are immunogenic. Therefore, the vaccine antigens identified must be evaluated in appropriate challenge models, and limited availability of good animal models may slow the progress to vaccine trials.

COMPARATIVE GENOMICS

The availability of genome sequence information now allows us to compare species and strains within species for the identification of conserved genes and putative virulence factors. In particular, the analysis of the genetic variability between pathogens and closely related nonpathogenic microorganisms leads to the rapid identification of the complete set of genes potentially responsible for acquisition of virulence. This offers a valuable guideline into the search for suitable proteins to use as purified antigens in subunit-based vaccines. Vaccines should ultimately target antigens that are conserved among pathogenic strains. Therefore, comparative genomics can be used to find antigens that are likely to confer broad protection. This approach can also provide the rational basis for a safe and stable attenuation of live vaccine candidates or vectors for vaccine delivery.

Comparisons can be performed either with genome sequence or by using microarray-based methods. Owing to the improvement of sequencing technologies and the consequent reduction of sequencing costs, multiple genome sequences have been completed for several species over the past few years that enable quantitative analyses of their genomic diversity through comparative genomic analyses. Some pathogens exhibit very little, while others have marked genetic variability. For example, *M. tuberculosis* is now recognized to be an intracellular clonal bacterium that harbors relatively little genetic diversity. Studies based on whole-genome comparisons use single-nucleotide polymorphisms (SNPs) to investigate *M. tuberculosis* evolution and phylogeny (53). However, there is increasing evidence that the interstrain variation that exists is biologically significant; for instance, underlying biological differences among clinical strains have been associated with an adaptation to a specific host range or a response to variations in vaccination practices.

On the other hand, *E. coli* represents a very wide group of organisms that have high levels of intraspecies diversity that vary in as much as 25% of their genome (54). *E. coli* are part of the natural gut flora as commensals but can also cause diverse infections in very different niches. Genomic variations that occur in the form of individual genes or larger genome islands contribute to differences in virulence potential. Isolates of the same species can be analyzed experimentally by subtractive hybridization and comparative genome hybridization (CGH). Microarrays spotted with predicted ORFs of a reference strain can be hybridized with labeled DNA from an experimental strain, allowing genes common to both, as well as those present in the reference strain but absent in the test strain, to be identified. CGH allows high-throughput, high-resolution global genome analysis without the need to sequence all strains tested. In a recent study, microarrays based on the genome sequence of CFT073 were utilized in CGH analysis of a panel of uropathogenic and fecal/commensal *E. coli* isolates. This approach resulted in the identification of 131 genes that were exclusively found in ueopathogenic *E. coli* (UPEC) relative to commensal and fecal isolates (55). However, this highlights one intrinsic technical limitation of microarray: detection is limited to the DNA spotted on the array.

The Pan-Genome

Multiple genomes of the same species and comparative genomics have led to an increased understanding of the intraspecies diversity, making clear that the sequence of one genome may not be sufficient to represent the genetic diversity of a microorganism. To overcome this limitation, the pan-genome concept was introduced (56) to define the global gene repertoire possibly pertaining to a given species. The unexpectedly high degree of intraspecies diversity 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. In the seminal work of Tettelin et al., authors set about answering the question of how many genomes are needed to fully describe a bacterial species using eight genomes representative of the diversity among group B *Streptococcus* (GBS) (56). Comparative analysis of the genomes enabled the estimation that 1806 genes are shared by all strains of *Streptococcus agalactiae*, and these genes form the species ''core genome.'' This represents approximately 80% of the average number of genes encoded in each strain and, in general, includes all genes responsible for the basic aspects of the biology of a given species. Instead, the ''dispensable genome'' is composed of genes absent or partially shared and strain-specific genes, and these genes are responsible for species diversity and might encode functions that can confer selective advantages. Surprisingly, mathematical extrapolation of the existing data predicted that, no matter how many strains have been sequenced, each new sequence would contain genes that have not been encountered before, leading to the counterintuitive conclusion that this species pan-genome continues to grow without bounds as the number of sequenced strains grows (Fig. 2), defined as an ''open'' pan-genome. It was estimated that the sampling of subsequent genomes would continue to reveal new genes, on average 33 per genome (56). However, the extent of intraspecies diversity is not always so vast, and a different behavior was observed in the study of eight independent *B. anthracis* isolates. In this case, the number of specific genes added to the pan-genome was found to rapidly converge to zero after the addition of only a fourth genome (56) (Fig. 2). Hence, the *B. anthracis* species has a ''closed'' pan-genome, and four genome sequences are sufficient to completely characterize this species, at least in terms of gene content (Fig. 2). A subsequent analysis of seven *E. coli* genomes has shown an extreme flexibility, with each new sequenced strain contributing 441 new genes to the core genome comprising 2865 genes, leading again to an open pan-genome (57). Mathematical modeling suggests that hundreds of genomes of other species will follow the same trend (58). Given that the number of unique genes is vast, the pan-genome of a bacterial species might be orders of magnitude larger than any single genome (58).

In conclusion, species can have an open or a closed pan-genome. An open pan-genome is typical of species that either colonize multiple niches or have efficient mechanisms, such as natural competence, of exchanging genetic material with unrelated species present within the same environment (e.g., *Helicobacter pylori*, *E. coli*, *N. meningitidis*, and *Streptococcus*). By contrast, other more clonal species (such as *B. anthracis*, *M. tuberculosis*, and *C. pneumoniae*), which are more conserved, live in isolated niches with limited access to the global microbial gene pool and, therefore, have a low capacity to acquire foreign genes. Data from multiple