

# Bacterial Genomics and Pathogen Evolution

David M. Raskin,<sup>1</sup> Rekha Seshadri,<sup>2</sup> Stefan U. Pukatzki,<sup>1</sup> and John J. Mekalanos<sup>1,\*</sup>

<sup>1</sup>Department of Microbiology & Molecular Genetics, Harvard Medical School, Boston, MA 02115, USA

<sup>2</sup>The Institute for Genomic Research, 9712 Medical Center Drive, Rockville, MD 20850, USA

\*Contact: [jmekalanos@hms.harvard.edu](mailto:jmekalanos@hms.harvard.edu)

DOI 10.1016/j.cell.2006.02.002

The availability of hundreds of bacterial genome sequences has altered the study of bacterial pathogenesis, affecting both design of experiments and analysis of results. Comparative genomics and genomic tools have been used to identify virulence factors and genes involved in environmental persistence of pathogens. However, a major stumbling block in the genomics revolution has been the large number of genes with unknown function that have been identified in every organism sequenced to date.

## Introduction

Over 250 bacterial genome sequences are currently available in public databases, representing hundreds of species as well as multiple strains of the same species. The study of these genomes by both computational and experimental approaches has significantly advanced our understanding of the physiology and pathogenicity of many microbes and provided insights into the mechanisms and history of genome evolution. This review will discuss specific examples, with an emphasis on how “postgenomic” methods have been utilized to identify genes that are essential for bacterial growth or pathogenesis. There are two general categories of postgenomic studies—those that are based primarily on analysis of sequence information itself and those that require some use of genomic sequence information but not based on direct sequence analysis. Direct sequence analysis has the advantage of enabling genome-level analysis of pathogens that are difficult to manipulate genetically and permitting examination of small differences such as single nucleotide polymorphisms (SNPs). However, the accuracy of bioinformatic genomic annotations remains unclear since there have been no reported attempts to confirm functional assignments of genes on a genome-wide scale. It is imperative not only to improve annotation quality but also to dovetail computational approaches with empirical studies, especially those that seek to unravel the functions of the numerous hypothetical genes still found in genomes.

Multiple genome-level empirical approaches have been developed for investigating gene function. Some do not require prior knowledge of genomic sequence, but this review will focus on those strategies that do, including a variety of microarray technologies, proteomic methods, and certain types of molecular genetic studies (e.g., genetic footprinting techniques). These strategies have the advantage of possessing some intrinsic experimental validation, even though data are collected on the genomic scale.

## Pathogenic Species in the Bacterial Genomic Era

Through a combination of critical insight and chance, pathogenic bacterial species were the first free-living organisms to be sequenced and thus have occupied

a unique place in the history of the genomic era. The genome of a laboratory strain of *Haemophilus influenzae* was the first bacterium to be sequenced and was followed by an isolate of *Mycoplasma genitalium* (Fleischmann et al., 1995; Fraser et al., 1995). In each of these cases, the genome sequences provided immediate insights into the biology of these bacterial species but less about pathogenicity in part because the strains selected were attenuated laboratory strains that had likely lost many virulence-related loci. Nonetheless, the genomes did provide some explanations for curious physiological properties. For instance, growth of *H. influenzae* in vitro requires large amounts of glutamate. The genome sequence showed the absence of three TCA cycle genes in this species. Glutamate can be converted to  $\alpha$ -ketoglutarate by glutamate dehydrogenase and then shunted into the remaining TCA cycle pathway, explaining the dependence on exogenously supplied glutamate. In the case of *M. genitalium*, the genomic sequence revealed an almost complete lack of enzymes involved in amino acid synthesis, de novo nucleotide biosynthesis, and fatty acid biosynthesis, providing an explanation for some of the undefined fastidious growth requirements of this species. Because both *H. influenzae* and *M. genitalium* grow in host tissues, these two examples illustrate that either the host milieu can be nutrient rich or the physiology of bacteria in vivo is very different from what is observed under laboratory growth conditions.

The field of comparative genomics next addressed questions that could only be approached after multiple genomes were available for analysis. For example, after the first two bacterial genomes were sequenced, Mushegian and Koonin (1996) tried to determine the minimum number of genes necessary for life by comparing them and determining which genes were present in both. *M. genitalium* contains only 468 protein-coding genes, while *H. influenzae* contains 1703 protein-coding genes. They identified 240 *M. genitalium* genes with orthologs in *H. influenzae* but found that there were still genes missing from several basic metabolic pathways. Based on the hypothesis that this could have resulted from nonorthologous gene displacement, they identified an additional 16 *M. genitalium*

genes to cover necessary functions and concluded that 256 genes were the minimum gene content necessary. More recently Gil et al. (2004) have used experimental data sets to eliminate more dispensable genes, while retaining the key biochemical pathways predicted to be essential for a free-living cell. They found 206 genes could make up a minimal gene set. Both of these approaches clearly vastly underestimate the number of genes required for growth in the “real world” of the host where all nutrients are seldom available and where host-mediated processes demand the specialized survival genes that are essential for pathogenicity.

The *H. influenzae* genome also provided the first example of how little we know about bacterial physiology. Of a total of 1743 annotated *H. influenzae* genes, 736 were of unknown function. Many of these hypothetical genes are now known to exist as orthologs in other bacterial genomes that have since been sequenced. In many bacteria, close to one-half of all putative genes have no known function. While the tools of comparative genomics can predict that some highly conserved “hypothetical” genes are likely to be important in microbial evolution, they cannot ascribe function per se without usually more computational or experimental information. Annotation programs can tell us whether the protein products of these genes contain membrane-spanning or DNA binding domains, nucleotide binding motifs, phosphorylation or kinase domains, or other functional information. Thus, these computational methods are essential first steps in formulating hypotheses about possible gene function across the expanse of bacterial genomic information. However, the integration of information obtained through genomic sequencing efforts, bioinformatics analysis, and experimental validation is still a significant challenge when trying to define virulence mechanisms in the post-genomic era.

### Virulence Gene Discovery in the Pre- and Postgenomic Eras

How has microbial genomics changed the way virulence genes are discovered? In the pregenomic era, virulence factors were typically identified either by biochemical approaches or through forward or reverse genetic screens. For example, a toxin active on mammalian cells might be purified and its gene then identified through its primary amino acid sequence. A transposon insertion mutant might be isolated that attenuated the replication of a pathogen in a given host (Leung and Finlay, 1991) or that encoded a gene coordinately regulated with a known virulence gene (Taylor et al., 1987). A virulence phenotype might be recapitulated by cloning a virulence gene into a heterologous organism (Isberg et al., 1987). Bacterial genomics has certainly facilitated these classical approaches but has also allowed virulence genes to be identified through other strategies. These employ computational methods, genetic signatures, physical linkage to accessory genetic elements, or biochemical and genetic approaches that depend on comprehensive genome-

sequence information. Several examples of these are discussed below.

### Computational Methods

The annotation phase of a genomic sequencing project frequently reveals putative genes coding for virulence factors with sequence similarity to known genes. Targeted genetic methods can then be used to inactivate the putative virulence gene and then develop an appropriate assay for its biochemical activity (Lin et al., 1999). For organisms that are not amenable to genetic manipulation, identifying sequence similarity to known virulence genes is one of the few available methods for understanding virulence, such as the identification of conserved type III secretion genes in *Chlamydia* species (Kalman et al., 1999). Sequence similarity to unique host proteins (e.g., tyrosine phosphatases) or host functional motifs (e.g., RGD integrin binding sequences) might also be used to identify new putative virulence genes. In some cases, the host-related protein revealed by genome sequencing can be confirmed to have the predicted biochemical activity despite the fact that a role in virulence cannot not be demonstrated with the available animal models (Vance et al., 2004).

### Genetic Signatures

Many virulence gene products are subject to positive selection at the level of the host through, for example, escape from protective immune responses, adaptive host-range, or functional mutations. Thus, genomic sequence information can identify putative virulence genes by identifying genetic signatures of such selective pressure. For example, polynucleotide repeats have been shown in *Haemophilus*, *Bordetella*, and *Neisseria* species to modulate the stochastic on or off expression of surface proteins and lipopolysaccharide-modifying enzymes through slip strand mutational events, a process important for phase variation of immunogenic factors leading to increased persistence in the host (van Belkum et al., 1999). Genome sequences have revealed numerous likely virulence genes that display evidence of slip strand control in their promoter and coding sequences. Other stochastic genetic mechanisms for alteration of gene expression (e.g., invertible elements encoding promoters) or coding sequence (e.g., retroelement diversity generating sequences) can be revealed through genomic sequence analysis and then used to implicate their associated genes in virulence related processes (Doulatov et al., 2004). Unique paralogous gene families that display high degrees of sequence variation (particularly nonsynonymous amino acid substitutions) are often revealed through genomic sequencing and later shown to correspond to virulence gene products (Fleischmann et al., 2002). Finally, the presence of predictive DNA motifs (e.g., iron regulation mediated by Fur) in the promoter region of a gene can be used to predict the gene's involvement in pathogenicity by coregulation with known virulence factors.

### Physical Linkage to Accessory Genetic Elements

Accessory genetic elements (plasmids, prophages, transposons, and islands) are known to frequently encode virulence factors. Such elements can exist as unlinked

replicons or as chromosomally integrated elements recognizable in part by the typical gene products they encode (replicases, transposases, phage structural proteins, integrases, etc.). These DNA elements also typically have a GC content and codon usage that is distinctly different from that of the rest of the genome, indicating their recent acquisition by horizontal genetic transfer. Thus, putative virulence genes can be recognized by looking for genes encoding, for example, secreted proteins that are associated with such elements (Perna et al., 2001). Occasionally, accessory elements are only recognized after comparative genomic methods, such as genomic microarray hybridization, reveal them as variable regions between different strains. Linkage of a hypothetical gene to a known virulence gene in a typical operon structure within the context of an accessory genetic element provides strong evidence for the hypothetical product's role in pathogenicity.

#### **Biochemical and Genetic Approaches for Identification of Virulence Genes Facilitated by Genome-Sequence Information**

The completion of a genome sequence provides an opportunity to employ postgenomic methods that can dramatically augment traditional methods of virulence-gene discovery. These include targeted genetic screens, proteomic analysis, and microarray-based methods of transcriptional analysis and mutant characterization. For example, virulence genes are, by their nature, expressed *in vivo*. A variety of methods have been described that seek to define *in vivo*-expressed genes (Rediers et al., 2005). Genomics has facilitated these studies and contributed the powerful methodologies such as gene chip or microarray technologies. Thus, genomic microarrays have been used to define virulence-gene regulons either by analyzing appropriate regulatory mutants (Bina et al., 2003; Zhu and Mekalanos, 2003) or *in vivo* expression during infection (Boyce et al., 2004; Merrell et al., 2002; Larocque et al., 2005). Proteomic analysis can be used to define all proteins produced by a pathogenic organism under a given growth condition once genomic sequence analysis has provided a database of all the peptides encoded by the organism. When proteomic analysis is combined with other genetic, biochemical, or immunological strategies, the result can be the accelerated discovery of virulence gene products (see below).

#### **Genomics and Gene Essentiality In Vitro and In Vivo**

A primary goal of genomics is to identify and define genes that are essential for growth or virulence. The historical distinction between these two categories is increasingly blurred, since genes that are required for growth *in vitro* are frequently required for growth inside a host. Indeed, some bacterial pathogens cannot be cultured at all outside a host.

There are various approaches other than direct sequence analysis or comparison for identifying essential or virulence genes. Prior to the availability of genome sequences, screens to identify these classes of genes were performed using methods such as isolating temper-

ature-sensitive mutants (Horowitz and Leupold, 1951), insertion-duplication mutagenesis (Mejean et al., 1981), *in vivo* expression technology (Mahan et al., 1993), signature-tagged transposon mutagenesis (Hensel et al., 1995), and TnAraOut (Judson and Mekalanos, 2000). These strategies are by now well known and will not be discussed further here. Genome sequencing has led to the development of other "high-throughput" approaches to defining essentiality of genes on the genomic scale. These include targeted genetic footprinting (Akerley et al., 1998; Posfai et al., 1999), antisense RNA screening (Ji et al., 2001), or saturating transposon mutagenesis (Hutchison et al., 1999). Genomic DNA microarrays can also be used in genetic footprinting techniques to define genes required for growth *in vitro* and *in vivo* (Sasseti et al., 2001; Sasseti and Rubin, 2003; Chan et al., 2005; Salama et al., 2004). A summary of a few examples of these approaches follows.

#### ***Mycobacterium Tuberculosis***

*M. tuberculosis* is a difficult organism to study and its sequence has provided information about genes important for virulence. A method of genetic footprinting termed "TraSH" (transposon site hybridization), has been a powerful new tool to understand which genes of this organism are essential *in vitro* and *in vivo* (Sasseti et al., 2001; Sasseti and Rubin, 2003). TraSH uses transposon mutagenesis to disrupt genes, and microarrays to compare mutants isolated from a host to the original mutant pool. Transposon mutants lost in the host are putatively involved in virulence. TraSH has been used *in vitro* and in macrophages. Perhaps unsurprisingly, the majority of the genes found essential *in vivo* were genes of unknown function. TraSH has been used to identify genes essential for survival in murine macrophage, in combination with transcriptional profiling to measure gene expression (Rengarajan et al., 2004; Schnappinger et al., 2003). Genes involved in lipid degradation were both upregulated *in vivo* and essential, indicating that fatty acids are necessary for a carbon source and for energy. Unknown genes in most organisms are likely to be involved in adaptation to different environments, and the environment of the host represents perhaps the most complex milieu to understand. For example, *M. tuberculosis* carries two slightly different copies of the *dnaE* gene, coding for the major replicative DNA polymerase. Boshoff et al. (2003) identified that the second copy, *dnaE2*, was an error-prone DNA polymerase involved in DNA repair. Further experiments showed that DnaE2 provides *M. tuberculosis* the ability to survive UV irradiation. Additionally, *dnaE2* expression is upregulated in mice and is important for survival *in vivo*. DnaE2 is now thought to be important for inducible mutagenesis and the emergence of drug resistance in *M. tuberculosis*.

#### ***Small and Large Genomes***

*P. aeruginosa* is an opportunistic pathogen with a large genome of over 6400 putative genes. The large number of genes is thought to be important for the variety of environments that *P. aeruginosa* can survive in, such as soil,

water, and many niches in a host organism. Presumably, the large gene content is important for uptake of nutrients or growth factors or to modify its environment through production of virulence factors. Saturating transposon mutagenesis was used to define all genes that could not be inactivated by insertion and thus were putatively essential in rich media (Jacobs et al., 2003). Six hundred seventy-eight genes were found to be essential, and these included most of the expected genes involved in transcription, translation, energy production, DNA replication, cell division, secretion, etc. One of the most intriguing results was just how many hypothetical genes showed up in this and other screens. After several decades of study by traditional genetic and biochemical techniques, a significant fraction of genes necessary for growth and survival on laboratory media is still not known. There were 261 genes of unknown function found to be essential, almost 40% of the total essential genes. Saturating screens for essential genes have also been done on other organisms and yielded numbers as low as 271 for *B. subtilis* to as high as 600–700 range for *S. aureus* and *E. coli* (Forsyth et al., 2002; Gerdes et al., 2003; Kobayashi et al., 2003). These screens for essential genes seem to have produced reasonable data; genes that were already known to be essential came up as essential in these screens, and genes known for virulence or pathogenesis functions did not turn up. Obligate intracellular bacteria such as *Chlamydia* or *Rickettsia* species have comparably small genomes (about 800–1300 ORFs) and are thought to represent organisms that contain near-minimal gene sets for a viable albeit not a free-living cell. The host environment may provide nutrients, growth factors, and other factors, and they will still require uptake systems and systems that allow survival and proliferation inside the host, making it likely that some genes involved in pathogenesis are essential. Therefore, a much higher percentage of their genomes are likely to be essential, as they have small genomes that have lost many of the genes that are dispensable. Similarly, in the small genome bacterium *M. genitalium*, 80% of the protein-coding genes are essential (Glass et al., 2006).

### Comparative Genomics and Virulence-Gene Identification

The goal of comparative genomics is to identify genetic differences across entire genomes, to correlate those differences to biological function, and to gain insight into selective evolutionary pressures and patterns of gene transfer or loss, particularly within the context of virulence in pathogenic species. Analysis of genomes from closely related species can also accelerate functional annotation of novel genes or other features (e.g., gene fusions, pseudogenes) that are apparent only in a comparative genomic context. However, identification of virulence genes through comparative genomics relies primarily on the identification of function of a gene in a given species or other genetic signature or characteristics consistent with a function in pathogenicity.

Comparisons can be performed either with sequence information or by using microarray-based methods to determine the presence or absence of specific genes contained on the array. Since the latter are much easier to apply to a large number of strains (e.g., clinical isolates), it has been more widely used than direct comparison of fully sequenced genomes. However, microarray-based analysis cannot detect genes that have not yet been identified by sequencing. If a gene is present in an organism but is not on the array, then no information can be gained about that gene. Similarly, only direct comparison of whole genomes can detect SNPs.

While the utility of cross-species genome comparisons depends on the evolutionary distance between the species, comparisons of closely-related species have found surprisingly high levels of diversity resulting from gross chromosomal changes (e.g., *Escherichia coli* strains vary in as much as 25% of their genome, and *Salmonella enterica* serovars vary by 10%–12%). These variations occur in the form of large genome islands that contribute to the acquisition of virulence factors or antibiotic resistance. The unexpectedly high degree of intraspecies diversity in these cases suggests that a single genome sequence is not representative of the genetic inventory of a given taxonomic group but is rather a sampling of genes characterizing members of a given population in the same gene pool. Multiple strains of *Streptococcus agalactiae* have led to the concept of a pangenome, with each species consisting of a core set of genes conserved in all strains, and a dispensable genome, consisting of partially shared or unique, strain-specific genes (Tettelin et al., 2005).

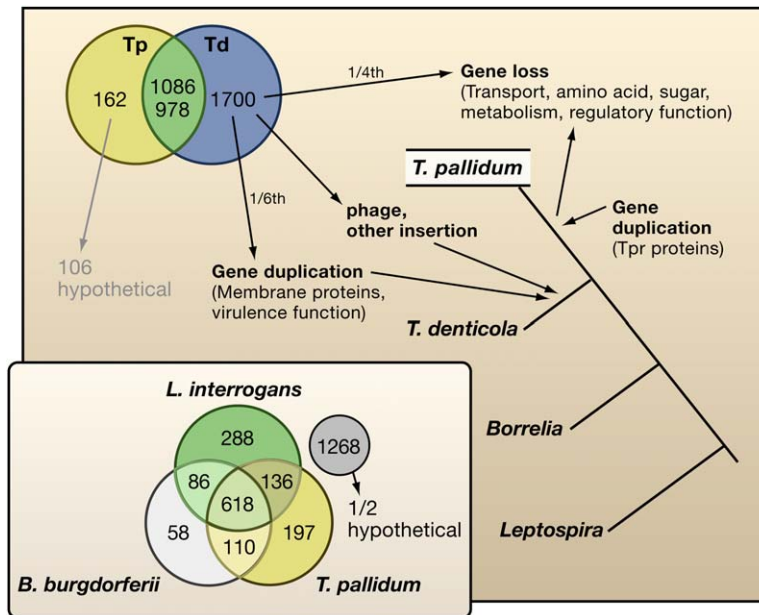
In other instances, genome differences are minimal and subtle (e.g., *Bacillus anthracis* strains that vary by a few SNPs or *Listeria monocytogenes* strains with extensive SNPs). This suggests that these species rely on a relatively small number of unique regions to generate variation within the host or other ecologically relevant settings. Without addressing the issues related to taxonomic discrimination, we will present instances of genome comparisons of closely related as well as more divergent species that have yielded interesting insights into their growth, virulence mechanisms, or evolution.

### Methods for Computational Sequence Analysis

The rapidly expanding quantity of sequence data has necessitated the evolution of software resources from analysis and functional annotation of a single genome toward simultaneous analysis of information across multiple genomes. Traditional genome annotation utilizes various software packages that optimize identification of predicted genes and intergenic regions (Chain et al., 2003; Field et al., 2005).

#### Example 1: Comparison of Sequenced Spirochete Species

Spirochetes comprise a monophyletic phylum that exhibits overall structural similarity but great variability in habitat, physiologic properties, and genome size and organization. Comparisons of the divergent sequenced



**Figure 1. Proposed Evolution of Spirochete Species Deduced through Comparative Genomics**

Comparison of *T. denticola* predicted ORFs with other completed Spirochete genomes. The circle in the inset represented *T. denticola* ORFs that had no significant matches ( $E < 10^{-5}$ ) in the other sequenced genomes. Td, *T. denticola*; Tp, *T. pallidum*.

spirochete pathogens, *Treponema denticola* (periodontal disease), *T. pallidum* (syphilis), *Borrelia burgdorferi* (Lyme disease), and *Leptospira interrogans* (leptospirosis), yielded insights into the basis for the differences and similarities in spirochete physiology and pathogenic potential (see Figure 1; Seshadri et al., 2004). Proteome comparison using BLASTP revealed that about one-fourth of the predicted *T. denticola* genes (618 out of 2786) were conserved in the other three spirochete genomes, consisting of mostly housekeeping functions. A small subset of these genes had no significant matches in other phyla and was predicted to relate to spirochete-specific properties, including morphology and host-pathogen interactions. These included characterized flagellar components that account for the unique internal periplasmically located flagellum and motility properties of spirochetes, allowing them to penetrate through high-viscosity matrices that inhibit the motility of most other bacteria. Almost half of the *T. denticola* genes had no significant matches in any of the other spirochete genomes. Of these, over half were hypothetical genes, while the remainder included various membrane proteins, ABC transporters, transcriptional regulators, and enzymes involved in amino acid metabolism and glycogen synthesis.

The *T. pallidum* and *T. denticola* genomes were found to share limited nucleotide identity and essentially no synteny, suggesting that their divergence from a common ancestor was an ancient event relative to the divergence of bacterial groups such as the *Brucella* and *Rickettsia* spp. and members of the Enterobacteriaceae. 1700 (~60%) *T. denticola* genes had no matches in *T. pallidum*, and 162 *T. pallidum* genes had no matches in *T. denticola*. The proposed reasons for disparity in genome size and predicted genes include (1) elimination of genes in one genome (reductive evolution), (2) acquisition of genes via

lateral transfer, (3) lineage-specific expansions, all contributing in varying degrees to the differences. About one-sixth of the 1700 *T. denticola*-specific predicted genes had their best matches to other *T. denticola* genes, thus indicating recent duplication or expansions within the *T. denticola* lineage. Many of these genes were implicated in virulence, such as internalin-like proteins, bacteriocin ABC transporters, and dipeptide YD-repeat proteins. Of the 1700 *T. denticola* genes that lack *T. pallidum* homologs, about one-sixth had homologs in *B. burgdorferi* or *L. interrogans*, suggesting that these may have been lost by gene reduction in the *T. pallidum* lineage. These corresponded to various transport, biosynthetic, or regulatory functions.

Of the 162 *T. pallidum* genes with no significant match in *T. denticola*, 106 were hypothetical proteins, and the remainder included transporters, components of a V-type ATPase, some proteases, and metabolic enzymes. A 12 member family (TprA–L) of putative membrane proteins in *T. pallidum* with a predicted role in immune response was specifically expanded in the *T. pallidum* lineage (Fraser et al., 1998). *T. denticola* possessed only a single gene related to this family.

Overall, genes lost due to reductive evolution appeared to be largely involved in basic metabolic functions and transport, while some genes that arose through lineage-specific expansions were implicated in pathogenesis. Genes acquired via horizontal gene transfer were phage-related or of unknown function. These results emphasize that the differing genome sizes of *T. pallidum* and *T. denticola* are not solely due to reductive evolution of *T. pallidum*. Like many other obligate parasites, *T. pallidum* had undergone some reduction in metabolic capabilities, indicating an increased dependence on the host for nutrients. Lineage-specific expansions and lateral gene

transfer reflect niche-specific adaptations and differences in pathogenic potential.

### Example 2: Comparison of Sequenced *Staphylococci* Species

Four strains of *Staphylococcus aureus* and two of *Staphylococcus epidermidis* were compared to delineate factors that contribute to increased pathogenicity in *S. aureus*, the evolution of *S. epidermidis* from a commensal to a more aggressive pathogen, and to the evolution of antibiotic resistance in these species (Gill et al., 2005). Both species are the major cause of varied nosocomial infections. Sequence comparisons revealed 454 species-specific genes in *S. aureus* (strains COL, N315, Mu50, and MW2) that were absent in *S. epidermidis* (RP62a or ATCC 12228) and a total of 286 species-specific genes in *S. epidermidis* strains not found in the *S. aureus* genomes. A total of 1681 genes were conserved among all strains of both species. The majority of the unique genes were found on prophage, genomic islands, and other mobile regions.

Genome islands and *Staphylococcus* cassette chromosome-like elements, integrated prophage, insertion sequence (IS) elements, composite transposons, and integrated plasmids were associated with virulence and antimicrobial resistance (Hiramatsu et al., 2001). Together, these integrated genomic elements made up ~7% of the *S. aureus* COL genome and ~9% of the *S. epidermidis* RP62a genome. Other studies have found that integrated elements comprised ~10% in group A *Streptococcus* (Beres et al., 2002) and as much as 24% in *Enterococcus faecalis* strains (Paulsen et al., 2003), corresponding in the latter case to increased virulence or vancomycin resistance. Seven pathogenicity genomic islands specific to *S. aureus* contained a variety of toxins or virulence factors (enterotoxins, exotoxins, and leukotoxins, etc.), and allelic variation of these genes may have contributed to the pathogenic potential of this species. Comparative analysis of these *S. aureus* isolates suggested that some genomic islands may have been mobilized by resident prophage into other members of this species.

Integrated plasmids in *S. epidermidis* contained genes with potential roles in virulence (e.g., species-specific surface proteins with an LPXTG motif required for anchoring to the cell wall envelope), and a genome island encoded multiple virulence-associated modulin proteins (Vuong et al., 2004). A family of small cytokine-stimulating peptides may have expanded within a genome island in *S. epidermidis* relative to the other staphylococci (including *S. aureus*). Acquisition of virulence factors may have also resulted from plasmid-mediated gene transfer between staphylococci and other low-GC-content gram-positive pathogens. For example, the *S. epidermidis* RP62a and ATCC 12228 genomes possessed a *cap* operon encoding a polyglutamate capsule, a major virulence factor in *Bacillus anthracis* (encoded on *B. anthracis* pX02 plasmid) but in *S. epidermidis* may play a role in biofilm formation.

Additional phenotypic differences may have arisen from single nucleotide polymorphisms, given that ~19,000–

22,000 SNPs were detected in *S. aureus* COL compared to other strains of *S. aureus*. Similarly, ~10,000 SNPs were found in *S. epidermidis* ATCC 12228 compared to RP62a. In both *S. aureus* and *S. epidermidis*, SNPs were most numerous in cell envelope proteins and resulted in nonsynonymous amino acid substitutions. This variation in immunodominant cell envelope or surface proteins may play a role in immune evasion.

These analyses illustrated the major role of gene transfer in the evolution of *Staphylococcal* virulence. While the *Staphylococcal* genomes were highly similar and syntenic, integrated elements bearing genes for virulence and resistance appear to have mediated a transition from an opportunistic or commensal to a more aggressive pathogenic lifecycle.

### Microarray-Based Comparative Genomic and Expression Studies

DNA microarrays are powerful tools that allow easy genomic comparisons between different strains of the same or similar species, allowing us to compare genomic sequence of unsequenced strains with a sequenced reference strain. Additionally, they are used for transcriptional profiling, monitoring gene expression of every gene in the genome, by determining levels or transcripts (Brown and Botstein, 1999). The use of microarrays has enhanced our understanding of evolution and pathogenesis by comparing genome content of pathogenic and nonpathogenic strains. They have also been instrumental in understanding virulence gene expression of pathogens during infection. However, there are drawbacks to these techniques. Transcriptional profiling of a strain generally involves isolating mRNA from a population of cells, so if a subpopulation of cells has a different gene-expression pattern, this critical subset of response will be missed. Genomic DNA comparisons can only identify what is absent from a given strain but cannot identify additional genes an organism may have. SNPs cannot not be accurately analyzed using most microarray-based techniques.

#### *Helicobacter pylori*

*H. pylori* causes gastritis and duodenal ulcers. Two strains have been sequenced so far (J99 and 26695) and these differ in approximately 100 genes which are localized in defined regions on the chromosome. Twenty-seven genes were confined to a 40 kb pathogenicity island and encode a type IV-like secretion system (Censini et al., 1996) and its substrate CagA (cytotoxin-associated gene A), which upon translocation into the host cytosol affects signaling pathways, cytoskeleton, and cell junctions of the infected host cell (Rieder et al., 2005). CagA turns out to be a major virulence determinant of *H. pylori* because strains that lack the *cagA*-containing pathogenicity island exhibit a less-aggressive gastroduodenal pathology (Nilsson et al., 2003). Microarray analysis allowed researchers to study how *H. pylori* adapts to the low-pH environment of the stomach. It turns out that the expression of approximately 120 genes is affected when cells are exposed to low pH (Merrell et al., 2003). Major responses include increased motility and

decreased virulence, suggesting that the bacterium is attempting to escape the low-pH environment to find a more suitable site for infection. Genomic analysis of infections with multiple *H. pylori* strains revealed that different strains colonize different parts of the stomach (Akada et al., 2003) and that strains undergo genetic changes during chronic infections (Israel et al., 2001). These findings suggest that *H. pylori* continuously evolves within its cognate host to adapt to a defined host environment.

### ***Pseudomonas Aeruginosa***

Wolfgang et al. (2003) compared 18 diverse clinical and environmental isolates and made a surprising discovery. All strains tested share 94% of their genes in common. Of 18 strains from different locales and infections, 368 genes found in the sequence PA01 strain were missing in at least one of the strains. These strain-specific genes were found in 90 discrete locations in blocks of one to four genes. Clearly, the genomic content does not define specialized lifestyles. It seems more likely that the majority of *P. aeruginosa* have the potential to occupy any niche described for *P. aeruginosa*.

Goodman et al. (2004) identified a regulator of genes involved in both acute infections as well as long-term persistence of *P. aeruginosa* using transcriptional profiling. They identified a response regulator, *retS*, that greatly induced biofilm formation. By comparing gene expression of the mutant to a wild-type strain, they found that *retS* regulates many virulence genes, including those involved in type III secretion, type IV pili, and exopolysaccharide production. Many transcriptional profiling studies have helped elucidate the role quorum sensing plays in virulence of *P. aeruginosa* (reviewed in Soberon-Chavez et al. [2005]). The *las* and *rhl* quorum-sensing systems are known to play a role in virulence, and transcriptional profiles were obtained comparing mutants in these systems to wild-type strains in the presence of quorum-sensing inducers. Hundreds of genes are regulated by these systems, including genes involved in motility, type II and type III secretion, drug efflux, and biofilm formation. Inhibitors of quorum sensing inhibit virulence by lowering expression of virulence factors, further stressing the importance of quorum sensing in virulence.

The availability of the genome sequence for *P. aeruginosa* has also been used to identify factors that may be involved in modulating host defense (Vance et al., 2004). The *P. aeruginosa* genome contained a predicted gene that had high homology to eukaryotic lipoxygenases, which are regulators of inflammation and host defense in eukaryotes. Lipoxygenases generally are involved in breaking down arachidonic acid into lipid mediators of leukocytes. Most bacteria including *P. aeruginosa* lack polyunsaturated fatty acids that these enzymes act on, making lipoxygenases rare in bacteria. The putative *P. aeruginosa* lipoxygenase was characterized and found to be able to metabolize arachidonic acid, implicating lipoxygenase as an anti-inflammatory agent that may partially explain the nonresolving chronic lung infection of cystic fibrosis patients caused by *P. aeruginosa*.

### ***Staphylococcus aureus***

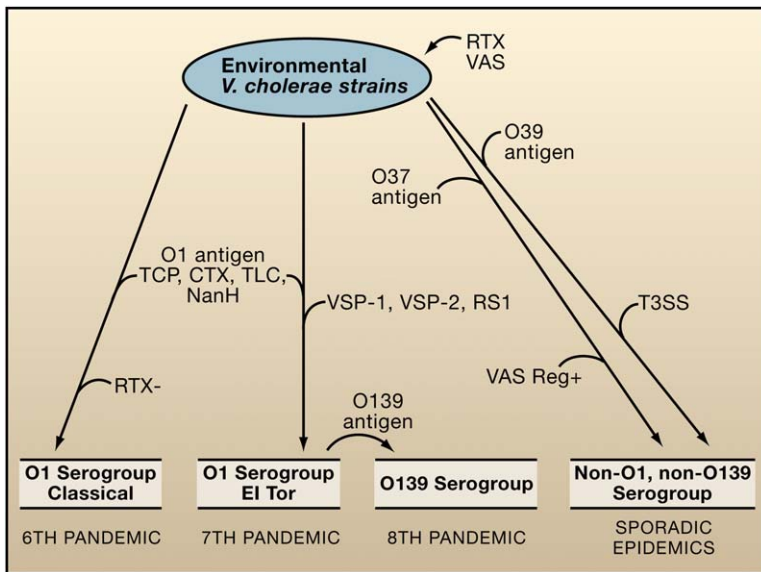
Microarray-based comparative genomic analysis of 36 *S. aureus* strains suggested that approximately 78% of the gene content is shared by all strains. Such genes are referred to as core genes since they seem to be essential to the life style of this bacterium. Twenty-two percent of the genes identified in *S. aureus* strains were missing in at least one of the strains examined. These genes are defined as accessory genes and are believed to be required for functions that are unique to the lifestyle of a particular strain and include virulence factors. Interestingly, accessory genes are not scattered throughout the genome but rather confined to certain regions of the chromosome. In *S. aureus*, 18 large regions of difference (called RD's) were identified and 10 of them carried putative virulence factors like enterotoxins, hypervirulence genes, and the methicillin-resistance gene *mec*. Phylogenetic analysis of the different genes showed that the *mec* gene was obtained by unrelated strains independently (Fitzgerald et al., 2001). This disproved the hypothesis that the toxic-shock syndrome epidemics in the 1970's and 1980's were caused by a newly emerging methicillin-resistant and hypervirulent strain, as *mec* was shown to have been introduced into *S. aureus* on at least five separate occasions.

### ***Neisseria meningitidis***

Another pathogen that maintains a conserved genome is *N. meningitidis*, which causes severe meningitis upon passing the blood-brain barrier. Using DNA microarrays from the commonly used laboratory strain Z2491, a virulent strain originally isolated from the 1983 epidemic in Gambia, 29 hyperinvasive strains and a clonal group of 20 strains not associated with human disease were compared. Not a single gene could be found that was present in all 29 invasive strains but missing in the noninvasive strains. However, an 8 kb fragment is present in all invasive strains but missing in 90% (18 of 20) of the noninvasive strains. It turns out that this island represents a filamentous phage that has the capacity to excise from the chromosome to be secreted via the type IV pilin secretin. Therefore, this element, by spreading among the population, may promote the development of new epidemic clones of *N. meningitidis* that are capable of breaking the normal commensal relationship with humans and causing invasive disease.

### ***Escherichia coli and Related Species***

To date, four *E. coli* strains have been sequenced and made publicly available, including three pathogenic strains. Welch et al. (2002) compared the genomes of the K12 strain MG1655 to the uropathogenic strain CFT073 and the enterohemorrhagic O157:H7 strain EDL933. They found that there is an *E. coli* backbone containing only about 40% of the total protein-coding gene content of the three strains. Each of the three strains contains islands not found in the others. Both nonpathogenic *E. coli* strains as well as the enterohemorrhagic pathogens are able to persist in the intestines. Extraintestinal strains, such as the uropathogenic CFT073, are able to colonize and grow elsewhere in the body. These differences are indicated in



**Figure 2. Comparative Genomics-Based Model for the Evolution of Toxigenic *V. cholerae* Strains**

The wealth of genomic data available for different *V. cholerae* strains suggests that toxigenic *V. cholerae* originated from nonpathogenic precursors by stepwise acquisition of virulence genes through horizontal transfer (see text for details). CTX, cholera toxin; TCP, toxin coregulated pilus; O1, O139, O37, and O39, various O antigen clusters; VSP-I, seventh pandemic island; T3SS; type III secretion system; VAS genes, virulence-associated secretion genes; TLC, toxin-linked cryptic.

the genome sequences. CFT073 is very enriched in fibrillar adhesions, secreted autotransporters, and phase-switch recombinases, allowing them to colonize and persist in environments much different than the other two strains. Only the enterohemorrhagic O157:H7 strain had type III secretion systems, differentiating it from the nonpathogenic strain. Interestingly, both O157:H7 and CFT073 strains contain about 1000 more genes than the nonpathogenic MG1655; however, the strains have not undergone much genome reduction of the backbone. This is somewhat surprising given that *E. coli* have lived as commensals in the animal gut for millions of years. The animal gut presumably is a rich environment that could have allowed bacteria to evolve with a reduced genome, including the loss of metabolic genes.

The idea of *E. coli* having a backbone core of genes, with individual strains containing islands of unique genes, has been borne out using comparative microarray studies (Dobryndt et al., 2003; Fukiya et al., 2004). A whole genome microarray used the K12 isolate MG1655 as a reference compared to many different *E. coli* strains, as well as several strains of the closely related *Shigella*. There is a backbone of about 2800 genes, with a mosaic structure of insertions and deletions throughout the *E. coli* and *Shigella* strains. Interestingly, it was found that the *Shigella* strains tested grouped phylogenetically with dysentery strains and enteroinvasive *E. coli* strains.

#### ***Vibrio cholerae***

There have been eight identified pandemics of *V. cholerae* since 1817. The first six were caused by the classical biotype of O1 serogroup strains. The seventh pandemic was caused by an O1 El Tor strain, which eventually replaced the classical biotype not just as cause of cholera, but in all environmental reservoirs worldwide. Another pandemic recently emerged, caused by strains of the O139 serogroup, and is occurring concurrently with the O1 El

Tor strains. Comparisons of the sixth, seventh, and eighth pandemic strains, as well as nonpandemic El Tor strains and environmental strains were performed (Dziejman et al., 2002). A whole genome microarray of a seventh pandemic strain was used as a reference for comparison. All strains tested were remarkably similar in gene content, as each strain contained at least 99% of the genes in the reference strain. It was found that O139 strains are very similar to O1 El Tor strains and are probably clonally derived from El Tor with changes only in the LPS structure. A seventh pandemic island was also identified, found only in pandemic El Tor strains and O139 strains, but not in classical or prepandemic El Tor strains. It is speculated that genes present in seventh pandemic strains but lacking in sixth pandemic strains may be involved in persistence in the environment, explaining the complete replacement of classical strains by El Tor strains. These genes may allow seventh pandemic strains to undergo nutrient depletion or withstand other environmental stresses, or they may be involved in interactions with nonhuman aquatic hosts.

While O1 and O139 strains are the only serogroups known to cause epidemic cholera, many non-O1, non-O139 strains are present in the environment and sporadically cause disease outbreaks. Some of these outbreaks are caused by strains lacking in the main *V. cholerae* virulence factors, cholera toxin (CTX) and toxin coregulated pilus (TCP). This led to a search for virulence factors present in these CTX, TCP strains (Dziejman et al., 2005). A strain of the O39 serogroup was sequenced, and a type III secretion system related to that of pandemic *Vibrio parahaemolyticus* was identified and found to be present in many environmental strains, including some that contain CTX and TCP. This indicates that the type III secretion system may be responsible for nonepidemic cholera disease and perhaps may improve environmental fitness as well. Further studies have shown how

pandemic strains originated from environmental isolates through acquisition of virulence elements (see Figure 2; Faruque and Mekalanos, 2003).

Many transcriptional profiling studies have been used to identify genes important for *V. cholerae* virulence in vivo, both in animal models and in humans (Larocque et al., 2005; Merrell et al., 2002; Xu et al., 2003). Stool samples from cholera epidemic patients were profiled, and it was found that genes involved in amino acid synthesis, iron uptake systems, and ribosomal proteins were upregulated. *V. cholerae* from stool was found to be more virulent than in vitro grown *V. cholerae*, possibly explaining the epidemic nature of the disease. *V. cholerae* isolated from patients' vomitus and stool were profiled to compare gene expression in the upper digestive tract to that in the lower digestive tract and differential expression of virulence regulators in the upper and lower digestive tract. It was found that colonization factors are predominately expressed in the upper digestive tract. To identify genes important for growth inside a host, *V. cholerae* grown in vivo in rabbit ileal loops were profiled. It was found that genes involved in virulence including motility, chemotaxis, and toxin production are upregulated in vivo when compared to in vitro cultures.

### Proteomics-Based Methods of Virulence Gene Discovery

Recently, proteomics was used in combination with genome-sequence analysis to define a new virulence-associated protein-secretion pathway that functions in *V. cholerae* strain V52. Mutants of V52 that are avirulent toward *Dictyostelium* amoebae and mouse J744 macrophages were found to carry insertion mutations in a gene cluster that included a homolog of *icmF*, a *Legionella* gene involved in secretion of proteins through the type IV secretion apparatus (Pukatzki et al., 2006). However, the genomic sequence of V52 indicated that this strain lacked other genes associated with type IV secretion. The gene cluster responsible was designated Vas, for virulence-associated secretion, because *vas* mutants were shown to be defective in the secretion of four proteins when analyzed by mass spectrometry. One of these secreted proteins was VgrG-1, a protein that previously had been reported to encode an actin-crosslinking domain (Sheahan et al., 2004). Recently, the cytotoxicity of strain V52 was shown to be associated with VgrG-1-mediated actin crosslinking in target cells demonstrating that the Vas system likely encodes a "molecular syringe" as part of a type VI secretion pathway to deliver effectors such as VgrG-1 to the cytosol of macrophages (S.U.P., A. Ma, and J.J.M., unpublished data).

A proteomic screen has identified secreted factors involved in virulence in *M. tuberculosis*. (Fortune et al., 2005) The ESX-1 locus contains two secreted proteins, ESAT-6 and CFP-10, as well as factors involved in their secretion. Culture supernatants of a wild-type strain and an ESX-1 deletion strain were compared, and a protein, EspA, not linked to ESX-1 on the chromosome, was iden-

tified as missing in the mutant. Mutants in each of the three secreted proteins blocked secretion of the other two proteins, suggesting a mutually-dependent secretion system. Additionally, both *espA* and ESX-1 deletions are attenuated for virulence.

### Conclusions

Genome sequencing has revolutionized how we approach the problem of understanding bacterial pathogenesis. In cases such as obligate intracellular bacteria and species that cannot be cultured in vitro, genomics has provided virtually the only path for efficient discovery of virulence genes. Genomics has also greatly benefited research on organisms that can be cultured by increasing the pace of identification and characterization of virulence factors. Genomics empowers the use of highly parallel methodologies that allow investigators to study all the genes or all the proteins of a pathogen in the context of its host or under various physiological or genetic states of interest. These "postgenomic" methods (e.g., genomic microarray-based methods and proteomics) have changed the way investigators approach the classical questions asked about the bacterial-host interaction. Which genes does a bacterium express in vivo during infection? Which proteins does this bacterium secrete? Which antigens of this pathogen are recognized by the host? Which gene products of this organism produce pathological changes in the host? In the future, the answers to these questions will not be for a single organism but rather for multiple species as well as multiple genes and proteins. We are getting closer to a time when vaccines will not be based on a single protein from a single organism but rather from a collection of conserved proteins that together protect against a group of organisms. Postgenomic studies have also moved us closer to understanding which genes are essential for growth and survival in different environments, including inside a host organism. Big challenges remain. Why are so many genes conserved in primary sequence yet have no known function? Which genes are essential for growth, and what do they do? It is clear that the identification of new antibiotic drug targets will be aided by identification of essential genes, particularly conserved genes of unknown function. It is clear that bacterial genomics will, in the end, drive a new era of therapeutic discovery by combining high-throughput methods of chemistry and screening with similarly high-throughput methods of genetic analysis and target discovery.

### ACKNOWLEDGMENTS

We thank Su Chiang for editing the manuscript. D.M.R. is an Ellison Medical Foundation Fellow of the Life Sciences Research Foundation.

### REFERENCES

- Akada, J.K., Ogura, K., Dailidene, D., Dailide, G., Cheverud, J.M., and Berg, D.E. (2003). *Helicobacter pylori* tissue tropism: mouse-colonizing strains can target different gastric niches. *Microbiol.* 149, 1901–1909.

- Akerley, B.J., Rubin, E.J., Camilli, A., Lampe, D.J., Robertson, H.M., and Mekalanos, J.J. (1998). Systematic identification of essential genes by *in vitro* mariner mutagenesis. *Proc. Natl. Acad. Sci. USA* **95**, 8927–8932.
- Beres, S.B., Sylva, G.L., Barbian, K.D., Lei, B., Hoff, J.S., Mammarella, N.D., Liu, M.Y., Smoot, J.C., Porcella, S.F., Parkins, L.D., et al. (2002). Genome sequence of a serotype M3 strain of group A *Streptococcus*: phage-encoded toxins, the high-virulence phenotype, and clone emergence. *Proc. Natl. Acad. Sci. USA* **99**, 10078–10083.
- Bina, J., Zhu, J., Dziejman, M., Faruque, S., Calderwood, S., and Mekalanos, J. (2003). ToxR regulon of *Vibrio cholerae* and its expression in vibrios shed by cholera patients. *Proc. Natl. Acad. Sci. USA* **100**, 2801–2806.
- Boshoff, H.I., Reed, M.B., Barry, C.E., 3rd, and Mizrahi, V. (2003). DnaE2 polymerase contributes to *in vivo* survival and the emergence of drug resistance in *Mycobacterium tuberculosis*. *Cell* **113**, 183–193.
- Boyce, J.D., Cullen, P.A., and Adler, B. (2004). Genomic-scale analysis of bacterial gene and protein expression in the host. *Emerg. Infect. Dis.* **10**, 1357–1362.
- Brown, P.O., and Botstein, D. (1999). Exploring the new world of the genome with DNA microarrays. *Nat. Genet.* **21**, 33–37.
- Censini, S., Lange, C., Xiang, Z., Crabtree, J.E., Ghiara, P., Borodovsky, M., Rappuoli, R., and Covacci, A. (1996). *cag*, a pathogenicity island of *Helicobacter pylori*, encodes type I-specific and disease-associated virulence factors. *Proc. Natl. Acad. Sci. USA* **93**, 14648–14653.
- Chain, P., Kurtz, S., Ohlebusch, E., and Slezak, T. (2003). An applications-focused review of comparative genomics tools: capabilities, limitations and future challenges. *Brief. Bioinform.* **4**, 105–123.
- Chan, K., Kim, C.C., and Falkow, S. (2005). Microarray-based detection of *Salmonella enterica* serovar Typhimurium transposon mutants that cannot survive in macrophages and mice. *Infect. Immun.* **73**, 5438–5449.
- Dobrindt, U., Agerer, F., Michaelis, K., Janka, A., Buchrieser, C., Samuelson, M., Svanborg, C., Gottschalk, G., Karch, H., and Hacker, J. (2003). Analysis of genome plasticity in pathogenic and commensal *Escherichia coli* isolates by use of DNA arrays. *J. Bacteriol.* **185**, 1831–1840.
- Doulatov, S., Hodes, A., Dai, L., Mandhana, N., Liu, M., Deora, R., Simons, R.W., Zimmerly, S., and Miller, J.F. (2004). Tropism switching in *Bordetella* bacteriophage defines a family of diversity-generating retroelements. *Nature* **431**, 476–481.
- Dziejman, M., Balon, E., Boyd, D., Fraser, C.M., Heidelberg, J.F., and Mekalanos, J.J. (2002). Comparative genomic analysis of *Vibrio cholerae*: genes that correlate with cholera endemic and pandemic disease. *Proc. Natl. Acad. Sci. USA* **99**, 1556–1561.
- Dziejman, M., Serruto, D., Tam, V.C., Sturtevant, D., Diraphat, P., Faruque, S.M., Rahman, M.H., Heidelberg, J.F., Decker, J., Li, L., et al. (2005). Genomic characterization of non-O1, non-O139 *Vibrio cholerae* reveals genes for a type III secretion system. *Proc. Natl. Acad. Sci. USA* **102**, 3465–3470.
- Faruque, S.M., and Mekalanos, J.J. (2003). Pathogenicity islands and phages in *Vibrio cholerae* evolution. *Trends Microbiol.* **11**, 505–510.
- Field, D., Feil, E.J., and Wilson, G.A. (2005). Databases and software for the comparison of prokaryotic genomes. *Microbiol.* **151**, 2125–2132.
- Fitzgerald, J.R., Sturdevant, D.E., Mackie, S.M., Gill, S.R., and Musser, J.M. (2001). Evolutionary genomics of *Staphylococcus aureus*: insights into the origin of methicillin-resistant strains and the toxic shock system epidemic. *Proc. Natl. Acad. Sci. USA* **98**, 8821–8826.
- Fleischmann, R.D., Adams, M.D., White, O., Clayton, R.A., Kirkness, E.F., Kerlavage, A.R., Bult, C.J., Tomb, J.F., Dougherty, B.A., Merrick, J.M., et al. (1995). Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* **269**, 496–512.
- Fleischmann, R.D., Alland, D., Eisen, J.A., Carpenter, L., White, O., Peterson, J., DeBoy, R., Dodson, R., Gwinn, M., Haft, D., et al. (2002). Whole-genome comparison of *Mycobacterium tuberculosis* clinical and laboratory strains. *J. Bacteriol.* **184**, 5479–5490.
- Forsyth, R.A., Haselbeck, R.J., Ohlsen, K.L., Yamamoto, R.T., Xu, H., Trawick, J.D., Wall, D., Wang, L., Brown-Driver, V., Froelich, J.M., et al. (2002). A genome-wide strategy for the identification of essential genes in *Staphylococcus aureus*. *Mol. Microbiol.* **43**, 1387–1400.
- Fortune, S.M., Jaeger, A., Sarracino, D.A., Chase, M.R., Sasseti, C.M., Sherman, D.R., Bloom, B.R., and Rubin, E.J. (2005). Mutually dependent secretion of proteins required for mycobacterial virulence. *Proc. Natl. Acad. Sci. USA* **102**, 10676–10681.
- Fraser, C.M., Gocayne, J.D., White, O., Adams, M.D., Clayton, R.A., Fleischmann, R.D., Bult, C.J., Kerlavage, A.R., Sutton, G., Kelley, J.M., et al. (1995). The minimal gene complement of *Mycoplasma genitalium*. *Science* **270**, 397–403.
- Fraser, C.M., Norris, S.J., Weinstock, G.M., White, O., Sutton, G.G., Dodson, R., Gwinn, M., Hickey, E.K., Clayton, R., Ketchum, K.A., et al. (1998). Complete genome sequence of *Treponema pallidum*, the syphilis spirochete. *Science* **281**, 375–388.
- Fukuya, S., Mizoguchi, H., Tobe, T., and Mori, H. (2004). Extensive genomic diversity in pathogenic *Escherichia coli* and *Shigella* strains revealed by comparative genomic hybridization microarray. *J. Bacteriol.* **186**, 3911–3921.
- Gerdes, S.Y., Scholle, M.D., Campbell, J.W., Balazsi, G., Ravasz, E., Daugherty, M.D., Somera, A.L., Kyrpides, N.C., Anderson, I., Gelfand, M.S., et al. (2003). Experimental determination and system level analysis of essential genes in *Escherichia coli* MG1655. *J. Bacteriol.* **185**, 5673–5684.
- Gil, R., Silva, F.J., Pereto, J., and Moya, A. (2004). Determination of the core of a minimal bacterial gene set. *Microbiol. Mol. Biol. Rev.* **68**, 518–537.
- Gill, S.R., Fouts, D.E., Archer, G.L., Mongodin, E.F., Deboy, R.T., Ravel, J., Paulsen, I.T., Kolonay, J.F., Brinkac, L., Beanan, M., et al. (2005). Insights on evolution of virulence and resistance from the complete genome analysis of an early methicillin-resistant *Staphylococcus aureus* strain and a biofilm-producing methicillin-resistant *Staphylococcus epidermidis* strain. *J. Bacteriol.* **187**, 2426–2438.
- Glass, J.I., Assad-Garcia, N., Alperovich, N., Yooseph, S., Lewis, M.R., Maruf, M., Hutchison, C.A., Smith, H.O., and Venter, J.C. (2006). Essential genes of a minimal bacterium. *Proc. Natl. Acad. Sci. USA* **103**, 425–430.
- Goodman, A.L., Kulasekara, B., Rietsch, A., Boyd, D., Smith, R.S., and Lory, S. (2004). A signaling network reciprocally regulates genes associated with acute infection and chronic persistence in *Pseudomonas aeruginosa*. *Dev. Cell* **7**, 745–754.
- Hensel, M., Shea, J.E., Gleeson, C., Jones, M.D., Dalton, E., and Holden, D.W. (1995). Simultaneous identification of bacterial virulence genes by negative selection. *Science* **269**, 400–403.
- Hiramatsu, K., Cui, L., Kuroda, M., and Ito, T. (2001). The emergence and evolution of methicillin-resistant *Staphylococcus aureus*. *Trends Microbiol.* **9**, 486–493.
- Horowitz, N.H., and Leupold, U. (1951). Some recent studies bearing on the one gene one enzyme hypothesis. *Cold Spring Harb. Symp. Quant. Biol.* **16**, 65–74.
- Hutchison, C.A., Peterson, S.N., Gill, S.R., Cline, R.T., White, O., Fraser, C.M., Smith, H.O., and Venter, J.C. (1999). Global transposon mutagenesis and a minimal *Mycoplasma* genome. *Science* **286**, 2165–2169.

- Isberg, R.R., Voorhis, D.L., and Falkow, S. (1987). Identification of invasion: A protein that allows enteric bacteria to penetrate cultured mammalian cells. *Cell* 50, 769–778.
- Israel, D.A., Salama, N., Krishna, U., Rieger, U.M., Atherton, J.C., Falkow, S., and Peek, R.M., Jr. (2001). *Helicobacter pylori* genetic diversity within the gastric niche of a single human host. *Proc. Natl. Acad. Sci. USA* 98, 14625–14630.
- Jacobs, M.A., Alwood, A., Thaipisuttikul, I., Spencer, D., Haugen, E., Ernst, S., Will, O., Kaul, R., Raymond, C., Levy, R., et al. (2003). Comprehensive transposon mutant library of *Pseudomonas aeruginosa*. *Proc. Natl. Acad. Sci. USA* 100, 14339–14344.
- Ji, Y., Zhang, B., Van Horn, S.F., Warren, P., Woodnutt, G., Burnham, M.K., and Rosenberg, M. (2001). Identification of critical staphylococcal genes using conditional phenotypes generated by antisense RNA. *Science* 293, 2266–2269.
- Judson, N., and Mekalanos, J.J. (2000). TnAraOut, a transposon-based approach to identify and characterize essential bacterial genes. *Nat. Biotechnol.* 18, 740–745.
- Kalman, S., Mitchell, W., Marathe, R., Lammel, C., Hyman, R.W., Ollinger, L., Grimwood, J., Davis, R.W., and Stephens, R.S. (1999). Comparative genomes of *Chlamydia pneumoniae* and *C. trachomatis*. *Nat. Genet.* 21, 385–389.
- Kobayashi, K., Ehrlich, S.D., Albertini, A., Amati, G., Andersen, K.K., Arnaud, M., Asai, K., Ashikaga, S., Aymerich, S., Bessieres, P., et al. (2003). Essential *Bacillus subtilis* genes. *Proc. Natl. Acad. Sci. USA* 100, 4678–4683.
- Larocque, R.C., Harris, J.B., Dziejman, M., Li, X., Khan, A.I., Faruque, A.S., Faruque, S.M., Nair, G.B., Ryan, E.T., Qadri, F., et al. (2005). Transcriptional profiling of *Vibrio cholerae* recovered directly from patient specimens during early and late stages of human infection. *Infect. Immun.* 73, 4488–4493.
- Leung, K.Y., and Finlay, B.B. (1991). Intracellular replication is essential for the virulence of *Salmonella typhimurium*. *Proc. Natl. Acad. Sci. USA* 88, 11470–11474.
- Lin, W., Fullner, K.J., Clayton, R., Sexton, J.A., Rogers, M.B., Calia, K.E., Calderwood, S.B., Fraser, C., and Mekalanos, J.J. (1999). Identification of a *vibrio cholerae* RTX toxin gene cluster that is tightly linked to the cholera toxin prophage. *Proc. Natl. Acad. Sci. USA* 96, 1071–1076.
- Mahan, M.J., Schlauch, J.M., and Mekalanos, J.J. (1993). Selection of bacterial virulence genes that are specifically induced in host tissues. *Science* 259, 686–688.
- Mejean, V., Claverys, J.P., Vasseghi, H., and Sicard, A.M. (1981). Rapid cloning of specific DNA fragments of *Streptococcus pneumoniae* by vector integration into the chromosome followed by endonucleolytic excision. *Gene* 15, 289–293.
- Merrell, D.S., Butler, S.M., Qadri, F., Dolganov, N.A., Alam, A., Cohen, M.B., Calderwood, S.B., Schoolnik, G.K., and Camilli, A. (2002). Host-induced epidemic spread of the cholera bacterium. *Nature* 417, 642–645.
- Merrell, D.S., Goodrich, M.L., Otto, G., Tompkins, L.S., and Falkow, S. (2003). pH-regulated gene expression of the gastric pathogen *Helicobacter pylori*. *Infect. Immun.* 71, 3529–3539.
- Mushegian, A.R., and Koonin, E.V. (1996). A minimal gene set for cellular life derived by comparison of complete bacterial genomes. *Proc. Natl. Acad. Sci. USA* 93, 10268–10273.
- Nilsson, C., Sillen, A., Eriksson, L., Strand, M.L., Enroth, H., Normark, S., Falk, P., and Engstrand, L. (2003). Correlation between *cag* pathogenicity island composition and *Helicobacter pylori*-associated gastro-duodenal disease. *Infect. Immun.* 71, 6573–6581.
- Paulsen, I.T., Banerjee, L., Myers, G.S., Nelson, K.E., Seshadri, R., Read, T.D., Fouts, D.E., Eisen, J.A., Gill, S.R., Heidelberg, J.F., et al. (2003). Role of mobile DNA in the evolution of vancomycin-resistant *Enterococcus faecalis*. *Science* 299, 2071–2074.
- Perna, N.T., Plunkett, G., 3rd, Burland, V., Mau, B., Glasner, J.D., Rose, D.J., Mayhew, G.F., Evans, P.S., Gregor, J., Kirkpatrick, H.A., et al. (2001). Genome sequence of enterohaemorrhagic *Escherichia coli* O157:H7. *Nature* 409, 529–533.
- Posfai, G., Kolisnichenko, V., Bereczki, Z., and Blattner, F.R. (1999). Markerless gene replacement in *Escherichia coli* stimulated by a double-strand break in the chromosome. *Nucleic Acids Res.* 27, 4409–4415.
- Pukatzki, S., Ma, A.T., Sturtevant, D., Krastins, B., Sarracino, D., Nelson, W.C., Heidelberg, J.F., and Mekalanos, J.J. (2006). Identification of a conserved bacterial protein secretion system in *Vibrio cholerae* using the *Dictyostelium* host model system. *Proc. Natl. Acad. Sci. USA* 103, 1528–1533.
- Rediers, H., Rainey, P.B., Vanderleyden, J., and De Mot, R. (2005). Unraveling the secret lives of bacteria: use of in vivo expression technology and differential fluorescence induction promoter traps as tools for exploring niche-specific gene expression. *Microbiol. Mol. Biol. Rev.* 69, 217–261.
- Rengarajan, J., Sasseti, C.M., Naroditskaya, V., Sloutsky, A., Bloom, B.R., and Rubin, E.J. (2004). The folate pathway is a target for resistance to the drug para-aminosalicylic acid (PAS) in mycobacteria. *Mol. Microbiol.* 53, 275–282.
- Rieder, G., Fischer, W., and Haas, R. (2005). Interaction of *Helicobacter pylori* with host cells: function of secreted and translocated molecules. *Curr. Opin. Microbiol.* 8, 67–73.
- Salama, N.R., Shepherd, B., and Falkow, S. (2004). Global transposon mutagenesis and essential gene analysis of *Helicobacter pylori*. *J. Bacteriol.* 186, 7926–7935.
- Sasseti, C.M., and Rubin, E.J. (2003). Genetic requirements for mycobacterial survival during infection. *Proc. Natl. Acad. Sci. USA* 100, 12989–12994.
- Sasseti, C.M., Boyd, D.H., and Rubin, E.J. (2001). Comprehensive identification of conditionally essential genes in mycobacteria. *Proc. Natl. Acad. Sci. USA* 98, 12712–12717.
- Schnappinger, D., Ehrt, S., Voskuil, M.I., Liu, Y., Mangan, J.A., Monahan, I.M., Dolganov, G., Efron, B., Butcher, P.D., Nathan, C., and Schoolnik, G.K. (2003). Transcriptional adaptation of *Mycobacterium tuberculosis* within macrophages: insights into the phagosomal environment. *J. Exp. Med.* 198, 693–704.
- Seshadri, R., Myers, G.S., Tettelin, H., Eisen, J.A., Heidelberg, J.F., Dodson, R.J., Davidsen, T.M., DeBoy, R.T., Fouts, D.E., Haft, D.H., et al. (2004). Comparison of the genome of the oral pathogen *Treponema denticola* with other spirochete genomes. *Proc. Natl. Acad. Sci. USA* 101, 5646–5651.
- Sheahan, K.L., Cordero, C.L., and Satchell, K.J. (2004). Identification of a domain within the multifunctional *Vibrio cholerae* RTX toxin that covalently cross-links actin. *Proc. Natl. Acad. Sci. USA* 101, 9798–9803.
- Soberon-Chavez, G., Aguirre-Ramirez, M., and Ordonez, L. (2005). Is *Pseudomonas aeruginosa* only “sensing quorum”? *Crit. Rev. Microbiol.* 31, 171–182.
- Taylor, R.K., Miller, V.L., Furlong, D.B., and Mekalanos, J.J. (1987). Use of *phoA* gene fusions to identify a pilus colonization factor coordinately regulated with cholera toxin. *Proc. Natl. Acad. Sci. USA* 84, 2833–2837.
- Tettelin, H., Massignani, V., Cieslewicz, M.J., Donati, C., Medini, D., Ward, N.L., Anguoli, S.V., Crabtree, J., Jones, A.L., Durkin, A.S., et al. (2005). Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial “pan-genome”. *Proc. Natl. Acad. Sci. USA* 102, 13950–13955.

- van Belkum, A., van Leeuwen, W., Scherer, S., and Verbrugh, H. (1999). Occurrence and structure-function relationship of pentameric short sequence repeats in microbial genomes. *Res. Microbiol.* *150*, 617–626.
- Vance, R.E., Hong, S., Gronert, K., Serhan, C.N., and Mekalanos, J.J. (2004). The opportunistic pathogen *Pseudomonas aeruginosa* carries a secretable arachidonate 15-lipoxygenase. *Proc. Natl. Acad. Sci. USA* *101*, 2135–2139.
- Vuong, C., Durr, M., Carmody, A.B., Peschel, A., Klebanoff, S.J., and Otto, M. (2004). Regulated expression of pathogen-associated molecular pattern molecules in *Staphylococcus epidermidis*: quorum-sensing determines pro-inflammatory capacity and production of phenol-soluble modulins. *Cell. Microbiol.* *6*, 753–759.
- Welch, R.A., Burland, V., Plunkett, G., 3rd, Redford, P., Roesch, P., Rasko, D., Buckles, E.L., Liou, S.R., Boutin, A., Hackett, J., et al. (2002). Extensive mosaic structure revealed by the complete genome sequence of uropathogenic *Escherichia coli*. *Proc. Natl. Acad. Sci. USA* *99*, 17020–17024.
- Wolfgang, M.C., Kulasekara, B.R., Liang, X., Boyd, D., Wu, K., Yang, Q., Miyada, C.G., and Lory, S. (2003). Conservation of genome content and virulence determinants among clinical and environmental isolates of *Pseudomonas aeruginosa*. *Proc. Natl. Acad. Sci. USA* *100*, 8484–8489.
- Xu, Q., Dziejman, M., and Mekalanos, J.J. (2003). Determination of the transcriptome of *Vibrio cholerae* during intrainestinal growth and midexponential phase in vitro. *Proc. Natl. Acad. Sci. USA* *100*, 1286–1291.
- Zhu, J., and Mekalanos, J.J. (2003). Quorum sensing-dependent biofilms enhance colonization in *Vibrio cholerae*. *Dev. Cell* *5*, 647–656.