OPEN ACCESS Freely available online

An Integrative Genomic Approach to Uncover Molecular Mechanisms of Prokaryotic Traits

Yang Liu[1,2], Jianrong Li[1,2], Lee Sam[1,2], Chern-Sing Goh[3], Mark Gerstein[3,4*], Yves A. Lussier[1,2,5*]

1 Section of Genetic Medicine, Department of Medicine, University of Chicago, Chicago, Illinois, United States of America, **2** Center for Biomedical Informatics, Department of Medicine, University of Chicago, Chicago, Illinois, United States of America, **3** Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, Connecticut, United States of America, **4** Department of Computer Science, Yale University, New Haven, Connecticut, United States of America, **5** Department of Biomedical Informatics, Columbia University, New York, New York, United States of America

With mounting availability of genomic and phenotypic databases, data integration and mining become increasingly challenging. While efforts have been put forward to analyze prokaryotic phenotypes, current computational technologies either lack high throughput capacity for genomic scale analysis, or are limited in their capability to integrate and mine data across different scales of biology. Consequently, simultaneous analysis of associations among genomes, phenotypes, and gene functions is prohibited. Here, we developed a high throughput computational approach, and demonstrated for the first time the feasibility of integrating large quantities of prokaryotic phenotypes along with genomic datasets for mining across multiple scales of biology (protein domains, pathways, molecular functions, and cellular processes). Applying this method over 59 fully sequenced prokaryotic species, we identified genetic basis and molecular mechanisms underlying the phenotypes in bacteria. We identified 3,711 significant correlations between 1,499 distinct Pfam and 63 phenotypes, with 2,650 correlations and 1,061 *anti*-correlations. Manual evaluation of a random sample of these significant correlations showed a minimal precision of 30% (95% confidence interval: 20%–42%; $n = 50$). We stratified the most significant 478 predictions and subjected 100 to manual evaluation, of which 60 were corroborated in the literature. We furthermore unveiled 10 significant correlations between phenotypes and KEGG pathways, eight of which were corroborated in the evaluation, and 309 significant correlations between phenotypes and 166 GO concepts evaluated using a random sample (minimal precision = 72%; 95% confidence interval: 60%–80%; $n = 50$). Additionally, we conducted a novel large-scale phenomic visualization analysis to provide insight into the modular nature of common molecular mechanisms spanning multiple biological scales and reused by related phenotypes (metaphenotypes). We propose that this method elucidates which classes of molecular mechanisms are associated with phenotypes or metaphenotypes and holds promise in facilitating a computable systems biology approach to genomic and biomedical research.

Citation: Liu Y, Li J, Sam L, Goh CS, Gerstein M, et al. (2006) An integrative genomic approach to uncover molecular mechanisms of prokaryotic traits. PLoS Comput Biol 2(11): e159. doi:10.1371/journal.pcbi.0020159

Introduction

With the completion of hundreds of prokaryotic genome sequences, computational methods in systems biology aimed at integrating genotypes and phenotypes are developed at an increasing speed. However, data integration and mining remain key challenges in bioinformatics as well as in cross-disciplinary research in biomedical informatics. In addition, a critical issue that remains unsolved is to derive meaningful general biological principles from predictions of statistically significant associations between phenotypes and different biological scales of molecular mechanisms (e.g., protein domains, cellular processes, and cellular pathways) to facilitate the understanding of a particular species. The availability of a large number of fully sequenced genomes and the relatively simple and well-characterized biological processes of prokaryotic organisms makes them ideal model organisms to demonstrate the feasibility of a computational systems biology approach to integrate, mine, and analyze genomic, phenotypic, and functional databases to derive general principles that govern the biology of prokaryotes.

Prokaryotic phenotypes defined by human observations (e.g., motility), living conditions of the organism (e.g., growth at high temperature), and experimental conditions (e.g., acid production in a medium containing D-mannose) are of great interest for post-genomics–era research [1] as well as systems biology research. In clinical microbiological practice, many of these phenotypes are used to discriminate human pathogens from other microorganisms. While a great amount of effort has been devoted to the analysis of prokaryotic phenotypes, prior technologies, operated in a semi-automatic fashion, can at best only analyze a handful of phenotypes at once to deduce their associations with genotypes. In the past, functional genomic approaches predicted prokaryotic genes associated to biochemical pathways [2–4]; however, these

Editor: Suzanna Lewis, University of California Berkeley, United States of America

Received May 5, 2006; **Accepted** October 10, 2006; **Published** November 17, 2006

A previous version of this article appeared as an Early Online Release on October 10, 2006 (doi:10.1371/journal.pcbi.0020159.eor).

Abbreviations: COGs, Clusters of Orthologous Groups; GO, Gene Ontology; KEGG, Kyoto Encyclopedia of Genes and Genomes; MKD, Microbiology Knowledge Dataset from the Global Infectious Diseases and Epidemiology Network database; NCBI, National Center for Biotechnology Information; Pfam, Protein Family Database; PTS, phosphotransferase system pathway

* To whom correspondence should be addressed. E-mail: mark.gerstein@yale.edu (MG); lussier@uchicago.edu (YAL)

Synopsis

A key challenge of the post-genomic era is to conceive large-scale studies of genomes and observable characteristics of organisms (phenotypes) and to interpret the data thus produced. The goal of this "phenomic" study is to improve our understanding of complex biological systems in terms of their molecular underpinnings. In this paper, Liu and colleagues present comprehensive computational and novel visualization methods for discovering biological knowledge spanning multiple scales of biology. The authors were able to predict and visualize new knowledge between clusters of microbiological phenotypes and their molecular mechanisms. To their knowledge, this is the first time this has been done. More specifically, the method integrates microbiological data with genomic-scale data from protein family databases, gene ontology, and biological pathways. Conducted over 59 fully sequenced bacteria, and including significantly more phenotypes than previous studies of its kind, this study enables a "systems biology" view across different classifications of genes and processes. This represents advancement over previous techniques, which are either limited in biological scale or analytical breadth. Visualization of the networks generated by this technique shows the common biological modules shared by related phenotypes. The results of this experiment demonstrate that the fusion of clinical data with genomic information is able to elucidate, in high throughput, a massive number of biological processes underlying phenotypes.

studies did not specifically look into systems properties of these genes and were not specifically focused on phenotypes associated to these pathways or the emerging multiscale properties of their molecular mechanisms. Recently, a few studies conducted semi-automatic analyses of associations between an individual prokaryotic phenotype (e.g., hyperthermophily, motility) and its clustering with genes that have similar nucleotide sequences [5] or with Clusters of Orthologuous Groups of proteins (COGs) [5,6]. These studies, limited by their need for manual curation (phenotypic annotations to species), were designed to predict linear relationships between only one biological scale of molecular functions and a limited number of manually annotated phenotypes.

To overcome the limitations of manual annotation in the creation of phenotypic datasets, others in the field conducted phenotype–genotype analyses by mining known knowledge on phenotype–genotype relationships from the scientific literature using high-throughput technologies. In this regard, Korbel et al. used a natural language processing approach to mine the MEDLINE literature and the genomic contents of prokaryotes, resulting in the identification of 2,700 statistically significant associations between COGs and words from the literature related to phenotypes [7]. In other approaches, researchers have built integrated systems to correlate phenotypes, pathways, and genes [3,8]. For example, the WIT system [8] used an integrated system to deduce metabolic systems using genomic data, genes, and pathways. Haft et al. built a Web-based system to query and display curated phenotypes and annotated prokaryotic genome properties, such as protein families, pathways, and phylogenies [9]; however, the system does not predict correlations between microbiological phenotypes and genome properties.

In addition to the work being done to integrate prokaryotic phenotypes to genotypes, researchers have also made significant advances in building large-scale phenotype–genotype networks in mice, rats, and humans. The Mouse Genome Database (MGD) has structured their mouse genomic data in terms of the Mammalian Phenotype Ontology [10]. Similarly, the Rat Genome Database (RGD) [11] also developed a phenome database, integrated with its genomic data. In humans, the GeneNetwork (WebQTL) provides a database of complex traits with mappings to quantitative trait loci [12]. And several studies have focused on integrating human phenome and genome resources. For example, Butte et al. created a large-scale phenome–genome network by integrating the Unified Medical Language System with human microarray gene expression data [13]; and Aerts et al. applied a prioritization method to associate genes with human diseases and pathways [14].

We hypothesized that by automatically and simultaneously merging and analyzing massive quantities of microbiological phenotypes and their molecular datasets, we could predict both the molecular underpinnings of prokaryotic phenotypes as well as the relationships between related groups of phenotypes. Thus, this study is designed to illustrate how the big picture emerges from the network of predictions between multiple scales of molecular mechanisms and their correlations to an individual phenotype or to clusters of phenotypes. We developed a high-throughput computational approach, and for the first time, demonstrate the feasibility of integrating a large quantity of prokaryotic phenotypes with genomic datasets from various sources for large-scale data mining across different scales of molecular biology (protein domains, pathways, molecular function, and cellular processes).

To analyze large quantities of prokaryotic phenotypes, we employed the Microbiology module of the Global Infectious Diseases and Epidemiology Network (GIDEON) that we refer to as the Microbiology Knowledge Dataset (MKD) as our source data on phenotypes [15,16]. MKD contains results from laboratory examinations though which users can distinguish different microorganisms. These laboratory results contain descriptions about the morphologic characteristics of microorganisms (e.g., Gram-positive, Gram-negative, motility, and cell wall deficiency), metabolic functions of microorganisms, (e.g., urea hydrolysis, acetate utilization, and gas production from glucose), and microorganisms' adaptation to extreme living conditions, (e.g., growth at 42 °C and growth in 6.5% sodium chloride). We regarded MKD laboratory test results as phenotypes or phenotypic traits, as they constitute observable physical or biochemical characteristics under certain experimental conditions that are determined by the microorganisms' genetic contents. MKD contains more than 100 phenotypic characterizations for more than 3,000 bacterial species, not only allowing us to conduct large-scale data mining on genomics data over phenotypic traits, but also enabling us to compare different phenotypic traits based on their correlations to their genetic contents. Of these 3,000 bacterial species, we included 59 species with fully sequenced genomes in our studies. To integrate phenotypes in MKD with genomic datasets, we chose to include the Protein Family Database (Pfam) [17], Clusters of Orthologous Groups (COGs) [18,19], Kyoto Encyclopedia of Genes and Genomes (KEGG) [20], and biological concepts found in the Gene Ontology (GO) [21,22] which span multiple scales of biology.

Applying our method of data integration and mining, we have identified the genetic basis and molecular mechanisms underlying the many bacterial phenotypes. We revealed 3,711 significant correlations and *anti*-correlations (p-value < 0.05) between 63 microbiological phenotypes and 1,499 Pfam families, and identified 17 and 506 significant molecular mechanisms of phenotypes according to our analyses of KEGG's biochemical pathways, and GO's biological concepts, respectively. In addition, for the first time, a novel phenomic analysis was conducted to compare phenotypes with each other on a large scale based on their genetic contents. The original visualization of the network of relationships between one cluster of phenotypes and its significant correlations to protein families, molecular pathways, processes, and function illustrates how clusters of phenotypes (metaphenotypes) share common molecular mechanisms. Such analysis could lead to a better understanding of the molecular relationships between microbial phenotypes on a genomic scale. We believe that this computational technology holds promise in facilitating a systems biology approach to biomedical research in the post-genomic era.

Results/Discussion

To address our hypothesis, we first describe the results from the high throughput mapping of phenotypes with multiple databases of molecular mechanisms: first the Pfam, followed by KEGG pathways and GO terms. Then, we present results from a combined phenomic analysis of the significant molecular mechanisms across multiple biological scales.

Mapping the MKD's Clinical Phenotypic Traits to Protein Families

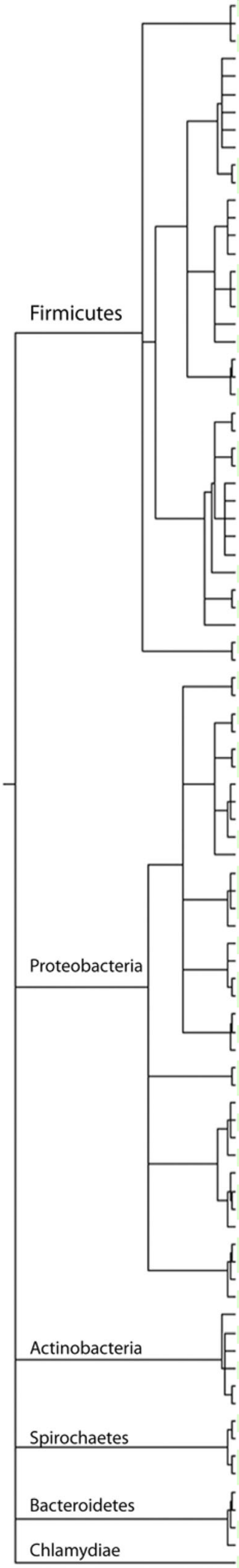
Currently, the availability of more than 208 microbial genome sequences in GenBank provides a rich source of information about the genetic contents of various microorganisms [23]. In addition, functional classification databases, such as COGs [18,19] and Pfam [17], enable us to compare conservation and divergence of functional genes across microorganisms. However, little has been done in the past to correlate the genomic data with phenotypic information. In this study, to uncover the underlying linkages between microorganism phenotypes and their genetic contents, we integrated and analyzed datasets of a microbiological phenotypic database (the MKD) and a genomic protein domains dataset (Pfam). In this study we have, by design, limited the analysis to complete genome sequences to avoid a selection bias toward genes coming from partial genomes that were preferentially sequenced. Methods to deal with organisms with partial genomic sequences will be explored in a future study. As a result, we selected each of the 59 species of microorganisms that exist in both the MKD and Pfam databases with fully deduced genome sequences (Figure 1). These species belong to six phylums, including 20 Firmicutes, 17 Proteobacteria, six Actinobacteria, four Spirochaetes, four Bacteroidetes, and one Chlamydiae, representing about 30% of the bacteria species that have been fully sequenced at the time of this study. Detailed information about their taxonomy in comparison with the fully sequenced bacteria at the time of this study is provided in Table 1. Out of the 208 fully sequenced bacteria available at the time of this study, the 59 species used in this study cover approximately 30% to 40% of available fully sequenced bacterial genomes at different taxonomic levels.

Taxonomical mapping between genomic and phenotypic databases. The 59 microorganisms were automatically mapped between datasets of MKD and Pfam using the National Center for Biotechnology Information's (NCBI) taxons as a reference, followed by manual examination by experts (Figure 1). Taxons assigned to fully sequenced microorganisms are all at the strain levels at the NCBI (marked as no rank), but those in the MKD are mostly available only at the level of species (57 species, one subspecies, and one no rank). Therefore, most of the organism mappings between MKD and Pfam were either exactly matched or within one taxonomical distance (e.g., a mapping between a species in MKD and a subspecies in fully sequenced bacteria). Given this limitation on data resources, our mapping approach between the phenotypic and genomic datasets is based on the principle that phenotypes for one species are valid for every subsumed strain (one taxonomical range). For example, the MKD contains microbial phenotypes documented as laboratory results for *B. anthracis,* which is defined as a species (Figure 1, Taxon 1392). Four fully sequenced strains of *B. anthracis* are defined as children of this species and categorized as no rank in the NCBI taxonomy. To control for overrepresentation of a species in the calculation of the hypergeometric distribution, we hence regarded this as a mapping, and all the Pfam families of the four subspecies were merged into one group to compare with the microbial phenotypes of *B. anthracis*. We took this approach to avoid excluding any Pfam families found in the annotations of the *B. anthracis* species classified as no rank (i.e., in the case that a sequencing error in one strain causes a gene being neglected, protein families associated to this gene would still be included in this study due to the annotation of the other strains). However, this approach also includes some additional Pfam families that belong to horizontally transferred genes from the different strains, which could introduce noise in the study. In addition, a sampling bias may have been introduced in our analysis because the phenotype database pertains to bacteria that are pathogenic or commensal to Homo sapiens and may therefore have more opportunities for horizontal gene transfer than a random set of prokaryotes. In future studies, we intend to combine new weighted statistical approaches that incorporate phylogenetic distance [24] and measurements of horizontal gene transfer [25] with the hypergeometric distribution to control for these potential biases. The number of Pfam families for all bacteria are also shown in Figure 1.

Identification of Correlations between Bacterial Protein Domains and Phenotypes

We applied a comprehensive statistical and visualization method based on the hypergeometric distribution to identify the correlations between phenotypic laboratory results and the genetic contents of bacteria. Details are described in Materials and Methods (Equations 1–3), and the procedure is illustrated in Figure 2. In total, we calculated the correlations of the co-occurrences between Pfam families and positive phenotypic laboratory results in the MKD across 59 bacteria species. The correlations can be defined within two categories: 1) correlation, in which the existence of a Pfam family correlates with positive laboratory results; 2) *anti*-correlation, in which existence of a Pfam family correlates with negative

List of fully-sequenced bacterial species used in this study



Species Group	Microorganism in the lab test	NCBI Taxon*	Strain sequenced	NCBI Taxon	Num. PFAMs
1	Mycoplasma genitalium	2097 (species)	Mycoplasma genitalium G-37	243273 (no rank)	400
2	Mycoplasma pneumoniae	2104 (species)	Mycoplasma pneumoniae M129	272634 (no rank)	424
3	Mycoplasma penetrans	28227 (species)	Mycoplasma penetrans HF-2	272633 (no rank)	510
4	Staphylococcus aureus	1280 (species)	Staphylococcus aureus subsp. aureus Mu50	158878 (no rank)	1343
4	Staphylococcus aureus	1280 (species)	Staphylococcus aureus subsp. aureus N315	158879 (no rank)	1343
4	Staphylococcus aureus	1280 (species)	Staphylococcus aureus subsp. aureus MW2	196620 (no rank)	1343
4	Staphylococcus aureus	1280 (species)	Staphylococcus aureus subsp. aureus MRSA252	282458 (no rank)	1343
4	Staphylococcus aureus	1280 (species)	Staphylococcus aureus subsp. aureus MSSA476	282459 (no rank)	1343
4	Staphylococcus aureus	1280 (species)	Staphylococcus aureus subsp. aureus COL	93062 (no rank)	1343
5	Staphylococcus epidermidis	1282 (species)	Staphylococcus epidermidis RP62A	176279 (no rank)	1198
5	Staphylococcus epidermidis	1282 (species)	Staphylococcus epidermidis ATCC 12228	176280 (no rank)	1198
6	Bacillus anthracis	1392 (species)	Bacillus anthracis str. A2012	191218 (no rank)	1470
6	Bacillus anthracis	1392 (species)	Bacillus anthracis str. Ames	198094 (no rank)	1470
6	Bacillus anthracis	1392 (species)	Bacillus anthracis str. Sterne	260799 (no rank)	1470
6	Bacillus anthracis	1392 (species)	Bacillus anthracis str. 'Ames Ancestor'	261594 (no rank)	1470
7	Bacillus cereus	1396 (species)	Bacillus cereus ATCC 10987	222523 (no rank)	1577
7	Bacillus cereus	1396 (species)	Bacillus cereus ATCC 14579	226900 (no rank)	1577
7	Bacillus cereus	1396 (species)	Bacillus cereus ZK	288681 (no rank)	1577
8	Bacillus licheniformis	1402 (species)	Bacillus licheniformis ATCC 14580	279010 (no rank)	152
9	Bacillus subtilis	1423 (species)	Bacillus subtilis subsp. subtilis str. 168	224308 (no rank)	1539
10	Listeria monocytogenes	1639 (species)	Listeria monocytogenes EGD-e	169963 (no rank)	1339
10	Listeria monocytogenes	1639 (species)	Listeria monocytogenes str. 4b F2365	265669 (no rank)	1339
11	Listeria innocua	1642 (species)	Listeria innocua Clip11262	272626 (no rank)	1298
12	Streptococcus agalactiae	1311 (species)	Streptococcus agalactiae 2603V/R	208435 (no rank)	147
12	Streptococcus agalactiae	1311 (species)	Streptococcus agalactiae NEM316	211110 (no rank)	147
13	Streptococcus pneumoniae	1313 (species)	Streptococcus pneumoniae TIGR4	170187 (no rank)	1073
13	Streptococcus pneumoniae	1313 (species)	Streptococcus pneumoniae R6	171101 (no rank)	1073
14	Streptococcus pyogenes	1314 (species)	Streptococcus pyogenes M1 GAS	160490 (no rank)	1023
14	Streptococcus pyogenes	1314 (species)	Streptococcus pyogenes MGAS8232	186103 (no rank)	1023
14	Streptococcus pyogenes	1314 (species)	Streptococcus pyogenes SSI-1	193567 (no rank)	1023
14	Streptococcus pyogenes	1314 (species)	Streptococcus pyogenes MGAS315	198466 (no rank)	1023
14	Streptococcus pyogenes	1314 (species)	Streptococcus pyogenes MGAS10394	286636 (no rank)	1023
15	Lactococcus lactis	1358 (species)	Lactococcus lactis subsp. lactis Il1403	272623 (no rank)	236
16	Lactobacillus plantarum	1590 (species)	Lactobacillus plantarum WCFS1	220668 (no rank)	1163
17	Lactobacillus johnsonii	33959 (species)	Lactobacillus johnsonii NCC 533	257314 (no rank)	854
18	Enterococcus faecalis	1351 (species)	Enterococcus faecalis V583	226185 (no rank)	1207
19	Clostridium perfringens	1502 (species)	Clostridium perfringens str. 13	195102 (no rank)	1259
20	Clostridium tetani	1513 (species)	Clostridium tetani E88	212717 (no rank)	1125
21	Pseudomonas aeruginosa	287 (species)	Pseudomonas aeruginosa PAO1	208964 (no rank)	1855
22	Pseudomonas putida	303 (species)	Pseudomonas putida KT2440	160488 (no rank)	1756
23	Escherichia coli	562 (species)	Escherichia coli K12	83333 (no rank)	1609
24	Escherichia coli O157:H7	83334 (no rank)	Escherichia coli O157:H7	83334 (no rank)	1747
25	Shigella flexneri	623 (species)	Shigella flexneri 2a str. 301	198214 (no rank)	1778
25	Shigella flexneri	623 (species)	Shigella flexneri 2a str. 2457T	198215 (no rank)	1778
26	Yersinia pestis	632 (species)	Yersinia pestis KIM	187410 (no rank)	1753
26	Yersinia pestis	632 (species)	Yersinia pestis CO92	214092 (no rank)	1753
26	Yersinia pestis	632 (species)	Yersinia pestis biovar Medievalis str. 91001	229193 (no rank)	1753
27	Yersinia pseudotuberculosis	633 (species)	Yersinia pseudotuberculosis IP 32953	273123 (no rank)	147
28	Photorhabdus luminescens	29488 (species)	Photorhabdus luminescens subsp. laumondii TTO1	243265 (no rank)	188
29	Legionella pneumophila	446 (species)	Legionella pneumophila str. Lens	297245 (no rank)	218
29	Legionella pneumophila	446 (species)	Legionella pneumophila str. Paris	297246 (no rank)	218
29	Legionella pneumophila	446 (species)	Legionella pneumophila subsp. pneumophila str. Philadelphia	272624 (no rank)	218
30	Coxiella burnetii	777 (species)	Coxiella burnetii RSA 493	227377 (no rank)	945
31	Vibrio cholerae	666 (species)	Vibrio cholerae O1 biovar eltor str. N16961	243277 (no rank)	1629
32	Vibrio parahaemolyticus	670 (species)	Vibrio parahaemolyticus RIMD 2210633	223926 (no rank)	1683
33	Vibrio vulnificus	672 (species)	Vibrio vulnificus YJ016	196600 (no rank)	1720
33	Vibrio vulnificus	672 (species)	Vibrio vulnificus CMCP6	216895 (no rank)	1720
34	Haemophilus influenzae	727 (species)	Haemophilus influenzae Rd KW20	71421 (no rank)	1209
35	Haemophilus ducreyi	730 (species)	Haemophilus ducreyi 35000HP	233412 (no rank)	992
36	Pasteurella multocida	747 (species)	Pasteurella multocida subsp. multocida str. Pm70	272843 (no rank)	1281
37	Agrobacterium tumefaciens	358 (species)	Agrobacterium tumefaciens str. C58	176299 (no rank)	1636
37	Agrobacterium tumefaciens	358 (species)	Agrobacterium tumefaciens str. C58	176299 (no rank)	1636
38	Bordetella bronchiseptica	518 (species)	Bordetella bronchiseptica RB50	257310 (no rank)	1494
39	Bordetella parapertussis	519 (species)	Bordetella parapertussis 12822	257311 (no rank)	1452
40	Bordetella pertussis	520 (species)	Bordetella pertussis Tohama I	257313 (no rank)	1335
41	Burkholderia pseudomallei	28450 (species)	Burkholderia pseudomallei K96243	272560 (no rank)	152
42	Neisseria gonorrhoeae	485 (species)	Neisseria gonorrhoeae FA 1090	242231 (no rank)	239
43	Neisseria meningitidis	487 (species)	Neisseria meningitidis MC58	122586 (no rank)	204
43	Neisseria meningitidis	487 (species)	Neisseria meningitidis Z2491	122587 (no rank)	204
44	Chromobacterium violaceum	536 (species)	Chromobacterium violaceum ATCC 12472	243365 (no rank)	1547
45	Helicobacter pylori	210 (species)	Helicobacter pylori 26695	85962 (no rank)	926
45	Helicobacter pylori	210 (species)	Helicobacter pylori J99	85963 (no rank)	926
46	Wolinella succinogenes	844 (species)	Wolinella succinogenes DSM 1740	273121 (no rank)	1012
47	Campylobacter jejuni	32022 (subspecies)	Campylobacter jejuni subsp. jejuni NCTC 11168	192222 (no rank)	1012
48	Corynebacterium diphtheriae	1717 (species)	Corynebacterium diphtheriae NCTC 13129	257309 (no rank)	1043
49	Mycobacterium avium	1764 (species)	Mycobacterium avium subsp. paratuberculosis str. k10	262316 (no rank)	111
50	Mycobacterium bovis	1765 (species)	Mycobacterium bovis AF2122/97	233413 (no rank)	989
51	Mycobacterium leprae	1769 (species)	Mycobacterium leprae TN	272631 (no rank)	927
52	Mycobacterium tuberculosis	1773 (species)	Mycobacterium tuberculosis CDC1551	83331 (no rank)	1179
52	Mycobacterium tuberculosis	1773 (species)	Mycobacterium tuberculosis H37Rv	83332 (no rank)	1179
53	Treponema denticola	158 (species)	Treponema denticola ATCC 35405	243275 (no rank)	955
54	Treponema pallidum	160 (species)	Treponema pallidum subsp. pallidum str. Nichols	243276 (no rank)	642
55	Leptospira interrogans	173 (species)	Leptospira interrogans serovar Copenhageni str. Fiocruz L1	267671 (no rank)	1152
55	Leptospira interrogans	173 (species)	Leptospira interrogans serovar Lai str. 56601	189518 (no rank)	1152
56	Bacteroides fragilis	817 (species)	Bacteroides fragilis NCTC 9343	272559 (no rank)	136
56	Bacteroides fragilis	817 (species)	Bacteroides fragilis YCH46	295405 (no rank)	136
57	Bacteroides thetaiotaomicron	818 (species)	Bacteroides thetaiotaomicron VPI-5482	226186 (no rank)	1208
58	Porphyromonas gingivalis	837 (species)	Porphyromonas gingivalis W83	242619 (no rank)	880
59	Chlamydia trachomatis	813 (species)	Chlamydia trachomatis D/UW-3/CX	272561 (no rank)	646

Figure 1. List of Bacterial Species with Full Genome Sequences Used in This Study

Bacteria with full genome sequences and laboratory tests used in this study are listed with their phylogenetic tree drawn according to the NCBI taxonomy (the NCBI Taxonomy database is widely used for taxonomy; however, it is not an authoritative source for nomenclature or classification, and an alternate dendogram could have been constructed using the Species2000 Bacteriology Insight Orienting System, http://www-sp2000ao.nies.go.jp/english/bios/). Many of the taxons of species in the laboratory tests are parents of the strains being fully sequenced. The numbers of Pfam families for all species are also shown.

doi:10.1371/journal.pcbi.0020159.g001

phenotypic laboratory results. Two statistical methods were employed to discover correlations and *anti*-correlations: 1) the conservative Šidák adjustment of the p-value for multiple a posteriori comparisons [26]; and 2) the calculation of error rates using statistical simulations based on permutation resampling without replacement. These results are available in files available at http://phenos.bsd.uchicago.edu/prok_phenotype/.

Overall, we have identified 3,711 significant correlations between 1,499 distinct Pfam and 63 phenotypes with an experiment-wide error rate of 5%, including 2,650 correlations and 1,061 *anti*-correlations. Here we weight the *anti*-correlations with the same importance as correlations, since the description of the opposite phenotypes would positively correlate with the same set of Pfam families. For example, the phenotype of vancomycin susceptiblity has an *anti*-correlation with the Pfam family of HlyD family secretion protein (PF00529); thus, we can also consider the converse relation to the opposite phenotype: the phenotype of vancomycin resistance has a positive correlation with the same Pfam family. We observed that while some phenotypes (i.e., motility and Gram-negative) correlate with a large number of Pfam families, others correlate with only a few families (i.e., Gelatin hydrolysis and urea hydrolysis). However, the number of correlations inferred by this method depends on both the limitations of the method [5] and the number of available phenotypic laboratory results for different species. For example, Pfam families that exist in all (or no) species would not be correlated with any laboratory results; neither would positive laboratory results that are lacking or existing in most species.

Adjusting results for multiple comparisons. The resulting p-values are adjusted for multiple a posteriori comparisons to reduce the experiment-wide error rate. One of the most commonly used methods to control for experiment-wide error rate is the conservative Bonferroni-type adjustment. We used the related Šidák method, as discussed in detail in Materials and Methods (Equation 3), which provides a conservative threshold to filter out false positive results. In this approach, we regarded each phenotypic laboratory test independently, and applied the Šidák adjustment according to the total number of comparisons analyzed. For example, when comparing phenotypes to Pfam families, the number of distinct comparisons with each phenotype is 2,665; the number of proteins that can be compared with each species. Additionally, we limited the comparisons to those phenotypes and Pfam families found in more than three and fewer than 56 species. This resulted in 478 correlations with corrected p-values of no more than 0.05 (shown in the file Phenotype_Sidak_Pfam_mapping.xls at http://phenos.bsd.uchicago.edu/prok_phenotype).

Since the Šidák adjustment provides a set of conservative results, many interesting correlations may be consequently filtered out due to its conservative criteria for genome-wide studies, as the variables under study are not entirely independent [27]. In this study, some laboratory tests and the organisms selected are not independent. For example, the laboratory tests Gram-negative and Gram-positive are *anti*-correlated. Organisms are phylogenetically related, of which Firmicutes and Proteobacteria are over-represented in the species used in this study (34% and 29% of the 59 species). Moreover, since the laboratory tests are designed to distinguish bacteria species, there is also a bias on laboratory tests being used to distinguish over-represented species. All of them are currently limitations in this study due to availability of prokaryotic phenotypes limited to MKD—a clinical microbiological database. Certainly, with more species being sequenced and more phenotypic data, we could explore using independent laboratory test results with a set of species more representative of overall prokaryotic diversity.

To overcome these limitations, we applied an additional method based on statistical simulation which can stratify predicted correlations as described in detail in Materials and Methods. With this method, we conducted a permutation resampling in which we compared the number of significant correlations inferred from the original data with those inferred from an experimental control consisting of the distributions of random permutations of the data with different statistical cutoffs for the hypergeometric distribution. Since this method predicts significant correlations in comparison with randomized samples, its results have less stringent cutoffs and cover more phenotypes. Figure 3 summarizes the results from the control experiment over random data, using cutoffs of uncorrected p-values, ranging from 0.0001 to 0.05 from the uncorrected hypergeometric test (details in Materials and Methods). As shown in Figure 3, we can expect about 5% of the predictions to be false positives if the uncorrected p-value of the hypergeometric distribution is equal to or less than 0.002. In our unadjusted dataset, using an uncorrected p-value of 0.002 or less, we identified 3,711 significant correlations in which we expect about 5% to be false positive predictions (data shown in the

Table 1. Phylogenetic Classification of Bacteria Used in This Study in Comparison with the Fully Sequenced Bacteria

Classification	Bacteria in This Study (59)	Bacteria with Full Genome Sequence (208)	Coverage
Phylum	6	17	35.3%
Class	12	28	42.9%
Order	18	57	31.6%
Family	28	81	34.6%
Genus	35	113	31.0%
Species	57	172	33.1%

The bacteria used in this study are classified into six taxonomy categories. They are compared with bacteria having fully sequenced genomes at the time of the study.

doi:10.1371/journal.pcbi.0020159.t001

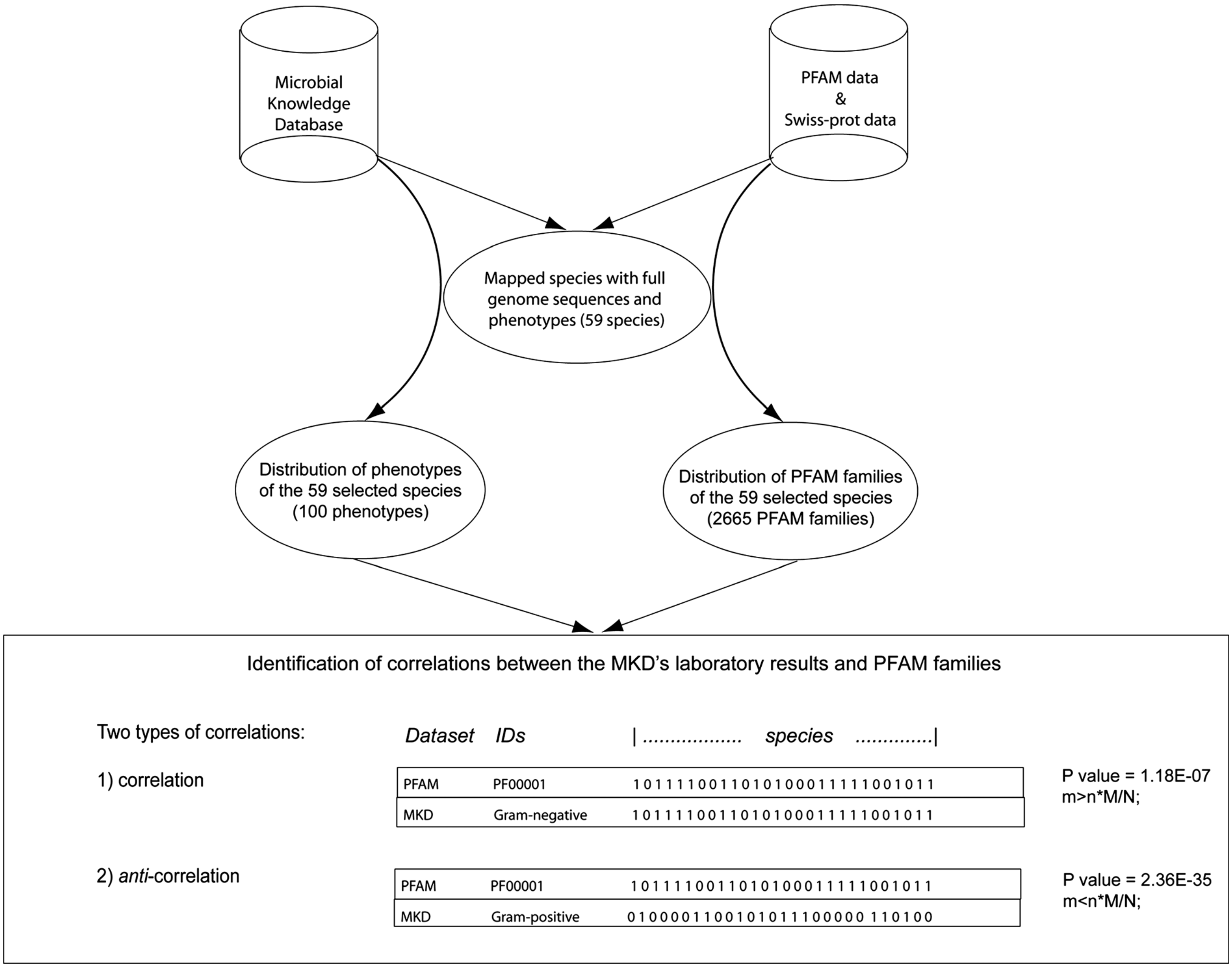


Figure 2. Procedure of Data Integration for Correlating Phenotypes with Pfam Families

This flowchart shows how the datasets have been integrated. The calculation of correlations between phenotypes and Pfam families is illustrated in the framed area at the bottom. The formula presented in the box is derived from the hypergeometric distribution and allows for a differentiation between correlation and *anti*-correlation.

N, the total number of species used in the study (59); M, the number of species that have a specific Pfam family, such as PF00001 illustrated; n, the number of species that have a specific phenotype, such as Gram-negative; m, the number of species that have both a specific Pfam family and a specific phenotype.

doi:10.1371/journal.pcbi.0020159.g002

file Phenotype_Sim_Pfam_mapping.xls at http://phenos.bsd.uchicago.edu/prok_phenotype). By carefully choosing complementary statistical methods for conducting data-mining calculations, we can provide researchers with accurate stratified information on the prediction, yet without filtering out potentially meaningful correlations.

Evaluation of the correlations between phenotypes and protein domains. We conducted an extensive manual evaluation of our predictions, which consists of five parts.

First, we manually examined all the phenotypes (21 phenotypes in total) with their 478 significantly correlated Pfam families based on the Šidák adjustment (data shown in the file Phenotype_Sidak_Pfam_mapping.xls at http://phenos.bsd.uchicago.edu/prok_phenotype). One hundred distinct predictions were manually assessed and 60 were corroborated and annotated with the supporting bibliographic references (Table S2). We then analyzed each of these manually curated sets and provide a summary of the analysis of these predictions for each of the 21 phenotypes (Table S2). Overall, 67% (14) of these phenotypes have at least one Pfam association that was corroborated as shown in Table S2.

Second, we randomly selected 50 positive correlations and 15 *anti*-correlations from the simulation method to evaluate the minimum precision of the predictions. In the evaluation process, we focused on evaluating the positively correlated phenotypes and Pfam families, since *anti*-correlations are often difficult to verify. Of the 50 positive correlations selected, 15 of them were confirmed by supporting literature. As future studies may provide additional corroborations, the precision of 30% (95% confidence interval: 20%–42%; $n = 50$) is a conservative estimate of the overall potential accuracy of the prediction method controlling for false positive rate (also known as false discovery rate) with permutation resampling. Of the 15 *anti*-correlations, two of them (13%) were supported by literature (95% confidence interval: 2%–40%; $n = 15$). A summary of this validation is provided in

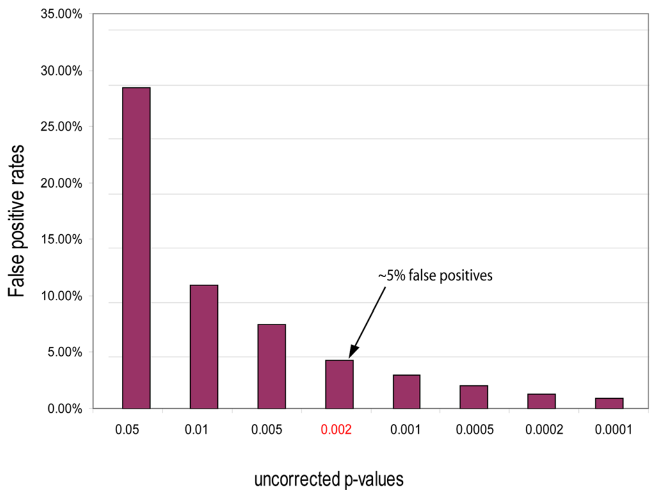


Figure 3. False Positive Error Rates Predicted from Random Datasets According to the Uncorrected Hypergeometric Distribution

The false positive error rate represents the ratio of the number of significant correlations from the randomized dataset (control experiment) to the number of significant correlations from the real dataset below a certain *p*-value. At different *p*-value cutoffs, we calculated the error rates from a sample of 1,000 random permutations of the relationship vectors within the dataset (permutation resampling method), and the cutoffs for the highest 1% of occurrences for each uncorrected *p*-value of the hypergeometric distribution (data presented). For uncorrected *p*-values of 0.002 or less, the correlations between phenotypes and Pfam families are predicted to have an error rate of approximately 5%. This cutoff is applied in this study to identify significant correlations.

doi:10.1371/journal.pcbi.0020159.g003

Table S3 with supportive references for each corroborated prediction.

Third, the false negative rate of the correlations that were regarded as *statistically insignificant* is also estimated. We evaluated 50 random samples, and only one of them has been shown to be correlated, resulting in a 2% false negative rate (95% confidence interval: 0.1%–10%). A summary of this validation is provided in Table S4.

Fourth, we conducted an in-depth evaluation of one phenotype (motility) and compared our results with those from a previously reported study [5], which used a different classification method to cluster full-length genes and interpreted the results using annotations of *E. coli* genes. We found the results of the two studies to be well-correlated, especially for the top-ranked genes (19 of the top 30, or about 63%, *E. coli* genes have corresponding Pfam families in the top 30 families in our study: see Table S1).

Fifth, to further evaluate the accuracy of the method in the well-studied phenotype of motility, we performed a manual validation of the predicted results pertaining to bacterial motility mediated through flagella. In this evaluation, significant correlations between phenotypes and Pfam families using the Šidák adjustment and the simulation methods are examined. Since the Šidák adjustment is more conservative, its predicted correlations are also included in those predicted by the simulation method. The results are shown in Table 2, where 18 and 58 Pfam families are predicted by the Šidák and simulation methods, respectively. By manual examination of the annotation of Pfam families, we identified those which participate in bacterial motility, including flagellar mediated motion and chemotaxis. We manually confirmed 12 (out of 18) and 27 (out of 58) Pfam family predictions from the Šidák and permutation resampling methods, respectively. These results confirmed that the Šidák method predicts relatively conservatively, and the data-mining method works equally well to provide accurate predictions. In addition, our results could help improve the functional understanding of current Pfam annotations. For example, we discovered one of the Pfam families, PF06429, described as Domain of unknown function (DUF1078), to be correlated with bacteria motility.

Overall, the results of these evaluations indicate that our approach can faithfully identify the most significantly correlated protein families as accurately as the other classification methods. However, our approach differs from the previous studies because we compared significantly more phenotypes and extended the phenotypic analyses to KEGG pathways and GO concepts which have not previously been analyzed in other studies to our knowledge (discussed below).

Limitations of the correlations of phenotypes to protein domain families and future work. In this study, we primarily used sequence-based classifications (Pfam and COGs) to correlate with phenotypes. The correlations identified by this method suggest hypothetical association based on statistical analysis. However, we limited our exploration of the converse, correlations that are not statistically significant, to the previously described one manual evaluation (Table S5). Though it is feasible to conduct studies to demonstrate that there is not a correlation between certain properties, this was not the design of this study, and therefore we cannot make conclusions about the absence of relationships between correlated elements that did not reach statistical significance in this study. Many factors could lead to statistically insignificant correlations in our approach, for example, the lack of available laboratory data could lead to poor correlations to Pfam families. In future work, it would be interesting to explore the use of structure-based classifications and databases, such as the Structural Classification of Proteins (SCOP) [28], CATH [29], or DALI [30], or using integrated structure and sequence-based classifications, such as classifications based on Pfam domains integrated with Structural Classification of Proteins domains, as studied by Pouliot et al. [31]. Furthermore, we could integrate the classified protein domains with a protein structure database, such as the Protein Data Bank (PDB) [32] or OCA (http://oca.ebi.ac.uk/oca-docs/oca-home.html), to further study their functions.

Mapping Phenotypes to KEGG Pathways and GO

We also applied the hypergeometric statistical and data-mining approaches to identify correlations of phenotypes with molecular pathways and GO concepts. Using existing bioinformatics resources, we integrated data using the following methods: 1) phenotypes with KEGG molecular pathways by mining their matching COG groups; and 2) phenotypes with GO concepts by mining their matching Pfam families. KEGG pathways and GO concepts significantly correlated with phenotypes were identified by their probabilities of occurrence (see Materials and Methods). This provided more correlations for the mapping, which are likely to reveal biological significance. The details of the procedure are described in Materials and Methods.

We unveiled ten significant correlations and seven signifi-

Table 2. Pfam Families Significantly Correlated with Bacterial Motility

PFAM ID	PFAM Description	Predictions (*p*-Values)			Related to Flagella or Chemotaxis
		Šidák-Adjusted	Simulation-Adjusted	Unadjusted	
PF00460.8	Flagella basal body rod protein	3.9E-04	3.7E-06	1.5E-07	Y
PF06429.1	Domain of unknown function (DUF1078)	3.9E-04	3.7E-06	1.5E-07	
PF00669.8	Bacterial flagellin N-terminus	6.0E-04	5.7E-06	2.3E-07	Y
PF01584.8	CheW-like domain	6.7E-04	6.3E-06	2.5E-07	Y
PF01514.7	Secretory protein of YscJ/FliF family	1.5E-03	1.4E-05	5.7E-07	Y
PF00771.7	FHIPEP family	2.7E-03	2.6E-05	1.0E-06	Y
PF02154.5	Flagellar motor switch protein FliM	3.8E-03	3.6E-05	1.5E-06	Y
PF02895.4	Signal transducing histidine kinase, homodimeric domain	3.8E-03	3.6E-05	1.5E-06	
PF00700.8	Bacterial flagellin C-terminus	5.2E-03	4.9E-05	2.0E-06	Y
PF00813.7	FliP family	5.2E-03	4.9E-05	2.0E-06	
PF01052.8	Surface presentation of antigens (SPOA) protein	5.2E-03	4.9E-05	2.0E-06	
PF01311.8	Bacterial export proteins, family 1	5.2E-03	4.9E-05	2.0E-06	
PF01312.8	FlhB HrpN YscU sugar porter activity S family	5.2E-03	4.9E-05	2.0E-06	
PF00015.10	Methyl-accepting chemotaxis protein (MCP) signaling domain	1.4E-02	1.4E-04	5.4E-06	Y
PF03963.3	Flagellar hook capping protein	1.4E-02	1.4E-04	5.4E-06	Y
PF01739.8	CheR methyltransferase, SAM binding domain	1.8E-02	1.7E-04	6.8E-06	Y
PF03705.5	CheR methyltransferase, all-alpha domain	1.8E-02	1.7E-04	6.8E-06	Y
PF01313.7	Bacterial export proteins, family 3	3.0E-02	2.9E-04	1.2E-05	Y
PF01706.6	FliG C-terminal domain	NS	5.9E-04	2.4E-05	Y
PF02465.7	Flagellar hook–associated protein 2 C-terminus	NS	5.9E-04	2.4E-05	Y
PF07195.1	Flagellar hook–associated protein 2 C-terminus	NS	5.9E-04	2.4E-05	Y
PF02743.6	Cache domain	NS	6.8E-04	2.7E-05	Y
PF03748.3	Flagellar basal body–associated protein FliL	NS	6.9E-04	2.7E-05	Y
PF07559.1	Flagellar basal body protein FlaE	NS	2.2E-03	8.9E-05	Y
PF02561.4	Flagellar protein FliS	NS	2.4E-03	9.6E-05	Y
PF01627.11	Hpt domain	NS	3.7E-03	1.5E-04	Y
PF03646.4	FlaG protein	NS	4.3E-03	1.7E-04	Y
PF00309.9	Sigma-54 factor, Activator interacting domain (AID)	NS	6.0E-03	2.4E-04	
PF01354.8	Antifreeze-like domain	NS	6.0E-03	2.4E-04	
PF03102.4	NeuB family	NS	6.0E-03	2.4E-04	
PF04509.2	CheC-like family	NS	7.0E-03	2.8E-04	Y
PF02120.6	Flagellar hook–length control protein	NS	7.1E-03	2.8E-04	Y
PF02049.6	Flagellar hook–basal body complex protein FliE	NS	7.3E-03	2.9E-04	Y
PF07196.1	Flagellin hook IN motif	NS	7.6E-03	3.0E-04	Y
PF07726.1	ATPase family associated with various cellular activities (AAA)	NS	7.6E-03	3.0E-04	
PF04552.2	Sigma-54, DNA binding domain	NS	1.4E-02	5.7E-04	
PF04963.2	Sigma-54 factor, core binding domain	NS	1.4E-02	5.7E-04	
PF01464.8	Transglycosylase SLT domain	NS	1.8E-02	7.1E-04	
PF03934.3	General secretion pathway protein K	NS	1.9E-02	7.8E-04	
PF07228.2	Stage II sporulation protein E (SpoIIE)	NS	2.0E-02	7.8E-04	
PF00704.14	Glycosyl hydrolases family 18	NS	2.0E-02	8.0E-04	
PF02107.5	Flagellar L-ring protein	NS	2.0E-02	8.0E-04	Y
PF01973.7	Protein of unknown function DUF115	NS	2.1E-02	8.3E-04	
PF02623.4	Uncharacterized BCR, COG1699	NS	2.1E-02	8.3E-04	
PF04187.2	Protein of unknown function, DUF399	NS	2.1E-02	8.3E-04	
PF05787.2	Bacterial protein of unknown function (DUF839)	NS	2.1E-02	8.3E-04	
PF07194.1	P2 response regulator binding domain	NS	2.1E-02	8.3E-04	Y
PF00032.7	Cytochrome b(C-terminal)/b6/petD	NS	3.3E-02	1.3E-03	
PF02433.5	Cytochrome C oxidase, mono-heme subunit/FixO	NS	3.3E-02	1.3E-03	
PF03597.3	Cytochrome oxidase maturation protein cbb3-type	NS	3.3E-02	1.3E-03	
PF04376.2	Arginine-tRNA-protein transferase, N terminus	NS	3.3E-02	1.3E-03	
PF04377.3	Arginine-tRNA-protein transferase, C terminus	NS	3.3E-02	1.3E-03	
PF00691.7	OmpA family	NS	3.9E-02	1.6E-03	
PF02719.4	Polysaccharide biosynthesis protein	NS	4.0E-02	1.6E-03	
PF02630.4	SCO1/SenC	NS	4.4E-02	1.8E-03	
PF02839.4	Carbohydrate binding domain	NS	4.4E-02	1.8E-03	
PF03186.3	CobD/Cbib protein	NS	4.4E-02	1.8E-03	
PF00785.12	PAC motif	NS	4.7E-02	1.9E-03	

The significant Pfam families (p-value < 0.05) predicted by the Šidák and data-mining methods are listed. Pfam families involved in flagellar motility and chemotaxis based on their Pfam annotations are marked.
NS, not significant.
doi:10.1371/journal.pcbi.0020159.t002

cant *anti*-correlations between phenotypes and KEGG pathways and 506 significant correlations between phenotypes and GO concepts. Complete results can be found at the study Web site, http://phenos.bsd.uchicago.edu/prok_phenotype, file Phenotype_KEGG_mapping_results.xls for KEGG mapping and file Phenotype_GO_mapping_results.xls for GO mapping.

Compared with the mapping of phenotypes to Pfam families, which provides the relationships of individual protein domain families to phenotypes, the mapping of phenotypes to GO and pathways provides a systematic view of the underlying molecular mechanisms (from multiple scales of biology) related to phenotypes.

Evaluation of the KEGG pathway mappings. To evaluate the accuracy of our mapping method, we conducted two evaluations: (i) we manually revised each of the 17 predictions, and eight correlations as well as two *anti*-correlations were found corroborated in the literature (Table S5), and (ii) we then pursued a deeper manual evaluation on the most significant mapping results in KEGG. Table 3 shows that two KEGG pathways, the Lipopolysaccharide [33] and Ubiquinone biosynthesis pathways, are significantly correlated with the Gram-negative phenotype, both of which are supported by the literature [34]. In theory, every gene family involved in the Lipopolysaccharide biosynthesis pathway should have significant correlations with the Gram-negative phenotype. Our method accurately identified 15 significantly correlated distinct COGs out of a total of 19 defined in the Lipopolysaccharide biosynthesis pathway. According to the phenotype–COG mapping described in Methods, the remaining four COGs that did not map to the phenotype are COG0438 (predicted glycosyltransferases), COG1442 (Lipopolysaccharide biosynthesis protein: glycosyltransferases), COG0451 (Nucleoside-diphosphate-sugar epimerases), and COG0515 (Serine/threonine protein kinases). This could be due to imprecise definitions in the classification method, resulting in diverse functions of the proteins in the families, as three COGs (COG0438, COG0451, and COG0515) participate in many other pathways; or it could also due to the limitation of our method by using hypergeometric function [5]. In contrast, of the 15 COGs mapped between the Lipopolysaccharide biosynthesis pathway and Gram-negative phenotype, 14 are well-defined and unique to only one pathway, with only one exception (COG0241) that exists in two pathways. This suggests that biases in classification method and gene annotation could reduce the signals for the correlations. Reduction of such biases could improve the accuracy of the prediction of correlations in future studies. Additionally, other data resources could be used in future

Table 3. KEGG Pathways and COGs Concepts That Are Significantly Correlated with the Gram-Negative Phenotype

Microbial Phenotype	Phenotype–KEGG Pathway Correlation			Phenotype–COG Correlation		
	KEGG ID	KEGG Pathway Description	*p*-Value	Mapped COGs	Šidák-Adjusted *p*-Value	Simulation-Adjusted *p*-Value
Gram-negative	ot00540	Lipopolysaccharide biosynthesis	6.6E-07	COG2877	4.8E-06	4.3E-08
				COG1044	4.8E-06	4.3E-08
				COG0763	4.8E-06	4.3E-08
				COG1043	4.8E-06	4.3E-08
				COG1519	4.8E-06	4.3E-08
				COG1663	4.8E-06	4.3E-08
				COG0774	4.8E-06	4.3E-08
				COG1212	4.8E-06	4.3E-08
				COG0859	1.6E-02	1.4E-04
				COG1560	4.5E-02	4.1E-04
				COG2908	5.0E-02	4.6E-04
				COG2870	5.0E-02	4.6E-04
				COG3307	5.3E-01	6.8E-03
				COG0241	1.0E+00	5.2E-02
				COG0279	1.0E+00	1.2E-01
	ot00130	Ubiquinone biosynthesis	5.9E-07	COG0043	4.1E-03	3.7E-05
				COG0163	4.1E-03	3.7E-05
				COG2227	1.4E-01	1.3E-03
				COG1008	9.7E-01	3.0E-02
				COG0838	9.7E-01	3.0E-02
				COG0377	9.7E-01	3.0E-02
				COG0852	9.7E-01	3.0E-02
				COG0839	9.7E-01	3.0E-02
				COG1007	9.7E-01	3.0E-02
				COG1143	9.7E-01	3.0E-02
				COG1005	9.7E-01	3.0E-02
				COG0713	9.7E-01	3.0E-02
				COG0649	9.7E-01	3.0E-02
				COG0382	9.8E-01	3.5E-02

doi:10.1371/journal.pcbi.0020159.t003

studies for assessing prokaryotic molecular pathways, such as Metacyc and Ecocyc [35,36].

Evaluation of GO mappings. In addition to the phenotype-pathway mapping, we mapped phenotypes to GO concepts (biological processes, molecular functions, cellular components) based on their correlated groups of Pfam families. Using the Šidák adjustment for a posteriori comparisons, there are 309 significant positive correlations with 33 distinct phenotypes within 166 distinct GO terms and 197 *anti*-correlations of 13 unique phenotypes within 142 distinct GO terms. We also provide two evaluations of the GO–phenotype predictions: (i) a random sample of 50 predictions were manually revised and showed a precision of 72% (95% confidence interval: 60%–82%; Table S6), and (ii) a manual evaluation of two phenotypes: Gram-negative and motility. Table 4 shows the GO concepts mapped to the Gram-negative phenotype. Lipopolysaccharide biosynthesis (GO:0009103) and lipid A biosynthesis (GO:0009245) are the top-ranking GO concepts mapped in the biological process branch of GO, while cell (GO:0005623), cell envelope (GO:0030313), and periplasmic space (sensu Gram-negative bacteria) (GO:0030288) are the top-ranking concepts mapped in the cellular component of GO (there are no mappings to the molecular functions of GO). In contrast to the phenotype-pathway mapping, phenotype-GO mapping provides characterizations of phenotypes using different aspects of GO. Though the mappings of phenotypes to pathways and GO concepts were conducted through differently classified gene families (COGs or Pfam), the results are strikingly comparable.

By applying a similar mapping to the motility phenotype (Tables 5 and 6), we identified four pathways and 27 GO concepts that are closely correlated with bacterial motility. The three pathways are 1) flagellar assembly, 2) type III secretion system, and 3) bacterial chemotaxis. Bacterial flagellar assembly and chemotaxis pathways are well-known to be important for bacterial motility [37,38], functioning together to guide bacteria's direction of movement. The type III secretion system is known to share many protein structure similarities with the flagellar assembly system in structure, function, and gene sequence [39,40]. Consequently, it is also shown to be significantly correlated with bacterial motility.

These case studies demonstrate that our high-throughput automated method for mapping phenotypes to pathways and GO concepts can faithfully recapitulate known knowledge. In addition, the method has the potential to predict new correlations between phenotypes and biological systems represented in GO as shown in the complete result datasets at http://phenos.bsd.uchicago.edu/prok__phenotype. While previous correlations studies had been completed on only four phenotypes [5,6], we present an additional 38 phenotype-to-GO correlations. We propose that this method potentially enables a systems-biology approach to analyze genomic datasets by providing a systematic view of the molecular mechanisms beneath phenotypes across different classifications of genes (protein families, pathways, molecular functions, and biological processes). In future studies, we intend to further explore the meaning of directionality of correlations between molecular mechanisms and phenotypes. Indeed, three types of significant correlations can be observed using the hypergeometric distribution: either the observed molecular mechanism is (i) disproportionably associated to a phenotype, or (ii) vice versa, or (ii) both are disproportionably associated to one another.

Phenomic Analysis and Visualization of Combined Genomic Information across Multiple Biological Scales

The results described above systematically provide significant correlations between classes of genes (protein families, pathways, molecular function, and biological processes) and prokaryotic phenotypes. To investigate how information from these classes of genes interacts together on groups of phenotypes, we conducted a cross-phenotype comparison using their correlations to genetic contents. This analysis is anchored on our previously described correlations between prokaryotic phenotypes and Pfam families. All the phenotypes were clustered using a hierarchical average-linkage method based on their correlation scores with Pfam families. Figure 4 shows a 2-D hierarchical clustering of both phenotypes and Pfam families, with green indicating correlation and red indicating *anti*-correlation. To our knowledge, this is the first large-scale cross-phenotype analysis of prokaryotic genomes. We will refer to it as a phenomic analysis, where phenotypes are compared based on their underlying genetic information. Our manual evaluation of two of the largest phenotypic clusters confirmed the results of

Table 4. GO Concepts That Are Significantly Correlated with the Gram-Negative Phenotype

Microbial Phenotype	GO ID	GO Description	GO Type	Šidák-Adjusted *p*-Value
Gram-negative	GO:0008653	Lipopolysaccharide metabolism	P	6.5E-04
	GO:0009103	Lipopolysaccharide biosynthesis	P	6.5E-04
	GO:0009245	Lipid A biosynthesis	P	1.5E-02
	GO:0046493	Lipid A metabolism	P	1.5E-02
	GO:0000271	Polysaccharide biosynthesis	P	8.1E-02
	GO:0043284	Biopolymer biosynthesis	P	8.1E-02
	GO:0008610	Lipid biosynthesis	P	1.0E-01
	GO:0005623	Cell	C	1.4E-02
	GO:0030313	Cell envelope	C	1.7E-02
	GO:0030288	Periplasmic space (sensu Gram-negative bacteria)	C	2.9E-02
	GO:0042597	Periplasmic space	C	6.0E-02

C, cellular component; P, biological process.
doi:10.1371/journal.pcbi.0020159.t004

Table 5. KEGG Pathways and COGs Concepts That Are Significantly Correlated with Bacterial Motility

Microbial Phenotype	Phenotype–KEGG Pathway Correlation		Phenotype–COGs Correlation		
	KEGG Pathway (KEGG ID)	*p*-Value	Mapped COGs	Šidák-Adjusted *p*-Value	Simulation-Adjusted *p*-Value
Motility	Flagellar assembly (ot02040)	2.5E-27	COG1516	1.4E-05	1.2E-07
			COG1345	1.4E-05	1.2E-07
			COG1815	1.2E-04	1.0E-06
			COG1843	1.2E-04	1.0E-06
			COG1766	1.2E-04	1.0E-06
			COG1677	1.2E-04	1.0E-06
			COG1256	1.2E-04	1.0E-06
			COG1684	1.2E-04	1.0E-06
			COG1987	1.2E-04	1.0E-06
			COG1360	1.2E-04	1.0E-06
			COG1344	1.2E-04	1.0E-06
			COG1868	1.2E-04	1.0E-06
			COG1558	1.2E-04	1.0E-06
			COG1291	1.2E-04	1.0E-06
			COG1580	2.6E-04	2.3E-06
			COG1749	2.0E-03	1.8E-05
			COG1377	3.4E-03	3.0E-05
			COG1157	3.4E-03	3.0E-05
			COG1536	3.4E-03	3.0E-05
			COG1338	3.4E-03	3.0E-05
			COG1886	3.4E-03	3.0E-05
			COG1298	3.4E-03	3.0E-05
			COG1419	9.2E-03	8.2E-05
			COG1317	1.1E-02	9.5E-05
			COG2882	3.4E-01	3.7E-03
			COG1706	4.0E-01	4.5E-03
			COG1191	4.0E-01	4.6E-03
			COG2747	7.9E-01	1.4E-02
			COG3144	7.9E-01	1.4E-02
			COG2063	8.7E-01	1.8E-02
			COG1261	8.7E-01	1.8E-02
	Type III secretion system (ot03070)	1.0E-07	COG1766	1.2E-04	1.0E-06
			COG1684	1.2E-04	1.0E-06
			COG1987	1.2E-04	1.0E-06
			COG1377	3.4E-03	3.0E-05
			COG1338	3.4E-03	3.0E-05
			COG1886	3.4E-03	3.0E-05
			COG1298	3.4E-03	3.0E-05
			COG1157	3.4E-03	3.0E-05
	Bacterial chemotaxis (ot02030)	1.4E-04	COG0840	1.4E-05	1.2E-07
			COG0643	1.4E-05	1.2E-07
			COG0835	1.4E-05	1.2E-07
			COG1352	2.6E-03	2.3E-05
			COG2201	1.7E-02	1.6E-04
			COG0784	4.6E-02	4.2E-04

KEGG pathways and COGs families that are the motility phenotype are listed with their *p*-values, respectively.
doi:10.1371/journal.pcbi.0020159.t005

this automated clustering, showing that biologically relevant phenotypes were generally grouped together. For example, the following phenotypic laboratory tests, Bacillus or cocco-bacillus, Growth on MacConkey agar, Catalase, Gram-negative, and Colistin-Polymyxin susceptible, are clustered together (highlighted in the red boxes in Figure 4). Within this cluster, the two phenotypes that have the shortest distance to the Gram-negative phenotype, Colistin-Polymyxin susceptible and Growth on MacConkey agar, are known to be closely related to the Gram-negative bacteria. Colistin–Polymyxin is an antibiotic specifically for Gram-negative bacteria [41], and the MacConkey agar test inhibits the growth of Gram-positive bacteria [42]. For the remaining two phenotypes within this cluster (Bacillus or coccobacillus, and Catalase), we were not able to find consistent associations with Gram-negative bacteria from the PubMED database. Gram-positive and Gram-negative prokaryotes are known to have baccillus or cocco-bacillus morphologies, thus the previous correlation could be a bias likely attributable to a disproportionate number of gram-negative species with bacillus morphologies in our dataset. In future studies, we intend to verify whether the same conclusion is generalizable to other bacterial species, and to explore the molecular underpinnings of these relations.

In the second cluster (highlighted in the blue boxes in Figure 4), the following phenotypes were clustered closely:

Table 6. GO Concepts That Are Significantly Correlated with Bacterial Motility

Microbial Phenotype	GO ID	GO Description	GO Type	Šidák-Adjusted *p*-Value
Motility	GO:0006928	Cell motility	P	1.4E-12
	GO:0007610	Behavior	P	1.4E-12
	GO:0007626	Locomotory behavior	P	1.4E-12
	GO:0040011	Locomotion	P	1.4E-12
	GO:0001539	Ciliary or flagellar motility	P	2.0E-09
	GO:0006935	Chemotaxis	P	2.0E-09
	GO:0042330	Taxis	P	2.0E-09
	GO:0009628	Response to abiotic stimulus	P	2.8E-06
	GO:0009605	Response to external stimulus	P	3.2E-05
	GO:0042221	Response to chemical substance	P	3.8E-05
	GO:0043064	Flagellum organization and biogenesis	P	4.7E-03
	GO:0030030	Cell projection organization and biogenesis	P	7.3E-03
	GO:0050896	Response to stimulus	P	1.6E-02
	GO:0007165	Signal transduction	P	1.6E-02
	GO:0000902	Cellular morphogenesis	P	2.1E-02
	GO:0009653	Morphogenesis	P	2.1E-02
	GO:0003774	Motor activity	F	1.5E-07
	GO:0004871	Signal transducer activity	F	2.1E-02
	GO:0004057	Arginyltransferase activity	F	2.6E-02
	GO:0004673	Protein histidine kinase activity	F	3.0E-02
	GO:0016775	Phosphotransferase activity, Nitrogenous group as acceptor	F	4.4E-02
	GO:0009288	Flagellum (sensu bacteria)	C	1.3E-17
	GO:0019861	Flagellum	C	2.1E-16
	GO:0042995	Cell projection	C	2.1E-16
	GO:0005623	Cell	C	3.5E-04
	GO:0009425	Flagellar basal body (sensu bacteria)	C	9.1E-04
	GO:0009347	Aspartate carbamoyltransferase complex	C	5.9E-03

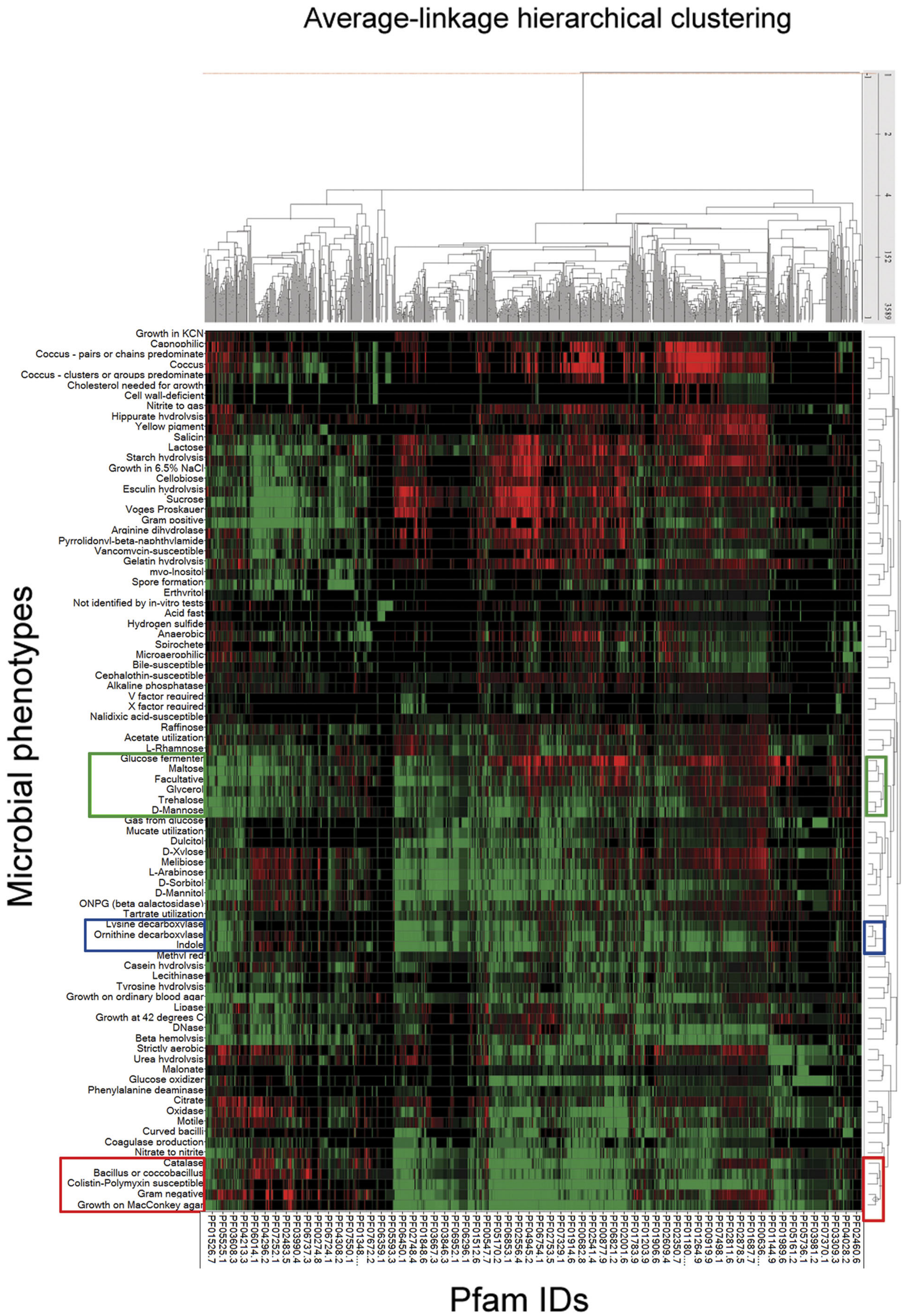
C, cellular component; F, molecular function; P, biological process.
doi:10.1371/journal.pcbi.0020159.t006

Lysine decarboxylase, Ornithine decarboxylase, and Indole. It is known that ornithine and indole are both involved in amino acid metabolism pathways; ornithine is a derivative of glutamate, and indole is the precursor of tryptophan [43]. Moreover, a protein has been identified in *Selenomonas ruminantium* that was shown to display the decarboxylating functions of both lysine and ornithine [44]. This is likely because the two functions are essential in this species, thus facilitating such evolution.

The third cluster of phenotypes within the green boxes contains six phenotypes related to the catabolism of carbohydrates clustered in the following order: Glucose fermenter (fermentation in a glucose medium), Maltose (production of acid in a medium containing maltose), Facultative anaerobic, Glycerol (production of acid in a medium containing glycerol), Trehalose (production of acid in a medium containing trehalose), and D-mannose (production of acid in a medium containing D-mannose). Every one of these phenotypes is also related to glycolysis [43]. We illustrated this cluster of phenotypes with their significantly correlated Pfam families, GO concepts, and KEGG pathways in detail (shown as a multiscale network in the Figure 5). To constrain the network of *cross*-scale relationships to the most relevant ones, the criteria for displaying a molecular class were the following: 1) GO terms significantly correlated with at least four phenotypes in the cluster, 2) a KEGG pathway with significant correlations to three phenotypes, and 3) Pfam significantly correlated with at least two phenotypes in the cluster (with the exception of one uncharacterized Pfam that has only one link to Glycerol, to illustrate the use of the integrated view for possible predictions). The *cross*-scale relationships between Pfam and GO terms (Figure 5, blue lines) were retrieved from public databases (discussed in Materials and Methods). Using these visualization criteria, we observe that this phenotypic cluster is particularly networked together, as many phenotypes share common KEGG pathways, GO concepts, and Pfam families based on our previous analyses. For example, facultative anaerobic bacteria with ability to metabolize D-mannose share one common KEGG pathway, phosphotransferase system pathway (PTS) and two GO concepts, phosphoenolpyruvate-dependent sugar phosphotransferase system, and sugar porter activity. In addition,

Figure 4. 2-D Hierarchical Clustering of Bacterial Phenotypes and Protein Families

Phenotypes are on the *x*-axis and Pfam families are on the *y*-axis. Correlation is represented in green, and red represents *anti*-correlation. The three clusters of laboratory tests that are discussed in the paper are highlighted (cluster 1 in a red box; cluster 2 in blue, and cluster 3 in green). We applied continuous coloring representing uncorrected *p*-values from 0 to 10^{-4} (red for *anti*-correlations with the value of –log(*p*-value) for color intensity, and green for correlations with the value of log(*p*-value) for color intensity) for displaying purposes. For details on the hierarchical clustering, see Equation 4 in Materials and Methods.
doi:10.1371/journal.pcbi.0020159.g004


Average-linkage hierarchical clustering
Microbial phenotypes
Growth in KCN
Capnophilic
Coccus - pairs or chains predominate
Coccus
Coccus - clusters or groups predominate
Cholesterol needed for growth
Cell wall-deficient
Nitrite to gas
Hippurate hydrolysis
Yellow pigment
Salicin
Lactose
Starch hydrolysis
Growth in 6.5% NaCl
Cellobiose
Esculin hydrolysis
Sucrose
Voges Proskauer
Gram positive
Arginine dihydrolase
Pyrrolidonyl-beta-naphthylamide
Vancomycin-susceptible
Gelatin hydrolysis
myo-Inositol
Spore formation
Erthyritol
Not identified by in-vitro tests
Acid fast
Hydrogen sulfide
Anaerobic
Spirochete
Microaerophilic
Bile-susceptible
Cephalothin-susceptible
Alkaline phosphatase
V factor required
X factor required
Nalidixic acid-susceptible
Raffinose
Acetate utilization
L-Rhamnose
Glucose fermenter
Maltose
Facultative
Glycerol
Trehalose
D-Mannose
Gas from glucose
Mucate utilization
Dulcitol
D-Xylose
Melibiose
L-Arabinose
D-Sorbitol
D-Mannitol
ONPG (beta galactosidase)
Tartrate utilization
Lysine decarboxylase
Ornithine decarboxylase
Indole
Methyl red
Casein hydrolysis
Lecithinase
Tyrosine hydrolysis
Growth on ordinary blood agar
Lipase
Growth at 42 degrees C
DNase
Beta hemolysis
Strictly aerobic
Urea hydrolysis
Malonate
Glucose oxidizer
Phenylalanine deaminase
Citrate
Oxidase
Motile
Curved bacilli
Coagulase production
Nitrate to nitrite
Catalase
Bacillus or coccobacillus
Colistin-Polymyxin susceptible
Gram negative
Growth on MacConkey agar
Pfam IDs
PF01526.7
PF05525.1
PF03608.3
PF04213.3
PF06014.1
PF04296.2
PF07252.1
PF02483.5
PF03990.4
PF06737.3
PF00274.8
PF06724.1
PF04308.2
PF07550.1
PF01348...
PF07672.2
PF06359.1
PF05593.3
PF06450.1
PF02748.4
PF01848.6
PF03667.3
PF03846.3
PF06952.1
PF06296.1
PF01512.6
PF00547.7
PF05170.2
PF06853.1
PF02556.4
PF04945.2
PF06754.1
PF02753.5
PF07329.1
PF01914.6
PF00682.8
PF02541.4
PF08877.9
PF06821.2
PF02001.6
PF01783.9
PF01203.9
PF01906.6
PF02609.4
PF02350.7
PF01180...
PF01264.9
PF07498.1
PF00919.9
PF02811.6
PF02878.5
PF01687.7
PF00636...
PF01144.9
PF01989.6
PF05161.2
PF05736.1
PF03981.2
PF07370.1
PF03309.3
PF04028.2
PF02460.6

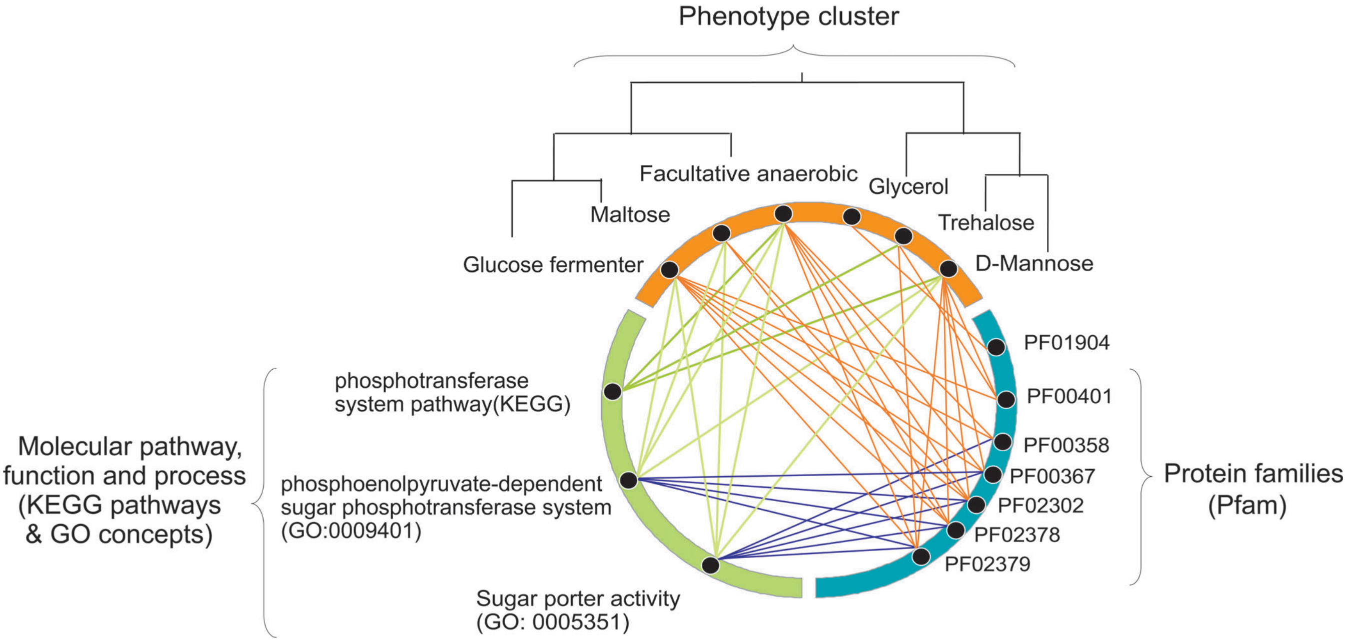


Figure 5. Scalar Network of Correlated Phenotypes, GO, Pathways, and Protein Families

As predicted by our study, six phenotypes, taken from a phenotypic cluster in Figure 4 (highlighted there in a green box) are shown highly connected with their significantly correlated biological scales: KEGG pathways, GO concepts, and Pfam families. Every relationship (orange and green lines between concept nodes) has been derived from our study with the exception of relationships between GO and Pfam (blue lines) that were taken from public databases.

D-mannose, acid production in a medium containing D-mannose; Facultative anaerobic, facultative anaerobic organism; Glucose fermenter, fermentation in a glucose medium; Glycerol, acid production in a medium containing glycerol; Maltose, acid production in a medium containing maltose; Trehalose, acid production in a medium containing trehalose; PF01904, unknown function; PF00401, ATP Synthase; PF00358, Phosphoenopyruvate-dependent sugar PTS (EIIA 1); PF00367, PTS (EIIB); PF02302, PTS Lactose/Cellobiose specific IIB subunit; PF02378, PTS (EIIC); PF02379, PTS system Fructose-specific IIB subunit.

doi:10.1371/journal.pcbi.0020159.g005

three molecular classes obviously related to the carbohydrate transport system in bacteria have been closely associated to the same phenotypic cluster: the KEGG pathway PTS, the cellular process phosphoenolpyruvate-dependent sugar phosphotransferase system PTS, and the molecular function sugar porter activity. Overall, five of the six phenotypes in this cluster share many common protein domain families (Pfam) intervening in the PTS system, as well as higher-level biological concepts, such as GO and KEGG pathways, strongly suggesting similarities or overlaps in their underlying molecular mechanism. In addition to the clustering of phenotypes, clustering of Pfam families based on their correlations to different phenotypes may also provide an informative view of the Pfam families, reflecting their activities in different phenotypes. Macroscopic phenotypes closely related to the catabolism of carbohydrates are thus also highly linked in this illustration with molecular classes closely related to the transport of carbohydrates. This visualization of *cross*-scale relationships, linked together across multiple biological scales and forming a multiscale nexus within the phenomic network, constitutes a proof of concept that the method could be applied to investigate less-understood regions of the network that we developed. We are in the process of further exploring this multiscale network in close collaboration with microbiologists. To our knowledge, this is the first phenomic study designed to predict and visualize *cross*-scale relationships between *clusters* of prokaryotic phenotypes (metaphenotypes) and their molecular mechanisms.

Conclusion

In this study, we developed a high throughput computational approach capable of automatically *integrating* clinical microbiological laboratory characterizations of bacterial phenotypes with various genomic databases spanning multiple scales of molecular biology (protein domains, pathways, molecular function, and cellular processes). To our knowledge, this is the first study demonstrating the feasibility of integrating a large quantity of prokaryotic phenotypes together with genomic datasets from various sources for large-scale data mining.

Furthermore, in contrast to previous *predictive* studies aimed at building large-scale phenotype–genotype networks, we have thoroughly elucidated systems properties involving multiple scales of molecular mechanisms underlying prokaryotic phenotypes. More specifically, we were able to achieve three objectives. First, we predicted and stratified previously unidentified and uncharacterized correlations (both positive and *anti*- correlations) between protein domain families (Pfam) and bacterial phenotypes using a comprehensive statistical data-mining and visualization method. Our evaluations attest that we faithfully recapitulated known biological knowledge between prokaryotic phenotypes and their molecular underpinnings, demonstrating the validity of our approach to integrate and analyze clinical and genomic datasets. Second, phenotypic information was correlated to additional biological scales such as cellular processes (GO), molecular functions (GO), and molecular pathways (KEGG). Third, the convergence of relationships in the phenomic visualization illustrates the nexus of specific biological systems shared within clusters of related phenotypes (metaphenotypes). This novel phenomic visualization analysis provides insight into the modular nature of common molecular mechanisms spanning multiple biological scales and reused by related phenotypes. We propose that this method, elucidating the relationship between classes of molecular mechanisms and their association with phenotypes or metaphenotypes, holds promise in facilitating a systems biology approach to genomic and biomedical research.

Materials and Methods

Datasets. In this study, we used the following six datasets. 1) Global Infectious Diseases and Epidemiology Network database (http://www.cyinfo.com) [15,16]. It contains an MKD, which contains 100 phenotypic microbiology laboratory results for more than 1,000 microorganism species (92 laboratory tests that contain test results in our 59 selected species were used in this study). The lack of data for some species laboratory tests in the MKD indicates that this knowledge is not useful in clinical bacteriology since MKD has been designed to satisfy the needs of clinical bacteriologists. It does not indicate that the knowledge does not exist elsewhere in the literature. We extracted the MKD data in December 2004. 2) Pfam dataset (release version 16, downloaded in April 2005) [17], of which the Pfam-A classifications were used in this study. 3) KEGG pathway data (KEGG Ontology file (KO), release version 31, downloaded in August 2004) [20,45]. 4) Gene Ontology Annotation (GOA) (downloaded in August 2005) [21,22]. 5) Pfam-GO mapping data, which is maintained by the Gene Ontology Consortium, (downloaded in August 2005 from http://www.geneontology.org/external2go/). 6) COGs data (downloaded in December 2004 from http://www.ncbi.nlm.nih.gov/COG/new/) [19].

Data integration. The laboratory results in the MKD are collected for bacterial species, which are primarily used for identifying bacterial strains for medical diagnostics. The MKD rarely has distinct annotations below the taxonomic level of the species according to the NCBI taxonomy. However, bacterial genomes are generally sequenced and annotated at the subspecies or strain levels according to the NCBI taxonomy. A complete list of fully sequenced prokaryotes, many of which have taxonomic annotation (NCBI Taxonomy ID) as no rank at present, was obtained from the NCBI (ftp://ftp.ncbi.nih.gov/genomes/Bacteria/). To map them, we first identified species taxons for the fully sequenced bacteria using the taxonomy tree from NCBI, and then mapped them to the bacteria species in the MKD through computational terminology mapping of text strings [46], followed by manual examination. As a result, we examined nearly 200 bacteria species that have complete genome sequences, and mapped 70 to the bacteria species in the MKD. In the case of a bacterial species having more than one strain with genome sequences, such as *B. anthracis,* all the strains were considered as one species and their genomic data were merged in a lossless way. Of the mapped species, we *merged* the data within each species group (refer to Figure 1 for the 89 fully sequenced genomes organized in species groups) that contained more than 100 Pfam families from the MKD and Pfam databases (altogether 59), and generated a table showing the presence and absence of Pfam families across species as shown in Figure 2. By design, we did not integrate partial genome sequences at the time the mapping started because of the possible bias it might also introduce. In addition, we used the dataset from our prior study [47], which was an integration of the MKD and COGs databases for mapping phenotypes to KEGG pathways.

Correlating the MKD's clinical diagnostic laboratory data of bacterial phenotypes with functional genomic data. To investigate whether there are correlations between the clinical diagnostic laboratory results (phenotypes) and the genomic data for bacteria species, we explored the functional classifications of genes. Based on the hypothesis that the existence of a family of genes (or the coexistence of families of genes) is responsible for a phenotype and leads to certain expressed phenotypes under controlled laboratory conditions, we calculated the probability of co-occurrence (by random chance) between a phenotypic laboratory result and presence of a certain cluster (family) of genes across species to uncover such correlations, according to the hypergeometric distribution shown and described below [5].

$$p(i>=m|N,M,n,m)=\sum_{i=m}^{n}\frac{\binom{M}{i}\binom{N-NM}{n-i}}{\binom{N}{n}} \tag{1}$$

The hypergeometric distribution takes into account the frequencies of species within a specific Pfam to a specific phenotype association and compares it with reference frequencies of species in the entire dataset for (i) the chosen Pfam and (ii) the chosen phenotype, independently of one another. It then calculates the probability (p-value) of obtaining these frequencies by chance assuming that the species are randomly distributed across phenotypes and Pfam. A p-value smaller than 5%, when corrected for multiple comparisons, indicates that the observed frequency of species sharing a specific Pfam and phenotypes are unlikely to have occurred by chance alone. In our study, there are 59 common species that have diagnostic laboratory results in the MKD and fully sequenced genomes. For instance, the MKD dataset contains 31 (n) positive species in the phenotypic class Gram-negative out of 56 species (N) for which there are some results for that laboratory data (there is some missing data in MKD because they are not relevant for microbiological characterizations); and the Lipid-A disaccharide synthetase family (Pfam ID: PF2684) has its member domains distributed in 25 (M) species. The number of common species between Gram-negative and PF2684 is 24 (m). The resulting p-value for calculating this co-occurrence distribution by random chance according to the above hypergeometric distribution expression is 1.2×10^{-8}.

The above-mentioned relationship could have two possible types of correlations: 1) a correlation, referring to a positive laboratory result correlated with the existence of a Pfam family; 2) an *anti*-correlation, referring to a positive lab result that correlates with the absence of a Pfam family. We believe that both correlation types could be equally important for inferring gene functions. To distinguish the two types of correlation, we used the mean value (μ) of hypergeometric distribution (shown below) as a reference.

$$\mu = n^{*}M/N \tag{2}$$

As illustrated in Figure 2, when m (Equation 1) is bigger than μ (Equation 2), the relationship is a correlation; on the other hand, if m is smaller than μ, it is an *anti*-correlation. The example above has a mean value of 11.4 (25*32/70), suggesting a correlation. However, if m in the above example had been equal to 2, the calculation would show an *anti*-correlation with a p-value of 2.5×10^{-11}.

To control for multiple comparisons, we applied two methods to identify and stratify significant correlations and *anti*-correlations: 1) the conservative Bonferroni-type method known as the Šidák single-step adjusted p-value for multiple comparisons [26], and 2) the calculation of error rates using a less conservative data-mining algorithm allowing finding correlation with p-value < 0.05.

Controlling for multiple comparison with a Bonferroni-type method. The Šidák adjustment for a posteriori comparisons, that was used to maintain an experiment-wide error rate of less than 5%, is calculated according to the following equation,

$$\alpha' = 1-(1-\alpha)^{k} \tag{3}$$

where α' and α represent the corrected and uncorrected p-values, respectively, and k represents the number of independent tests. However, since the laboratory dataset contains missing values for some species in different tests, applying the Šidák adjustment for multiple comparisons could be overly conservative or biased toward the laboratory tests with more data.

Controlling for multiple comparison with a simulation method. Therefore, to stratify our results with a less conservative method which can predict more correlations, albeit with a higher error rate, we also applied a simulation method to the datasets. Using established statistical resampling principles [48], we created random datasets for a control experiment by generating 1,000 random distributions for each combination of the laboratory results and Pfam families (keeping the total number of occurrences of each lab and Pfam constant in the datasets, while randomizing their distributions in the species—permutation resampling without replacement). For each random distribution, we then calculated the number of statistically correlated laboratory results and Pfam families from these random datasets using the previously described hypergeometric method, with different cutoffs (uncorrected p-value $<= 0.05$, 0.01, 0.005, 0.002, 0.001, 0.0005, and 0.0001). Rather than controlling for multiple comparisons with a statistical test, we used the statistically significant results from the random datasets to predict the number of false positive errors that we should expect in the real dataset when analyzed under the same conditions and subjected to uncorrected multiple comparisons. Since each of the 1,000 random datasets provides a slightly different interpretation using the hypergeometric statistic, we chose a threshold for the calculated hypergeometric statistics that would be observed as the worst case 99% of the time (i.e., 99% confidence). A distribution of the number of errors has been generated for each cutoff, and the numbers that are greater than 99% of the total numbers were selected as references for confidence levels.

Evaluating the results. A manual examination was conducted on the predicted results of correlated phenotypes and Pfam families using the two methods. For the Šidák corrected result, we examined the correlated Pfam families for all phenotypes and summarized the results for each phenotype (Table S2). To estimate the false positive rate, we randomly selected 50 predicted phenotype–Pfam correlations from the result of the simulation method, whose Pfam families have biological annotation (i.e., Pfam families annotated as domain with unknown function are not included in this evaluation). Correlations with literature supports were identified as correct predictions from the random set. The false negative rates were also estimated by evaluating a random selection of 50 phenotype–Pfam correlations from all possible combinations between phenotypes and Pfam excluding the significant correlations predicted by the simulated method.

Correlating MKD's laboratory data with KEGG's molecular pathways and GO concepts. In a previous study, we calculated correlations between COGs and phenotypes using the hypergeometric and Bonferroni-type methods [47]. In the current study, we also applied the previously described data-mining method to generate a less conservative estimate of phenotypes related to COGs (phenotype–COG dataset), which we have used as intermediary results to compare KEGG's pathways to phenotypes. We also applied the previously described hypergeometric function and Šidák adjustment (Equations 1 and 3) for a posteriori comparisons to identify significant correlations between phenotypes, and either KEGG's molecular pathways or phenotypes and GO concepts. To correlate phenotypes and KEGG molecular pathways, we integrated the correlation of COGs and pathway data from the KEGG ontology file and assigned the following numbers to the hypergeometric function: 1) the number of COGs families in the KEGG ontology file (N); 2) the number of correlated COG families for each microbial phenotype from the phenotype–COG dataset (n); 3) the number of unique COGs families in each pathway that are also used in this analysis (M); 4) the number of common COGs between 2) and 3) (m).

To further identify significant correlations between phenotypes and GO concepts, we used a GO term finder software [49] to correlate phenotypes with GO using the Pfam to GO mapping data from the Gene Ontology Consortium. The GO term finder, designed for correlating genes to GO, also exploits the hypergeometric distribution function for identification of significant correlations and provides Šidák-adjusted p-values. A set of common Pfam families between the two datasets (Pfam–phenotype and Pfam–GO) was retrieved. Relevant subsets of these two datasets were generated for this study, and subsequently used in calculating phenotypes and GO correlations. The availability of data resources at the time of this study limited our method. Though we first thought to map KEGG through Pfam families, we could not find reliable resources that provide a mapping between them. However, we found a good resource for mapping KEGG and COGs and therefore used it for the study as a convenient alternative.

Hierarchical clustering of Pfam families and phenotypes for phenomics analysis. We conducted hierarchical clustering using unweighted average linkage and Euclidean distance of all the phenotypes and Pfam families using normalized correlation p-values [50]. The Euclidean distance is defined as:

$$\sqrt{\left((p_1+q_1)^2+(p_2+q_2)^2+\ \ldots\ +(p_n+q_n)^2\right)}=\sqrt{\sum_{i=1}^{n}(p_i+qi)^2} \quad (4)$$

where P and Q represent series of p-values of two phenotypes. To normalize the p-values for display purposes, we used the absolute logarithmic value of the p-value, and assigned + for positive correlations and − for negative correlations. For example, a p-value of 1.0E-07 would be converted to $7 = -(\log(1.0E\text{-}07))$ for positive correlation, and −7 for negative correlation. Therefore, the correlations between Pfam families and phenotypes would be properly represented. We then used the Spotfire software [51] to cluster Pfam families and phenotypes based on the normalized data.

Supporting Information

Table S1. Comparison of Pfam and *E. coli* Genes Significantly Correlated to Flagellar-Mediated Motility

The top 30 *E. coli* genes identified by a previous study [5] that correlated with flagellar-mediated motility are compared with their corresponding Pfam families identified in this study. The significance of the correlations defined by the two studies, including uncorrected p-values, and p-values adjusted according to the Šidák and the resampling methods, are shown.

Found at doi:10.1371/journal.pcbi.0020159.st001 (85 KB DOC).

Table S2. Evaluation of Phenotypes with Their Significantly Correlated Pfam Families by the Šidák Adjustment Method

We manually evaluated 21 phenotypes that have significantly correlated Pfam families by the Šidák adjustment method. Descriptions of the phenotypes are provided with summary and references.

Found at doi:10.1371/journal.pcbi.0020159.st002 (298 KB DOC).

Table S3. Manual Evaluation of a Random Sample of Correlations and Anti-Correlations of Phenotypes and Pfam Families

Fifty positive correlations of phenotype–Pfam were randomly selected from the 3,711 significant correlations by the simulation method. Manual examination indicated that 15 of them have strong literature support (provided), suggesting that they are true positives.

Found at doi:10.1371/journal.pcbi.0020159.st003 (187 KB DOC).

Table S4. Manual Evaluation of Randomly Selected Correlations, Which Were Statistically Insignificant, to Estimate False Negative Rates

A random selection of 50 phenotype–Pfam correlations from all possible combinations between phenotypes and Pfam excluding the significant correlations predicted by the simulated method was evaluated to estimate false negative rate.

Found at doi:10.1371/journal.pcbi.0020159.st004 (112 KB DOC).

Table S5. Manual Evaluation of Every Statistically Significant Correlation and Anti-Correlation between Phenotypes and KEGG Pathways

Found at doi:10.1371/journal.pcbi.0020159.st005 (68 KB DOC).

Table S6. Manual Evaluation of 50 Randomly Selected Significant Correlations of Phenotype and GO

A random selection of 50 phenotype–GO significant correlations was evaluated.

Found at doi:10.1371/journal.pcbi.0020159.st006 (185 KB DOC).

Acknowledgments

We thank H. Rosie Xing and Tara Borlawsky for their critical comments.

Author contributions. YL, JL, MG, and YAL conceived and designed the experiments. YL, JL, and YAL performed the experiments. YL, JL, LS, and YAL analyzed the data. YL, JL, CSG, and YAL contributed reagents/materials/analysis tools. YL, LS, MG, and YAL wrote the paper.

Funding. We acknowledge the support of the following grants: US National Institutes of Health (NIH)/National Library of Medicine 1K22 LM008308 (Semantic Approaches to Phenotypic Database Analysis), and NIH/National Cancer Institute 1U54CA121852-01A1 (National Center for the Multiscale Analysis of Genomic and Cellular Networks (MAGNet), NIH/National Institute for Allergies and Infectious Diseases 5U54 AI057158–02, and the Ruth L. Kirschtein NIH Postdoctoral fellowship.

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

References

1. Bork P, Dandekar T, Diaz-Lazcoz Y, Eisenhaber F, Huynen M, et al. (1998) Predicting function: From genes to genomes and back. J Mol Biol 283: 707–725.
2. Su Z, Mao F, Wu H, Dam P, Chen X, et al. (2006) Understanding microbial genomic structures and applications to biological pathway inference. pp. 23–24.
3. Olman V, Peng H, Su Z, Xu Y (2004) Mapping of microbial pathways

through constrained mapping of orthologous genes. Proc IEEE Comput Syst Bioinform Conf: 363–370.
4. Arakawa K, Yamada Y, Shinoda K, Nakayama Y, Tomita M (2006) GEM System: Automatic prototyping of cell-wide metabolic pathway models from genomes. BMC Bioinformatics 7: 168.
5. Jim K, Parmar K, Singh M, Tavazoie S (2004) A cross-genomic approach for systematic mapping of phenotypic traits to genes. Genome Res 14: 109–115.
6. Makarova KS, Wolf YI, Koonin EV (2003) Potential genomic determinants of hyperthermophily. Trends Genet 19: 172–176.
7. Korbel JO, Doerks T, Jensen LJ, Perez-Iratxeta C, Kaczanowski S, et al. (2005) Systematic association of genes to phenotypes by genome and literature mining. PLoS Biol 3 (5): e134.
8. Overbeek R, Larsen N, Pusch GD, D'Souza M, Selkov E Jr, et al. (2000) WIT: Integrated system for high-throughput genome sequence analysis and metabolic reconstruction. Nucleic Acids Res 28: 123–125.
9. Haft DH, Selengut JD, Brinkac LM, Zafar N, White O (2005) Genome properties: A system for the investigation of prokaryotic genetic content for microbiology, genome annotation and comparative genomics. Bioinformatics 21: 293–306.
10. Eppig JT, Bult CJ, Kadin JA, Richardson JE, Blake JA, et al. (2005) The Mouse Genome Database (MGD): From genes to mice—A community resource for mouse biology. Nucleic Acids Res 33: D471–D475.
11. de la Cruz N, Bromberg S, Pasko D, Shimoyama M, Twigger S, et al. (2005) The Rat Genome Database (RGD): Developments towards a phenome database. Nucleic Acids Res 33: D485–D491.
12. Wang J, Williams RW, Manly KF (2003) WebQTL: Web-based complex trait analysis. Neuroinformatics 1: 299–308.
13. Butte AJ, Kohane IS (2006) Creation and implications of a phenome–genome network. Nat Biotechnol 24: 55–62.
14. Aerts S, Lambrechts D, Maity S, Van Loo P, Coessens B, et al. (2006) Gene prioritization through genomic data fusion. Nat Biotechnol 24: 537–544.
15. Edberg SC (2005) Global Infectious Diseases and Epidemiology Network (GIDEON): A World Wide Web–based program for diagnosis and informatics in infectious diseases. Clin Infect Dis 40: 123–126.
16. Berger SA (2005) GIDEON: A comprehensive Web-based resource for geographic medicine. Int J Health Geogr 4: 10.
17. Bateman A, Coin L, Durbin R, Finn RD, Hollich V, et al. (2004) The Pfam protein families database. Nucleic Acids Res 32: D138–D141.
18. Tatusov RL, Koonin EV, Lipman DJ (1997) A genomic perspective on protein families. Science 278: 631–637.
19. Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, et al. (2003) The COG database: An updated version includes eukaryotes. BMC Bioinformatics 4: 41.
20. Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, et al. (1999) KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res 27: 29–34.
21. Camon E, Magrane M, Barrell D, Binns D, Fleischmann W, et al. (2003) The Gene Ontology Annotation (GOA) project: Implementation of GO in SWISS-PROT, TrEMBL, and InterPro. Genome Res 13: 662–672.
22. Camon E, Barrell D, Lee V, Dimmer E, Apweiler R (2004) The Gene Ontology Annotation (GOA) Database—An integrated resource of GO annotations to the UniProt Knowledgebase. In Silico Biol 4: 5–6.
23. Peterson JD, Umayam LA, Dickinson T, Hickey EK, White O (2001) The Comprehensive Microbial Resource. Nucleic Acids Res 29: 123–125.
24. Pagel M (1999) Inferring the historical patterns of biological evolution. Nature 401: 877–884.
25. Kunin V, Goldovsky L, Darzentas N, Ouzounis CA (2005) The net of life: Reconstructing the microbial phylogenetic network. Genome Res 15: 954–959.
26. Sokal R, Rohlf F (1995) Biometry: The principles and practice of statistics in biological research. New York: Freeman W. H. pp. 236–240.
27. Storey JD, Tibshirani R (2003) Statistical significance for genomewide studies. Proc Natl Acad Sci U S A 100: 9440–9445.
28. Hubbard TJ, Ailey B, Brenner SE, Murzin AG, Chothia C (1999) SCOP: A Structural Classification of Proteins database. Nucleic Acids Res 27: 254–256.
29. Pearl FM, Bennett CF, Bray JE, Harrison AP, Martin N, et al. (2003) The CATH database: An extended protein family resource for structural and functional genomics. Nucleic Acids Res 31: 452–455.
30. Dietmann S, Park J, Notredame C, Heger A, Lappe M, et al. (2001) A fully automatic evolutionary classification of protein folds: Dali Domain Dictionary version 3. Nucleic Acids Res 29: 55–57.
31. Pouliot Y, Gao J, Su QJ, Liu GG, Ling XB (2001) DIAN: A novel algorithm for genome ontological classification. Genome Res 11: 1766–1779.
32. Sussman JL, Lin D, Jiang J, Manning NO, Prilusky J, et al. (1998) Protein Data Bank (PDB): Database of three-dimensional structural information of biological macromolecules. Acta Crystallogr D Biol Crystallogr 54: 1078–1084.
33. Skurnik M, Bengoechea JA (2003) The biosynthesis and biological role of lipopolysaccharide O-antigens of pathogenic Yersiniae. Carbohydr Res 338: 2521–2529.
34. Soballe B, Poole RK (1999) Microbial ubiquinones: Multiple roles in respiration, gene regulation and oxidative stress management. Microbiology 145 (Part 8): 1817–1830.
35. Kar PD, Riley M, Saier M, Paulsen IT, Paley SM, et al. (2000) The EcoCyc and MetaCyc databases. Nucleic Acids Res 28: 56–59.
36. Kar PD, Riley M, Paley SM, Pellegrini-Toole A (2002) The MetaCyc Database. Nucleic Acids Res 30: 59–61.
37. Alberts B, Johnson A, Lewis J, Raff M, Roberts K, et al. (2002) Molecular biology of the cell. London: Garland Science. pp. 966–968.
38. Alberts B, Johnson A, Lewis J, Raff M, Roberts K, et al. (2002) Molecular biology of the cell. London: Garland Science. pp. 890–892.
39. Galan JE, Ginocchio C, Costeas P (1992) Molecular and functional characterization of the *Salmonella* invasion gene invA: Homology of InvA to members of a new protein family. J Bacteriol 174: 4338–4349.
40. Blocker A, Komoriya K, Aizawa S (2003) Type III secretion systems and bacterial flagella: Insights into their function from structural similarities. Proc Natl Acad Sci U S A 100: 3027–3030.
41. Gum WC, Walsh JW (2005) Intrathecal colistin for treatment of highly resistant Pseudomonas ventriculitis. Case report and review of the literature. J Neurosurg 102: 915–917.
42. Saffer M, Lubianca Neto JF, Piltcher OB, Petrillo VF (1996) Chronic secretory otitis media: Negative bacteriology. Acta Otolaryngol 116: 836–839.
43. Berg J, Tymoczko J, Stryer L (2002) Biochemistry. New York: Freeman W. H. pp. 648–680.
44. Takatsuka Y, Tomita T, Kamio Y (1999) Identification of the amino acid residues conferring substrate specificity upon Selenomonas ruminantium lysine decarboxylase. Biosci Biotechnol Biochem 63: 1843–1846.
45. Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M (2004) The KEGG resource for deciphering the genome. Nucleic Acids Res 32: D277–D280.
46. Lussier YA, Li J (2004) Terminological mapping for high throughput comparative biology of phenotypes. Pac Sym Biocomput: 202–213.
47. Goh C, Liu Y, Li J, Lussier Y, Gerstein M (2006) Integration of curated databases to identify genotype–phenotype associations. BMC Genomics 7: e257.
48. Kantardzic M (2003) Data mining: Concepts, models, methods, and algorithms. New York: John Wiley. 360 p.
49. Boyle EI, Weng S, Gollub J, Jin H, Botstein D, et al. (2004) GO::TermFinder—Open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. Bioinformatics 20: 3710–3715.
50. Lussier YA, Sarkar IN, Cantor M (2002) An integrative model for in-silico clinical-genomics discovery science. Proc AMIA Symp: 469–473.
51. Ahlberg C (1996) Spotfire: An information exploration environment. SIGMOD Record (ACM Special Interest Group on Management of Data) 25: 25–29.