GENOMIC EVOLVABILITY AND THE ORIGIN OF NOVELTY: STUDYING THE PAST, INTERPRETING THE PRESENT, AND PREDICTING THE FUTURE

Jonathan A. Eisen, Ph.D.[5]
University of California, Davis

Introduction

One of the unifying goals of this workshop, as well as of the Forum on Microbial Threats, has been to promote the study of microbes, not only to enhance our understanding of their present roles in the world but also, we hope, to predict their future changes (e.g., the emergence of new infectious diseases). This was, of course, one of the life missions of Joshua Lederberg, who helped create the Forum and who this workshop is honoring.

Studies of evolution are central to these goals because "nothing in biology makes sense except in the light of evolution" (Dobzhansky, 1973). Evolutionary studies help us understand the past and interpret the present, and from a combination of those two we have some possibility of being able to predict the future. Since Lederberg was also keen on evolutionary studies (Lederberg, 1997, 1998), it is appropriate for a workshop in his honor to focus on *Microbial Evolution and Co-Adaptation*.

I would like to note that I feel a personal connection to Joshua Lederberg, as I received much of my microbiology training from Ann Ganesan who had been a Ph.D. student in his lab. However, anyone, with or without a specific connection to Lederberg, can learn a great deal about him and his work through a wonderful website made available by the National Library of Medicine.[6] There you can find many of his scientific papers, as well as his letters, scientific articles he wrote, articles he read, columns he wrote for *The Washington Post* on science policy, and more. In addition, I would like to point out that Lederberg was an ardent supporter of "open access" to scientific publications. He was on the board of PubMed when PubMed Central was created. PubMed Central is a centralized archive of freely available, full-text versions of scientific publications. Although not all of his papers are in PubMed Central,[7] most are available at the National Library of Medicine site.

In this paper, I am focusing on one key aspect of evolution: the *origin of novelty*, or how new forms, functions, processes, and properties originate. In addition, I consider some of the factors that influence the likelihood that novelty

[5] University of California, Davis Genome Center; Department of Medical Microbiology and Immunology and Section of Evolution and Ecology, Davis, CA 95616; E-mail: jaeisen@ucdavis.edu; Website: http://phylogenomics.blogspot.com.

[6] See http://profiles.nlm.nih.gov/BB.

[7] See http://www.pubmedcentral.nih.gov.

will originate—something generally referred to as *evolvability*. Why do some organisms invent new functions readily while others are "novelty challenged"? I note that I focus here on work from my lab and am not attempting to review the entire field.

I have been interested in the origin and novelty and evolvability, particularly as they occur in microbes, since I was introduced to microbes as an undergraduate through studies of hydrothermal vent ecosystems. Actually, I had written a paper on this back in high school, but it was not until college that I truly focused on the topic. A bit later, in 1995, my career—and that of most other microbiologists and evolutionary biologists—was changed forever with the publication of the first complete genome of any free-living organism (Fleischmann et al., 1995). It was then that I shifted my research to the integration of evolutionary analyses with studies of genome sequences. For better or for worse, I coined the term for this field: *phylogenomics* (Eisen et al., 1997). Note that the way I use this term is a bit different than some others in the community. Many people use the term phylogenomics to refer to the use of genome-scale data (e.g., genome sequences) for phylogenetic studies. With that introduction, I will now relate some "phylogenomic tales" as examples of how phylogenomic analysis can help us understand the origin of novelty. This will also demonstrate the usefulness of this approach for understanding the past, interpreting the present, and—maybe—predicting the future.

Phylogenomics and Novelty I: Predicting Gene Functions Using Evolutionary Trees

Throughout this workshop, we have seen many examples of genome sequencing leading to wonderful insights about the microbial world. Indeed, it can be said that genome sequence data have sparked a renaissance in microbiology. It is important to realize, however, that much of this renaissance rests on one particular step in the analysis—the prediction of gene function based on gene sequence. This step is critical because typically one generates the genome sequence of a particular organism, most of whose genes will not have been studied experimentally. Prediction of gene function adds value to the genome sequence data because such predictions can guide further computational and experimental studies of the organism. My first phylogenomic tale illustrates how, in the course of a genome-sequencing project, the evolutionary analysis of a particular gene can enable us to make more accurate predictions about the function of that gene in a particular organism and, in some cases, can also provide insight into the evolutionary processes in that organism, as well.

This is the story of one such organism, *Helicobacter pylori*, a bacterium that dwells in the stomachs of humans and some other mammals. For many years, these stomach dwellers were generally ignored. However, thanks in a large part to the work of people like Barry Marshall, it is now known that *H. pylori* is a causative agent of stomach ulcers as well as gastric cancers (Marshall, 2002). Due to its medical importance as well as its novel ability to tolerate very high acidity, this species was one of the first targeted for genome sequencing. In 1997, the genome of one strain was published (Tomb et al., 1997).

At that time, as a Ph.D. student at Stanford, I was relentlessly badgering everyone I knew, attempting to convince them that evolutionary analysis could help in the prediction of gene function. I had become convinced of this myself through analysis of the trickle of genome sequence data for humans, yeast, and other organisms that had already begun to flow before the first complete genome was published. Back in 1995, I had even published a paper showing the benefits of evolutionary reconstructions in studies of one family of proteins, the SNF2 family (Eisen et al., 1995). Although the benefits were clear to me, others were not so sure. Fortunately, at the time I was teaching a class with Rick Myers, a professor in the genetics department and the head of the Stanford Human Genome Center. He had been asked to write a "News and Views" piece for *Nature Medicine* commenting on the recent papers reporting the sequencing and analysis of the genomes of *H. pylori* (Tomb et al., 1997) and *Escherichia coli* K12 (Blattner et al., 1997). Also, since he was one of the people I had been badgering, he suggested I try to come up with an example of where the inclusion of evolutionary analysis could have benefited their work.

Luckily for me, there was a claim in the *H. pylori* paper that was a perfect candidate for evolutionary analysis. The authors reported (Tomb et al., 1997):

> The ability of *H. pylori* to perform mismatch repair is suggested by the presence of methyl transferases, mutS and uvrD. However, orthologues of MutH and MutL were not identified.

This was right up my alley because I was working on the evolution of DNA mismatch repair at the time.

A DNA mismatch can be created when the wrong base is put into a newly synthesized strand by the enzyme carrying out DNA replication (i.e., DNA polymerase). Thus, a mismatch indicates a replication error. Mismatch repair is a process whereby, immediately following DNA replication, repair enzymes scan for mismatches between the template and newly synthesized DNA strands. When the mismatch repair machinery finds one, it removes a section of DNA containing the mismatch from the newly synthesized strand. That section is then resynthesized using the original (and presumably accurate) template strand as a guide. Mismatch repair is vital. It greatly reduces the mutation rate by correcting many of the replication errors made by DNA polymerase.

It was because of my knowledge of the evolution of mismatch repair that the report in the *H. pylori* paper caught my attention. I knew that every time a mismatch repair system had been found in an organism, regardless of whether that organism was from the bacteria, mammals, plants, yeast, or a variety of other groups, and regardless of whether it was found by genetics, by biochemistry, or even by targeted cloning, the pattern was the same. The system always required at least one member of the MutS family of proteins and one member of the MutL family. Yet, according to the paper, *H. pylori* did not encode a MutL homolog. So I decided to look at this in more detail.

My first step was to recheck the genome sequence analysis. I did this by first using the Basic Local Alignment Search Tool (BLAST), which compares a given DNA, RNA, or protein sequence with corresponding sequences in a database and determines if there are similar sequences therein and, if so, generates a list of the closest matches. First, I took all known MutL-like proteins and searched them against the *H. pylori* genome data and found, as Tomb et al. did, that there were no close matches. Given that they had determined the complete genome of this strain, the absence of BLAST match suggested there was indeed no MutL encoded in the genome. I note that this "determining the absence" of something from a genome is one of the key benefits of determining *complete* genome sequences (Fraser et al., 2002).

Then I did some BLAST searches with MutS-like proteins as "queries" and found, as the authors had reported, that there was one, and only one, protein encoded in the genome that was similar to proteins in the MutS family. So I took this protein and then used it to search against all known sequences from other organisms, to see to what it was most similar. This in essence was mimicking the searches done in the analysis of the genome, and the result seemed quite convincing (Table 5-5). All of the proteins that were most similar to the *H. pylori* protein were described in the database as "Mismatch repair protein MutS" or something similar. This description of the related proteins, also known as their annotation, was clearly what led the authors to conclude that this protein was involved in mismatch repair. This left me with a conundrum. There was no MutL protein encoded in the genome, yet there was, apparently, a MutS protein. Many possible explanations came to mind, all of which were interesting. *H. pylori* might have been the first species to be found with a mismatch repair system that did not require a MutL homolog. Or it might have recently lost its MutL, as had been seen in many strains of *E. coli* and *Salmonella* (LeClerc et al., 1996). Alternatively, perhaps the MutS-like protein was not a normal MutS involved in mismatch repair, but rather was used for a different function in this organism.

Although these, along with yet other explanations, seemed plausible, one observation suggested to me that the latter explanation—that the MutS-like protein was doing something else—might be the correct one. In the list of BLAST matches, I had noticed that members of the MutS family that I knew to have documented roles in mismatch repair were not high on the list, indicating the *H. pylori* protein was not as similar to these as it was to some other MutS-like proteins that might have a novel function (Table 5-5). In addition, I knew from my prior work (Eisen et al., 1995), and from the work of others (e.g., Tatusov et al., 2000), that BLAST scores were not a reliable indicator of evolutionary relat-

TABLE 5-5 BLAST Search Results as They Were Seen in 1997 Using the MutS-like Protein from *Helicobacter pylori* as a Query

	Score	E
Sequences producing significant alignments	(bits)	Value
sp|P73625|MUTS_SYNY3 DNA MISMATCH REPAIR PROTEIN	117	3e-25
sp|P74926|MUTS_THEMA DNA MISMATCH REPAIR PROTEIN	69	1e-10
sp|P44834|MUTS_HAEIN DNA MISMATCH REPAIR PROTEIN	64	3e-09
sp|P10339|MUTS_SALTY DNA MISMATCH REPAIR PROTEIN	62	2e-08
sp|O66652|MUTS_AQUAE DNA MISMATCH REPAIR PROTEIN	57	4e-07
sp|P23909|MUTS_ECOLI DNA MISMATCH REPAIR PROTEIN	57	4e-07

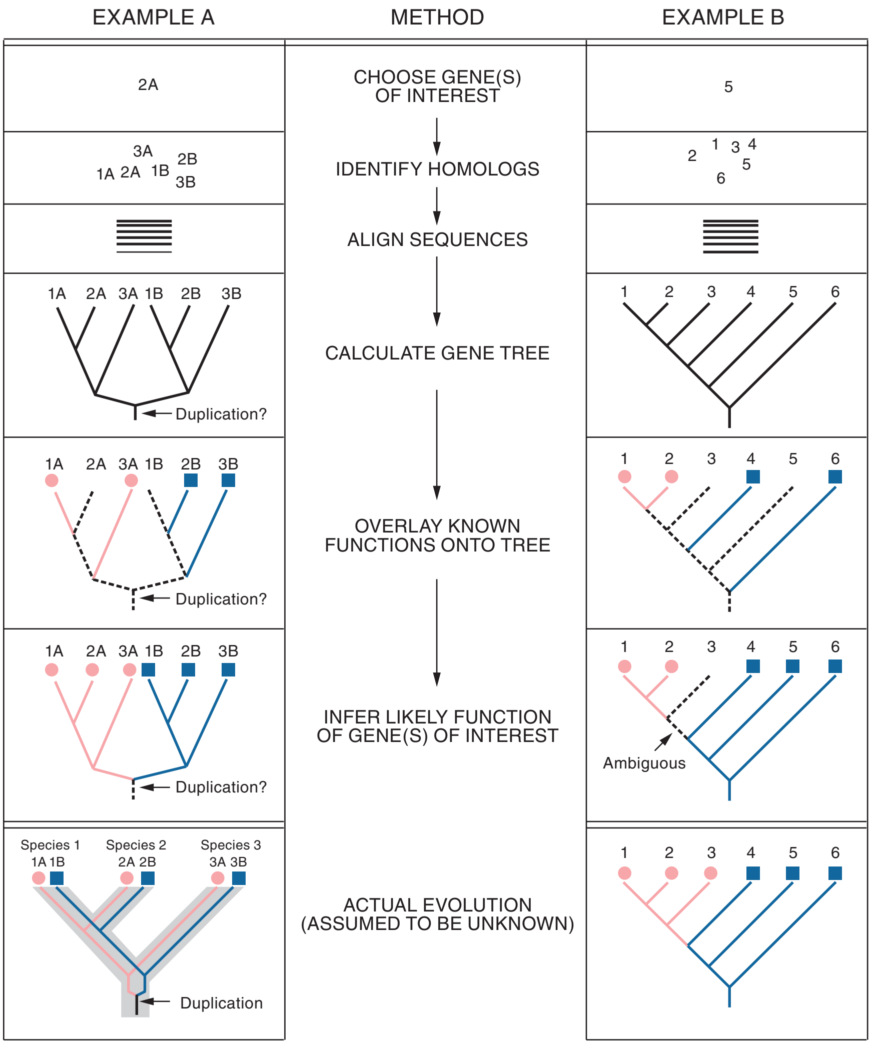
edness. So my next step was to investigate the evolutionary history of the MutS proteins, including the new one from *H. pylori*. I did this by generating a multiple sequence alignment of all available MutS sequences and inferring an evolutionary tree from that alignment. The evolutionary tree revealed that there were two subfamilies of MutS homologs in bacteria, one containing the "normal" MutS-like proteins known to be involved in mismatch repair and the other containing the *H. pylori* protein along with a few others. None of the proteins in this second subfamily had ever been studied experimentally and all were only distantly related to the "normal" MutS subfamily. Given this finding along with the observation that *H. pylori* lacked a MutL homolog, we wrote in our *Nature Medicine* article (Eisen et al., 1997) that it was premature to predict that mismatch repair would be found in this species. I followed this up with a more comprehensive evolutionary study (Eisen, 1998b) that came to the same conclusion.

I would like to point out that this was not simply an esoteric exercise. Mismatch repair has great significance due to its role in modifying the mutation rate. Without mismatch repair, an organism's mutation rate usually goes way up (and in addition the rate of acquisition of DNA from other organisms also tends to go up). This has important implications for the evolution of virulence, pathogenicity, and drug resistance. Many papers published since this initial analysis have confirmed that *H. pylori* actually does have a high baseline mutation rate (Kang et al., 2006). In fact, the entire group of epsilon proteobacteria (of which *H. pylori* is a member) does not have a normal MutS homolog. Thus, the question arises: Do all of these organisms have high mutation rates? Or have they evolved some compensatory process that reduces mutation rate even without mismatch repair? At least from current data, it seems that many members of this group do have somewhat elevated mutation rates. For example, when the Sanger Center was sequencing the genome of a close relative of *H. pylori*, *Campylobacter jejuni* (which also does not encode a normal MutS homolog), even in the few generations required to grow up a sample of this strain for sequencing, many mutations were acquired (Parkhill et al., 2000). This suggests that the mutation rate for this strain is quite high. Awareness of this dynamic is vitally important when designing therapeutics to target organisms that lack mismatch repair. This example illustrates how evolutionary analysis of a gene found in a genome can not only tell us something about the biology of that organism, but can also help us to predict its evolvability.

This *H. pylori* story is but one of many that demonstrate the value of including evolutionary analysis when predicting gene function. In this regard, I must point out that I am far from unique in holding this view. For example, while I was working on the use of phylogenetic trees, multiple groups were showing how classifying proteins into families and subfamilies was critical for predicting function (Sonnhammer et al., 1997; Tatusov et al., 2000). My approach to this functional prediction was somewhat different from these subfamily- or ortholog-focused approaches in that I have argued that one needs to actively use the tree itself by using an approach known as character state reconstruction (Figure 5-10). Character state reconstruction is a commonly used method in phylogenetics whereby one can infer for particular traits (also known as characters) the history of change between different forms of those traits (also known as states). Normally, character state reconstruction is used to infer information about ancestral nodes in a tree (e.g., the common ancestor of two extant organisms), but it can also be used to infer the likely state of modern organisms. It is relatively straightforward to use these methods to infer information about protein function by treating each protein much as you would treat different organisms. Importantly, not only can one infer likely functions for proteins using this approach, but this has a benefit over subfamily classification approaches in that it is less likely to make incorrect predictions of function (such as, when function changes rapidly; Eisen, 1998a). It is worth noting that this adaptation of character state reconstruction methods for predicting the functions of uncharacterized genes is analogous to predicting the biology of a species based on the position of that organism in the tree of life. Such predictions tend to work better for gene function, in a large part because organism-level biology can change much more rapidly than the function of specific genes.

Regardless of whether one uses my character state-based approach or one of the subfamily-based approaches, it is clear that adding information about the evolutionary history of a gene can help predict its functions. Last, and perhaps most important, methods that make use of evolutionary information have been automated (Brown et al., 2007; Haft et al., 2001; Tatusov et al., 2000; Zmasek and Eddy, 2002) and thus can be employed more readily and on larger genomic data sets.

FIGURE 5-10 Phylogenetic prediction of gene function. Outline of a phylogenomic methodology. In this method, information about the evolutionary relationships among genes is used to predict the functions of uncharacterized genes (see text for details). Two hypothetical scenarios are presented and the path of trying to infer the function of two uncharacterized genes in each case is traced. (A) A gene family has undergone a gene duplication that was accompanied by functional divergence. (B) Gene function has changed in one lineage. The true tree (which is assumed to be unknown) is shown at the bottom. The genes are referred to by numbers (which represent the species from which these genes come) and letters (which in A represent different genes within a species). The thin branches in the evolutionary trees correspond to the gene phylogeny and the thick gray branches in A (bottom) correspond to the phylogeny of the species in which the duplicate genes evolve in parallel (as paralogs). Different colors (and symbols) represent different gene functions; gray (with hatching) represents either unknown or unpredictable functions.
SOURCE: Reprinted from Eisen (1998a) with permission from Cold Spring Harbor Laboratory Press.


EXAMPLE A
METHOD
EXAMPLE B
2A
CHOOSE GENE(S) OF INTEREST
5
3A 2B 1A 2A 1B 3B
IDENTIFY HOMOLOGS
2 1 3 4 5 6
ALIGN SEQUENCES
1A 2A 3A 1B 2B 3B
Duplication?
CALCULATE GENE TREE
1 2 3 4 5 6
1A 2A 3A 1B 2B 3B
Duplication?
OVERLAY KNOWN FUNCTIONS ONTO TREE
1 2 3 4 5 6
1A 2A 3A 1B 2B 3B
Duplication?
INFER LIKELY FUNCTION OF GENE(S) OF INTEREST
1 2 3 4 5 6
Ambiguous
Species 1 1A 1B
Species 2 2A 2B
Species 3 3A 3B
Duplication
ACTUAL EVOLUTION (ASSUMED TO BE UNKNOWN)
1 2 3 4 5 6

Phylogenomics and Novelty II: Recent Evolution

The methods for predicting function outlined above focus on making use of known information about some genes to predict the functions of uncharacterized genes. These methods do not work well, or even at all, if completely novel functions have arisen in an organism over short evolutionary time scales. Fortunately, over the last few years researchers have developed suites of methods to scan through genomic data for evidence of recent evolutionary diversification. Thus, my second phylogenomic tale relates to how knowledge about the origin of novelty helps us both carry out and interpret these scans.

The key to leveraging information about recent evolutionary events is to first get an understanding of how new functions arise on short time scales. Fortunately, we know a decent amount about this and have heard a great deal of recent new insights at this meeting. Examples include clustered regularly interspaced short palindromic repeats (CRISPRs) loci, which appear to be immune-system analog in bacteria and archaea that provides for immunity from phage, the rapid loss of genes that are not under strong positive selection, the use of contingency loci to rapidly change the sequence of a protein, and so forth. In fact, many of these phenomena have been either discovered or characterized in detail through comparative genomic analysis of closely related organisms.

Based upon this we can design a relatively simple process for taking a genome and identifying recent events in its history: sequence the genome and the genomes of some close relatives; compare the genomes to each other (including documentation of gene order conservation, gene gain and loss, gene duplication, and generation of simple polymorphisms); and then catalog the variation into different classes that correspond to different mechanisms of novelty generation. For example, polymorphisms in protein coding regions can be classified into synonymous (do not change protein sequence) and non-synonymous (change amino acid sequence) and then the pattern of synonymous versus non-synonymous substitutions can be used to screen a genome for the selective pressure different genes are under. Similarly, one can build evolutionary trees of all genes in a genome and look for those with longer branches in one lineage over another as evidence for an acceleration of evolutionary rate (Pollard et al., 2006). This sort of logic can be applied to just about any type of recent evolutionary event in genomes. Here I go into a bit more detail about how one can use this approach focusing on recent gene duplication events.

We know from the classic work of Ohta (2000) and others that gene duplication followed by subsequent divergence of the duplicates is a very important mechanism for the generation of novelty in virtually all organisms. Thus, to identify those genes within a lineage that are most likely to have recently diversified functions, we can turn this around and look for recent duplications. We did this by scanning complete genomes, looking for gene families that are expanded in one lineage compared to related lineages. As far as I know, we were the first to use this method when we applied it to the *Deinococcus radiodurans* genome (White et al., 1999). Subsequently, this general approach has been used in the analysis of many genomes and developed into a robust tool for characterizing them (Jordan et al., 2001).

The work I am going to describe here involves analysis of the genome of *Vibrio cholerae*. John Heidelberg had led a project to sequence this genome at The Institute for Genomic Research (TIGR) and asked me for help in carrying out some analyses (Heidelberg et al., 2000). One thing I did was to scan the genome for gene families that had undergone lineage-specific duplications (i.e., duplications that occurred since the organism last shared a common ancestor with any other organism for which we also had the complete genome sequence available). This was done "function blind"—meaning we simply analyzed the raw sequence data and not the known or predicted functions of genes. We found something very striking. In one gene family, the number of genes in this species was much greater than that in other related species. More importantly, the "extra" genes in *V. cholerae* were apparently the result of multiple rounds of gene duplication that occurred in the evolutionary branch leading up to this species (i.e., since it diverged from other lineages for which genomes were available). This family encoded the methyl-accepting chemotaxis proteins (MCPs) which were predicted to be involved in sensing and responding to chemical gradients in the environment (Figure 5-11).

Given the known biology of *V. cholerae* as an aquatic microbe, it seemed even more likely that this protein family might indeed have experienced recent evolutionary adaptations. Of course, not all duplications are related to evolutionary diversification, but with a genome encoding more than 4,000 proteins, identifying a candidate subset to pursue with more careful informatics and with experimental studies was definitely helpful.

Phylogenomics III: Uncharacterized Genes

Both of the approaches described above predict the function of particular genes by making use of experimental information about homologs of those genes. Unfortunately, this does not always work well, for many reasons. For instance, much of the time a gene of interest will have homologs in other species but none of those homologs have been studied experimentally. Such genes, known as "conserved hypothetical" genes, pose a significant challenge for function prediction. Fortunately, over the last 10 years, many new methods have been developed that are particularly useful for characterizing their functions (see Marcotte, 2000, for review). Since these methods make use of other types of experimental information (such as coexpression patterns, protein-protein interaction networks) or computational analysis (including chromosomal location, shared promoter sequences, protein domain patterns), they are generally known as "nonhomology" methods.

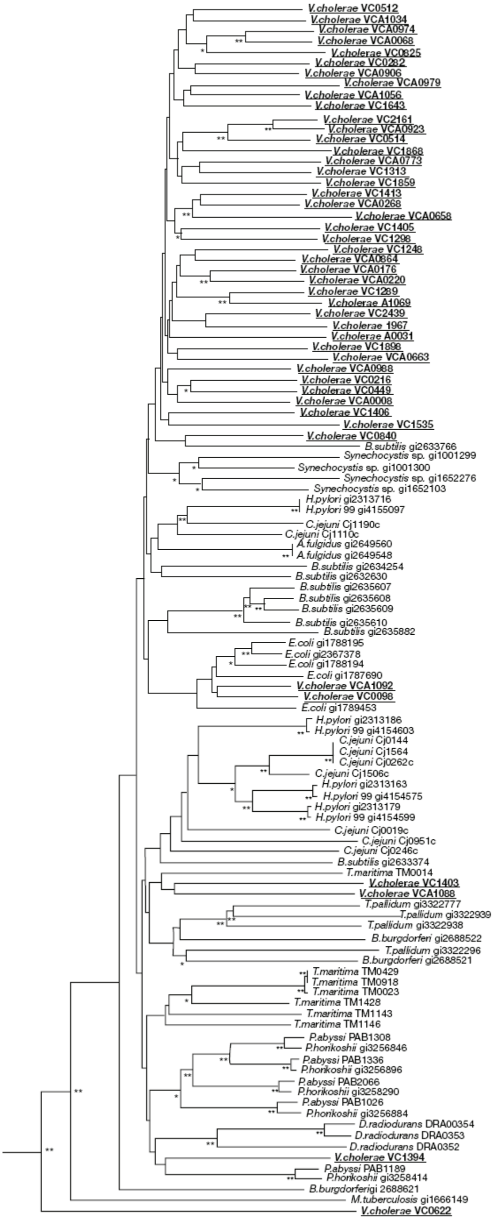


FIGURE 5-11 Phylogenetic tree of methyl-accepting chemotactic protein (MCP) homologs in completed genomes. Homologs of MCPs were identified by FASTA3 searches of all available complete genomes. Amino acid sequences of the proteins were aligned using CLUSTALW, and a neighbor-joining phylogenetic tree was generated from the alignment using the PAUP* program (using a PAM-based distance calculation). Hypervariable regions of the alignment and positions with gaps in many of the sequences were excluded from the analysis. Nodes with significant bootstrap values are indicated: two asterisks, .70 percent; asterisk, 40 ± 70 percent.

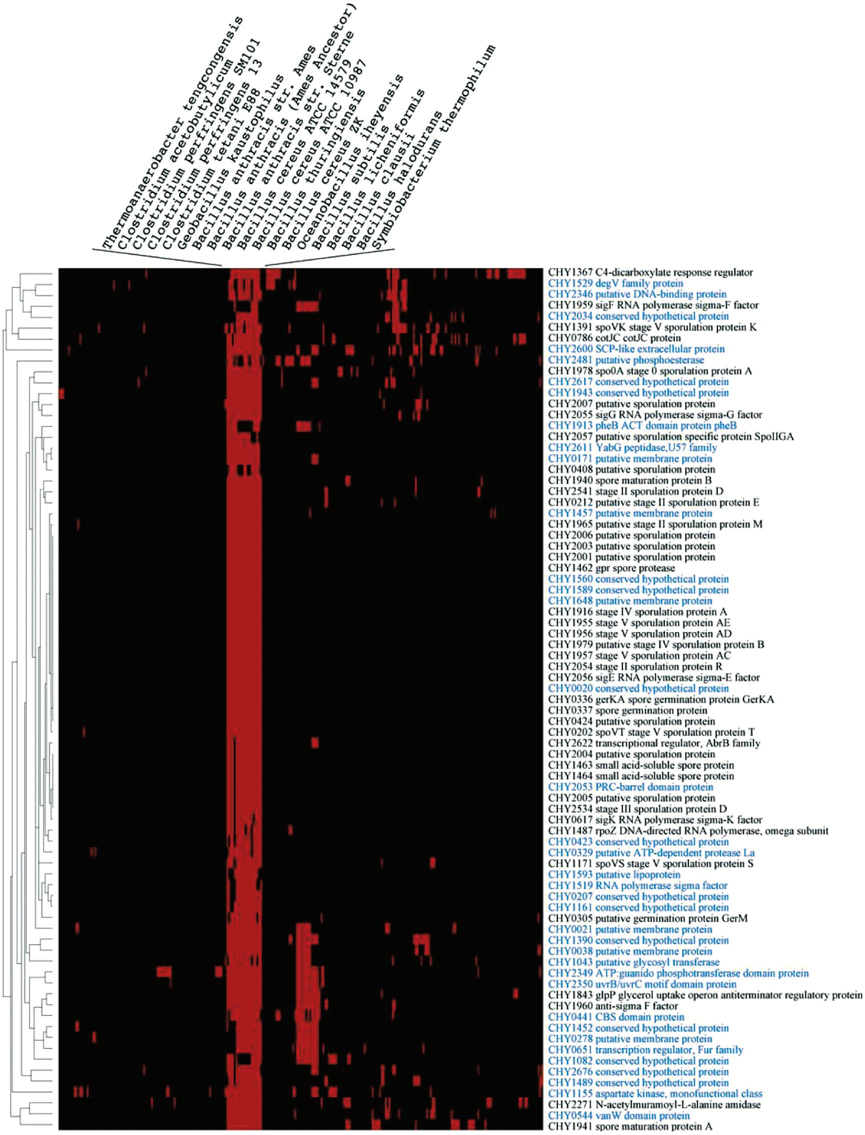
SOURCE: Adapted from Heidelberg et al. (2002), with permission from Macmillan Publishers Ltd. Copyright 2002.

I'm going to introduce you to one of them, my favorite: phylogenetic profiling (Pellegrini et al., 1999).

In phylogenetic profiling, we first determine the distribution of genes of interest across many species. Genes with similar patterns of distribution are then grouped together. The underlying idea here is that often several genes interact in some way, for example, all being subunits of a complex protein or being involved in carrying out a particular process such as methanogenesis. For one gene to be functional, all must be present in an organism. Such genes would thus tend to be found in groupings that have similar patterns of distribution across species. It is important to point out that when interpreting these profiles, one must take into account two key processes in the evolution of microbial genomes. First, unless genes are used or are under strong selection to be maintained, they tend to disappear. Second, microbes don't just inherit genes vertically within a lineage; they also acquire genes from other organisms by horizontal gene transfer. Significantly, when genes that work together are acquired horizontally, they tend to all get added or deleted simultaneously, or nearly simultaneously, with the result that when we compare genomes, we see that all members of such a group are either present or absent.

Here's how one actually carries out phylogenetic profiling. You start with a set of genes in which you are interested, perhaps all the genes in the complete genome sequence of "your" organism. You then compare them against each complete genome sequence in a genome database by asking a simple yes-or-no question: For each gene in your organism, is there a homolog in the other genome? After you have done this for every gene in your genome, you create a profile for each gene by plotting its presence or absence across all the species. With such profiles in hand, one can then identify genes with similar profiles. One way to do this is to simply cluster genes by their profiles and look for tight clusters of genes with highly similar distribution patterns. An example of this is shown in Figure 5-12 in which each row corresponds to a gene and each column represents one species. Conceptually, this is analogous to microarray clustering of gene expression patterns. In fact, microarray clustering software is often used for analyzing phylogenetic profiles.

Once you have such groupings of genes with similar cross-species distribution patterns, you can then use them to aid in predicting gene functions. For example, we used phylogenetic profiling to analyze the genome of the bacterium *Carboxydothermus hydrogenoformans* (Wu et al., 2005). There we found a very tight cluster of genes shared among many sporulating species (e.g., *Bacillus subtilis*) but absent from species that did not sporulate—even if closely related. Many of the gene families in this cluster were known to be involved in sporulation in other species. Based on this information, we predicted that *C. hydrogenoformans* had the ability to sporulate, and indeed we subsequently confirmed this experimentally. Moreover, our analysis revealed that there were also many other gene families of unknown function that were shared by sporulating species


Thermoanaerobacter tengcongensis
Clostridium acetobutylicum
Clostridium perfringens SM101
Clostridium perfringens 13
Clostridium tetani E88
Geobacillus kaustophilus
Bacillus anthracis str. Ames
Bacillus anthracis (Ames Ancestor)
Bacillus anthracis str. Sterne
Bacillus cereus ATCC 14579
Bacillus cereus ATCC 10987
Bacillus thuringiensis
Bacillus cereus ZK
Oceanobacillus iheyensis
Bacillus subtilis
Bacillus licheniformis
Bacillus clausii
Bacillus halodurans
Symbiobacterium thermophilum
CHY1367 C4-dicarboxylate response regulator
CHY1529 degV family protein
CHY2346 putative DNA-binding protein
CHY1959 sigF RNA polymerase sigma-F factor
CHY2034 conserved hypothetical protein
CHY1391 spoVK stage V sporulation protein K
CHY0786 cotJC cotJC protein
CHY2600 SCP-like extracellular protein
CHY2481 putative phosphoesterase
CHY1978 spo0A stage 0 sporulation protein A
CHY2617 conserved hypothetical protein
CHY1943 conserved hypothetical protein
CHY2007 putative sporulation protein
CHY2055 sigG RNA polymerase sigma-G factor
CHY1913 pheB ACT domain protein pheB
CHY2057 putative sporulation specific protein SpoIIGA
CHY2611 YabG peptidase,U57 family
CHY0171 putative membrane protein
CHY0408 putative sporulation protein
CHY1940 spore maturation protein B
CHY2541 stage II sporulation protein D
CHY0212 putative stage II sporulation protein E
CHY1457 putative membrane protein
CHY1965 putative stage II sporulation protein M
CHY2006 putative sporulation protein
CHY2003 putative sporulation protein
CHY2001 putative sporulation protein
CHY1462 gpr spore protease
CHY1560 conserved hypothetical protein
CHY1589 conserved hypothetical protein
CHY1648 putative membrane protein
CHY1916 stage IV sporulation protein A
CHY1955 stage V sporulation protein AE
CHY1956 stage V sporulation protein AD
CHY1979 putative stage IV sporulation protein B
CHY1957 stage V sporulation protein AC
CHY2054 stage II sporulation protein R
CHY2056 sigE RNA polymerase sigma-E factor
CHY0020 conserved hypothetical protein
CHY0336 gerKA spore germination protein GerKA
CHY0337 spore germination protein
CHY0424 putative sporulation protein
CHY0202 spoVT stage V sporulation protein T
CHY2622 transcriptional regulator, AbrB family
CHY2004 putative sporulation protein
CHY1463 small acid-soluble spore protein
CHY1464 small acid-soluble spore protein
CHY2053 PRC-barrel domain protein
CHY2005 putative sporulation protein
CHY2534 stage III sporulation protein D
CHY0617 sigK RNA polymerase sigma-K factor
CHY1487 rpoZ DNA-directed RNA polymerase, omega subunit
CHY0423 conserved hypothetical protein
CHY0329 putative ATP-dependent protease La
CHY1171 spoVS stage V sporulation protein S
CHY1593 putative lipoprotein
CHY1519 RNA polymerase sigma factor
CHY0207 conserved hypothetical protein
CHY1161 conserved hypothetical protein
CHY0305 putative germination protein GerM
CHY0021 putative membrane protein
CHY1390 conserved hypothetical protein
CHY0038 putative membrane protein
CHY1043 putative glycosyl transferase
CHY2349 ATP:guanido phosphotransferase domain protein
CHY2350 uvrB/uvrC motif domain protein
CHY1843 glpP glycerol uptake operon antiterminator regulatory protein
CHY1960 anti-sigma F factor
CHY0441 CBS domain protein
CHY1452 conserved hypothetical protein
CHY0278 putative membrane protein
CHY0651 transcription regulator, Fur family
CHY1082 conserved hypothetical protein
CHY2676 conserved hypothetical protein
CHY1489 conserved hypothetical protein
CHY1155 aspartate kinase, monofunctional class
CHY2271 N-acetylmuramoyl-L-alanine amidase
CHY0544 vanW domain protein
CHY1941 spore maturation protein A

and absent from nonsporulating ones. Such genes were likely candidates for carrying out novel sporulation-associated activities. A bit of confirmation came just as we were finishing our paper. Richard Losick at Harvard published a set of studies on sporulation in *B. subtilis* that identified a few new sporulation genes (Eichenberger et al., 2004; Silvaggi et al., 2004), and many of our candidates were in their list of novel sporulation genes. Perhaps most interestingly, many of our candidates were still not identified as likely sporulation genes and likely represent novel sporulation-associated functions yet to be characterized.

I note that the approach of phylogenetic profiling can be strengthened by modifying the basic yes-or-no question. Instead of asking if there is a *homolog* of your gene present in another species, you want to ask if there is an *ortholog*, thus using some evolutionary information to improve your clustering (Eisen and Wu, 2002). With either method, phylogenetic profiling is a powerful tool for finding sets of genes that function in related processes or in a pathway. Although it does not characterize their biochemical activity well, it can provide insight into the process in which they participate (e.g., sporulation) and thus guide experimental studies. As we sequence more and more genomes, this method will become more and more informative.

Phylogenomics IV: Acquisition of Function from Others

There are two basic strategies by which organisms evolve new functions. One option is through modification of their own genome (e.g., mutation, gene

FIGURE 5-12 Phylogenetic profile analysis of sporulation in *Carboxydothermus hydrogenoformans*. For each protein encoded by the *C. hydrogenoformans* genome, a profile was created of the presence or absence of orthologs of that protein in the predicted proteomes of all other complete genome sequences. Proteins were then clustered by the similarity of their profiles, thus allowing the grouping of proteins by their distribution patterns across species. Examination of the groupings showed one cluster consisting mostly of homologs of sporulation proteins. This cluster is shown with *C. hydrogenoformans* proteins in rows (and the predicted function and protein ID indicated on the right) and other species in columns with the presence of an ortholog indicated in red and its absence in black. The tree to the left represents the portion of the cluster diagram for these proteins. Note that most of these proteins are found only in a few species represented in red columns near the center of the diagram. The species corresponding to these columns are indicated. We also note that though most of the proteins in this cluster, for which functions can be predicted, are predicted to be involved in sporulation and some have no predictable functions (highlighted in blue). This indicates that functions of these proteins' homologs have not been characterized in other species. Since these proteins show similar distribution patterns to so many proteins with roles in sporulation, we predict that they represent novel sporulation functions.
SOURCE: Wu et al. (2005).

duplication, domain swapping, invention of new genes), but these processes can sometimes be quite slow. In many cases, it is much easier instead to acquire the function from another organism that already has it. How is this done? By acquisition or affiliation. In other words, they can acquire the requisite genes via sex or lateral gene transfer, or they can gain access to the products of those genes through some type of affiliation with organisms that have those functions. Such affiliations include long-term symbioses. Symbioses are categorized as being parasitic (when one partner obviously benefits and the other is harmed), commensal (where one benefits and the other is unaffected), or mutualistic (where both benefit), but often we do not actually know the full extent of mutual impacts.

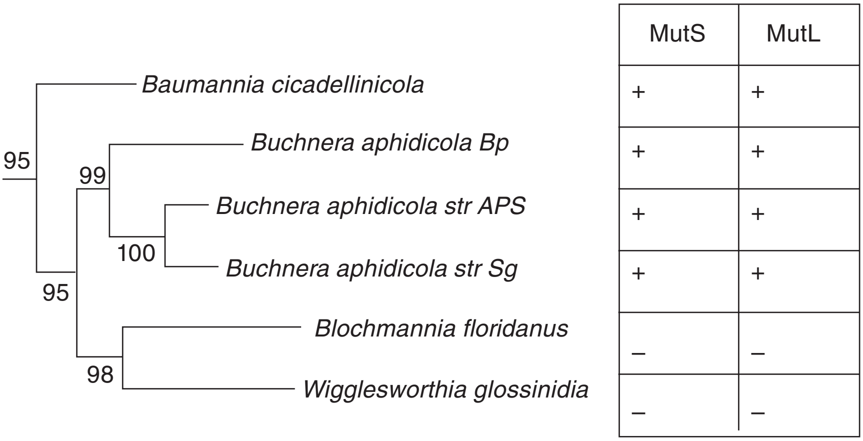
I am going to give an example of function acquisition by symbiosis, and demonstrate how genomic studies, combined with an understanding of the biology and evolution of the symbiosis, can aid in functional predictions. One partner in this symbiosis is the glassy-winged sharpshooter. This insect, like other sharpshooters, is an obligate xylem feeder that makes its living by feeding on the fluids in the xylem portion of the circulatory system of the host plant. This particular species has received special attention because it is a vector for Pierce's disease, a nasty problem in grape vineyards. The disease agent is a bacterium, *Xylella fastidiosa*, that infects the xylem and can be transmitted between plants by the sharpshoorters, much like bloodborne pathogens are transmitted between animal hosts (see Chatterjee et al., 2008, for a review).

Obligate sap-feeding insects face a serious challenge. As part of their defenses, many plants make their sap less useful to sap-feeding insects by removing some nutrients that are essential for animals. For example, the essential amino acids (that all animals cannot synthesize and thus require in their diet) tend to be present in very low concentration in phloem sap. To counter this, many obligate phloem-feeding insects have bacterial symbionts living inside specialized cells in their gut. The insect provides the bacteria with sugars from sap, and the bacteria, in turn, make amino acids for their hosts. Xylem sap, which moves from the roots to the rest of the plant, tends to be even more nutrient-poor than phloem sap, and obligate xylem feeders also have bacterial symbionts living inside specialized cells in the gut (see Moran et al., 2008, for a review on heritable symbionts).

When we started our project, obligate phloem-feeding insects, such as aphids, had already been studied extensively, but much less was known about the obligate xylem-feeders. We (especially our collaborator, Nancy Moran) thought there might be a different twist to the story for the bacterial symbionts living in xylem-feeding hosts. At that time, all the species of sharpshooters examined had been found to host *Baumannia cicadellinicola*, a close relative of the symbionts that make amino acids for the aphids (Moran et al., 2003). Our first step was to apply shotgun genome sequencing methods to the DNA obtained from endosymbiont-containing tissue dissected from this sharpshooter. Using this approach we were able to determine the complete genome of *B. cicadellinicola*. Examination of the genome revealed many very interesting things (Wu et al., 2006).

First, we found that this organism had many of the hallmarks typical of intracellular symbionts: a small genome, low G+C content, and high evolutionary rates. As an aside, the high evolutionary rates often seen in intracellular symbionts are thought to be due, in large part, to the small effective population sizes for intracellular organisms. However, we found significant variation in the rate of evolution among endosymbionts, with the highest rates tending to be found in those that lack homologs of mismatch repair genes (Figure 5-13). Adaptation to an intracellular existence is typically accompanied by marked reduction in genome size, probably due to random forces. When important DNA repair genes are lost, mutation rates may go up—an evolutionarily significant consequence of their small genome size. Furthermore, whereas free-living species would have opportunities to reacquire the repair genes from other organisms, this is very unlikely for intracellular ones, isolated as they are. Thus another evolutionary consequence of an intracellular existence is reduced evolvability by means of lateral gene transfer.

Secondly, further examination of the genome and prediction of gene functions revealed pathways for synthesizing diverse vitamins and cofactors, suggesting that this symbiont was also helping its xylem-feeding host to deal with a very nutrient-poor diet. Based on our prior knowledge of these types of symbioses, we expected to find pathways for the synthesis of the essential amino acids required by the sharpshooter—but we could not find any. In thinking about the mechanisms for the evolution of novelty, it seemed unlikely to us that this host, the



	MutS	MutL
Baumannia cicadellinicola	+	+
Buchnera aphidicola Bp	+	+
Buchnera aphidicola str APS	+	+
Buchnera aphidicola str Sg	+	+
Blochmannia floridanus	–	–
Wigglesworthia glossinidia	–	–

FIGURE 5-13 There is significant variation in the rate of evolution among endosymbionts, with the highest rates tending to be found in those that lack homologs of mismatch repair genes.
SOURCE: Adapted from Wu et al. (2006).

glassy-winged sharpshooter, would have evolved the ability to synthesize essential amino acids, given that this capability has never been found, as far as I know, in any animal species. Nevertheless, the observations were that the sharpshooter eats only xylem sap, xylem does not contain the essential amino acids, and the genes for essential amino acid synthesis pathways were not present in the genome of either the sharpshooter or its *Baumannia* endosymbiont. We were vexed.

There were three possibilities we considered that could reasonably explain this conundrum. One was that the sharpshooter was acquiring amino acids from other food sources. This seemed unlikely as sharpshooters are generally considered to be obligate xylem sap feeders. A second possibility was that the glassy-winged sharpshooter was getting the essential amino acids from the xylem sap. Though we could not rule this out, it seemed unlikely because there should be strong selection on the plants to keep essential amino acids out of the xylem sap and because xylem generally was not known to have such amino acids. A third possibility was that another organism in the sharpshooter system was making amino acids. This seemed to be the most likely possibility especially since our collaborator Nancy Moran had just recently shown that there was a second type of bacterial symbiont living inside the guts of all sharpshooters (Moran et al., 2005). We had not paid much attention to this second type of symbiont since the *Baumannia* symbionts were so closely related to the *Buchnera* symbionts of aphids that provided all the nutritional supplements needed by their host to feed on phloem sap (and since these new symbionts were from a completely different phylum of bacteria).

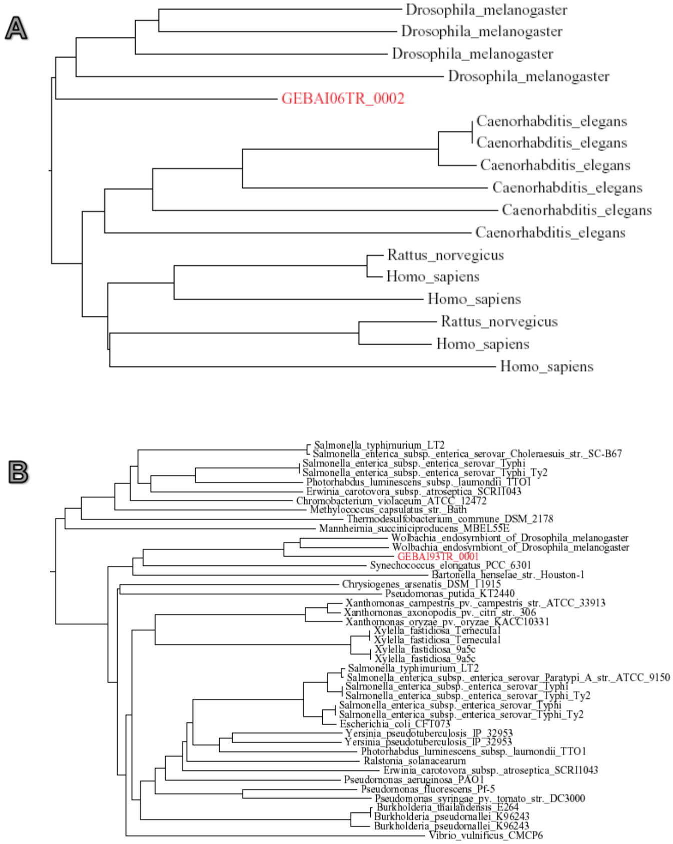
Fortunately, we had a quick, though somewhat dirty, way to test for the possibility that another organism in the system was making essential amino acids. To sequence the *Baumannia* genome we did not use a pure culture since these symbionts had never been grown in the lab. Instead, we had done a "metagenomics" project in which Nancy Moran's lab had dissected hundreds of sharpshooters and removed as carefully as possible the tissue that was known to contain the *Baumannia* symbionts. We then extracted DNA from this material and used it for whole-genome shotgun sequencing during which we sheared the DNA into moderately small pieces, cloned these pieces into a plasmid library, and then sequenced the ends of these plasmid clones. From these data we were able to generate a good assembly of the *Baumannia* genome, which we then finished with PCR and primer walking methods. The key for us was that not all of the sequence reads that we obtained were found to map to the *Baumannia* genome. Some came from other organisms in the sample. So the first thing we did was to look in these other data for genes that might be involved in synthesizing essential amino acids—and we immediately found a few.

So the next question was: From what organism did these genes originate? We knew there should be host DNA in the sample (although we thought it was unlikely that the host would be synthesizing essential amino acids since no animals are known to do so) and that there might also be DNA from the second symbiont as well as from other resident microbes. So what we needed to do was to sort the DNA sequence reads into which came from which organism. This sorting is commonly known as "binning" in metagenomic studies. We tried every binning method in use at the time including genome assembly, analysis of DNA base composition and word frequencies, examination of depth of coverage, and others. Unfortunately none of them worked well, most likely because we had very little coverage of the genomes from these other organisms. This is where phylogenomic approaches came in handy.

We decided to try to sort the sequence reads by phylogenetic analyses. So we took all the reads, identified all possible proteins or protein fragments that they could encode, then for these identified which had apparent homologs in sequence databases, and for those built phylogenetic trees. We then sorted the phylogenetic trees by which organism's genes showed up in the tree as the nearest neighbor of the protein or fragment.

Overall the trees showed only a few major patterns. In some (Figure 5-14A) the nearest neighbor was something from an animal. Thus, we concluded that the sequence reads in this "bin" likely corresponded to fragments of the host genome. In other trees, the nearest neighbor was a *Wolbachia* or some close relative (Figure 5-14B). Since *Wolbachia* (a type of bacteria related to *Rickettsia*) are common intracellular parasites of insects, we concluded that these reads came from *Wolbachia* that infected at least some of the insects that Nancy had dissected. Then there was a large collection of reads for which the trees showed a grouping with species in the Bacteroidetes phylum (Figure 5-14C). Because *Sulcia* was in this phylum, we concluded these were likely from the second symbiont.

We then asked: Of the potential essential amino acid synthesizing genes we had identified in some of the reads, to which of the bins did they belong? The answer was clear as day—all belonged to the *Sulcia* bin. We thus concluded that it was likely that the second symbiont was the provider of essential amino acids for the host and so we spent another year or so trying to finish the genome of this symbiont. Though we did not quite finish the genome, from the 130 or so kilobase pairs (kbp) of DNA we mapped to this organism, we found that it encoded in essence all the essential amino acid synthesis pathways (Wu et al., 2006); this was later confirmed by the complete genome (McCutcheon and Moran, 2007). What we had discovered was a dual symbioses where one symbiont (*Baumannia*) makes vitamins and cofactors and the other (*Sulcia*) makes essential amino acids, and together they supplement the nutrient-poor diet of the glassy-winged sharpshooter contributing to this organism's annoying ability to spread Pierce's disease. Most importantly for this article here, we would not have been able to sort out the data from the different organisms (and thus would not have discovered the dual symbioses) without phylogenetic analysis of the metagenomic data.


A
Drosophila_melanogaster
Drosophila_melanogaster
Drosophila_melanogaster
Drosophila_melanogaster
GEBAI06TR_0002
Caenorhabditis_elegans
Caenorhabditis_elegans
Caenorhabditis_elegans
Caenorhabditis_elegans
Caenorhabditis_elegans
Caenorhabditis_elegans
Rattus_norvegicus
Homo_sapiens
Homo_sapiens
Rattus_norvegicus
Homo_sapiens
Homo_sapiens
B
Salmonella_typhimurium_LT2
Salmonella_enterica_subsp._enterica_serovar_Choleraesuis_str._SC-B67
Salmonella_enterica_subsp._enterica_serovar_Typhi
Salmonella_enterica_subsp._enterica_serovar_Typhi_Ty2
Photorhabdus_luminescens_subsp._laumondii_TTO1
Erwinia_carotovora_subsp._atroseptica_SCRI1043
Chromobacterium_violaceum_ATCC_12472
Methylococcus_capsulatus_str._Bath
Thermodesulfobacterium_commune_DSM_2178
Mannheimia_succiniciproducens_MBEL55E
Wolbachia_endosymbiont_of_Drosophila_melanogaster
Wolbachia_endosymbiont_of_Drosophila_melanogaster
GEBAI93TR_0001
Synechococcus_elongatus_PCC_6301
Bartonella_henselae_str._Houston-1
Chrysiogenes_arsenatis_DSM_11915
Pseudomonas_putida_KT2440
Xanthomonas_campestris_pv._campestris_str._ATCC_33913
Xanthomonas_axonopodis_pv._citri_str._306
Xanthomonas_oryzae_pv._oryzae_KACC10331
Xylella_fastidiosa_Temecula1
Xylella_fastidiosa_Temecula1
Xylella_fastidiosa_9a5c
Xylella_fastidiosa_9a5c
Salmonella_typhimurium_LT2
Salmonella_enterica_subsp._enterica_serovar_Paratypi_A_str._ATCC_9150
Salmonella_enterica_subsp._enterica_serovar_Typhi
Salmonella_enterica_subsp._enterica_serovar_Typhi_Ty2
Salmonella_enterica_subsp._enterica_serovar_Typhi
Salmonella_enterica_subsp._enterica_serovar_Typhi_Ty2
Escherichia_coli_CFT073
Yersinia_pseudotuberculosis_IP_32953
Yersinia_pseudotuberculosis_IP_32953
Photorhabdus_luminescens_subsp._laumondii_TTO1
Ralstonia_solanacearum
Erwinia_carotovora_subsp._atroseptica_SCRI1043
Pseudomonas_aeruginosa_PAO1
Pseudomonas_fluorescens_Pf-5
Pseudomonas_syringae_pv._tomato_str._DC3000
Burkholderia_thailandensis_E264
Burkholderia_pseudomallei_K96243
Burkholderia_pseudomallei_K96243
Vibrio_vulnificus_CMCP6

Phylogenomics V: Knowing What We Do Not Know

As has often been heard at this workshop, Lederberg was very fond of emphasizing that we need to know what we do not know. In that spirit, I want to discuss how knowing what we do not know can help with functional predictions. One aspect of what we do not know that influences our ability to make useful functional predictions is that genome-sequencing projects are highly biased in terms of what types of organisms have been sequenced. For example, I and many others noticed a few years ago (Eisen, 2000; Hugenholtz, 2002) that most of the genomes of bacteria were coming from just three of the 40+ phyla of bacteria (Figure 5-15). The same trend was seen in Archaea and microbial eukaryotes. So based on this we applied for, and in 2002 received, a grant from the "Assembling the Tree of Life" program at the National Science Foundation (NSF) to sequence

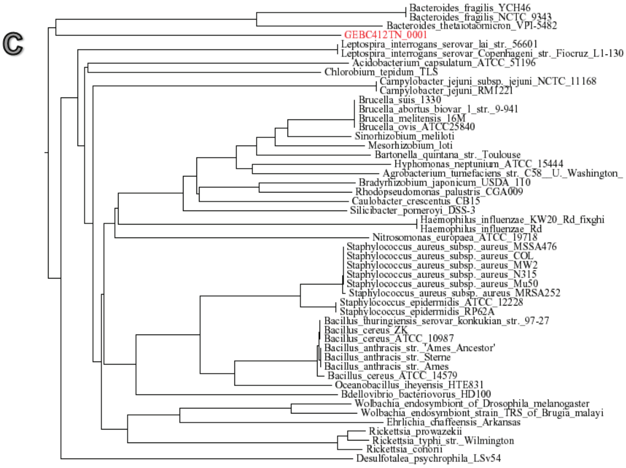

C
Bacteroides_fragilis_YCH46
Bacteroides_fragilis_NCTC_9343
Bacteroides_thetaiotaomicron_VPI-5482
GEBC412TN_0001
Leptospira_interrogans_serovar_lai_str._56601
Leptospira_interrogans_serovar_Copenhageni_str._Fiocruz_L1-130
Acidobacterium_capsulatum_ATCC_51196
Chlorobium_tepidum_TLS
Campylobacter_jejuni_subsp._jejuni_NCTC_11168
Campylobacter_jejuni_RM1221
Brucella_suis_1330
Brucella_abortus_biovar_1_str._9-941
Brucella_melitensis_16M
Brucella_ovis_ATCC25840
Sinorhizobium_meliloti
Mesorhizobium_loti
Bartonella_quintana_str._Toulouse
Hyphomonas_neptunium_ATCC_15444
Agrobacterium_tumefaciens_str._C58__U._Washington_
Bradyrhizobium_japonicum_USDA_110
Rhodopseudomonas_palustris_CGA009
Caulobacter_crescentus_CB15
Silicibacter_pomeroyi_DSS-3
Haemophilus_influenzae_KW20_Rd_fixghi
Haemophilus_influenzae_Rd
Nitrosomonas_europaea_ATCC_19718
Staphylococcus_aureus_subsp._aureus_MSSA476
Staphylococcus_aureus_subsp._aureus_COL
Staphylococcus_aureus_subsp._aureus_MW2
Staphylococcus_aureus_subsp._aureus_N315
Staphylococcus_aureus_subsp._aureus_Mu50
Staphylococcus_aureus_subsp._aureus_MRSA252
Staphylococcus_epidermidis_ATCC_12228
Staphylococcus_epidermidis_RP62A
Bacillus_thuringiensis_serovar_konkukian_str._97-27
Bacillus_cereus_ZK
Bacillus_cereus_ATCC_10987
Bacillus_anthracis_str._'Ames_Ancestor'
Bacillus_anthracis_str._Sterne
Bacillus_anthracis_str._Ames
Bacillus_cereus_ATCC_14579
Oceanobacillus_iheyensis_HTE831
Bdellovibrio_bacteriovorus_HD100
Wolbachia_endosymbiont_of_Drosophila_melanogaster
Wolbachia_endosymbiont_strain_TRS_of_Brugia_malayi
Ehrlichia_chaffeensis_Arkansas
Rickettsia_prowazekii
Rickettsia_typhi_str._Wilmington
Rickettsia_conorii
Desulfotalea_psychrophila_LSv54

FIGURE 5-14 Phylogenetic trees of putative proteins encoded by single sequence reads of DNA isolated from symbiont-containing tissue of the glassy-winged sharpshooter. Trees were constructed by aligning putative proteins encoded by the reads to homologs from complete genome sequences. (A) Phylogenetic tree for read GEBA106TR, translated in the second reading frame (thus the label *002*). Note how the encoded protein groups in the tree with proteins from *Drosophila*. This read is likely from a piece of DNA from the host genome. (B) Phylogenetic tree for read GEBA193TR_001. Note how the encoded protein groups with sequences from *Wolbachia*; thus, this read likely corresponds to DNA from a *Wolbachia* infecting one of the dissected sharpshooters. (C) Phylogenetic tree of read GEBA412TN_001. Note how the encoded protein groups in trees with *Bacteroidetes* species which are relatives of the second (*Sulcia)* symbiont.
SOURCE: Based on data in Wu et al. (2006).

the first genomes from representatives of 8 phyla of bacteria. We have now finished this project and are in the process of writing up a series of papers on our findings. Yet even from the initial analyses, what was abundantly clear was that a single genome from these phyla was simply not enough. Each phylum represents something on the order of 1 billion to 2 billion years of evolution, and a lot happens in that time in bacteria. So a single genome cannot do justice to the diversity of genes and features of each phylum.

Based on this, in collaboration with the Joint Genome Institute (a Department of Energy [DOE]-funded genome center), we have started a new initiative to really fill in the genomes from across the tree of life. This Genomic Encyclopedia of Bacteria and Archaea (GEBA)[8] is just getting started, with 100 genomes

[8] See http://www.jgi.doe.gov/programs/GEBA/ for more detail.

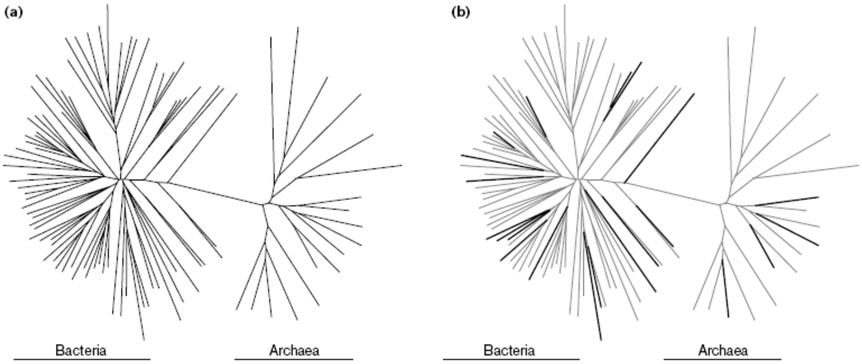


FIGURE 5-15 The diversity of Bacterial and Archaeal species for which complete genomes are available is still poor. (a) Phylogenetic tree, based on rRNA sequences, of representatives of many major Bacterial and Archaeal lineages. (b) Lineages for which complete genomes are available are highlighted. The rRNA tree for representative species was downloaded from the Ribosome Database Project (available at http://www.cme.msu.edu/RDP/html/index.html).
SOURCE: Reprinted from Eisen (2000) with permission from Elsevier. Copyright 2000.

being sequenced from across the tree in the first year. Already the results are quite convincing that sampling from across the tree leads to enormous benefits. For example, the phylogenetic profiling method outlined above works best when you have sampling of diverse genomes from different phylogenetic groups. Adding these GEBA genomes to the mix makes phylogenetic profiling work much better.

Sampling from across the tree will take some effort because there are many, many, major groups of Bacteria, Archaea, and microbial Eukaryotes, and many of these do not have any cultured representatives. However, the benefits will likely be enormous.[9]

It is important to point out however that just having genome sequences from across the tree is not sufficient. Functional information from diverse organisms is also critical. I give one example here of why. When I first went to TIGR, Owen White was in charge of a project to sequence the genome of the bacterium *Deinococcus radiodurans*, the most radiation-resistant organism known. This was very exciting to me since I did my Ph.D. research in part on the evolution of radiation resistance. So I volunteered to help Owen analyze the genome. Since there was some experimental evidence that active DNA repair processes contributed to the

[9] I note it will be good to sample from across viral diversity, although since there is no phylogenetic tree linking all viruses it is unclear exactly how to do this sampling.

resistance of this organism, I spent some time looking for likely DNA repair genes in the genome, making use of the phylogenomic approaches I had been advocating. Indeed we were able to find many genes that appeared likely to be involved in DNA repair processes.

The problem was that the list we came up with was very similar to the list that we could make for nonradiation-resistant organisms such as *E. coli* and *B. subtilis*. However, a little thinking about what we did not know helps explain this. Imagine if sometime in the recent history of *D. radiodurans* a novel DNA repair gene evolved. The method we were using to look for DNA repair genes would not have found this since we were looking primarily for homologs of genes that were shown in other species to be involved in DNA repair processes. Even using novel methods such as phylogenetic profiling would not necessarily help if the new genes in this species were not connected in any way to known DNA repair pathways. The problem here is that most experimental studies of repair genes in bacteria were done in two phyla and we would have a hard time identifying novel repair genes if they had been invented anywhere else in the bacterial tree of life.

This is a general lesson for all functional predictions. Such predictions rely upon some functional characterizations done in some organism. The more these functional studies are done across the tree of life, the better will our functional predictions become. Similarly, functional predictions rely in part upon comparative genome analysis; thus, the more genomes we have from across the tree, the better functional predictions we will get. Thus, knowing what we do not know is critical in guiding experimental and sequencing studies to get the most out of the diversity of organisms.

Summary: A Call for a Field Guide to the Microbes

Overall, what I have tried to do here is present examples of how evolutionary and genome analysis can be integrated into "phylogenomic" studies. I have focused on predicting functions of genes, but the benefits of phylogenomics extend to all aspects of the biology of microbes.

I should emphasize that this is not some radical or overly novel concept as the integration of phylogeny and function is well known to be critical for understanding the diversity of life. What I have tried to show here is that this is as true for genomics and micreobes as it is for physiology, behavior, genes, ecosystems and other arenas in which evolution has been shown to be a powerful tool. I should note that there is a third piece of information that is useful in addition to integrating phylogeny and function—the biogeographical patterns of the distributions of organisms. For microbes, figuring out the distribution patterns of organisms and the rules determining these patterns is one of the final frontiers. If we are able to integrate phylogeny, function and genomes, and biogeography, we will have something for microbes that is known to be useful in many other organisms—a field guide. A field guide to microbes would no doubt be useful in many arenas, and I am certain it would be a book that Josh Lederberg would have carried with him wherever he went.

Acknowledgments

I thank Merry Youle for assistance in the editing of this manuscript and Martin Wu and Dongying Wu for help in many of the genome analyses reported here. In addition, I thank the many people at TIGR and UC Davis who provided help in generating and analyzing the genome and metagenomes under discussion. This work was supported by the Defense Advanced Research Projects Agency under grant HR0011-05-1-0057.