



Published in final edited form as:

*Trends Microbiol.* 2002 December ; 10(12): 571–574.

## Information overload: assigning genetic functionality in the age of genomics and large-scale screening

D. Scott Merrell\* and

Stanford University School of Medicine, Dept of Microbiology and Immunology, 299 Campus Drive, Fairchild D051, Stanford, CA 94305, USA.

Andrew Camilli

Tufts University School of Medicine, Dept of Molecular Biology and Microbiology, 136 Harrison Avenue, Boston, MA 02111, USA.

### Abstract

As more and more genome sequences are completed, it is becoming increasingly evident that our understanding of the function of most bacterial gene products is lacking. This is frustrating, particularly in the study of pathogens, where an understanding of the role of individual gene products would probably facilitate the development of novel antimicrobials and vaccines. Recently, we devised a technique known as virulence-attenuated pool (VAP) screening to help assign genetic functionality to gene products that the pathogen *Vibrio cholerae* requires for colonization. This screen and potential new applications of the VAP technique are discussed here.

---

Just a few years ago, the way that we do science changed drastically with the development of high-throughput technologies that made it feasible to determine rapidly the complete genomic sequence of microorganisms. For those of us who work with one of the >60 organisms whose genetic code has been unraveled [1], rather than spending weeks or months screening, selecting or attempting to clone theoretical genes with degenerate primers, we can now simply go to the database and search for genes of potential interest. The glut of information available from these many sequencing projects, although making our day-to-day research lives more efficient, also highlights how naive our understanding of the functional roles of most of the gene products is (Table 1). Elucidating the function, and functional interacting networks, of the gene products of microorganisms is a major goal of scientific research. This is particularly true for pathogens, where this understanding could provide valuable insights that would help potentiate the development of antimicrobials and vaccines.

### High-throughput screening techniques

In addition to the advent of genome-sequencing projects, researchers have also developed multiple high-throughput screening techniques within the past decade, with the goal of enhancing our understanding of host–pathogen interactions. These techniques are probably best exemplified by *in vivo* expression technology (IVET) [2]; signature-tagged mutagenesis (STM) [3]; *in vivo*-induced antigen technology (IVIAT) [4]; and microarray technology [5], each of which has now been employed with great success in multiple pathogens and host systems to elucidate genes that are induced and/or required within the context of a host infection (reviewed in [6–8]).

---

\*dmerrell@stanford.edu.

IVET relies on the use of various gene reporters to identify transcriptional units that are induced within specific environmental conditions. The major advantage of IVET over other techniques is that a live host, with intact tissue barriers and immune system, can be used to identify the 'natural' induction of virulence genes. Microarray technology, where DNA samples that correspond to every predicted gene of an organism are spotted onto a suitable surface (such as a glass slide), can be used for transcriptional profiling. By direct comparison of the levels of mRNA in bacteria exposed to various environmental conditions, researchers can obtain a genome-wide view of the organism's transcriptional response to a particular environment. Microarrays have successfully been used to investigate changes in gene expression upon exposure of host to pathogen in various tissue-culture systems, but have remained largely unsuccessful within the context of a host infection with one exception: *Vibrio cholerae* [9]. This lack of success results from the inability to obtain large enough quantities of relatively pure bacterial mRNA from infected animals. The principle of IVIAT relies on the use of sera from actual patients, rather than from animal models, to probe expression libraries to identify genes that are expressed *in vivo*. This technique is advantageous in that it does not require direct genetic manipulation of the pathogen of interest, and thus bacteria that have poorly developed genetic systems can still be analyzed. A limitation of the technique, however, is that it requires that the expressed gene product elicit a strong antibody response in its host to be identified, and this is not the case for many such expressed factors. Finally, STM combines the strength of insertional mutagenesis with *in vivo* negative selection of attenuated strains in an animal model. The presence or loss of each mutant from an animal model can be assessed by following the unique signature tags (STs) present in the insertional element. Thus, the strength of STM lies in the fact that one can screen many different mutants in a single animal to identify those factors that are essential for colonization.

One limitation of all of these techniques is that the function of many of the identified gene products within the context of the infection is often not immediately evident. In some cases, a function can be predicted based on homology of the gene product, or on prior *in vitro* studies concerning the gene and gene product. However, in most cases the gene identified is hypothetical, conserved hypothetical or homologous to genes of ill-defined function(s). Even in instances where a function has been ascribed through *in vitro* studies (e.g. an enzymatic or regulatory function), the gene product's precise role during infection will often depend upon the context in which it is expressed and in which tissue or host cell compartment its target lies. This additional level of complexity stems from the fact that pathogenicity is a multifactorial, dynamic process and that pathogens must adapt to changing environments and express virulence factors in an orchestrated manner within the context of the host. Thus, a factor required at one stage of infection might not necessarily be required at later stages. As the host environment largely remains a black box, elucidation of the involvement of multiple genes at certain points in infection becomes a daunting task when each is considered individually. It is for this reason that we developed virulence-attenuated pool (VAP) screening [10].

## VAP screening

VAP screening is a process whereby mutants that have previously been determined to be avirulent are systematically screened for function in conditions designed to mimic *in vivo* environments. Overall, the power of VAP screening relies on the fact that one is testing only for functions of factors already shown to be important for *in vivo* survival; thus, factors that might be important *in vitro* but which are not necessarily relevant to *in vivo* survival are eliminated from the screening process. In its first application, VAP screening was used to identify colonization genes of *V. cholerae* that are also important for survival upon exposure to acid stress [10], as this is one environmental condition encountered by the bacterium during the course of colonizing its human host. Pools of avirulent signature-tagged transposon mutants of *V. cholerae* were assembled into VAPs and exposed to acid stress. ST probes were then

generated from the surviving bacteria and subsequently used for hybridization to ST master dot blots (Fig. 1). Mutations that affect survival in this condition can then be easily identified by the lack of hybridization to the blots. Of the 95 avirulent *V. cholerae* mutants tested, nine were also found to be defective for survival in acidic environments. Among these were factors predicted to be important for maintenance of Na<sup>+</sup> and K<sup>+</sup> homeostasis, DNA repair and transcriptional regulation. Of particular note, a role for survival in acid stress conditions had not previously been attributed to the majority of these identified factors. Thus, VAP screening was able to assign a putative role for these genes within the context of an actual infection.

In other work, STM-based VAP screening has been used to begin to delineate the requirement for *Streptococcus pneumoniae* virulence genes in multiple host tissues (D. Hava and A. Camilli, unpublished). *S. pneumoniae* has the ability to colonize the nasopharyngeal cavity and is often carried asymptotically. If, however, there is dissemination from this site, *S. pneumoniae* can cause severe pneumonia, bacteremia and/or meningitis. Owing to its ability to colonize or multiply within diverse host tissues, it is easy to imagine that factors that are essential within one host site might not necessarily be required within the context of other host sites. To investigate this, lung-infection-defective mutants of *S. pneumoniae* were assembled into VAPs and subsequently screened for defects in nasopharyngeal carriage and/or bacteremia. In this way, mutants of different tissue-trophic classes were preliminarily assigned (Fig. 2). Thus, VAP screening, in addition to being used to assign potential *in vivo* functionality for virulence genes, can be used to screen avirulent mutants *en masse* within the context of different animal models to define tissue-trophic virulence factors.

### IVET-based VAP screens

The two previous uses of VAP screening relied on STM-based screens to identify virulence genes affected in the condition of interest. It is not difficult to envision that similar VAP screens could be conducted where IVET or microarray technologies were used for the screening process. As IVET does not use pools of defined mutant strains like STM does, the *en masse* screening process would need to be conducted in a 96-well format where individual strains could be assayed for transcriptional activity in the condition of interest. The major drawback to this type of approach lies in the fact that expression does not always reflect the requirement for a particular gene [11]. So, although *in vivo*-induced genes could be screened in a large-scale format by a VAP-based strategy, the expression pattern within the host or the *in vitro* test condition says nothing about the absolute requirement or function for this gene within the host or in the test condition. For this reason, recently developed microarray techniques are probably better suited to VAP-based screens.

### Microarray-based VAP screens

In addition to being used for transcriptional profiling, microarrays have recently been used as a tool to map large numbers of transposon insertion sites ([12,13] and N. Salama, pers. commun.). Each of the groups referenced uses a different acronym to describe their technology and a different method to amplify the chromosomal DNA flanking the site of the transposon insertion. Each technique, however, relies on the basic principles outlined here. A general scheme for microarray transposon tagging (MATT) is depicted in Fig. 3 and involves PCR amplification, using a transposon-specific primer, of junctional fragments adjacent to the transposon. Pools of transposon mutants are exposed to different conditions, and direct comparisons between mutant pools are made by labeling the resulting pool's DNA with different fluorophores and subsequent hybridization to the microarray. Thus, genes that are conditionally required can be identified by the presence of labeling in only one channel on the microarray. This technique has been used to identify essential genes of *Helicobacter pylori* (N. Salama, pers. commun.) and *Escherichia coli* and *Mycobacterium tuberculosis* genes required

for growth in minimal media [12,13]. It is easy to envision how MATT could be used first to identify virulence genes essential for colonization within an animal model and subsequent VAP screening could assign functions to these genes. Sassetti and Rubin have used their own version of MATT, termed TraSH, to begin to define the set of genes that are essential for growth and/or survival of *M. tuberculosis* in a murine model after intravenous infection (C. Sassetti and E. Rubin, pers. commun.).

Additionally, they are beginning to perform microarray-based VAP screens to assign putative function to these genes by determining the subsets that are essential for survival under *in vitro* conditions that mimic different aspects of the environment encountered by *M. tuberculosis in vivo*.

## Conclusion

Genomic information and the advent of large-scale screening techniques such as IVET, STM, IVIAT and microarray technology have made it easier for us to dissect the intricate relationship between bacterial pathogens and their hosts by determining subsets of genes that are specifically induced or required within the context of the host milieu. Unfortunately, these techniques do not often assign functionality to the gene products they identify. The uses of the VAP screening technology described here, however, should begin to allow preliminary assignment of functionality to not only the many hypothetical genes but also to the known genes that could have previously unappreciated roles in diverse environmental conditions found within the host. Potential VAP screens that might shed valuable insight into host–pathogen interactions might include: (1) identification of virulence genes of *Listeria monocytogenes* that are required for mice systemic infection followed by a VAP-based screen to identify the subset defective for invasion, survival or growth in macrophages in tissue culture; (2) identification of *S. pneumoniae* genes that are required for bacteremia followed by a VAP-based screen to identify those strains sensitive to complement-mediated killing *in vitro*; and (3) identification of gastric colonization genes of *H. pylori* followed by VAP-identification of those genes essential for surviving oxidative, nitrosative or acid stress. Overall, the VAP technique should be widely applicable to many bacterial pathogens and the environments that can be explored are virtually limited only by the imagination of the investigator.

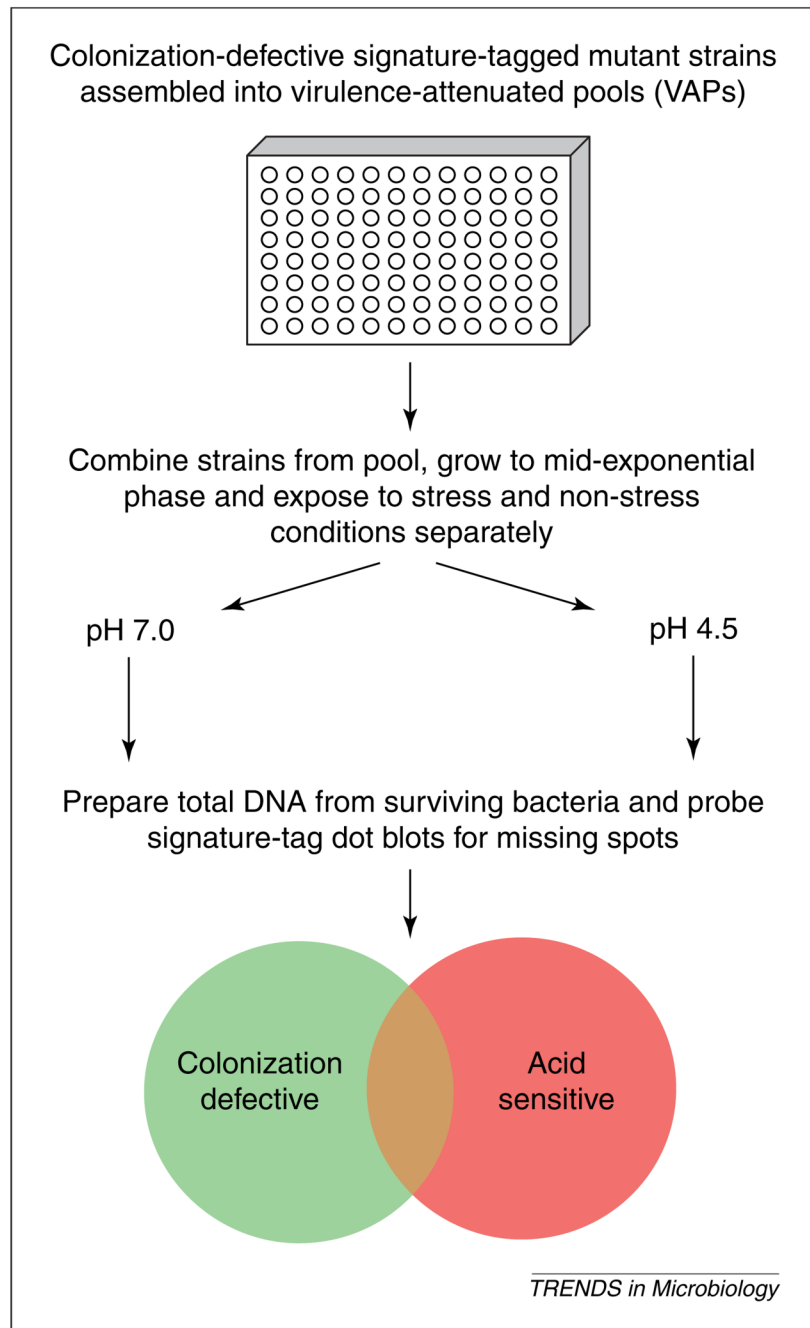
## Acknowledgments

We are thankful to D. Hava, E. Rubin, C. Sassetti and N. Salama for sharing data and information before publication. D.S.M. is supported by funds from the Damon Runyon Foundation and research in the laboratory of A.C. is supported through funds from the NIH.

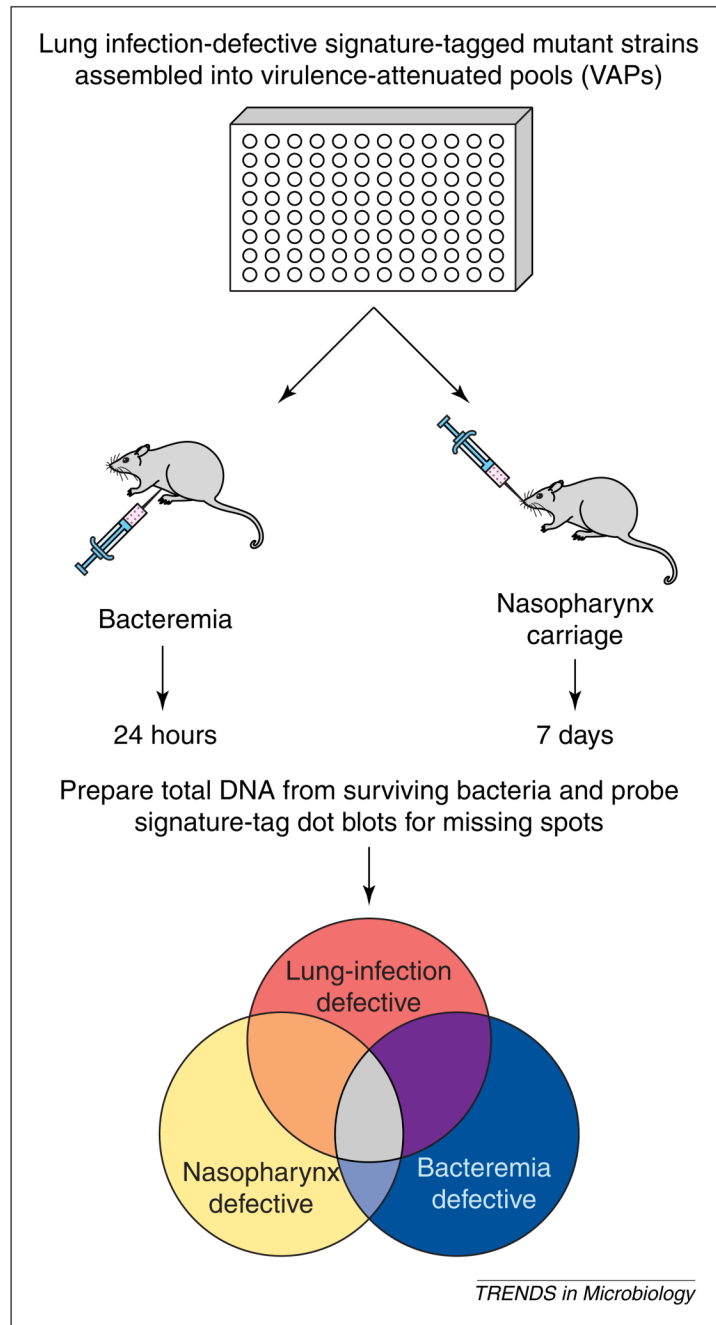
## References

1. Zhou, J.; Miller, JH. Microbial genomics – challenges and opportunities: the 9th international conference on microbial genomes; J. Bacteriol; 2002. p. 4327–4333.
2. Mahan MJ, et al. Selection of bacterial virulence genes that are specifically induced in host tissues. Science 1993;259:686–688. [PubMed: 8430319]
3. Hensel M, et al. Simultaneous identification of bacterial virulence genes by negative selection. Science 1995;269:400–403. [PubMed: 7618105]
4. Handfield M, et al. IVIAT: a novel method to identify microbial genes expressed specifically during human infections. Trends Microbiol 2000;8:336–339. [PubMed: 10878769]
5. Schena M, et al. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science 1995;270:467–470. [PubMed: 7569999]
6. Mecsas J. Use of signature-tagged mutagenesis in pathogenesis studies. Curr. Opin. Microbiol 2002;5:33–37. [PubMed: 11834366]

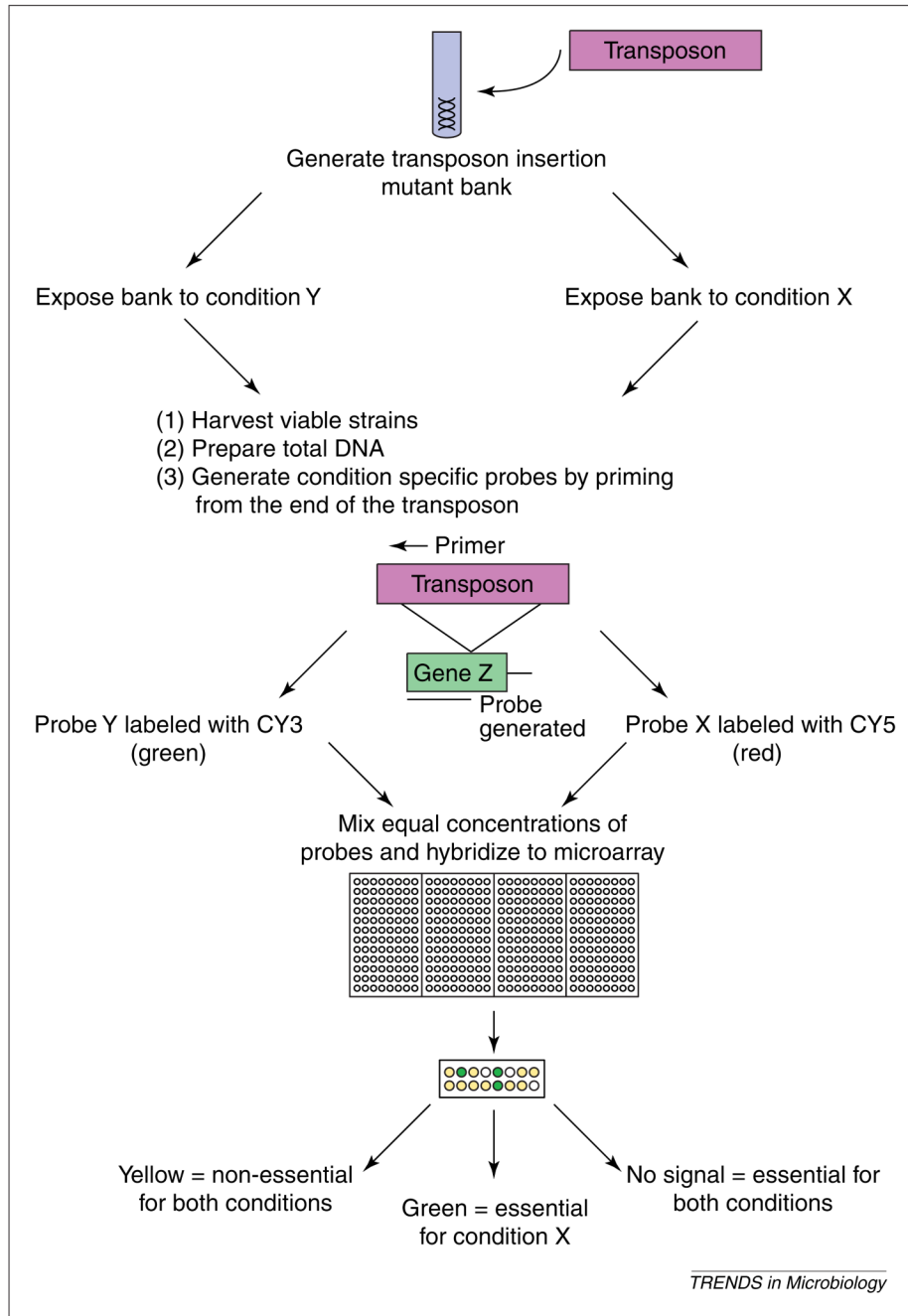
7. Mahan MJ, et al. Assessment of bacterial pathogenesis by analysis of gene expression in the host. *Annu. Rev. Genet* 2000;34:139–164. [PubMed: 11092824]
8. Kato-Maeda M, et al. Microarray analysis of pathogens and their interaction with hosts. *Cell. Microbiol* 2001;3:713–719. [PubMed: 11696031]
9. Merrell DS, et al. Host-induced epidemic spread of the cholera bacterium. *Nature* 2002;417:642–645. [PubMed: 12050664]
10. Merrell DS, et al. Identification of novel factors involved in colonization and acid tolerance of *Vibrio cholerae*. *Mol. Microbiol* 2002;43:1471–1491. [PubMed: 11952899]
11. Birrell GW, et al. Transcriptional response of *Saccharomyces cerevisiae* to DNA-damaging agents does not identify the genes that protect against these agents. *Proc. Natl. Acad. Sci. U. S. A* 2002;99:8778–8783. [PubMed: 12077312]
12. Badarinarayana V, et al. Selection analyses of insertional mutants using subgenic-resolution arrays. *Nat. Biotechnol* 2001;19:1060–1065. [PubMed: 11689852]
13. Sassetti CM, et al. Comprehensive identification of conditionally essential genes in mycobacteria. *Proc. Natl. Acad. Sci. U. S. A* 2001;98:12712–12717. [PubMed: 11606763]



**Fig. 1.** *In vitro* virulence-attenuated pool (VAP) screening to assign putative *in vivo* functionality. Pools of signature-tagged (ST) mutant bacteria that were previously determined to be attenuated for colonization in an animal model can be systematically exposed to various *in vitro* conditions that mimic environs that would be encountered within the context of a natural infection. In the example shown, *Vibrio cholerae* mutants were exposed to acidic conditions that mimic the low pH environment encountered by the bacteria during the course of transit through the human stomach. The theoretical classes of mutants that can be determined are represented by the Venn diagram at the bottom of the figure, where some genes that are required for colonization are also required for acid resistance.



**Fig. 2.** Signature-tagged mutagenesis (STM)-based virulence-attenuated pool (VAP) screening to identify tissue-trophic requirements for *Streptococcus pneumoniae* genes. Lung infection-defective mutants are assembled into pools and screened by STM to determine the requirement for each gene in nasopharyngeal carriage and bacteremia. The theoretical classes of mutants that can be determined are represented by the Venn diagram at the bottom of the figure, where some genes that are required for infection of the lung are also required for colonization of the nasopharynx and/or bacteremia.



**Fig. 3.** Microarray-based virulence-attenuated pool (VAP) screening using microarray transposon tagging (MATT) to determine *in vivo* functionality. Transposon libraries of bacteria are passed through an animal model as well as exposed to *in vitro* conditions that mimic host environments. Genomic DNA prepared from each pool of bacteria is subsequently used to amplify the chromosomal DNA flanking the transposon insertion. These flanking products are then labeled with different fluorophores and hybridized to the microarray to reveal the requirement for each gene in the conditions being tested.



**Table 1**

Percentages of different functional classes of genes in various bacterial pathogens

Pathogen	% Unknown function <sup>a</sup>	% Known function <sup>b</sup>
<i>Helicobacter pylori</i>	44.41	55.59
<i>Pseudomonas aeruginosa</i>	64.67	35.33
<i>Salmonella typhimurium</i>	46.53	53.47
<i>Staphylococcus aureus</i>	42.17	57.83
<i>Streptococcus pneumoniae</i>	36.57	63.43
<i>Vibrio cholerae</i>	45.37	54.63
<i>Yersinia pestis</i>	35.75	64.25
All <sup>c</sup>	52.59	47.41

<sup>a</sup>The percentage of genes of unknown function comprises the functional classes defined by the Institute for Genomic Research (TIGR) within the Comprehensive Microbial Resource (CMR) database (<http://www.tigr.org/tigr-scripts/CMR2/CMRHomePage.spl>). Those classifications that are defined here as 'unknown' function are the hypothetical, conserved hypothetical, unclassified and unknown function.

<sup>b</sup>The percentage of genes of known function comprises the remainder of the functional classes defined by TIGR. It should be noted that many of the genes in this class are categorized according to conserved motifs or predicted function (i.e. transporters, membrane proteins, regulatory functions and signal transduction) and might not be of actual known function. Therefore, the percentages listed are vast overestimates of the actual number of genes whose real roles are defined.

<sup>c</sup>All represents the cumulative percentage of each functional class for all 87 of the completed genomes contained within the CMR.