Molecular Systems Biology (2006) doi:10.1038/msb4100047
© 2006 EMBO and Nature Publishing Group All rights reserved 1744-4292/06
www.molecularsystemsbiology.com
Article number: 2006.0005

molecular systems biology

Ab initio genotype–phenotype association reveals intrinsic modularity in genetic networks

Noam Slonim[1,3], Olivier Elemento[2,3] and Saeed Tavazoie[2,*]

[1] Department of Physics, Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton, NJ, USA and [2] Department of Molecular Biology, Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton, NJ, USA
* Corresponding author. Department of Molecular Biology, Lewis-Sigler Institute for Integrative Genomics, Princeton University, Washington Street, Carl Icahn, Room 245, Princeton, NJ 08544, USA. Tel.: +1 609 258 0331; Fax: +1 609 258 3565; E-mail: tavazoie@genomics.princeton.edu
[3] These authors contributed equally to this work

Received 30.8.05; accepted 5.12.05

Microbial species express an astonishing diversity of phenotypic traits, behaviors, and metabolic capacities. However, our molecular understanding of these phenotypes is based almost entirely on studies in a handful of model organisms that together represent only a small fraction of this phenotypic diversity. Furthermore, many microbial species are not amenable to traditional laboratory analysis because of their exotic lifestyles and/or lack of suitable molecular genetic techniques. As an adjunct to experimental analysis, we have developed a computational information-theoretic framework that produces high-confidence gene–phenotype predictions using cross-species distributions of genes and phenotypes across 202 fully sequenced archaea and eubacteria. In addition to identifying the genetic basis of complex traits, our approach reveals the organization of these genes into generic preferentially co-inherited modules, many of which correspond directly to known enzymatic pathways, molecular complexes, signaling pathways, and molecular machines.
Molecular Systems Biology 31 January 2006; doi:10.1038/msb4100047
Subject Categories: computational methods; microbiology & pathogens
Keywords: comparative genomics; genotype–phenotype association; information theory; microbiology; modularity

Introduction

Since the time of Gregor Mendel, a central focus of biology has been to understand the hereditary basis of organismal traits and their variation. The field of genetics approached this problem by identifying heritable variation in a phenotype of interest within a *single species*. Using recombination, the genetic basis of this variation could be mapped to individual genes. This simple approach, coupled with biochemical and cell biological analysis of gene products and their interactions, is at the core of our modern molecular understanding of life (Alberts *et al*, 1994).

Today, one might pose the same genotype–phenotype question in a different way: can we understand the genetic basis of a trait by differential inheritance of genetic elements across *many species* that show variation in the expression of that trait? Several recent works have used the availability of complete microbial genomes, along with their phenotype annotations, to demonstrate the feasibility of such a program (Huynen *et al*, 1998; Levesque *et al*, 2003; Makarova *et al*, 2003; Jim *et al*, 2004; Korbel *et al*, 2005). In all these works, the basic output is essentially a list of genes that are predicted to be associated with a particular trait. However, the expression of a complex phenotype often involves the coordinated activity of multiple functional modules, such as signaling pathways or molecular complexes. Here, we propose an information-theoretic computational approach that uses complete genome sequences and their phenotypic annotations, to recover this rich structure.

We pose a simple question: are there generic *modules* that are preferentially inherited for the expression of a specific microbial trait? The common ancestry of extant species, and widespread horizontal gene transfer among prokaryotes (Lerat *et al*, 2005), would facilitate the sharing of such modules. In fact, one would expect that once optimized, such modules would be preferentially employed for expressing common phenotypic traits, and that the statistical signature of this differential coinheritance would be detectable across a large enough number of diverse species. Here, we have applied our approach to a set of 202 complete microbial genomes, focusing on diverse phenotypic characteristics such as behavior, cellular morphology, physiological capacity, cellular differentiation, and pathogenicity. In all these cases, our approach identifies the known genes involved, and reveals their organization into robust modules. Our observations support the notion that modularity in molecular networks is a native property of biological systems, and not an artifact of biases historically imposed by biologists (Hartwell *et al*, 1999).

© 2006 EMBO and Nature Publishing Group

Results

Recovering generic gene (GG) modules underlying observed microbial traits

Given the sequenced genomes of many species, a gene can be described by a vector of ones and zeros, indicating the presence or absence of homologs of this gene in each of the available genomes. This representation, first suggested in (Pellegrini *et al*, 1999), is often referred to as the gene phylogenetic profile. Certain phenotypes like motility in bacteria can be described by similar vectors, indicating the presence (e.g., motile) or absence (non-motile) of this phenotype across the same set of species. The underlying assumption in our study is that genes whose phylogenetic profiles closely correlate with the phenotypic descriptions are likely to be involved in some aspects of the corresponding phenotypes. For example, it is natural to expect that the phylogenetic profile of a gene encoding a flagellar apparatus component will be positively correlated with the pattern of motility/non-motility across many bacterial genomes.

Our approach can be summarized as follows (Figure 1). We created phylogenetic profiles for all the ~600 000 genes in 202 fully sequenced prokaryotic genomes. Next, we estimated the statistical correlation between each gene and a given phenotype through their mutual information (Cover and Thomas, 1991). We then collected only genes with a statistically significant correlation with the phenotype, from all the organisms having the phenotype, and merged them into a single cross-genome list. Within this list, we identified groups of homologs, termed here phenotype generic genes (GGs) that are predicted to constitute the genetic basis of the examined phenotype. Finally, we found robust modules among these phenotype GGs with phylogenetic profiles that are highly informative about each other. Specifically, each module corresponds to a set of GGs that are consistently placed by a clustering algorithm in the same cluster, in numerous runs with different initial conditions (see Materials and methods). In other words, these are *generic modules*, required in a relatively conserved form by most microbes expressing the phenotype, while typically unnecessary for the remaining organisms. As a concrete example, we first present our results for a well-studied bacterial behavior, that of motility. Then, we detail our results for phenotypes like Gram-negativity, endospore formation, oxygen respiration, and intracellular pathogenicity. The complete data, results, and relevant software are available at our Web site (http://tavazoielab.princeton.edu/genphen/).

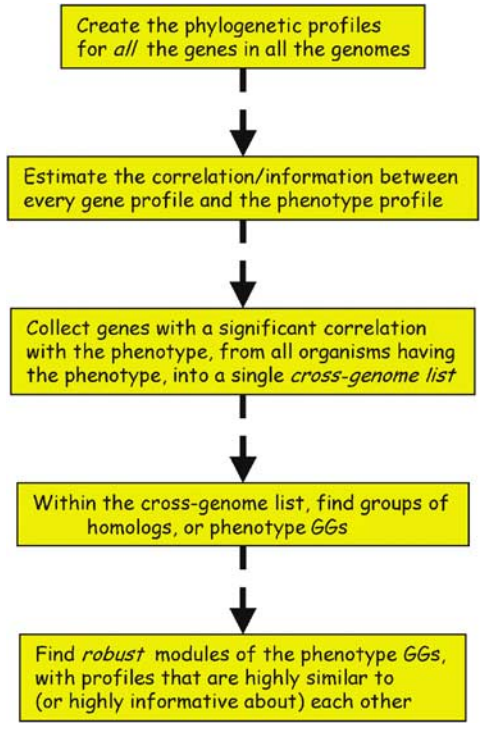


Figure 1 A schematic overview of the approach.

Behavior: motility

Our cross-genome list for motility consists of 3833 genes, collected from 92 motile eubacteria and archaea, based on their significant positive correlation with this phenotype. These genes correspond to only 75 groups of homologs, or motility GGs. These 75 motility GGs were then clustered based on their phylogenetic profiles, yielding 14 robust modules (see Materials and methods).

In Figure 2A, we present the phylogenetic profiles of five of these 14 modules, with the maximal average correlation with the motility phenotype. The rows correspond to the motility GGs, the columns to the different organisms in our data, and each entry indicates whether a particular GG is represented in the genome of a specific organism. Rows in Figure 2A are organized according to the obtained modules, and columns are grouped based on broad phylogenetic classifications. As seen in Figure 2A, the observed motility GGs induce a remarkably accurate dichotomy between motile versus non-motile organisms. Note that in a few cases, we detect many motility-related genes in the genomes of non-motile organisms (e.g., for *Burkholderia mallei*), in agreement with previous studies (Nierman *et al*, 2004).

Our method uncovers the main modules underlying microbial motility with impressive precision. The first two modules consists solely of genes encoding bacterial flagella components; the third module corresponds to the chemotaxis pathway (receptors and signal transduction); the next module consists of two GGs associated with the flagella outer pair of rings (L- and P-ring) that are present only in Gram-negative bacteria, and support the proximal rod through the outer membrane (OM). Finally, the fifth module includes σ^{54}, a known regulator of nitrogen metabolism in *Escherichia coli* that also regulates motility genes in other species (Jagannathan *et al*, 2001; Wolfe *et al*, 2004). The second motility GG in this module is a σ^{54}-dependent transcriptional regulator, suggesting its potential role in motility.

In Figure 2B, we further explore the relations between these five modules. Every entry in this matrix-figure indicates the probability of two GGs to be placed in the same cluster by the clustering algorithm (see Materials and methods). Evidently, the distinction between the two main flagellar modules is rather weak, and is likely due to the fact that the second module is less dominant in α-proteobacteria; however, the

© 2006 EMBO and Nature Publishing Group

distinction between the other modules is extremely sharp. For example, a chemotaxis GG and a flagellar GG almost never end up in the same cluster. Supplementary Figures S1 and S2 present a similar analysis for all 14 modules obtained by our approach.

To further validate our results, in Figure 3A, we illustrate the *E. coli* chemotactic pathway and flagellar apparatus (Kanehisa and Goto, 2000); our approach recovers most of the genes in the *E. coli* motility system, and moreover partitions these genes into biologically meaningful modules. The genes that are not detected by our procedure correspond to two opposite scenarios: some genes are too specific, while others are too abundant (Figure 3B). For example, *fliO* is too specific, as we detect its variants only in *E. coli* and seven other closely related species. On the other hand, *fliI* is too abundant, as we find *fliI* variants in motile as well as non-motile microbes (Dreyfus *et al*, 1993). Interestingly, all three scenarios can occur within the same operon, as *fliP* (motility GG-5), *fliO*, and *fliI*, all reside next to each other in the *E. coli* genome. In addition, the *E. coli* flagella-related chaperones (FliT, FliJ, FlgN) and transcription factors (FlhC, FlhD, FlgM) are all too specific, present only in some of the γ/β-proteobacteria in our data. Within the chemotaxis system, *cheY* is too abundant while *cheZ* is too specific. Indeed, it was previously suggested that in the absence of CheZ, the CheV protein (motility GG-38) could fulfill a similar function (Pittman *et al*, 2001).

Morphology: Gram-negativity

Gram-negative bacteria are in general characterized by the presence of an additional membrane layer, the OM that serves as a permeability barrier to prevent the entry of toxic compounds while allowing the influx of nutrient molecules (Nikaido, 2003). The biogenesis of the OM is only partially understood, and methods to probe the assembly process are only starting to emerge (Ruiz *et al*, 2005; Wu *et al*, 2005). Our approach provides an appealing alternative for gaining novel insights regarding this phenotype.

Our cross-genome list for Gram-negativity consists of 4678 genes collected from 105 Gram-negative bacteria based on their significant correlation with this phenotype. These genes correspond to 117 groups of homologs, or Gram-negative GGs. Cluster analysis of these 117 GGs yields 19 robust modules (Supplementary Figures S3 and S4), four of which are presented in Figure 4A. As for the motility phenotype, the Gram-negative GGs induce a pronounced dichotomy between Gram-negative and Gram-positive bacteria.

The first two modules in Figure 4A have the maximum average correlation with the Gram-negative phenotype, and are strongly related to each other (Supplementary Figure S4). These two modules are present in almost all Gram-negative bacteria in our data, with the notable exception of *Mycoplasma* genomes; indeed, *Mycoplasma* stain Gram-negative but lack a cell wall and are therefore expected to have special genetic characteristic with respect to this phenotype (Hegermann *et al*, 2002). In addition, both modules are absent from a few intracellular proteobacteria, such as *Buchnera*, consistent with previous reports that detected very few genes encoding lipoproteins and OM proteins in these species (Shigenobu *et al*, 2000). The first module encapsulates most of the lipid-A biosynthesis pathway. Lipid-A is the hydrophobic anchor of lipopolysaccharide (LPS), which is present in high abundance in the OM of Gram-negative bacteria. As shown in Figure 4B, this module captures many of the components of the lipid-A biosynthesis pathway in *E. coli*. The only genes that are not detected by our analysis are *lpxH* and *kdsA*: the first is too specific, found mainly in γ/β-proteobacteria, while the latter is too abundant, that is, a variant of this gene is present in many Gram-positive bacteria (Figure 4C). Another member in this first module is GG-1, which corresponds to the *yaeT* gene from *E. coli*. This gene was recently found to be involved in OM assembly (Wu *et al*, 2005); moreover, the homolog of yaeT in *Neisseria meningitides*, also in GG-1, was found to be required for LPS and phospholipid transport to the OM (Genevrois *et al*, 2003). This further reinforces our prediction that GG-1 genes are functionally related to the lipid-A biosynthesis pathway in many organisms.

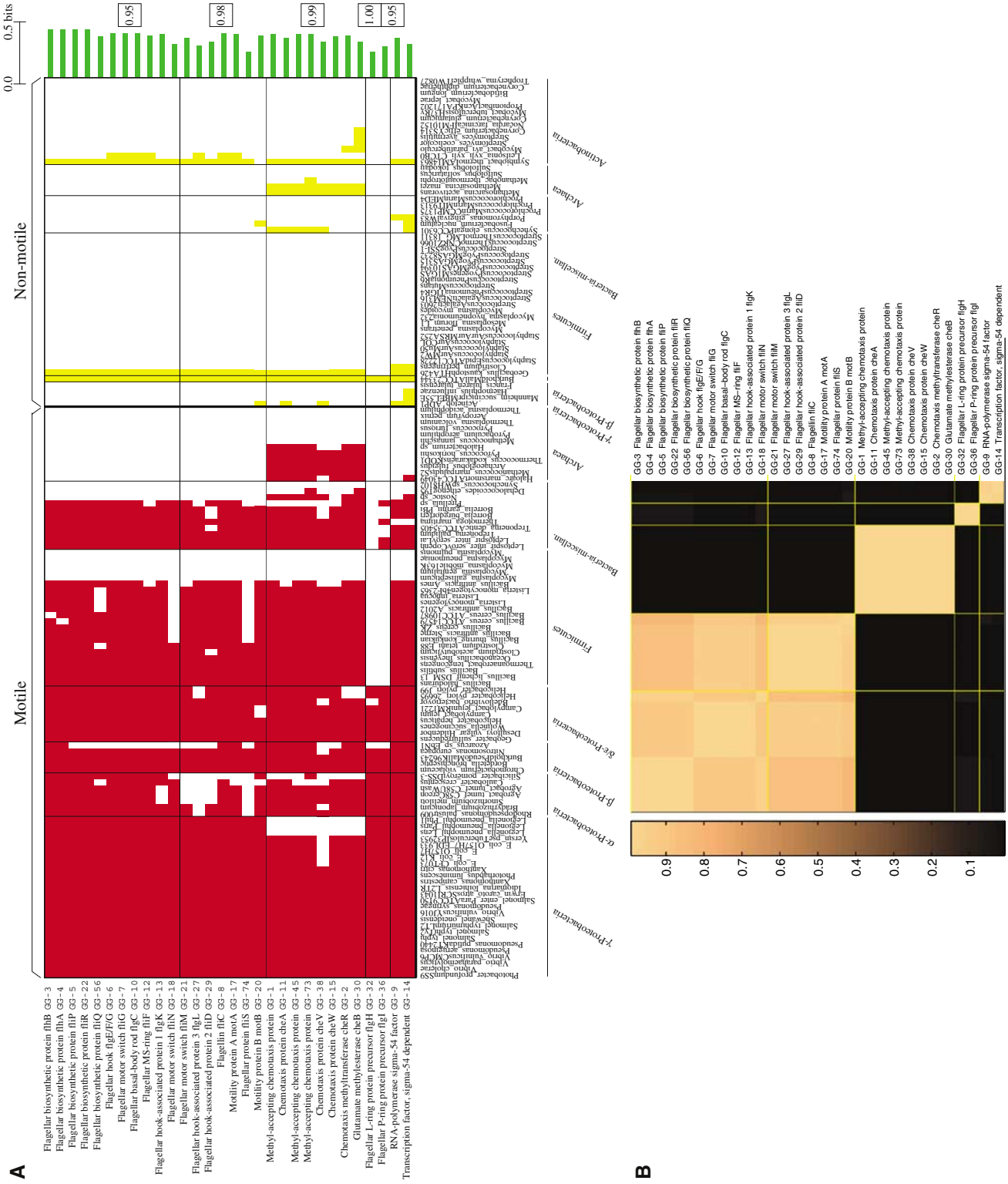
The second module includes several components (GG-9, GG-10, GG-12) associated with a single—and apparently quite generic—system involved in stabilizing the OM organization in an energy-dependent manner (Nikaido, 2003). The third module captures several members of the type I export pathway of proteins across the OM, for example, GG-20, GG-53, GG-96 and GG-111 (Nikaido, 2003). In the fourth module, we find mainly GGs that are glutaredoxin and glutathione related. Glutaredoxins are a family of proteins that catalyze a variety of redox reactions, particularly the reduction of protein disulfides. They are reduced by glutathione that is found primarily in Gram-negative bacteria (Copley and Dhillon, 2002). Both glutaredoxin and glutathione are involved in the resistance of Gram-negative cells to arsenic, and possibly also in the cellular response to oxidative stress (Berardi and Bushweller, 1999). Interestingly, this entire module is typically absent from δ/ε-proteobacteria.

Physiology: oxygen respiration

Oxygen respiration is a fundamentally important bioenergetic process in bacteria. Microorganisms respond differently to oxygen; for strict aerobes, it is essential, while for strict anaerobes, it is toxic; facultative microbes can grow either in the presence or absence of oxygen. Below, we discuss the application of our method to these three phenotypic classes.

Our cross-genome list for the aerobic phenotype consists of 2828 genes collected from 71 strictly aerobic organisms; these 2828 genes correspond to 130 groups of homologs, or aerobic GGs, that were clustered as before (Supplementary Figures S5 and S6). Two of the resulting modules are presented in the upper part of Figure 5. The first module captures almost perfectly the NADH dehydrogenase I complex, a common component of the electron transport chain. Specifically, among the 14 *E. coli* genes that participate in this complex, only two are missing from this module; *nuoG* is too abundant, while *nuoJ* is in a different module. Unlike previously considered phenotypes, the members of this module, while highly enriched in aerobic organisms, also show significant presence in the facultative and anaerobic species. On the other hand, the second module includes several GGs associated with the cytochrome *C*-oxidase complex, which are typically absent

© 2006 EMBO and Nature Publishing Group



© 2006 EMBO and Nature Publishing Group

from anaerobic species, since this complex carries out the final step in O_2 respiration.

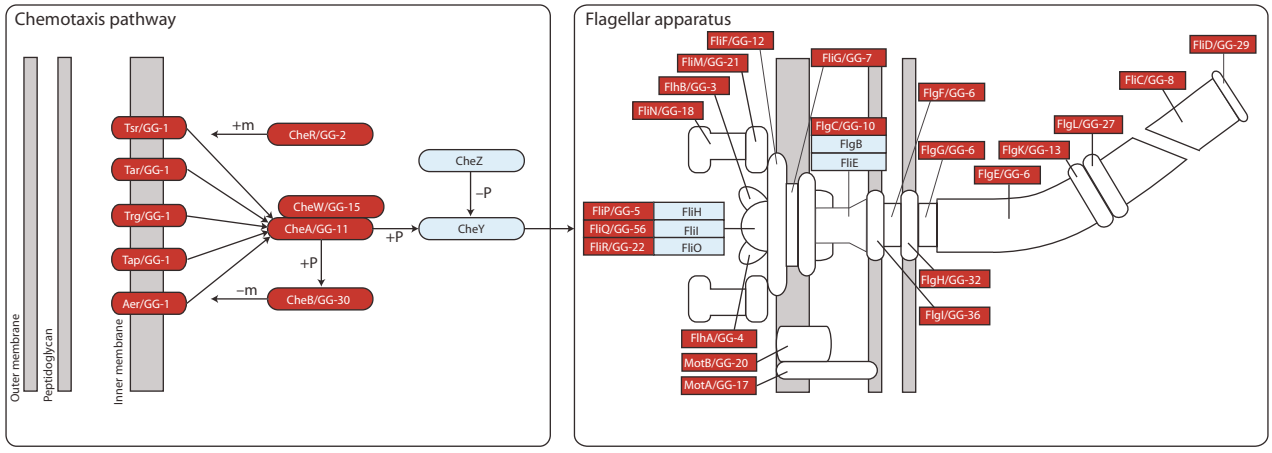
Similar analysis of the anaerobic phenotype yields 20 modules (Supplementary Figures S7 and S8), three of which are presented in the lower part of Figure 5. The first includes four GGs, all associated with ferrodoxin oxidoreductase activity; the second consists of three subunits of V-type sodium ATP synthase; and the third includes two iron-sulfur flavoprotein GGs, which are present almost exclusively in strict anaerobes.

The analysis of the facultative phenotype yields 22 modules (Supplementary Figures S9 and S10), two of which—with a perfect average joint-assignment probability of 1.0—are presented in the middle part of Figure 5. The first module corresponds to specific components in a phosphotransferase system (PTS), involved in uptake and phosphorylation of carbohydrates cellobiose and lactose. The second module corresponds to a different complex in the PTS system, involved in the uptake of other carbohydrates like galactosamin, mannose, and fructose. Several other modules of facultative GGs are also enriched in PTS system components (Supplementary Figure S9). The lack of any known mechanistic link between PTS-mediated carbohydrate transport and the examined facultative phenotype may reflect an indirect relationship. For example, it may correspond to the observed tendency of facultative bacteria in our data to inhabit carbohydrate-rich environments within multicellular hosts.

Cellular differentiation: endospore formation

Some bacteria are capable of producing endospores—resting structures that are formed by an unusual asymmetric cell division, followed by engulfment of the smaller cell (the prespore) by the mother cell (see, Errington, 2003, for a review). Endospores are the most resistant biological structures known; they can survive high temperatures, radiation, and chemical solvents, and can remain dormant for many years (Vreeland *et al*, 2000). Only 17 bacteria in our data were annotated as capable of forming endospores, most of them from the genus *Bacillus*. Nevertheless, our analysis reveals

A



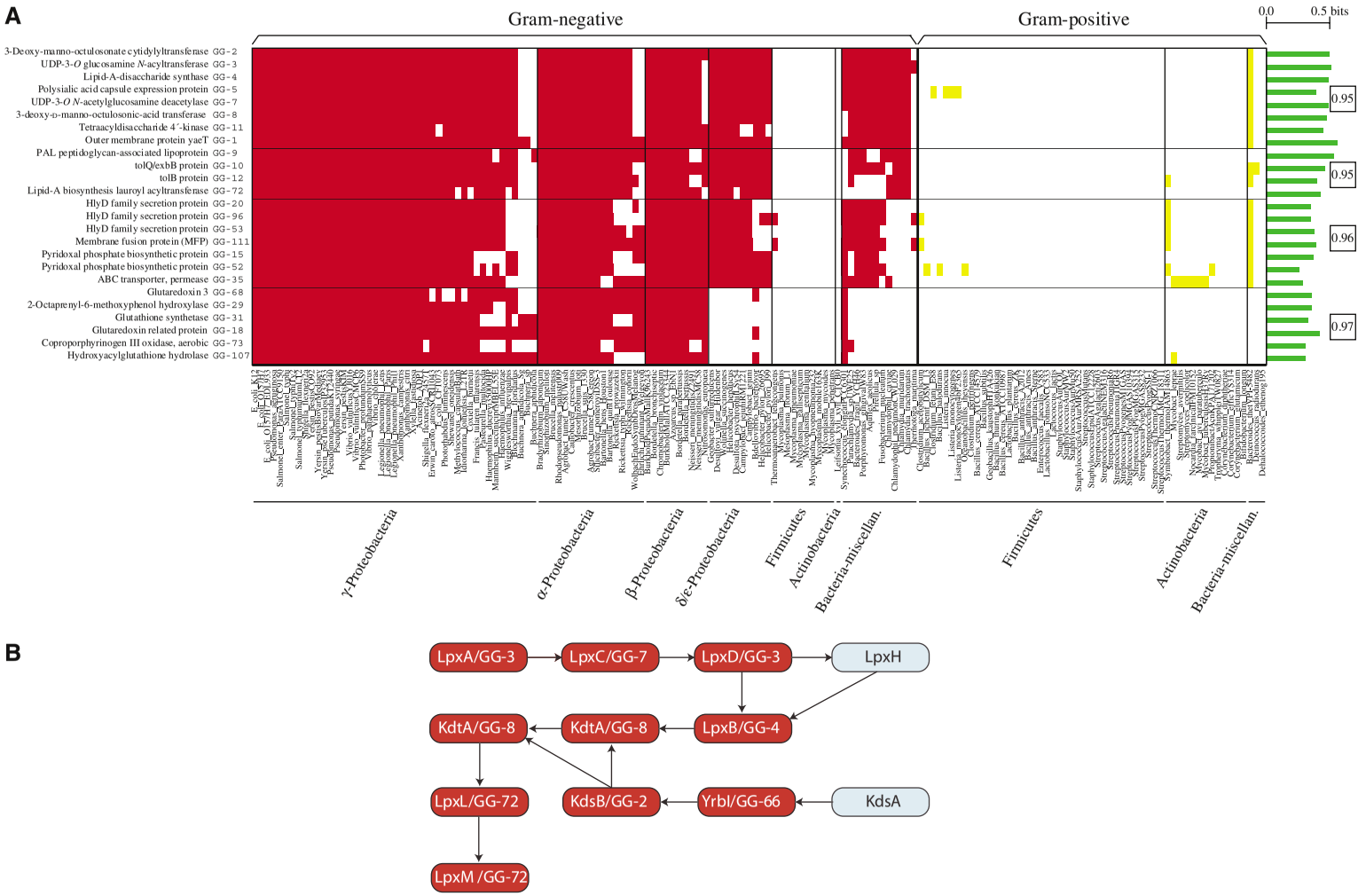
B

Name	GG	p− g−	p+ g−	p− g+	p+ g+	MI	Function
FliP	5	50	22	3	71	0.36	Protein export
FliE	--	51	70	2	23	0.05	Rod MS-ring junction
FliH	--	52	76	1	17	0.04	Protein export
FliO	--	53	85	0	8	0.04	Protein export
FliI	--	0	0	53	93	0.00	Protein export ATPase
FlgC	10	49	22	4	71	0.31	Proximal rod
FlgB	--	51	54	2	39	0.11	Proximal rod
CheA	11	46	19	7	74	0.30	Adds P to CheB/CheY
CheZ	--	52	68	1	25	0.08	Removes P from CheY-P
CheY	--	18	22	35	71	0.00	Promotes CW rotation

Figure 3 Motility GGs in the *E. coli* genome. (**A**) Depiction of the chemotaxis pathway and flagellar apparatus in *E. coli*. Genes detected by our approach are highlighted in red. (**B**) Count matrices of known motility genes in *E. coli*, not detected by our approach; red rows correspond to genes that were detected by our analysis, and are shown here for comparison; the third 'p− g−' column indicates the number of organisms in our data, without the phenotype and without the gene; the next three columns are defined similarly; the seventh column indicates the gene–phenotype correlation (in bits); some genes (*fliO*, *cheZ*) are too specific, while others (*fliI*, *cheY*) are too abundant.

Figure 2 Results for motility GGs. (**A**) Phylogenetic profiles of the five most informative modules. Rows correspond to motility GGs and columns to organisms; each entry indicates whether a particular motility GG is represented in the genome of a specific organism (red for motile and yellow for non-motile); the green bars indicate the correlation (in bits) between every motility GG and the motility phenotype; the average joint-assignment probability in each module is specified on the right (see Materials and methods). (**B**) Every entry in this matrix indicates the probability of two motility GGs to be placed in the same cluster by the clustering algorithm, that is, their joint-assignment probability.

© 2006 EMBO and Nature Publishing Group



C

Name	GG	p- g-	p+ g-	p- g+	p+ g+	MI	Function
LpxD	3	53	15	2	90	0.56	UDP-3-*O*--glucosamine *N*-acyltransferase
LpxH	137	54	51	1	54	0.22	UDP-2,3-diacylglucosamine hydrolase
KdsA	--	21	16	34	89	0.04	2-dehydro-3-deoxyphosphooctulonate aldolase

Figure 4 Results for Gram-negative GGs. (**A**) Phylogenetic profiles of four Gram-negative GG modules. Each entry indicates whether a particular Gram-negative GG is represented in the genome of a specific organism (red for Gram-negatives and yellow for Gram-positives). (**B**) Depiction of the lipid-A biosynthesis pathway in *E. coli*. Genes detected by our approach are highlighted in red. (**C**) Count matrices of known lipid-A biosynthesis *E. coli* genes, not detected by our approach (red row for *lpxD* is presented for comparison). See Figure 2B for columns descriptions.

© 2006 EMBO and Nature Publishing Group

Figure 5 Examples of GG modules obtained for the three respiration phenotypes. The two modules at the upper part (with suffix 'ae') correspond to two aerobic GG modules; the two modules in the middle (with suffix 'fa') correspond to two facultative GG modules; the three modules at the bottom (with suffix 'an') correspond to three anaerobic GG modules. Each entry indicates whether a GG is represented in the genome of a specific organism (red for strict aerobes, orange for facultatives, and yellow for strict anaerobes).

© 2006 EMBO and Nature Publishing Group

genes associated with this phenotype with an encouragingly high precision.

The cross-genome list for this phenotype consists of 850 genes collected from all 17 endospore-forming bacteria. These 850 genes correspond to 145 groups of homologs, or endospore GGs. These GGs were clustered as before, yielding 13 robust modules (Supplementary Figures S11 and S12). The three most robust modules, namely with the maximal average joint-assignment probability, are presented in Figure 6.

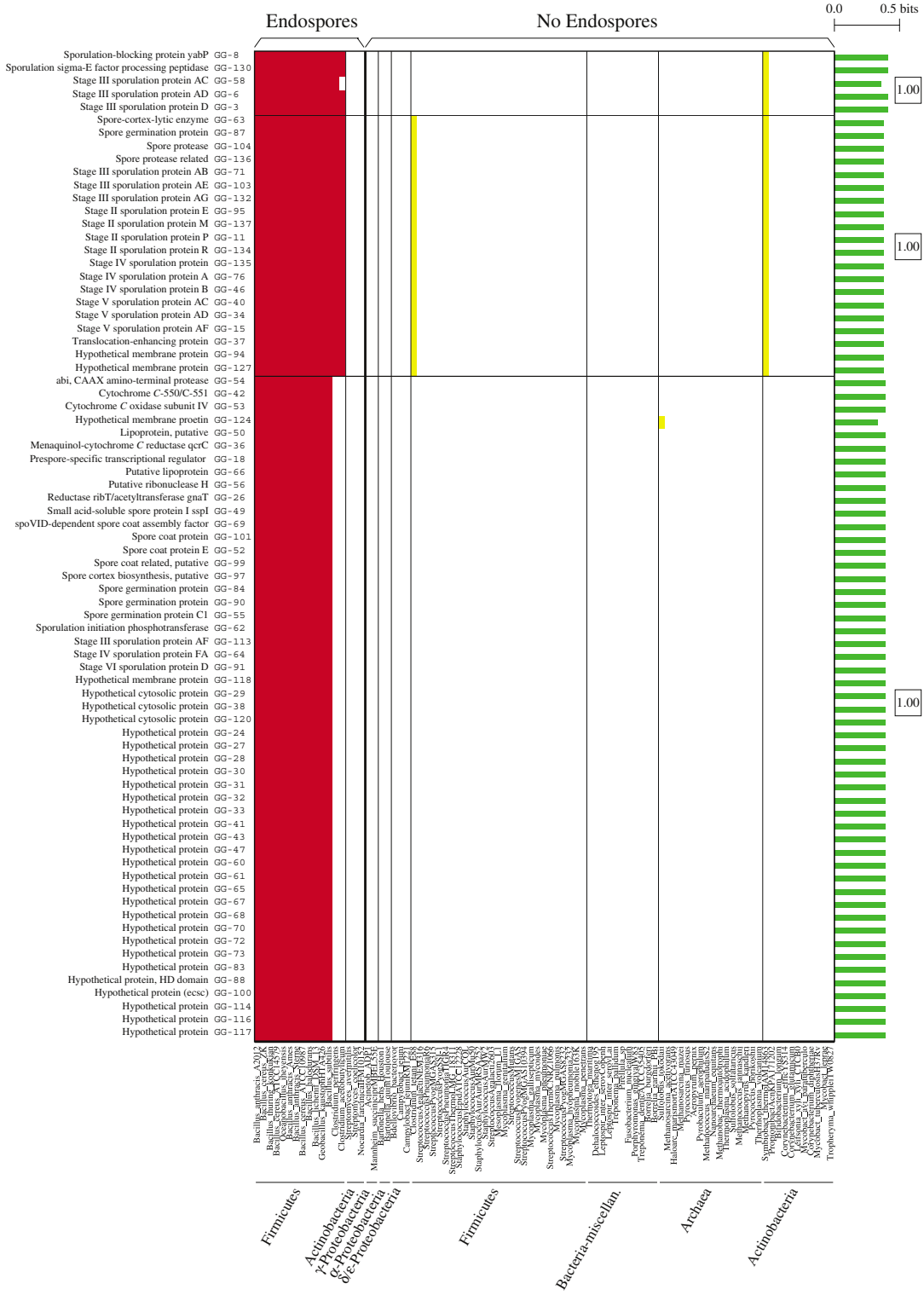


Figure 6 Phylogenetic profiles of the three most robust modules of endospore GGs. Each entry indicates whether an endospore GG is represented in the genome of a specific organism (red for sporulating organisms and yellow for nonsporulating ones).

© 2006 EMBO and Nature Publishing Group

The first two modules are dominant among the genera of *Bacillus* and the anaerobic *Clostridia*. Remarkably, 22 out of the 25 GGs in these two modules are known to play a role in sporulation, where two of the three remaining GGs are poorly characterized. Note that the *Clostridium tetani* species is known to be capable of forming spores, but the particular sequenced strain in our data is a nonsporulating variant, and thus was annotated as such by NCBI. Therefore, it is not surprising that this nominally nonsporulating species is represented in one of these two modules. In addition, both modules are found in the *Symbiobacterium thermophilum* genome, which according to NCBI is not capable of forming endospores. However, it was recently shown that this organism indeed forms endospore-like cellular structures (Ueda *et al*, 2004), in agreement with our results.

The third module is limited to the *Bacillus* genera. Nevertheless, 14 out of the 19 characterized GGs in this module are also known to be associated with sporulation; in addition, 29 GGs in this module correspond to groups of homologous genes where all group members are poorly characterized to date. Finally, we note that the three spore-forming *Actinobacteria* are not represented in these three modules; in fact, these three species are represented only in a few modules in Supplementary Figure S11. Although the genetic basis of this phenotype has not been extensively studied in these species, it is known to be different than the canonical pathways in *Bacillus subtilis* (Sonenshein, 2002). Thus, recovering the genetic basis of sporulation in this class will likely require more than just the three complete genome sequences that were available in our data.

Pathogenicity: intracellular

Intracellular pathogenicity is a complex, largely unexplored phenotype of great medical interest. It involves multiple interactions between the bacteria and the host eukaryotic cells (e.g., cell invasion, host-bacteria small-molecule transfer, and immune evasion), and is therefore likely to correspond to a diverse set of molecular mechanisms. As shown below, the modules predicted by our approach capture the diversity of these interactions, and in several cases allow us to make functional predictions.

The cross-genome list for this phenotype consists of 1178 genes collected from 47 intracellular pathogenic bacteria. These 1178 genes correspond to 224 groups of homologs, or intracellular pathogenesis GGs. These GGs were clustered as before, yielding 19 robust modules (Supplementary Figures S13 and S14), six of which are presented in Figure 7.

The first two modules in Figure 7 correspond to different members of the type III secretion system. The type III secretion system was thoroughly studied in *Salmonella enterica* (Galan, 2001); it directs the translocation of bacterial proteins, termed effector proteins, into the host cell. These effector proteins carry out several distinct roles like modulating the actin cytoskeleton to facilitate bacterial entry into non-phagocytic cells (Fu and Galan, 1999). Moreover, the type III secretion system remains active after internalization in order to deliver proteins into the cytosol of the host cell (Collazo and Galan, 1997). The first of these two modules, present in many intracellular pathogenic γ-proteobacteria and in two β-proteobacteria, but in none of the extracellular pathogens of the same phylogenetic groups, contains three GGs associated with components of the type III secretion system, and two GGs that correspond to effector proteins. The second module is more generic and includes components of the type III secretion system that are also present in *Chlamidiae* genomes. Interestingly, one of the two GGs in this module corresponds to a low calcium response chaperone. Such chaperones have been shown to bind effector proteins in the bacterial cytosol, and may be involved in stabilizing these molecules or preventing their interactions with other proteins (Mecsas and Strauss, 1996).

The presence of several phage-related GGs in the third and fourth modules of Figure 7 is intriguing, as it has been shown that a type III effector protein is encoded within the genome of a cryptic phage present in the *Salmonella typhimurium* genome (Hardt *et al*, 1998). It is possible that the phage-related GGs in Figure 7 are not directly involved in intracellular pathogenicity, but have been cotransferred along with certain type III effector proteins. Nevertheless, the significant association of several of these GGs with the phenotype provides corroborating evidence that phage-mediated transfer of genetic material may play an important role in host cell invasion by pathogenic intracellular bacteria (Hardt *et al*, 1998).

The fifth module in Figure 7 consists of two GGs that are present in all the obligate intracellular *Rickettsia* and *Chlamidia* in our data. One of these two GGs is uncharacterized, while the other is associated with ADP/ATP carrier proteins. In obligate intracellular species, these proteins are known to take up ATP in exchange for ADP within the cytosol of their eukaryotic hosts.

Finally, the sixth module in Figure 7 consists of GGs present in almost all pathogenic intracellular *Actinobacteria*, but in none of the extracellular pathogens. This module contains several membrane-associated GGs, but almost all of them are uncharacterized in all the respective genomes. Interestingly, one of the GGs in this module, GG-195, includes the lprG lipoprotein in *Mycobacterium tuberculosis*; lprG was recently shown to inhibit MHC-II antigen processing in human macrophages, thus possibly contributing to immune evasion of *M. tuberculosis* (Gehring *et al*, 2004). It will be interesting to experimentally test whether the other members of this module also have similar molecular functions, as predicted by our approach.

Discussion

Our capacity to sequence genomes has far outpaced our ability to understand their biology (Margulies *et al*, 2005; Shendure *et al*, 2005). Although functional genomic strategies show enormous promise when it comes to model organisms, adapting them to the vast majority of species of biomedical and industrial interest presents a formidable challenge. Here, we have shown that our computational approach provides an effective alternative in revealing the genetic basis of a variety of phenotypic traits and behaviors. Specifically, our approach recovers distinct gene modules that are associated with the

© 2006 EMBO and Nature Publishing Group

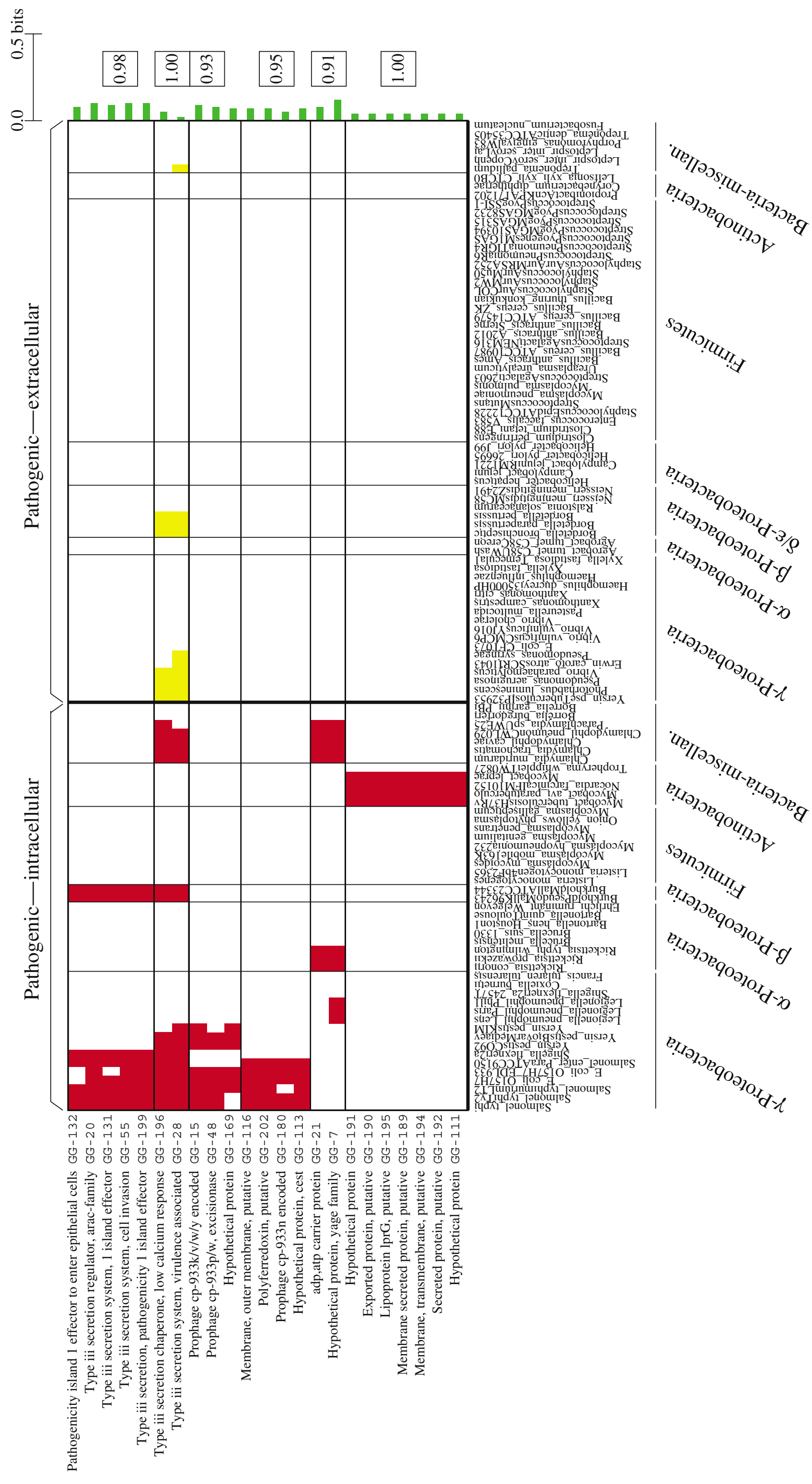


Figure 7 Phylogenetic profiles of six robust modules of intracellular pathogenicity GGs. Each entry indicates whether an intracellular pathogenicity GG is represented in the genome of a specific organism (red for intracellular pathogens, yellow for extracellular pathogens).

© 2006 EMBO and Nature Publishing Group

considered phenotype in a generic sense, that is, across most of the species expressing the phenotype.

The modules we find map directly to enzymatic pathways (lipid-A biosynthesis), molecular complexes (NADH dehydrogenase), signaling pathways (chemotaxis system), and molecular machines (bacterial flagellum, type III secretion system). It is important to emphasize that these modules are the products of an unsupervised clustering process. The phenotype-associated GGs are clustered numerous times under different initial conditions, and the reported modules correspond to sets of GGs that are consistently placed by the algorithm in the same cluster, regardless of its initialization.

The phylogenetic signatures of these modules provide insights into dominant evolutionary trends in their utilization. For example, many chemotaxis and flagella genes have phylogenetic profiles that naturally correlate with motility, but our analysis separates them into two distinct and robust modules, implying that they actually correlate with motility in different ways (Figure 2). In fact, as we see, the generic chemotaxis module—as a module—may couple with other types of flagella (e.g., those of motile archaea) or motility mechanisms, while the generic flagella module may couple with other types of chemotaxis systems (e.g., in the *Legionella* genomes).

Similar modularity patterns were observed essentially in all the phenotypes that we examined, most of which are only partially understood at present. Thus, our results support the notion of modularity in molecular systems (Hartwell *et al*, 1999; Ravasz *et al*, 2002) and demonstrate how such modules have acquired distinguishable co-inheritance patterns throughout evolution. In cases where the molecular mechanisms underlying a phenotype are relatively uncharacterized (e.g., pathogenic intracellular bacteria), the modularity revealed by our approach provides concrete hypotheses that can be used to design more focused experiments. For example, the uncharacterized membrane proteins of the last module in Figure 7 should be tested first for their ability to inhibit MHC-II antigen processing in human macrophages, as the only characterized member of this module was shown to have this property. On the other hand, experiments designed to test the involvement of these genes in type III secretion may be less fruitful, as the members of this system are dominant in two other modules in Figure 7, both of which have very different phylogenetic signatures.

At the time this report was submitted (October 2005), 266 complete microbial genomes were already available through NCBI and the sequencing of 549 others is reported as being in progress. As more whole genome sequences become available, we expect that methods like the one presented here will allow the detection of increasingly precise modularity, as large cohorts of new bacterial genomes will be shown to have lost or gained entire groups of genes involved in similar functions. Meanwhile, methodologies for high-throughput phenotypic annotation are being actively investigated, and are being applied to phenotypes of which very little is understood at present. One promising such methodology is the comprehensive mapping of microbial communities, either through rDNA sequencing (Eckburg *et al*, 2005) or low-coverage shotgun sequencing (Venter *et al*, 2004; Tringe *et al*, 2005). The computational framework presented here should be instrumental in revealing modules of genes shared by members of these communities (as opposed to non-members), which are broadly used to sustain life and harness environmental resources in their natural habitats. Systematic phenotype annotations, such as those already available at NCBI, have been created only recently, and typically for phenotypes that have been studied for decades, and that are relatively well understood. It is plausible to assume that this trend will soon change, as phenotype and genotype will be almost immediately associated, and the corresponding genes will be automatically grouped into functionally coherent modules, using methods such as ours.

While our approach can be applied to arbitrary phenotypes, one should bear in mind that association does not necessarily reflect causality. In particular, the lifestyles and native habitats of free-living microbes induce strong phenotype–phenotype correlations that could potentially confound the interpretations of studies such as ours. The observed association of the PTS sugar transporter modules with the facultative phenotype (Figure 5) may correspond to this type of correlation, where the ability to thrive in both aerobic and anaerobic settings may reflect the dominance of facultative bacteria in carbohydrate-rich environments. Indeed, most of the facultative microbes in our data are commensal or pathogenic bacteria that can inhabit nutrient-rich environments within the host. In principle, such phenotype correlations can be explicitly modeled once they are measured. However, measuring these correlations will require considering many more phenotypes, in a much larger sample of species, than those currently available. Interestingly, a recent study demonstrated how phenotypic data for a diverse set of species can be automatically and successfully gathered from the literature (Korbel *et al*, 2005).

A related issue of concern is the degree to which the uneven phylogenetic distribution of a particular trait may confound the interpretation of our results. In a severe scenario, the approach may identify species-distinguishing modules, rather than those that underlie a particular trait. For example, the *Bacillus* species constitute most of the endospore-forming species in our data. Thus, *a priori*, one might be concerned that the *Bacillus* phylogenetic signature may overwhelm the endospore phenotype; nonetheless, as our results show, we identify a large number of components known to be involved in sporulation with an encouraging precision. For example, 22 out of the 25 endospore GGs in the first two modules in Figure 6 correspond to known sporulation genes that are also present in non-*Bacillus*-sporulating organisms. While these observations may not necessarily generalize to all phenotypes, rapid advances in sequencing efficiency (Margulies *et al*, 2005; Shendure *et al*, 2005) are expected to even out phylogenetic coverage well beyond any conceivable concern here.

We have presented a computational framework for revealing the underlying genetic architecture of a trait by characterizing its expression at the organism level, across many species. Our complete set of results, available at our Web site, provides a wealth of experimentally testable hypotheses that associate genes with complex traits via the simultaneous analysis of more than 200 complete genome sequences. Beyond its utility for generating hypotheses, our approach reveals an intrinsic modularity in genetic networks, and highlights the extensive

© 2006 EMBO and Nature Publishing Group

and broad sharing of optimized genetic modules across the tree of life. The utility of this approach depends crucially on the success of efforts to systematically characterize the vast variety of diverse phenotypic traits throughout the microbial biosphere.

Materials and methods

Genome sequences and phenotype annotations

We downloaded the 214 complete microbial genome sequences that were available at the NCBI Web site (http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi) on January 18, 2005. To avoid redundancy, for each pair of highly similar genomes, we only retained the genome that was sequenced first. This led to a total of 202 genomes. Phenotype annotations were also downloaded from the NCBI site. The intracellular pathogenicity phenotype was generated by manual curation, based on literature search, data available at NCBI, and data reported at the *IslandPath* Web site at http://pathogenomics.sfu.ca/islandpath/current/IPindex.pl (Hsiao *et al*, 2003). All these phenotype annotations are available at our Web site, http://tavazoielab.princeton.edu/genphen/.

Creating the gene phylogenetic profiles

For every gene in every genome, we applied a BLAST search (Altschul *et al*, 1990) against all the remaining 201 genomes to identify possible homologs. Similarly to previous work (Jim *et al*, 2004), a genome was considered as containing a homolog when one of its proteins had an alignment to the query protein sequence with an e-value smaller than 10^{-10}. Proteins with less than 50 amino acids were ignored. This resulted in 591 640 phylogenetic profiles where each profile is a binary vector with 202 elements. Preliminary tests indicated that using the raw BLAST e-values in the phylogenetic profiles (Date and Marcotte, 2003) yields similar results in the analysis that follows. Therefore, all the subsequent analysis used this binary representation. The entire phylogenetic profile collection is available at our Web site.

Estimating gene–phenotype mutual information

Given a gene phylogenetic profile and a phenotype profile, we can define a count matrix, N, where $N_{1,1}$ is the number of species with the phenotype and the gene, $N_{1,2}$ is the number of species with the phenotype but without the gene, $N_{2,1}$ is the number of species without the phenotype but with the gene, and $N_{2,1}$ is the number of species without the phenotype and without the gene. The *empirical* mutual information between the profiles is given by

$$I(gene;phen) = \sum_{i,j} P_{i,j} \log P_{i,j}/(P_i \times P_j)$$

where $P_{i,j}=N_{i,j}/\sum_{i,j} N_{i,j}$, $P_i=P_{i,1}+P_{i,2}$, and $P_j=P_{1,j}+P_{2,j}$ (Cover and Thomas, 1991). This information is naturally normalized between 0 and 1 bits, where 0 bits means no dependency while high information values imply strong correlation between the gene and the phenotype profile. All the gene–phenotype information relations were estimated through the *direct* method (Strong *et al*, 1998; Slonim *et al*, 2005a) in order to correct for finite sample effects, using the software available at http://www.genomics.princeton.edu/biophysics-theory/DirectMI/webcontent/index.html with its default parameters. The same procedure was applied for randomly shuffled gene phylogenetic profiles and the maximum information value obtained was used as a threshold for significance. That is, the association of a gene with a phenotype was considered significant if and only if their mutual information was found to be greater than the maximal value obtained in the shuffled data. Importantly, in contrast to previously used correlation measures (Huynen *et al*, 1998; Levesque *et al*, 2003; Makarova *et al*, 2003; Jim *et al*, 2004; Korbel *et al*, 2005), the mutual information can be equally applied to continuous phenotypes like optimal temperature growth, and to measure the correlations between sets of genes and phenotypes, as we plan to investigate in a subsequent study.

Finding the phenotype GGs

The construction of the phenotype GGs consists of two phases. In the first phase, for every phenotype, we collected the 50 genes with the strongest positive correlation with the phenotype from every organism having the phenotype (as long as this correlation was significant), and joined them into a single *cross-genome* list. We then defined a similarity graph among all the genes in this list where two genes were connected by an edge if their corresponding BLAST e-value was smaller than 10^{-10}; next, an agglomerative merging process was applied to find strongly connected components in this graph. Specifically, the algorithm first assigns every gene in a singleton group, and then recurrently performs the merger with the maximal 'score', where the score of merging two groups of genes is defined as the probability of having an edge between two genes chosen independently from both groups. More formally, denoting both groups by c_1 and c_2, the corresponding merger score is

$$\frac{1}{|c_1||c_2|} \sum_{g_1 \in c_1} \sum_{g_2 \in c_2} B(g_1, g_2)$$

where $B(g_1,g_2)$ is 1 for homologous genes and 0 otherwise, and $|c_i|$ denotes the number of genes in each group. If more than one merger attained the maximal score, the one resulting with the largest new group was preferred. We chose a score threshold of 0.5 as a stopping criterion for the merging process. Our analysis was highly robust with respect to this parameter, where using score thresholds of up to 0.7 gave identical results. The merging process results in groups of homologous genes, in which most gene pairs have a BLAST e-value below 10^{-10}. The average edge density in the resulting groups was around 99%, that is, most groups corresponded to almost fully connected components in the afore-mentioned BLAST similarity graph. To further validate the robustness of these results, we used the BLASTClust software (available as part of the BLAST package) over the same cross-genome lists. For all phenotypes, this resulted with groups of homologous genes that were highly similar to those extracted by our merging algorithm.

In the second phase of our construction, each group of homologous genes was further expanded to include additional homologs that were not detected through the first phase. Specifically, a gene from a species having the phenotype was added to a group if it had a BLAST e-value below 10^{-10} with at least one-third of the original group members (again, different values of this parameter gave very similar results). As a simple example, let us consider the case of the *flgL* gene in *E. coli*. This gene, involved in flagellar biosynthesis, obtained an information score of $\sim$0.16 bits over the motility phenotype, which was not sufficient for it to be included among the 50 *E. coli* genes that were most informative about motility. As a result, this gene was not included in the cross-genome list, out of which we constructed the groups of homologous genes in the first phase. Nevertheless, several homologs of this gene in other species (e.g., *flgL* in *Bacillus subtilis)* obtained higher information scores that placed them among the 50 most informative genes about motility in their respective genomes. This gave rise to a group of *flgL* homologous genes that was constructed in the first phase, to which the *flgL* gene in *E. coli* was added as an expansion in this second phase. A summary file, describing all the genes in every group, along with relevant details from NCBI annotation, is available at our Web site. The NCBI gene textual descriptions were used to determine a concise textual title for every group. In principle, the genes in each group correspond to different reflections of the same ancestral entity, with phylogenetic profiles that strongly correlate with the examined phenotype profile. Therefore, these groups are termed here phenotype generic genes (GGs).

Finding robust GG modules

A phenotype GG corresponds to a group of homologous genes, all taken only from species that have the phenotype. However, constructing the GG phylogenetic profile (e.g., for the purpose of identifying GG

© 2006 EMBO and Nature Publishing Group

modules) requires that *all* species be considered. To that end, we applied the following procedure. If more than one-third of the original GG members had a BLAST e-value smaller than 10^{-10} in a particular genome, the profile entry of the GG for this genome was set to 1. Otherwise, it was set to 0.

Given the GG phylogenetic profiles, we estimated the mutual information between every pair of GGs, and used these information relations as input to the *Iclust* clustering algorithm (Slonim *et al*, 2005b) (manuscript and software available at http://www.genomics.princeton.edu/biophysics-theory/Clustering/web-content/index.html). This algorithm finds a partition of the GGs into clusters such that GGs in the same cluster are highly informative about each other, that is, have highly similar phylogenetic profiles.

The *Iclust* algorithm corresponds to a fully principled clustering methodology. However, as with any other clustering algorithm, it may produce suboptimal solutions in a single run, depending on the (random) partition used in its initialization. To address this issue, we applied the *Iclust* algorithm with default parameter values 1000 times, each with a different initial random partition, yielding potentially 1000 (slightly) different clustering solutions. Next, for every pair of GGs we defined the *joint-assignment probability* as the number of solutions in which the pair was placed by the algorithm in the same cluster, divided by 1000. Thus, two GGs placed very often in the same cluster by the algorithm will have a relatively high joint-assignment probability. Finally, we defined a graph between all the GGs where two GGs were connected by an edge if their corresponding joint-assignment probability was greater than 0.9, and used the merging process described earlier to find *fully* connected components in this graph. The resulting connected components with at least two GGs correspond to the robust GG modules that we report and analyze in this study. By definition, each such module corresponds to GGs consistently placed by the clustering algorithm in the same cluster, almost regardless of the initial random partition used. Our analysis was relatively insensitive to variations in the threshold used in this stage. For example, using a joint-assignment probability threshold of 0.8 gave similar results for all phenotypes.

Supplementary information

Supplementary information is available at the *Molecular Systems Biology* website (www.nature.com/msb).

Acknowledgements

We are grateful to CS Chan, H Girgis, W Bialek, and the two anonymous referees for many insightful comments on preliminary versions of the manuscript. This work was supported in part by NIH, NSF, and DARPA.

References

Alberts B, Bray D, Lewis J, Raff M, Roberts K, Watson J (1994) *Molecular Biology of the Cell*. New York: Garland Science Publishing

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol* **215:** 403–410

Berardi MJ, Bushweller JH (1999) Binding specificity and mechanistic insight into glutaredoxin-catalyzed protein disulfide reduction. *J Mol Biol* **292:** 151–161

Collazo CM, Galan JE (1997) The invasion-associated type III system of *Salmonella typhimurium* directs the translocation of Sip proteins into the host cell. *Mol Microbiol* **24:** 747–756

Copley SD, Dhillon JK (2002) Lateral gene transfer and parallel evolution in the history of glutathione biosynthesis genes. *Genome Biol* **3:** research0025.1–0025.16

Cover TM, Thomas JA (1991) *Elements of Information Theory*. New York: John Willey & Sons

Date SV, Marcotte EM (2003) Discovery of uncharacterized cellular systems by genome-wide analysis of functional linkages. *Nat Biotechnol* **21:** 1055–1062

Dreyfus G, Williams AW, Kawagishi I, Macnab RM (1993) Genetic and biochemical analysis of *Salmonella typhimurium* FliI, a flagellar protein related to the catalytic subunit of the F0F1 ATPase and to virulence proteins of mammalian and plant pathogens. *J Bacteriol* **175:** 3131–3138

Eckburg PB, Bik EM, Bernstein CN, Purdom E, Dethlefsen L, Sargent M, Gill SR, Nelson KE, Relman DA (2005) Diversity of the human intestinal microbial flora. *Science* **308:** 1635–1638

Errington J (2003) Regulation of endospore formation in *Bacillus subtilis*. *Nat Rev Microbiol* **1:** 117–126

Fu Y, Galan JE (1999) A salmonella protein antagonizes Rac-1 and Cdc42 to mediate host-cell recovery after bacterial invasion. *Nature* **401:** 293–297

Galan JE (2001) Salmonella interactions with host cells: type III secretion at work. *Annu Rev Cell Dev Biol* **17:** 53–86

Gehring AJ, Dobos KM, Belisle JT, Harding CV, Boom WH (2004) *Mycobacterium tuberculosis* LprG (Rv1411c): a novel TLR-2 ligand that inhibits human macrophage class II MHC antigen processing. *J Immunol* **173:** 2660–2668

Genevrois S, Steeghs L, Roholl P, Letesson JJ, van der Ley P (2003) The Omp85 protein of *Neisseria meningitidis* is required for lipid export to the outer membrane. *EMBO J* **22:** 1780–1789

Hardt WD, Urlaub H, Galan JE (1998) A substrate of the centisome 63 type III protein secretion system of *Salmonella typhimurium* is encoded by a cryptic bacteriophage. *Proc Natl Acad Sci USA* **95:** 2574–2579

Hartwell LH, Hopfield JJ, Leibler S, Murray AW (1999) From molecular to modular cell biology. *Nature* **402** (Suppl)**:** C47–C52

Hegermann J, Herrmann R, Mayer F (2002) Cytoskeletal elements in the bacterium *Mycoplasma pneumoniae*. *Naturwissenschaften* **89:** 453–458

Hsiao W, Wan I, Jones SJ, Brinkman FS (2003) IslandPath: aiding detection of genomic islands in prokaryotes. *Bioinformatics* **19:** 418–420

Huynen M, Dandekar T, Bork P (1998) Differential genome analysis applied to the species-specific features of *Helicobacter pylori*. *FEBS Lett* **426:** 1–5

Jagannathan A, Constantinidou C, Penn CW (2001) Roles of *rpoN*, *fliA*, and *flgR* in expression of flagella in *Campylobacter jejuni*. *J Bacteriol* **183:** 2937–2942

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

Kanehisa M, Goto S (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* **28:** 27–30

Korbel JO, Doerks T, Jensen LJ, Perez-Iratxeta C, Kaczanowski S, Hooper SD, Andrade MA, Bork P (2005) Systematic association of genes to phenotypes by genome and literature mining. *PLoS Biol* **3:** e134

Lerat E, Daubin V, Ochman H, Moran NA (2005) Evolutionary origins of genomic repertoires in bacteria. *PLoS Biol* **3:** e130

Levesque M, Shasha D, Kim W, Surette MG, Benfey PN (2003) Trait-to-gene: a computational method for predicting the function of uncharacterized genes. *Curr Biol* **13:** 129–133

Makarova KS, Wolf YI, Koonin EV (2003) Potential genomic determinants of hyperthermophily. *Trends Genet* **19:** 172–176

Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z, Dewell SB, Du L, Fierro JM, Gomes XV, Godwin BC, He W, Helgesen S, Ho CH, Irzyk GP, Jando SC, Alenquer ML, Jarvie TP, Jirage KB, Kim JB, Knight JR, Lanza JR, Leamon JH, Lefkowitz SM, Lei M, Li J, Lohman KL, Lu H, Makhijani VB, McDade KE, McKenna MP, Myers EW, Nickerson E, Nobile JR, Plant R, Puc BP, Ronan MT, Roth GT, Sarkis GJ, Simons JF, Simpson JW, Srinivasan M, Tartaro KR, Tomasz A, Vogt KA, Volkmer GA, Wang SH, Wang Y, Weiner MP, Yu P, Begley RF, Rothberg JM (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437:** 376–380

© 2006 EMBO and Nature Publishing Group

Mecsas JJ, Strauss EJ (1996) Molecular mechanisms of bacterial virulence: type III secretion and pathogenicity islands. *Emerg Infect Dis* **2:** 270–288

Nierman WC, DeShazer D, Kim HS, Tettelin H, Nelson KE, Feldblyum T, Ulrich RL, Ronning CM, Brinkac LM, Daugherty SC, Davidsen TD, Deboy RT, Dimitrov G, Dodson RJ, Durkin AS, Gwinn ML, Haft DH, Khouri H, Kolonay JF, Madupu R, Mohammoud Y, Nelson WC, Radune D, Romero CM, Sarria S, Selengut J, Shamblin C, Sullivan SA, White O, Yu Y, Zafar N, Zhou L, Fraser CM (2004) Structural flexibility in the *Burkholderia mallei* genome. *Proc Natl Acad Sci USA* **101:** 14246–14251

Nikaido H (2003) Molecular basis of bacterial outer membrane permeability revisited. *Microbiol Mol Biol Rev* **67:** 593–656

Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc Natl Acad Sci USA* **96:** 4285–4288

Pittman MS, Goodwin M, Kelly DJ (2001) Chemotaxis in the human gastric pathogen *Helicobacter pylori:* different roles for CheW, the three CheV paralogues, and evidence for CheV2 phosphorylation. *Microbiology* **147** (Part 9)**:** 2493–2504

Ravasz E, Somera AL, Mongru DA, Oltvai ZN, Barabasi AL (2002) Hierarchical organization of modularity in metabolic networks. *Science* **297:** 1551–1555

Ruiz N, Falcone B, Kahne D, Silhavy TJ (2005) Chemical conditionality: a genetic strategy to probe organelle assembly. *Cell* **121:** 307–317

Shendure J, Porreca GJ, Reppas NB, Lin X, McCutcheon JP, Rosenbaum AM, Wang MD, Zhang K, Mitra RD, Church GM (2005) Accurate multiplex polony sequencing of an evolved bacterial genome. *Science* **309:** 1728–1732

Shigenobu S, Watanabe H, Hattori M, Sakaki Y, Ishikawa H (2000) Genome sequence of the endocellular bacterial symbiont of aphids *Buchnera* sp. *APS*. *Nature* **407:** 81–86

Slonim N, Atwal GS, Tkacik G, Bialek W (2005a) Estimating mutual information and multi-information in large networks. http://arxiv.org/abs/cs.IT/0502017

Slonim N, Atwal GS, Tkacik G, Bialek W (2005b) Information based clustering. *Proc Natl Acad Sci USA* **102:** 18297–18302

Sonenshein AL (2002) Developmental biology: regulation by selective gene localization. *Curr Biol* **12:** R90–R92

Strong SP, Koberle R, de Ruyter van Steveninck RR, Bialek W (1998) Entropy and information in neural spike trains. *Phys Rev Lett* **80:** 197–200

Tringe SG, von Mering C, Kobayashi A, Salamov AA, Chen K, Chang HW, Podar M, Short JM, Mathur EJ, Detter JC, Bork P, Hugenholtz P, Rubin EM (2005) Comparative metagenomics of microbial communities. *Science* **308:** 554–557

Ueda K, Yamashita A, Ishikawa J, Shimada M, Watsuji TO, Morimura K, Ikeda H, Hattori M, Beppu T (2004) Genome sequence of *Symbiobacterium thermophilum,* an uncultivable bacterium that depends on microbial commensalism. *Nucleic Acids Res* **32:** 4937–4944

Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA, Wu D, Paulsen I, Nelson KE, Nelson W, Fouts DE, Levy S, Knap AH, Lomas MW, Nealson K, White O, Peterson J, Hoffman J, Parsons R, Baden-Tillson H, Pfannkoch C, Rogers YH, Smith HO (2004) Environmental genome shotgun sequencing of the Sargasso Sea. *Science* **304:** 66–74

Vreeland RH, Rosenzweig WD, Powers DW (2000) Isolation of a 250 million-year-old halotolerant bacterium from a primary salt crystal. *Nature* **407:** 897–900

Wolfe AJ, Millikan DS, Campbell JM, Visick KL (2004) *Vibrio fischeri* sigma54 controls motility, biofilm formation, luminescence, and colonization. *Appl Environ Microbiol* **70:** 2520–2524

Wu T, Malinverni J, Ruiz N, Kim S, Silhavy TJ, Kahne D (2005) Identification of a multicomponent complex required for outer membrane biogenesis in *Escherichia coli*. *Cell* **121:** 235–245

© 2006 EMBO and Nature Publishing Group