergey's Manual (6), the online catalogs of the American Type Culture Collection (www.atcc.org) and German Collection of Microorganisms and Cell Cultures (www.dsmz.de), and a previous large-scale description of bacterial phenotype and habitat (57). In the interest of clarity for Fig. 1, we grouped certain habitats: "Extreme environment" in the figure legend corresponds to Habitat-Lite categories "hot spring," "hydrothermal vent," or "extreme environment." "Sediment/sludge" in the figure legend corresponds to Habitat-Lite categories "sediment," "sludge," or "biofilm."

Calculating homologs and protein abundances in metagenomes. Metagenome sequence data from five metagenomes (6,109,937 ORFs from surface sea water from the Global Ocean Survey (70) including the Sargasso Sea (64), 46,771 ORFs from northern California acidic mine drainage (63), 121,927 ORFs from deep-sea Pacific whalefall (62), 183,159 ORFs from Minnesota farm soil (62), and 135,756 ORFs from a Mexican hypersaline microbial mat (39))

were BLASTed against the entire set of 1,510,991 proteins (representing 373 sequenced organisms) in the STRING 7.0 database (67) using wu-blastall with the parameters: –a 1 –p blastp mformat 2 –filter seg –E 1 –V 17000000 –B 17000000. From this dataset, the number of hits to each of the 20 query light-mediated proteins (Table 1) were counted, discarding hits less than 60 bits, which has been previously estimated to correspond roughly to an *e*-value < 10^8 (30). The abundances of genome orthologs was counted based on the size of Clusters of Orthologous Groups (COGs) and Non-supervised Orthologous Groups (NOGs) previously identified by the STRING database. Because each dataset has a different total number of ORFs, protein abundances were normalized in Matlab (The Mathworks, Natick, MA), as follows. For each row (genome or metagenome) of Fig. 3, the absolute number of hits to each query protein was divided by the total number of predicted ORFs in that dataset. This percentage is reported in Fig. 3B.

Constructing alignments and phylogenetic trees. All metagenome hits were then filtered for length (>250 amino acids) and diversity (<80% identity to any other hit in the orthologous family). Amino acid sequences were aligned using MUSCLE (21) with clustal-strict output, and the alignments were manually spot-checked. Further, 100 bootstrap replicates of each alignment were generated using seqboot from the Phylip package (23) with default parameters, and these replicates were used to estimate phylogenetic trees with PHYML (27) using 8 Gamma-estimated rate categories and the Jones-Taylor-Thornton (JTT) rate matrix. A consensus of the 100 resulting trees was obtained with the consense package of Phylip (extended majority rule), and branch lengths for the consensus were calculated using tree-puzzle (54) with 8 Gamma-estimated rate categories and the JTT rate matrix. To check the taxonomic diversity of the metagenome hits

and exclude the possibility that novel discoveries were the result of errors in sequence assembly, phylogenetic placement of each of the metagenome fragments was calculated, as follows. Metagenome fragments were used as query proteins and BLASTed against all STRING 7.0 genome proteins with a bitscore cutoff of 60 bits. All hits within 5% of the maximum bitscore were retained, and then mapped to a phylogenetic tree of sequenced genomes (16) as previously described (66). Not all best hits were to genomes; some were to internal nodes of the tree. For cases where metagenome hits originated from closely related species, each gene was mapped to the assembled reads to conclusively exclude the possibility of assembly errors. All phylogenetic trees were visualized on iTOL (40).

Searching for neighborhoods and domains. Gene neighborhoods for each of the metagenome hits to the 20 query proteins were calculated as previously described (30), only counting genes as neighbors if they were adjacent on the contig in the same transcription direction.We used cotranscribed gene neighbors (as opposed to bidirectional or convergently transcribed gene neighbors) because their existence was previously established to be most predictive of related function (38). Protein domains for metagenome hits were obtained by searching against the SMART version 5 database (41) with default parameters.

All analyses were carried out on a dedicated 256-node supercomputing cluster with 1,320 CPU-cores communicating via a Gigabit-Ethernet network, each running a 64-bit Linux operating system with 1G of memory.

166 **Results**

167

168 To illustrate the application of methods for discovering functional novelty, we focused on light-
169 mediated microbial pathways. Although light is an important abiotic factor impacting all the
170 major biological processes (growth, sensing, maintenance, and reproduction), its utilization by
171 microbes remains poorly understood. We began by constructing a taxonomy of light-mediated
172 processes (Fig. 2) guided by the GO biological process ontology (31) with the broad categories
173 of growth (photosynthesis and circadian rhythms), sensing (phytochromes), maintenance (DNA
174 repair and pigment synthesis), and adaptation (phototaxis, chromatic adaptation, and
175 bioluminescence). We omitted from further analysis photosynthetic bacterial pathways, whose
176 evolution and variation have been described extensively elsewhere (5, 14, 52, 69), and instead
177 focused on sensing, repair, and adaptation. Next we conducted a literature search (1, 3, 14, 15,
178 20, 22, 45, 46, 58-60, 65) to identify representative bacterial proteins and their orthologs
179 involved in those processes. We searched this candidate list of 20 proteins (Table 1) in five
180 environmental metagenomes comprising 59 sample sites of surface sea water (64, 70), acidic
181 mine runoff from an abandoned gold mine in northern California (63), three sample sites of
182 deep-sea Pacific and Antarctic whalefall carcass (62), 5g of Minnesota farm topsoil (62), and ten
183 layers of a 41.5mm-thick Mexican hypersaline microbial mat (39). For the purpose of analysis,
184 we denoted the surface sea water and top two layers of the microbial mat to be "high-light"
185 environments, and the others to be "variable-light" environments. We note here that our choice
186 of proteins is by no means an exhaustive survey of all light-mediated proteins and pathways in
187 bacteria, but rather a selected list to illustrate metagenomic data mining techniques.

188

A quantitative estimate of light-related proteins in the environment. We first counted both the absolute and relative amounts of 20 metagenome proteins (Fig. 3) functioning in light-mediated growth (7 protein families participating in photosynthesis and the circadian clock), sensing (6 protein families of blue- and red-light sensors), repair and defense (3 protein families consisting of photolyases, water-soluble carotenoids as intracellular UV sunscreens, and scytonemin as extracellular sunscreen), and adaptation (4 protein families participating in phototaxis and complementary chromatic adaptation). As expected, light-mediated growth and repair proteins are found predominantly in high-light environments in both absolute and relative terms (30,435 or 94% of all growth-, defense- or repair-related proteins) rather than variable-light environments (2,021 or 6% of all proteins). Interestingly, sensing and adaptation proteins are over-represented in variable-light environments (13,786 or 57% of all sensing proteins, 17,960 or 60% of all adaptation proteins) as compared to high-light environments (10,273 or 43% of all sensing proteins, 11,736 or 40% of all adaptation proteins). Further, unlike sensing proteins, adaptation proteins are present in high amounts even in deeper (darker) layers of the salt mat. Below we discuss three examples of novel light-mediated function in each of these categories, discovered using analysis of gene neighborhood, protein domains, and protein subfamilies respectively.

Novel light-mediated sensing: neighborhood approach. For a candidate sensing process mediated by light, we chose proteins containing the BLUF (blue-light FAD-binding) domain (26, 45) as it is rather rare in genomes and we expected a limited variety in operon organization. BLUF-domain proteins are part of the larger family of blue light photosensors that use flavin chromophores, which together with the phytochromes, rhodopsins, and UV receptors, make up

212 the four major classes of bacterial light-sensing proteins (14). Proteins containing a BLUF
213 domain have been shown to function as sensors upstream of phototaxis (24), nucleotide
214 metabolism (35), and repression of anoxygenic photosynthesis (28).The domain is extremely
215 well-conserved among Proteobacteria and Cyanobacteria, absent from Archaea, and absent from
216 Eukaryotes except for the protist *Euglena gracilis*. As expected, BLUF domain containing
217 proteins are not only relatively rare in the genomes (34 instances, or 0.002% of all proteins; Fig.
218 3) but also in the metagenomes (73 instances, of which 46 are from surface sea water, accounting
219 for 0.0008% of that dataset; Fig. 3).
220
221 The vast majority of BLUF-containing proteins in the metagenomes do not contain additional
222 domains, which precludes a domain-based analysis as above. Further, the BLUF domain is short
223 (98 amino acids) and highly conserved in sequence (70 % identity of the multiple sequence
224 alignment), so that constructing phylogenetic trees with robust statistical support is practically
225 impossible. Thus a tree- or subfamily-based analysis is ruled out as well. However, since BLUF-
226 domain proteins are known to function in sensing and stress response, we surmised that either
227 their expression or the expression of their functional partners would be inducible, and thus
228 correlated with the expression of nearby genes on the chromosome. This made it a good
229 candidate for gene neighborhood analysis.
230
231 For the 73 environmental BLUF-domain proteins, we identified 36 functionally characterizable
232 neighborhoods (32 neighborhoods from surface sea water, 4 from deep-sea whalefall; Fig. 4).
233 We rediscovered the known functions of BLUF in phototaxis (2 neighborhoods), nucleotide
234 metabolism (5 neighborhoods), and repression of anoxygenic photosynthesis (5 neighborhoods).

235 Interestingly, we also discovered neighborhoods of BLUF with novel function, including
236 luciferase synthesis (4 neighborhoods), nitrate metabolism (3 neighborhoods) quorum sensing (3
237 neighborhoods). These neighbors are promising candidates for experimental elucidation of
238 BLUF's cellular role.
239
240 **Novel light-mediated adaptation: domain approach.** For a candidate adaptation process
241 mediated by light, we chose the protein RcaE (regulator of chromatic adaptation), best
242 characterized in the filamentous cyanobacterium *Fremyella diplosiphon* (7, 37). This protein
243 regulates the ability to radically alter cell pigmentation in response to changes in ambient light
244 wavelength, particularly across the green-red range, and has been shown to optimize light
245 antennae for photosynthesis (65). In the sequenced genomes, RcaE is a relatively rare protein
246 (212 homologs primarily in cyanobacteria and plants, accounting for 0.01% of total proteins). In
247 the metagenomes, however, RcaE orthologs are over-represented in variable-light environments
248 (723 proteins or 0.53% in the microbial mat, 1155 proteins or 0.97% in deep-sea whalefall; Fig.
249 3) and under-represented in high-light environments (5,434 proteins or 0.08% in surface sea
250 water; Fig. 3). Because the known cyanobacterial homologs of this protein have an unusual
251 domain composition (GAF-PAS-PAC-HisKA-HATPase-REC) that has been modified among
252 plants and non-photosynthetic bacteria (Fig. 5), we hypothesized that additional modular
253 arrangements of this protein must exist in the wild. Further, we expected that these novel domain
254 arrangements, especially the associated receiver/output domains, would provide clues as to the
255 downstream cellular function being adapted.
256

257 Finding "true" domain hits within the metagenomes, however, proved to be less than
258 straightforward. Several of RcaE's domains, such as the kinase (HisKA-HAPTase), redox
259 sensing (PAS-PAC (61)), and response regulator receiver (REC (49)), are extremely widespread
260 and promiscuous, and the metagenome datasets typically contain fragments of genes without the
261 key light-sensing GAF domain, together leading to many spurious hits. Of the 11,456
262 environmental sequences initially obtained at >60bits BLAST score (corresponding roughly to a
263 stringent *e*-value $<10^{-8}$ (30)), only 762 sequences were longer than 250 amino acids and less than
264 80% identical to one another. Of these, 112 sequences contained the GAF domain and aligned to
265 the query at greater than 60% of their length, a typical cutoff for excluding single-domain hits
266 (34). However, since this cutoff entirely eliminated proteins from the hypersaline microbial mat,
267 we relaxed it to 40% alignment length, which yielded 650 environmental sequences (632 surface
268 sea water, 18 deep-sea whalefall, 26 soil, 11 salt mat, and 2 acid mine).
269
270 This sample of 650 environmental sequences contained 50 unique domain arrangements, of
271 which 16 were novel, and not seen before in any genome. All 16 novel arrangements preserve
272 the pattern of "specific sensing domain(s) – PAS-PAC – kinase – receiver", but vary in the
273 number and type of domain repeats. Two arrangements include repeats in PBPb (periplasmic
274 solute-binding) as one of the sensing domains and another has PBPp with PAS repeats without a
275 PAC domain. Another seven arrangements include 3–6 repeats of PAS-PAC; three arrangements
276 have duplicated REC domains. This is a surprising result because domain repeats are generally
277 less common in bacteria than in eukaryotes, where they are thought to encode increased
278 variability to compensate for longer eukaryotic generation times (4). However, the conservation
279 of the overall domain pattern of the protein, together with the remarkable number of PAS-PAC

repeats, allows us to speculate that this domain architecture provides increased substrate affinity and a tuning switch on the sensitivity of the response.

Novel light-mediated repair: a subfamily approach. For a candidate repair process mediated by light, we focused on photolyases, an intriguing family of light-activated DNA repair enzymes that are virtually ubiquitous in bacterial species. Photolyases reverse T<>T cyclobutane dipyrimidine dimers (CPDs) formed by UV damage to DNA, using a photon of light to transfer electrons from a catalytic flavin chromophore to the damaged DNA (53). While the structure and function of photolyases was being characterized, an additional family of homologs, the cryptochromes, were discovered (12, 42, 53, 56). Cryptochromes are similar to photolyases in sequence and three-dimensional structure, but lack catalytic activity for DNA repair, and have unclear function. To date, two kinds of photolyases (CPD-I and CPD-II) and three kinds of cryptochromes have been identified (plant cryptochromes, animal cryptochromes, and CRY-DASH proteins—named after the representative four genera in which they were identified: Drosophila Arabidopsis Synechocystis Homo). Thus the photolyase-cryptochrome family in sequenced genomes is quite large, spanning the inclusive gene family COG0415 (328 proteins in 209 species), but also including COG3046 (56 genes in 53 species), COG4338 (35 genes in 33 species) and NOG16378 (22 proteins in 19 species). In the metagenomes, the photolyase-cryptochrome homologs are over-represented in surface sea water (9,703 proteins, or 0.1% of the total) and the top two layers of the microbial mat (8 proteins, or 0.6% of the total) as compared to all other environments together (84 proteins). This is consistent with the high amount of UV radiation incident on surface waters of the open ocean or the top layers of the microbial mat, but does not account for the other possible functions of cryptochromes in remaining niches.

To tease apart the functional diversity of this protein family, we undertook a subfamily analysis by constructing high-quality alignments, feeding them to a hidden Markov model (HMM), and using the resulting HMM profile to refine the alignment and construct a phylogenetic tree. Although this approach is now standard practice when analyzing small gene families, it foundered when fed with roughly 10,000 sequences, and our phylogenetic tools of choice (phyml (27) and tree-puzzle (54)) often took weeks to estimate a tree when running on dedicated supercomputing clusters, even without statistical bootstraps. We therefore added several filtering steps to our protocol. First, we removed sequences shorter than 250 amino acids as well as sequences that were >80% identical to any other sequence in the dataset. This approximately halved the number of photolyase hits from 9,703 to 4,828. Next we constructed phylogenetic trees by randomly sub-sampling the 4,828 sequences in batches of 1,000 sequences, and compared the resulting trees for topology and grouping. Finally, we combined the genomic and metagenomic sequences, constructed trees again, and checked whether the same groupings resulted.

Because bootstraps on the photolyase trees could not be calculated on the entire dataset of approximately 10,000 sequences, we report here 1,196 photolyase-cryptochrome orthologs from the sequenced genomes and four metagenomes: surface sea water from the Sargasso sea samples 1–4, farm soil, acidic mine runoff, and deep-sea whalefall (Fig. 6). Although the bootstrap at the deeper branches is somewhat low (<25%), it is consistently high near the leaves (>80%), indicating that the relationships between the subfamilies are poorly resolved, but the clustering within subfamilies is strong. Most notably, our tree recovers the four known groups of

photolyases (CPD-I, CPD-II, DASH cryptochromes, and animal cryptochromes), and additionally identifies two novel deep-branching groups of photolyases/cryptochromes. The deepest branching "novel family I" represents a new family of 34 photolyases/cryptochromes of unknown function never seen before in the genomes. Because the tree covers photolyases from all known species from all three domains of life, the novel family must include newly detected enzymes of unknown function related to the cryptochrome superfamily. The taxonomic origins of these enzymes are a mixture of *Pelagibacter*/SAR11-like species and other Alpha-Proteobacteria (69%), and Cyanobacteria dominated by *Prochlorococcus* (31%). "Novel family II," which is clearly grouped between CPD-II photolyases and animal DASH cryptochromes, is an additional uncharacterized diverse subfamily with 54 sequences from Alpha-Proteobacteria (80%) and Cyanobacteria (20%). The species compositions are as expected, since both Alpha-Proteobacteria and Cyanobacteria are the dominant marine microbial species. Both newly discovered photolyase/cryptochrome families are exciting candidates for further computational and experimental characterization.

Discussion

We have analyzed the distribution and molecular diversity of light-mediated proteins from five diverse environments receiving varying ambient light. Instead of assigning genes to functions as is usually done with current metagenomics data sets using BLAST-like procedures, we sought to identify novel protein functions. Using gene neighborhood, domain, and subfamily analysis, we have attempted to characterize functional novelty in proteins sensing light, adapting to changes in light color, and repairing UV-damaged DNA. We found new functional partners for blue-light sensors, new domain architectures of chromatic adaptation proteins, and new subfamilies of DNA repair enzymes. Our results represent the first quantification of these cellular processes and provide an early insight into their spectacular diversity.

While these results serve as a proof of principle for the possibility to infer novel functionality by using the three different concepts described above, and represent the opportunities inherent in those huge datasets, they implicitly also illustrate the challenges of mining environmental sequence data to discover novel function. The difficulty is due to the nature of the data itself (vast amount, fragmented, uniform coverage, shotgun sequence); the lack of appropriate methods and analysis tools together with bottlenecks in CPU, memory, and network bandwidth; and ongoing conceptual difficulties with defining homology/paralogy and novel function. Indeed, while the sequencing of environment after environment continues to generate gigabytes of data, there has been little corresponding investment in the analysis of these data, pointing to an urgent and immediate need for methods and tool development. For example, we would have been unable to derive bootstrap values for some of the phylogenetic trees had we included more

363 environments, not to mention the enormous challenges on the CPU to compute all the data. Our
364 previous work demonstrated that even a slightly better function assignment protocol could lead
365 to a near-doubling of number of functional annotations for gene fragments from 40% to 70%
366 (30)—suggesting that with improved analysis, perhaps only half the sequence data are really
367 needed. The saved effort could be redirected at gathering time-series and spatial data, which
368 would help to interpret functional novelty, and allow the development of dynamical models to
369 explore larger concepts in ecology and evolution such as species succession, pathway evolution,
370 or metabolic flux.
371
372 In summary, we have demonstrated the use of computational analysis techniques for discovering
373 molecular functional novelty in environmental snapshots of bacterial communities. Our results
374 indicate that information on gene neighborhood, protein domains, and subfamilies can all be
375 successfully used to discover functional novelty, although various challenges considerably
376 hamper the analysis and will continue to do so as more data are generated in the future.
377
378 **Acknowledgments**
379
380 We thank Chris Creevey and Jean Muller for suggesting alternate phylogenetic analysis methods
381 when the existing tools crashed, Mani Arumugam for helping with domain analysis, and Yan
382 Yuan for excellent technical assistance. Thanks to members of the Bork group for useful
383 discussions and feedback.

Figure Legends

Figure 1. Trends in the increase of genomic data and represented habitats. The number of sequenced open reading frames continues to increase exponentially, accompanied by an increase in the number and complexity of represented habitats. In 1995, the two sequenced organisms (*Haemophilus influenzae* and *Mycoplasma pneumoniae*) contributed just a few thousand genes to the public databases and represented a single habitat (organism-associated). By 2008, well over ten million genes have been sequenced from over 150 distinct habitats. Raw habitat and sequence data were collected from the Genomes Online database (43) and habitats were classified into the categories above using the Habitat-Lite terms of the Environment Ontology (32). When an organism was reported to have multiple habitats, the primary one was used. "Extreme environment" corresponds to EO categories "hot spring," "hydrothermal vent," and "extreme environment." "Sediment/sludge" corresponds to EO categories "sediment," "sludge," and "biofilm." Note that *(i)* dates reported for each sequence are publication dates, even if the genome was released to the public earlier in database form; *(ii)* the numbers for 2008 represent the available data until June 2008.

Figure 2. An overview of light-mediated processes in biology. Organisms sense visible and UV light that they use for growth, adaptation, and defense/repair. Light sensing is carried out by antenna molecules with a photoactive pigment, such as carotenoids, phycocyanin, phycoerythrin, or rhodopsins. Photosynthetic bacteria can process the light energy through a reaction center and store it as ATP via a proton gradient. Bacteria living in high-light environments must also protect against and repair UV damage. Extracellular and intracellular UV-absorbing compounds such as

scytonemin and mycosporine-like amino acids act as a natural sunscreen, while photolyase enzymes reverse point mutations in UV-damaged DNA by using a photon of blue light to catalyze the repair reaction. Finally, bacteria living in variable-light environments can adapt to the changing light conditions in a number of ways, e.g. by moving to a more favorable environment via phototaxis, reconfiguring the wavelength specificity of light sensing antennae via adaptation proteins, or providing their own light via luminescence.

Figure 3. Abundances of light-sensing proteins in metagenomes. (A) Total number of proteins orthologous to 20 query proteins (columns) in 373 sequenced genomes (top row) and five metagenomes (remaining rows). (B) Number of proteins as a percentage of the total number of predicted proteins per environment. Rows labeled in gray are subsamples. Columns are labeled as follows (please see Table 1 for details): *psa*, photosystem I subunits ABC; *psb*, photosystem II subunits ABDEHIJKLF; *pet*, photosynthetic electron transfer subunits A123; *apc*, allophycocyanin; *cpc*, phycocyanin; *kaiAB*, circadian clock regulators; *bluf*, blue-light FAD-binding domain containing proteins; *slr0359/plpA*, blue-light absorbing phototropins; *cph1/2*, red and far-red absorbing phytochromes; *taxD1*, photoreceptor for phototaxis; *cry*, DNA photolyase and cryptochrome families; *carot*, water-soluble carotenoids as intracellular UV sunscreen; *scyto*, scytonemin as extracellular sunscreen; *taxP1*, phototaxis putative regulatory element; *taxY1*, phototaxis CheY-like protein; *taxAY1*, phototaxis histidine kinase; *rcaE*, complementary chromatic adaptation protein. Growth and repair proteins are more abundant in high-light environments than variable-light ones, whereas sensing and adaptation proteins are more abundant in variable-light environments than high-light ones. In particular, photolyase DNA repair proteins are over-represented in the high-UV environment of surface sea water as

compared to all other environments. BLUF-domain blue-light sensing proteins are extremely rare in both genomes and environments, although the majority are found in surface rather than deep water. Red-light sensors Cph1/2 are over-represented in deep water rather than primarily blue surface water. RcaE chromatic adaptation proteins are over-represented in variable-light environments, such as the deep sea and lower (darker) layers of the microbial mat.

Figure 4. BLUF operons from genomes and metagenomes. BLUF-domain proteins are shown in blue (center) with none containing additional domains. Genome neighbors include genes that function in phototaxis, nucleotide metabolism, repression of anoxygenic photosynthesis, and virulence, primarily from the Alpha-Proteobacteria (*Rhodopseudomonas* and *Rhodobacter*), Beta-Proteobacteria (*Ralstonia* and *Chromobacterium*), and Gamma-Proteobacteria (*Shewanella* and *Psychrobacter*). Novel metagenome neighbors include genes that function in luciferase synthesis, nitrate metabolism, and quorum sensing, primarily from *Rhodopseudomonas* and *Comamonaceae*.

Figure 5. RcaE domain variations in sequenced bacteria and plants. Whereas the majority of bacterial proteins have the conserved domain architecture of GAF-PAS-PAC-HISKA-HATPase-REC, many additional architectures exist with different signal transduction domains, multiple sensing domains, and multiple receiver domains.

Figure 6. Photolyase/cryptochrome subfamilies representing 1,196 sequences from sequenced genomes and four metagenomes. The tree recovers the known groupings of CPD-I and CPD-II photolyases, and animal and DASH cryptochromes (plant cryptochromes, the fifth known group,

are not shown here). In addition, we discovered two novel subfamilies of photolyases with 38 (family I) and 54 members (family II) that appear to originate from diverse Alpha-Proteobacteria and Cyanobacteria.

Function	Gene	Genbank ID	Gene ID	COG/NOG Assignment	Description
Growth: photosynthesis and circadian rhythms	psa	16330029	slr1834	NOG04762	P700 apoprotein subunit Ia. psaA
		16330030	slr1835	NOG04763	P700 apoprotein subunit Ib; psaB
		16331238	ssl0563	COG1145	Photosystem I iron-sulfur center; psaC
	psb	16332289	slr1181	-	photosystem II D1 protein; psbA1
		16329178	slr1311	NOG06868	photosystem II D1 protein; psbA2
		16330822	sll1867	NOG06868	photosystem II D1 protein; psbA3
	pet	16331429	sll0199	COG3794	plastocyanin; petE
		16330840	sll1382	COG0633	ferredoxin; petF
		16331144	slr0150	COG0633	ferredoxin
		16330020	slr1828	COG0633	ferredoxin
		16331399	ssl0020	COG0633	ferredoxin
		16331051	slr1643	COG0369	ferredoxin-NADP oxidoreductase; petH
		16329946	sll1796	COG2010	cytochrome c6 precursor; petJ
	apc	16330466	slr2067	NOG09444	allophycocyanin a chain; apcA
		16330467	slr1986	NOG08465	allophycocyanin b chain; apcB
		16330468	ssr3383	NOG13001	phycobilisome LC linker polypeptide; apcC
		16329478	sll0928	NOG10841	allophycocyanin-B; apcD
		16331244	slr0335	NOG04733	phycobilisome LCM core-membrane linker; apcE
		16332118	slr1459	NOG09429	phycobilisome core component; apcF
	cpc	16329823	sll1578	NOG09446	phycocyanin a subunit; cpcA
		16329824	sll1577	NOG09445	phycocyanin b subunit; cpcB
		16329822	sll1579	NOG09475	phycocyanin associated linker protein; cpcC
		16329821	sll1580	NOG07680	phycocyanin associated linker protein
		16329820	ssl3093	COG0369	phycocyanin associated linker protein; cpcD
		16330275	slr1878	COG1413	phycocyanin alpha phycocyanobilin lyase; cpcE
		16329246	sll1051	COG1413	phycocyanin alpha phycocyanobilin lyase; cpcF
		16332194	sll1471	NOG10782	phycobilisome rod-core linker polypeptide; cpcG
		16329710	slr2051	NOG09477	phycobilisome rod-core linker polypeptide
	kaiA	16332220	slr0756	NOG10854	circadian clock protein; kaiA
	kaiB	16332221	slr0757	COG0526	circadian clock protein; kaiB
	puc	90422812	RPC_1301	COG2204	putative PAS/PAC sensor protein
		90422813	RPC_1302	COG0382	bacteriochlorophyll/chlorophyll a synthase
		90422814	RPC_1303	COG0477	formation of the LHII complex
		90422815	RPC_1304	COG0644	geranylgeranyl reductase
		90422816	RPC_1305	COG3476	TspO and MBR like proteins
	puf	77463830	RSP_0259	-	Protein pufQ
		77463829	RSP_6109	-	Transcriptional regulatory protein pufK
		77463828	RSP_6108	-	LHI beta, Light-harvesting B875 subunit
		77463827	RSP_0258	-	LHI alpha, Light-harvesting B875 protein

Function	Gene	Genbank ID	Gene ID	COG/NOG Assignment	Description
		77463826	RSP_0257	-	PufL, Photosynthetic reaction center L subunit
		77463825	RSP_0256	-	PufM, photosynthetic reaction center M subunit
		77463824	RSP_0255	-	Intrinsic membrane pufX protein
Sensing	bluf	16330981	slr1694	NOG16599	FAD-binding domain protein (blue)
	slr0359	16331282	slr0359	COG2199, COG2200	uncharacterized signaling protein (blue)
	plpA	16329960	sll1124	COG0642, COG2202	sensory transduction histidine kinase; plpA
	cph1	16331509	slr0473	COG4251	bacteriophytochrome (red/far red); cph1
	cph2	16331738	sll0821	COG2199, COG2200, COG2203	bacteriophytochrome (red/far red); cph2
	taxD1	16331988	sll0041	COG0840, COG2203	photoreceptor aka pixJ1 (blue); taxD1
Defense and repair	cry	-	-	COG0415, COG3046, COG4338, NOG16378	photolyase/cryptochrome families
	carot	16330780	slr1963	NOG04725	water-soluble carotenoid
		17230641	all3149	NOG04725	orange carotenoid-binding protein
		75910047	Ava_3843	NOG04725	orange carotenoid-binding protein
		33865901	SYNW1367	NOG04725	carotenoid binding protein
		33865435	SYNW0901	COG1233	carotenoid isomerase; crtH
		37519619	glr0050	NOG04725	carotenoid isomerase
		37523504	glr3935	NOG04725	water-soluble carotenoid protein
	scyto	17227918	all0422	COG3391	hypothetical protein (scytonemin synthesis)
		17227919	all0423	NOG19292	hypothetical protein (scytonemin synthesis)
		17227920	all0424	NOG19292	hypothetical protein (scytonemin synthesis)
		17227921	all0425	-	hypothetical protein (scytonemin synthesis)
		17227922	all0426	COG0334	leucine dehydrogenase
		17227923	all0427	COG0028	acetolactate synthase large subunit
Adaptation	taxP1	16331991	sll0038	COG0784	phototaxis putative regulatory element
	taxY1	16331990	sll0039	COG0784	phototaxis CheY-like protein
	taxAY1	16331987	sll0042	COG0840	phototaxis methyl-accepting protein; tar
	rcaE	75908636	Q47897	COG5002	sensor hybrid histidine kinase (red/green)

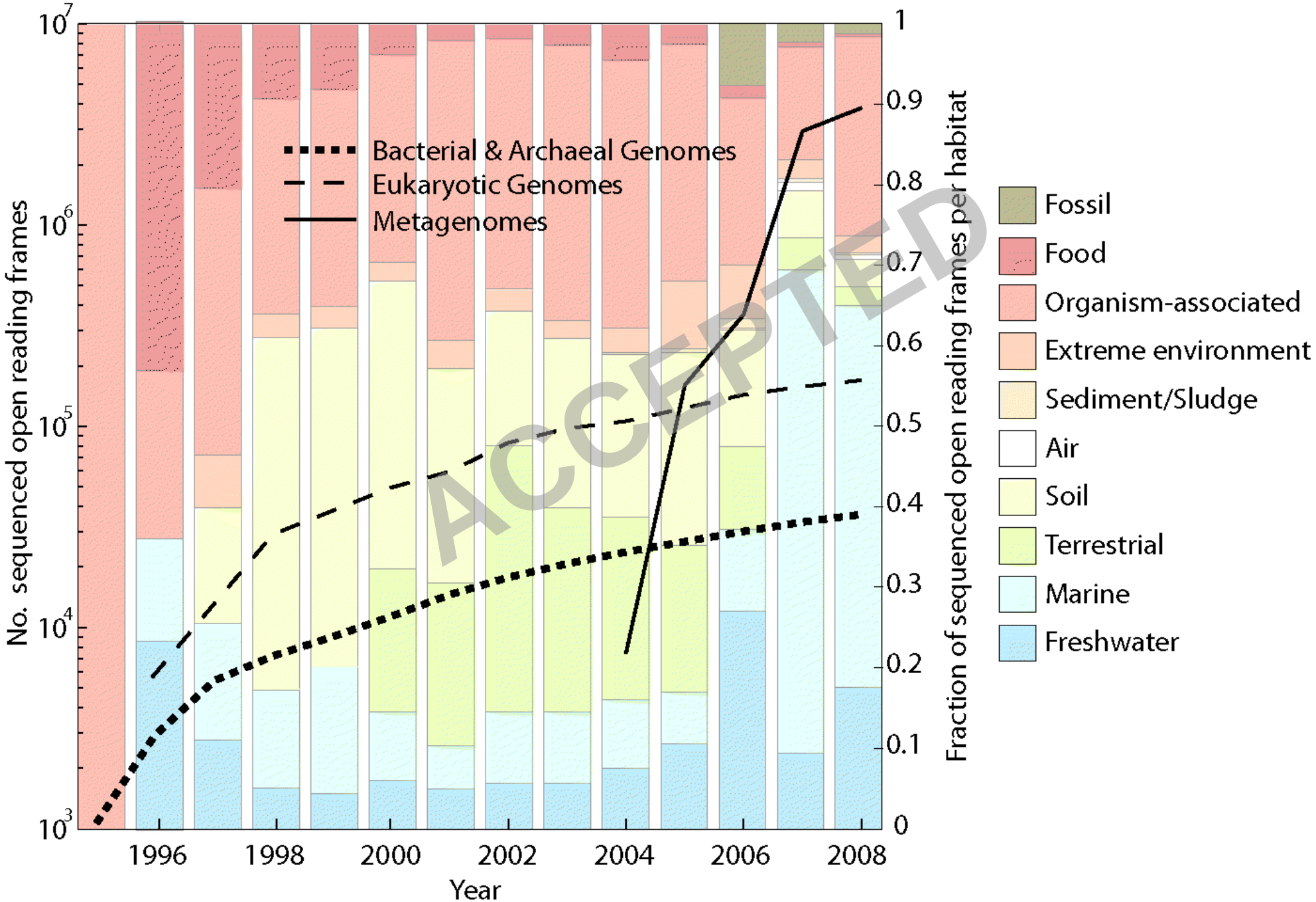
Table 1. Query genes used in metagenome searches. Absolute and relative abundances of these genes in the environments are presented in Fig. 3.

References

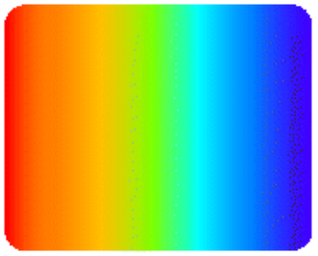
1. **Ashby, M. K., and J. Houmard.** 2006. Cyanobacterial two-component proteins: structure, diversity, distribution, and evolution. Microbiol Mol Biol Rev **70:**472-509.
2. **Beja, O., L. Aravind, E. V. Koonin, M. T. Suzuki, A. Hadd, L. P. Nguyen, S. B. Jovanovich, C. M. Gates, R. A. Feldman, J. L. Spudich, E. N. Spudich, and E. F. DeLong.** 2000. Bacterial rhodopsin: evidence for a new type of phototrophy in the sea. Science **289:**1902-6.
3. **Bhaya, D.** 2004. Light matters: phototaxis and signal transduction in unicellular cyanobacteria. Mol Microbiol **53:**745-54.
4. **Bjorklund, A. K., D. Ekman, and A. Elofsson.** 2006. Expansion of protein domain repeats. PLoS Comput Biol **2:**e114.
5. **Blankenship, R. E.** 1992. Origin and early evolution of photosynthesis. Photosynth Res **33:**91-111.
6. **Boone, D. R., R. W. Castenholz, and G. M. Garrity.** 2001. Bergey's manual of systematic bacteriology, 2nd ed. Springer, New York.
7. **Bordowitz, J. R., and B. L. Montgomery.** 2008. Photoregulation of cellular morphology during complementary chromatic adaptation requires sensor-kinase-class protein RcaE in Fremyella diplosiphon. J Bacteriol **190:**4069-74.
8. **Bork, P., T. Dandekar, Y. Diaz-Lazcoz, F. Eisenhaber, M. Huynen, and Y. Yuan.** 1998. Predicting function: from genes to genomes and back. J Mol Biol **283:**707-25.
9. **Bork, P., and L. Serrano.** 2005. Towards cellular systems in 4D. Cell **121:**507-9.
10. **Boucher, Y., C. J. Douady, R. T. Papke, D. A. Walsh, M. E. Boudreau, C. L. Nesbo, R. J. Case, and W. F. Doolittle.** 2003. Lateral gene transfer and the origins of prokaryotic groups. Annu Rev Genet **37:**283-328.
11. **Brown, D. P., N. Krishnamurthy, and K. Sjolander.** 2007. Automated protein subfamily identification and classification. PLoS Comput Biol **3:**e160.
12. **Brudler, R., K. Hitomi, H. Daiyasu, H. Toh, K. Kucho, M. Ishiura, M. Kanehisa, V. A. Roberts, T. Todo, J. A. Tainer, and E. D. Getzoff.** 2003. Identification of a new cryptochrome class. Structure, function, and evolution. Mol Cell **11:**59-67.
13. **Bryant, D. A., A. M. Costas, J. A. Maresca, A. G. Chew, C. G. Klatt, M. M. Bateson, L. J. Tallon, J. Hostetler, W. C. Nelson, J. F. Heidelberg, and D. M. Ward.** 2007. Candidatus Chloracidobacterium thermophilum: an aerobic phototrophic Acidobacterium. Science **317:**523-6.
14. **Bryant, D. A., and N. U. Frigaard.** 2006. Prokaryotic photosynthesis and phototrophy illuminated. Trends Microbiol **14:**488-96.
15. **Buchanan, B. B., and Y. Balmer.** 2005. Redox regulation: a broadening horizon. Annu Rev Plant Biol **56:**187-220.
16. **Ciccarelli, F. D., T. Doerks, C. von Mering, C. J. Creevey, B. Snel, and P. Bork.** 2006. Toward automatic reconstruction of a highly resolved tree of life. Science **311:**1283-7.
17. **Ciccarelli, F. D., C. von Mering, M. Suyama, E. D. Harrington, E. Izaurralde, and P. Bork.** 2005. Complex genomic rearrangements lead to novel primate gene function. Genome Res **15:**343-51.
18. **Copley, R. R., T. Doerks, I. Letunic, and P. Bork.** 2002. Protein domain analysis in the era of complete genomes. FEBS Lett **513:**129-34.
19. **Dandekar, T., B. Snel, M. Huynen, and P. Bork.** 1998. Conservation of gene order: a fingerprint of proteins that physically interact. Trends Biochem Sci **23:**324-8.
20. **Dvornyk, V., O. Vinogradova, and E. Nevo.** 2003. Origin and evolution of circadian clock genes in prokaryotes. Proc Natl Acad Sci U S A **100:**2495-500.
21. **Edgar, R. C.** 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res **32:**1792-7.
22. **Ehling-Schulz, M., W. Bilger, and S. Scherer.** 1997. UV-B-induced synthesis of photoprotective pigments and extracellular polysaccharides in the terrestrial cyanobacterium Nostoc commune. J Bacteriol **179:**1940-5.
23. **Felsenstein, J.** 1993. PHYLIP -- Phylogeny Inference Package version 3.2. Cladistics**:**164.
24. **Fiedler, B., T. Borner, and A. Wilde.** 2005. Phototaxis in the cyanobacterium Synechocystis sp. PCC 6803: role of different photoreceptors. Photochem Photobiol **81:**1481-8.
25. **Fraser, C., W. P. Hanage, and B. G. Spratt.** 2007. Recombination and the nature of bacterial speciation. Science **315:**476-80.
26. **Gomelsky, M., and G. Klug.** 2002. BLUF: a novel FAD-binding domain involved in sensory transduction in microorganisms. Trends Biochem Sci **27:**497-500.

27. **Guindon, S., and O. Gascuel.** 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol **52:**696-704.
28. **Han, Y., S. Braatsch, L. Osterloh, and G. Klug.** 2004. A eukaryotic BLUF domain mediates light-dependent gene expression in the purple bacterium Rhodobacter sphaeroides 2.4.1. Proc Natl Acad Sci U S A **101:**12306-11.
29. **Hannenhalli, S. S., and R. B. Russell.** 2000. Analysis and prediction of functional sub-types from protein sequence alignments. J Mol Biol **303:**61-76.
30. **Harrington, E. D., A. H. Singh, T. Doerks, I. Letunic, C. von Mering, L. J. Jensen, J. Raes, and P. Bork.** 2007. Quantitative assessment of protein function prediction from metagenomics shotgun sequences. Proc Natl Acad Sci U S A **104:**13913-8.
31. **Harris, M. A., J. Clark, A. Ireland, J. Lomax, M. Ashburner, R. Foulger, K. Eilbeck, S. Lewis, B. Marshall, C. Mungall, J. Richter, G. M. Rubin, J. A. Blake, C. Bult, M. Dolan, H. Drabkin, J. T. Eppig, D. P. Hill, L. Ni, M. Ringwald, R. Balakrishnan, J. M. Cherry, K. R. Christie, M. C. Costanzo, S. S. Dwight, S. Engel, D. G. Fisk, J. E. Hirschman, E. L. Hong, R. S. Nash, A. Sethuraman, C. L. Theesfeld, D. Botstein, K. Dolinski, B. Feierbach, T. Berardini, S. Mundodi, S. Y. Rhee, R. Apweiler, D. Barrell, E. Camon, E. Dimmer, V. Lee, R. Chisholm, P. Gaudet, W. Kibbe, R. Kishore, E. M. Schwarz, P. Sternberg, M. Gwinn, L. Hannick, J. Wortman, M. Berriman, V. Wood, N. de la Cruz, P. Tonellato, P. Jaiswal, T. Seigfried, R. White, and C. Gene Ontology.** 2004. The Gene Ontology (GO) database and informatics resource. Nucleic acids research **32:**D258.
32. **Hirschman, L., C. Clark, K. B. Cohen, S. Mardis, J. Luciano, R. Kottmann, J. Cole, V. Markowitz, N. Kyrpides, and D. Field.** 2008. Habitat-Lite: A GSC Case Study Based on Free Text Terms for Environmental Metadata. Omics.
33. **Huson, D. H., A. F. Auch, J. Qi, and S. C. Schuster.** 2007. MEGAN analysis of metagenomic data. Genome Res **17:**377-86.
34. **Huynen, M., T. Dandekar, and P. Bork.** 1998. Differential genome analysis applied to the species-specific features of Helicobacter pylori. FEBS Lett **426:**1-5.
35. **Jung, A., T. Domratcheva, M. Tarutina, Q. Wu, W. H. Ko, R. L. Shoeman, M. Gomelsky, K. H. Gardner, and I. Schlichting.** 2005. Structure of a bacterial BLUF photoreceptor: insights into blue light-mediated signal transduction. Proc Natl Acad Sci U S A **102:**12350-5.
36. **Kannan, N., S. S. Taylor, Y. Zhai, J. C. Venter, and G. Manning.** 2007. Structural and functional diversity of the microbial kinome. PLoS Biol **5:**e17.
37. **Kehoe, D. M., and A. R. Grossman.** 1996. Similarity of a chromatic adaptation sensor to phytochrome and ethylene receptors. Science **273:**1409-12.
38. **Korbel, J. O., L. J. Jensen, C. von Mering, and P. Bork.** 2004. Analysis of genomic context: prediction of functional associations from conserved bidirectionally transcribed gene pairs. Nature biotechnology **22:**911.
39. **Kunin, V., J. Raes, J. K. Harris, J. R. Spear, J. J. Walker, N. Ivanova, C. von Mering, B. M. Bebout, N. R. Pace, P. Bork, and P. Hugenholtz.** 2008. Millimeter-scale genetic gradients and community-level molecular convergence in a hypersaline microbial mat. Mol Syst Biol **4:**198.
40. **Letunic, I., and P. Bork.** 2007. Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. Bioinformatics **23:**127-8.
41. **Letunic, I., R. R. Copley, B. Pils, S. Pinkert, J. Schultz, and P. Bork.** 2006. SMART 5: domains in the context of genomes and networks. Nucleic Acids Res **34:**D257-60.
42. **Lin, C., and T. Todo.** 2005. The cryptochromes. Genome Biol **6:**220.
43. **Liolios, K., K. Mavromatis, N. Tavernarakis, and N. C. Kyrpides.** 2008. The Genomes On Line Database (GOLD) in 2007: status of genomic and metagenomic projects and their associated metadata. Nucleic Acids Res **36:**D475-9.
44. **Lozupone, C. A., and R. Knight.** 2007. Global patterns in bacterial diversity. Proc Natl Acad Sci U S A **104:**11436-40.
45. **Montgomery, B. L.** 2007. Sensing the light: photoreceptive systems and signal transduction in cyanobacteria. Mol Microbiol **64:**16-27.
46. **Moran, M. A., and W. L. Miller.** 2007. Resourceful heterotrophs make the most of light in the coastal ocean. Nat Rev Microbiol **5:**792-800.
47. **Oren, A.** 2004. Prokaryote diversity and taxonomy: current status and future challenges. Philos Trans R Soc Lond B Biol Sci **359:**623-38.
48. **Overbeek, R., M. Fonstein, M. D'Souza, G. D. Pusch, and N. Maltsev.** 1999. The use of gene clusters to infer functional coupling. Proc Natl Acad Sci U S A **96:**2896-901.

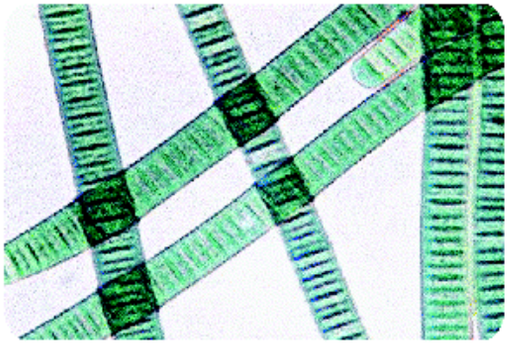
49. **Pao, G. M., and M. H. Saier, Jr.** 1995. Response regulators of bacterial signal transduction systems: selective domain shuffling during evolution. J Mol Evol **40:**136-54.
50. **Pignatelli, M., G. Aparicio, I. Blanquer, V. Hernandez, A. Moya, and J. Tamames.** 2008. Metagenomics reveals our incomplete knowledge of global diversity. Bioinformatics.
51. **Raes, J., E. D. Harrington, A. H. Singh, and P. Bork.** 2007. Protein function space: viewing the limits or limited by our view? Curr Opin Struct Biol **17:**362-9.
52. **Raymond, J., O. Zhaxybayeva, J. P. Gogarten, S. Y. Gerdes, and R. E. Blankenship.** 2002. Whole-genome analysis of photosynthetic prokaryotes. Science **298:**1616-20.
53. **Sancar, A.** 2004. Photolyase and cryptochrome blue-light photoreceptors. Adv Protein Chem **69:**73-100.
54. **Schmidt, H. A., K. Strimmer, M. Vingron, and A. von Haeseler.** 2002. TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. Bioinformatics **18:**502-4.
55. **Schultz, J., F. Milpetz, P. Bork, and C. P. Ponting.** 1998. SMART, a simple modular architecture research tool: identification of signaling domains. Proc Natl Acad Sci U S A **95:**5857-64.
56. **Selby, C. P., and A. Sancar.** 2006. A cryptochrome/photolyase class of enzymes with single-stranded DNA-specific photolyase activity. Proc Natl Acad Sci U S A **103:**17696-700.
57. **Singh, A. H., D. M. Wolf, P. Wang, and A. P. Arkin.** 2008. Modularity of stress response evolution. Proc Natl Acad Sci U S A **107:**7500-7505.
58. **Singh, S. P., M. Klisch, R. P. Sinha, and D. P. Hader.** 2008. Effects of Abiotic Stressors on Synthesis of the Mycosporine-like Amino Acid Shinorine in the Cyanobacterium Anabaena variabilis PCC 7937. Photochem Photobiol.
59. **Sinha, R. P., N. K. Ambasht, J. P. Sinha, M. Klisch, and D. P. Hader.** 2003. UV-B-induced synthesis of mycosporine-like amino acids in three strains of Nodularia (cyanobacteria). J Photochem Photobiol B **71:**51-8.
60. **Soule, T., V. Stout, W. D. Swingley, J. C. Meeks, and F. Garcia-Pichel.** 2007. Molecular genetics and genomic analysis of scytonemin biosynthesis in Nostoc punctiforme ATCC 29133. J Bacteriol **189:**4465-72.
61. **Taylor, B. L., and I. B. Zhulin.** 1999. PAS domains: internal sensors of oxygen, redox potential, and light. Microbiol Mol Biol Rev **63:**479-506.
62. **Tringe, S. G., C. von Mering, A. Kobayashi, A. A. Salamov, K. Chen, H. W. Chang, M. Podar, J. M. Short, E. J. Mathur, J. C. Detter, P. Bork, P. Hugenholtz, and E. M. Rubin.** 2005. Comparative metagenomics of microbial communities. Science **308:**554-7.
63. **Tyson, G. W., J. Chapman, P. Hugenholtz, E. E. Allen, R. J. Ram, P. M. Richardson, V. V. Solovyev, E. M. Rubin, D. S. Rokhsar, and J. F. Banfield.** 2004. Community structure and metabolism through reconstruction of microbial genomes from the environment. Nature **428:**37-43.
64. **Venter, J. C., K. Remington, J. F. Heidelberg, A. L. Halpern, D. Rusch, J. A. Eisen, D. Wu, I. Paulsen, K. E. Nelson, W. Nelson, D. E. Fouts, S. Levy, A. H. Knap, M. W. Lomas, K. Nealson, O. White, J. Peterson, J. Hoffman, R. Parsons, H. Baden-Tillson, C. Pfannkoch, Y. H. Rogers, and H. O. Smith.** 2004. Environmental genome shotgun sequencing of the Sargasso Sea. Science **304:**66-74.
65. **Vierstra, R. D., and S. J. Davis.** 2000. Bacteriophytochromes: new tools for understanding phytochrome signal transduction. Semin Cell Dev Biol **11:**511-21.
66. **von Mering, C., P. Hugenholtz, J. Raes, S. G. Tringe, T. Doerks, L. J. Jensen, N. Ward, and P. Bork.** 2007. Quantitative phylogenetic assessment of microbial communities in diverse environments. Science **315:**1126-30.
67. **von Mering, C., L. J. Jensen, M. Kuhn, S. Chaffron, T. Doerks, B. Kruger, B. Snel, and P. Bork.** 2007. STRING 7--recent developments in the integration and prediction of protein interactions. Nucleic Acids Res **35:**D358-62.
68. **Wilson, E. O.** 1999. The diversity of life, New ed. W. W. Norton, New York.
69. **Xiong, J., K. Inoue, and C. E. Bauer.** 1998. Tracking molecular evolution of photosynthesis by characterization of a major photosynthesis gene cluster from Heliobacillus mobilis. Proc Natl Acad Sci U S A **95:**14851-6.
70. **Yooseph, S., G. Sutton, D. B. Rusch, A. L. Halpern, S. J. Williamson, K. Remington, J. A. Eisen, K. B. Heidelberg, G. Manning, W. Li, L. Jaroszewski, P. Cieplak, C. S. Miller, H. Li, S. T. Mashiyama, M. P. Joachimiak, C. van Belle, J. M. Chandonia, D. A. Soergel, Y. Zhai, K. Natarajan, S. Lee, B. J. Raphael, V. Bafna, R. Friedman, S. E. Brenner, A. Godzik, D. Eisenberg, J. E. Dixon, S. S. Taylor, R. L. Strausberg, M. Frazier, and J. C. Venter.** 2007. The Sorcerer II Global Ocean Sampling expedition: expanding the universe of protein families. PLoS Biol **5:**e16.

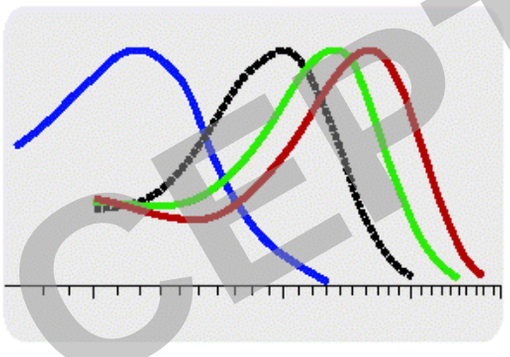

No. sequenced open reading frames
Fraction of sequenced open reading frames per habitat
Year
Bacterial & Archaeal Genomes
Eukaryotic Genomes
Metagenomes
Fossil
Food
Organism-associated
Extreme environment
Sediment/Sludge
Air
Soil
Terrestrial
Marine
Freshwater
ACCEPTED

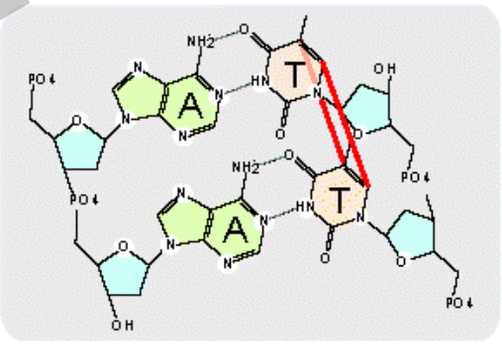
visible light



UV light



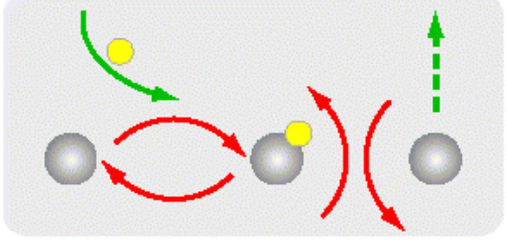




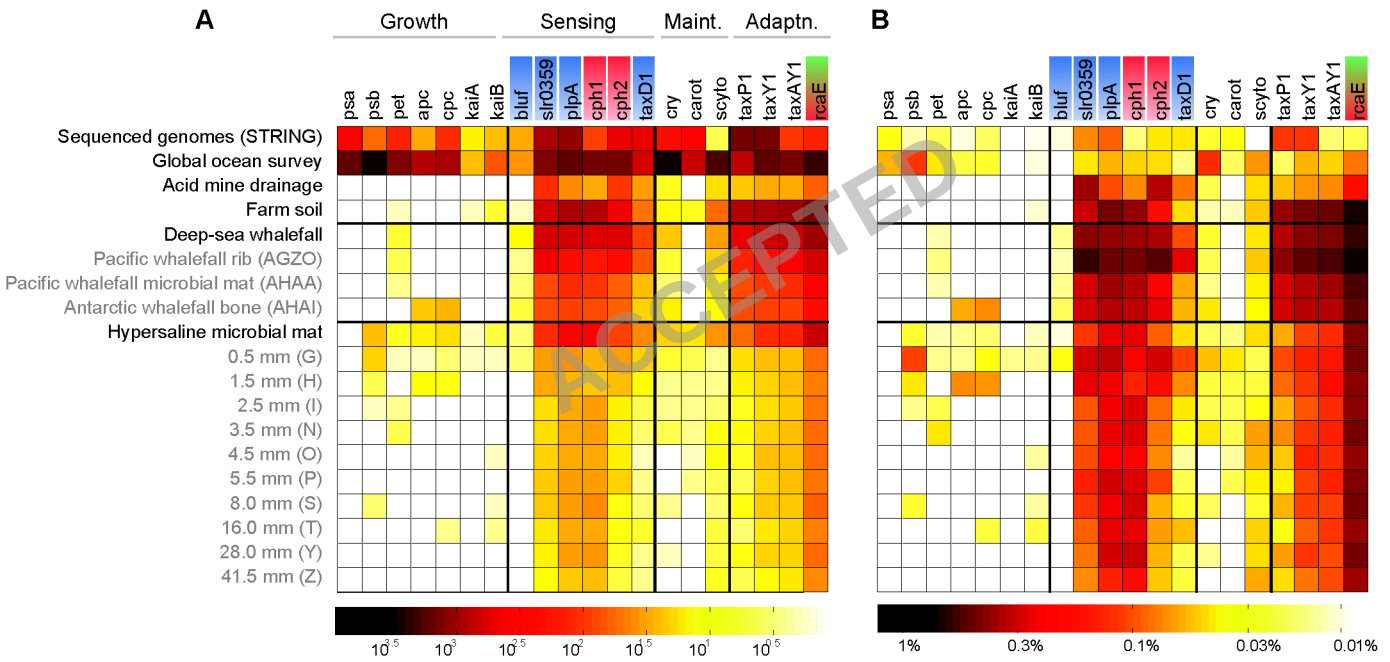
growth
photosynthesis
circadian rhythms

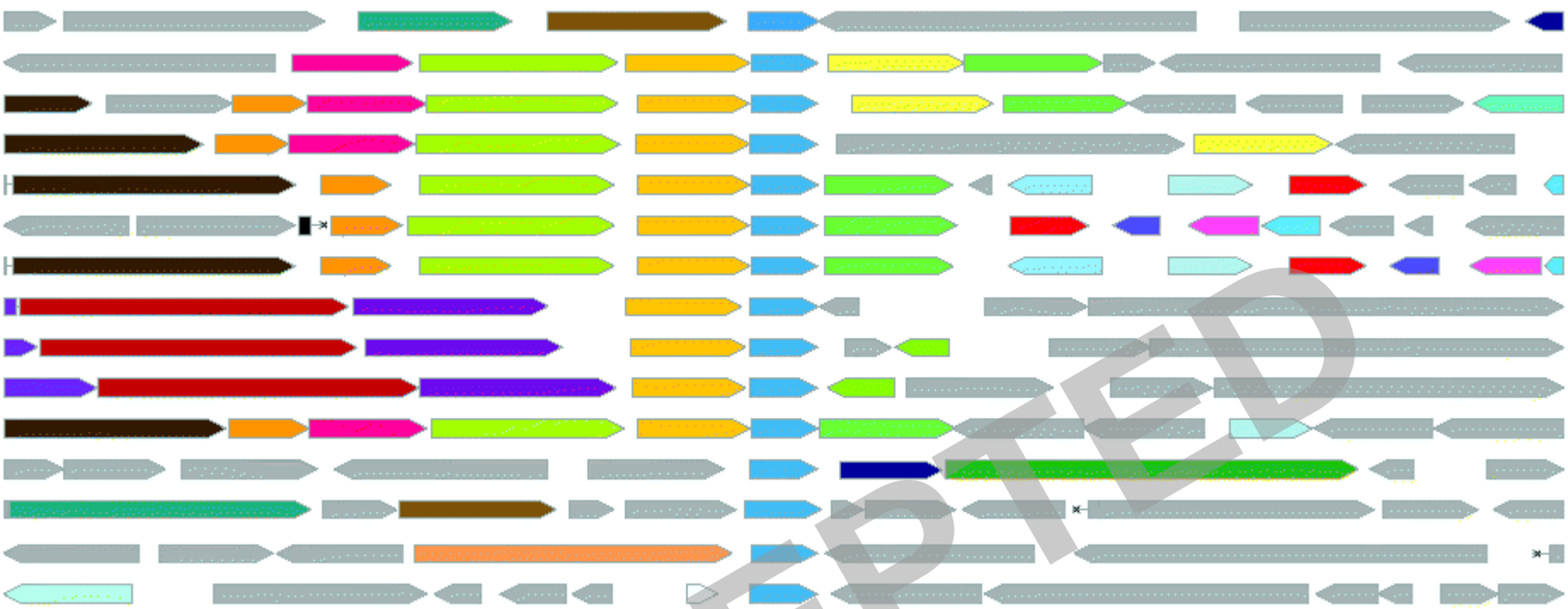
sensing
phytochromes
pineal gland

defense, repair
melanin/pigment
photolyases

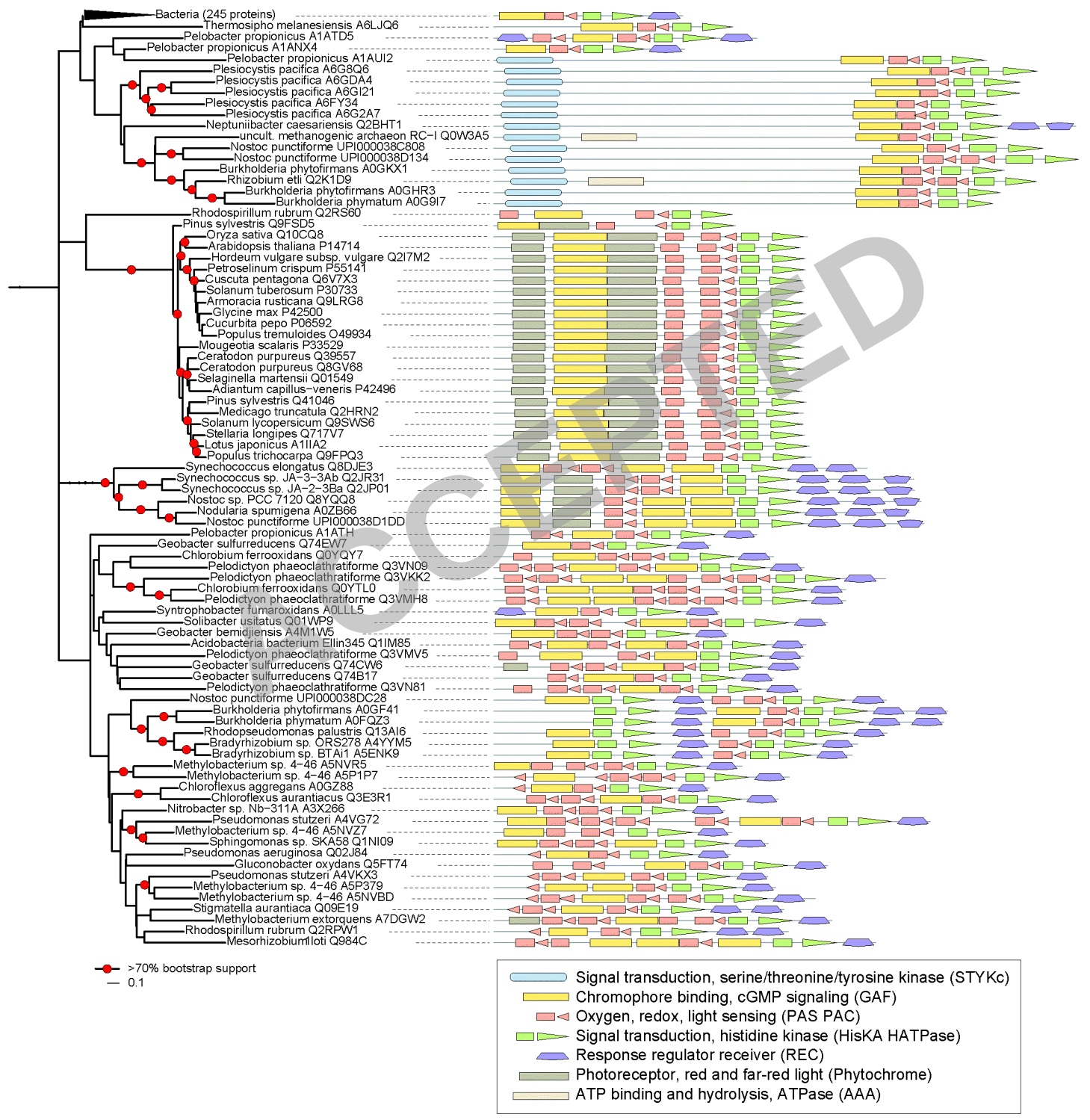


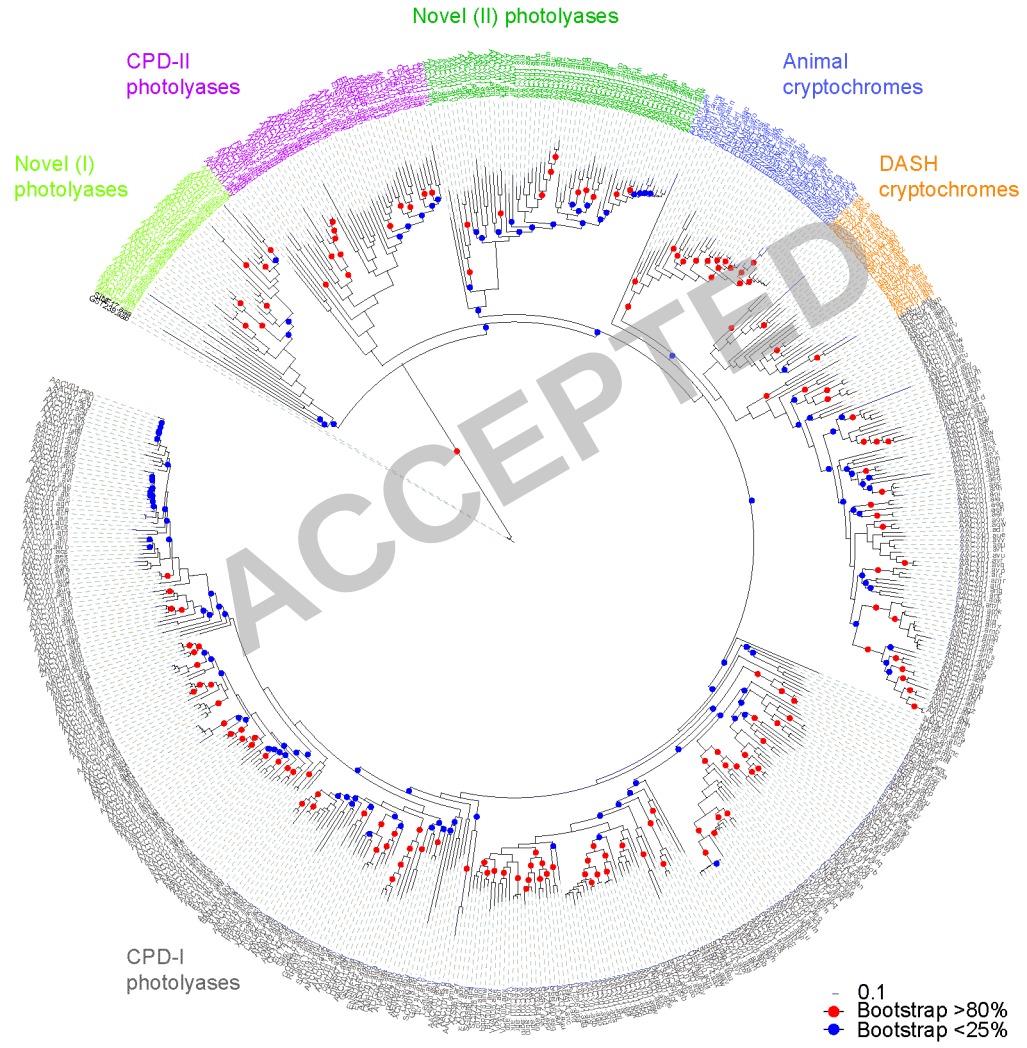
adaptation
phototaxis
bioluminescence


A
Growth
Sensing
Maint.
Adaptn.
psa
psb
pet
apc
cpc
kaiA
kaiB
bluf
slr0359
plpA
cph1
cph2
taxD1
cry
carot
scyto
taxP1
taxY1
taxAY1
rcaE
Sequenced genomes (STRING)
Global ocean survey
Acid mine drainage
Farm soil
Deep-sea whalefall
Pacific whalefall rib (AGZO)
Pacific whalefall microbial mat (AHAA)
Antarctic whalefall bone (AHAI)
Hypersaline microbial mat
0.5 mm (G)
1.5 mm (H)
2.5 mm (I)
3.5 mm (N)
4.5 mm (O)
5.5 mm (P)
8.0 mm (S)
16.0 mm (T)
28.0 mm (Y)
41.5 mm (Z)
$10^{-3.5}$
10^{-3}
$10^{-2.5}$
10^{-2}
$10^{-1.5}$
10^{-1}
$10^{-0.5}$
B
psa
psb
pet
apc
cpc
kaiA
kaiB
bluf
slr0359
plpA
cph1
cph2
taxD1
cry
carot
scyto
taxP1
taxY1
taxAY1
rcaE
1%
0.3%
0.1%
0.03%
0.01%
ACCEPTED



Phototaxis	EAL-domain sensing protein
	CheY-family respose regulator
Nucleotide metabolism	GTP cyclohydrolase
	Transphophatidylase
	Metal-dependent hydrolase
	NUDIX hydrolase
Anoxygenic photosynthesis	Arsenic-pump permease
	Periplasmic transporter
Nitrogen metabolism	NIF3 / nitroreductase
	DUF619 urea cycle enzyme
Luciferase synthesis	LuxR-family response regulator
	Sensor histidine kinase
Quorum sensing	Extracytoplasmic solute receptor / uptake
	Propeptide Pep54 / peptidase M4


Bacteria (245 proteins)
Thermosipho melanesiensis A6LJQ6
Pelobacter propionicus A1ATD5
Pelobacter propionicus A1ANX4
Pelobacter propionicus A1AUI2
Plesiocystis pacifica A6G8Q6
Plesiocystis pacifica A6GDA4
Plesiocystis pacifica A6GI21
Plesiocystis pacifica A6FY34
Plesiocystis pacifica A6G2A7
Neptuniibacter caesariensis Q2BHT1
uncult. methanogenic archaeon RC-I Q0W3A5
Nostoc punctiforme UPI000038C808
Nostoc punctiforme UPI000038D134
Burkholderia phytofirmans A0GKX1
Rhizobium etli Q2K1D9
Burkholderia phytofirmans A0GHR3
Burkholderia phymatum A0G9I7
Rhodospirillum rubrum Q2RS60
Pinus sylvestris Q9FSD5
Oryza sativa Q10CQ8
Arabidopsis thaliana P14714
Hordeum vulgare subsp. vulgare Q2I7M2
Petroselinum crispum P55141
Cuscuta pentagona Q6V7X3
Solanum tuberosum P30733
Armoracia rusticana Q9LRG8
Glycine max P42500
Cucurbita pepo P06592
Populus tremuloides O49934
Mougeotia scalaris P33529
Ceratodon purpureus Q39557
Ceratodon purpureus Q8GV68
Selaginella martensii Q01549
Adiantum capillus-veneris P42496
Pinus sylvestris Q41046
Medicago truncatula Q2HRN2
Solanum lycopersicum Q9SWS6
Stellaria longipes Q717V7
Lotus japonicus A1IIA2
Populus trichocarpa Q9FPQ3
Synechococcus elongatus Q8DJE3
Synechococcus sp. JA-3-3Ab Q2JR31
Synechococcus sp. JA-2-3Ba Q2JP01
Nostoc sp. PCC 7120 Q8YQQ8
Nodularia spumigena A0ZB66
Nostoc punctiforme UPI000038D1DD
Pelobacter propionicus A1ATH
Geobacter sulfurreducens Q74EW7
Chlorobium ferrooxidans Q0YQY7
Pelodictyon phaeoclathratiforme Q3VN09
Pelodictyon phaeoclathratiforme Q3VKK2
Chlorobium ferrooxidans Q0YTL0
Pelodictyon phaeoclathratiforme Q3VMH8
Syntrophobacter fumaroxidans A0LLL5
Solibacter usitatus Q01WP9
Geobacter bemidjiensis A4M1W5
Acidobacteria bacterium Ellin345 Q1IM85
Pelodictyon phaeoclathratiforme Q3VMV5
Geobacter sulfurreducens Q74CW6
Geobacter sulfurreducens Q74B17
Pelodictyon phaeoclathratiforme Q3VN81
Nostoc punctiforme UPI000038DC28
Burkholderia phytofirmans A0GF41
Burkholderia phymatum A0FQZ3
Rhodopseudomonas palustris Q13AI6
Bradyrhizobium sp. ORS278 A4YYM5
Bradyrhizobium sp. BTAi1 A5ENK9
Methylobacterium sp. 4-46 A5NVR5
Methylobacterium sp. 4-46 A5P1P7
Chloroflexus aggregans A0GZ88
Chloroflexus aurantiacus Q3E3R1
Nitrobacter sp. Nb-311A A3X266
Pseudomonas stutzeri A4VG72
Methylobacterium sp. 4-46 A5NVZ7
Sphingomonas sp. SKA58 Q1NI09
Pseudomonas aeruginosa Q02J84
Gluconobacter oxydans Q5FT74
Pseudomonas stutzeri A4VKX3
Methylobacterium sp. 4-46 A5P379
Methylobacterium sp. 4-46 A5NVBD
Stigmatella aurantiaca Q09E19
Methylobacterium extorquens A7DGW2
Rhodospirillum rubrum Q2RPW1
Mesorhizobium loti Q984C
>70% bootstrap support
0.1
Signal transduction, serine/threonine/tyrosine kinase (STYKc)
Chromophore binding, cGMP signaling (GAF)
Oxygen, redox, light sensing (PAS PAC)
Signal transduction, histidine kinase (HisKA HATPase)
Response regulator receiver (REC)
Photoreceptor, red and far-red light (Phytochrome)
ATP binding and hydrolysis, ATPase (AAA)


Novel (II) photolyases
CPD-II photolyases
Animal cryptochromes
Novel (I) photolyases
DASH cryptochromes
CPD-I photolyases
0.1
Bootstrap >80%
Bootstrap <25%