ELSEVIER

'The avalanche of genomic data will provide innovative tools to prevent and cure bacterial infections'

The pan-genome: towards a knowledge-based discovery of novel targets for vaccines and antibacterials

Alessandro Muzzi, Vega Masignani and Rino Rappuoli

Novartis Vaccines, Via Fiorentina 1, 53100 Siena, Italy

During the past decade, sequencing of the entire genome of pathogenic bacteria has become a widely used practice in microbiology research. More recently, sequence data from multiple isolates of a single pathogen have provided new insights into the microevolution of a species as well as helping researchers to decipher its virulence mechanisms. The comparison of multiple strains of a single species has resulted in the definition of the species pan-genome, as a measure of the total gene repertoire that can pertain to a given microorganism. This concept can be exploited not only to study the diversity of a species, but also, as we discuss here, to provide the opportunity to use a knowledge-based approach for the development of novel vaccine candidates and new-generation targets for antimicrobials.

Introduction

After the first complete genome sequence of a free-living organism, *Haemophilus influenzae*, was determined in 1995 [1], whole-genome sequencing became a standard, rapid method for the study of the biological processes in living forms. During the past decade, the number of available complete genome sequences has grown exponentially and, currently, the most up-to-date genomic databases [2] contain 433 published sequences (42 eukaryotic and 391 prokaryotic); there are 1683 ongoing projects (634 eukaryotic and 1049 prokaryotic), with the chromosomal sequence of pathogenic organisms being the most represented. This increasing amount of sequence data can be used to find new strategies for the discovery of drug targets, diagnostic markers and vaccine candidates. One of the most popular applications of genomic technologies is the reverse vaccinology approach, where the whole genome of a bacterial pathogen is analyzed by computational approaches to predict *in silico* previously undiscovered vaccine candidates.

At the beginning of the genomic era, it was thought that a single representative isolate was sufficient to describe the genetic complexity of a species, and the use of 'comparative genomics' was restricted to investigating the diversity among different yet closely related bacteria. More recently, isolates of the same species have been analyzed by subtractive hybridization and comparative genome hybridization (CGH). These studies have shown that intraspecies variation can be as significant as interspecies diversity [3–5]. Comparative genomics studies have revealed that the microbial genome is a dynamic entity shaped by multiple forces, including genome reduction, gene duplication and loss, genome rearrangements, and acquisition of new genes

Corresponding author. Rappuoli, R. (rino.rappuoli@novartis.com)

ALESSANDRO MUZZI

Alessandro Muzzi studied physics at the University of Florence, Italy. In 2000, he joined the Cellular Microbiology and Bioinformatics Unit at Novartis Vaccines & Diagnostics in Siena, Italy, as a bioinformatics researcher. His research interests focus on bacterial genomics and microarray data analysis as applied to the discovery of targets for vaccine development.

VEGA MASIGNANI

Vega Masignani received her PhD in Biotechnology with a thesis on the computational approach to the development of a novel *Neisseria meningitidis* protein-based vaccine. She is currently a senior scientist at the Cellular Microbiology and Bioinformatics Unit of Novartis Vaccines & Diagnostics, carrying out research in the field of computer analysis as applied to microbial pathogenesis.

RINO RAPPUOLI

Rino Rappuoli is the Global Head of Vaccines Research at Novartis Vaccines & Diagnostics. He is member of the US National Academy of Sciences. He developed the first recombinant bacterial vaccine (against pertussis) and a conjugate vaccine against meningococcus C. Currently, he is involved in the development of a vaccine against group B meningococcus using a genome-based approach termed 'reverse vaccinology,' and in the development of a vaccine against avian influenza.

1359-6446/06/$ - see front matter doi:10.1016/j.drudis.2007.04.008

through lateral gene transfer [6]. For example, comparison of the genetic repertoire of multiple pathogenic strains of *Helicobacter pylori* isolated from a single patient showed that a single clone colonizing a restricted environment such as the stomach can generate an appreciable amount of genetic diversification even over the course of less than a decade [7,8]. As a further example, the sequencing of *Escherichia coli* O157:H7 in 2001 [9] revealed that this new isolate contained >1300 strain-specific genes compared with the *E. coli* laboratory strain K12, reinforcing the idea that sequencing a single strain is not sufficient to cover fully the genomic structure of a bacterial pathogen. For this reason, genomic data from multiple strains of the same species are required to answer general questions about bacterial physiology; for example, identifying the minimum set of essential genes required for basic metabolism of that species, and to form a global picture of fundamental processes such pathogenesis, drug resistance, environment adaptability and evolution [10,11]. Consequently, during the past few years, multiple complete genomes have been made available for single species, such as *Bacillus anthracis*, *Streptococcus agalactiae*, *Staphylococcus aureus*, *E. coli* and *Streptococcus pyogenes*, paving the way to a better understanding of species complexity and to a re-evaluation of current strategies to combat these pathogens.

Our increasing understanding of the intraspecies diversity challenged the present definition of species, which is substantially defined as a group of strains with at least one classifiable phenotypic trait and whose genomic DNA molecules show at least 70% of

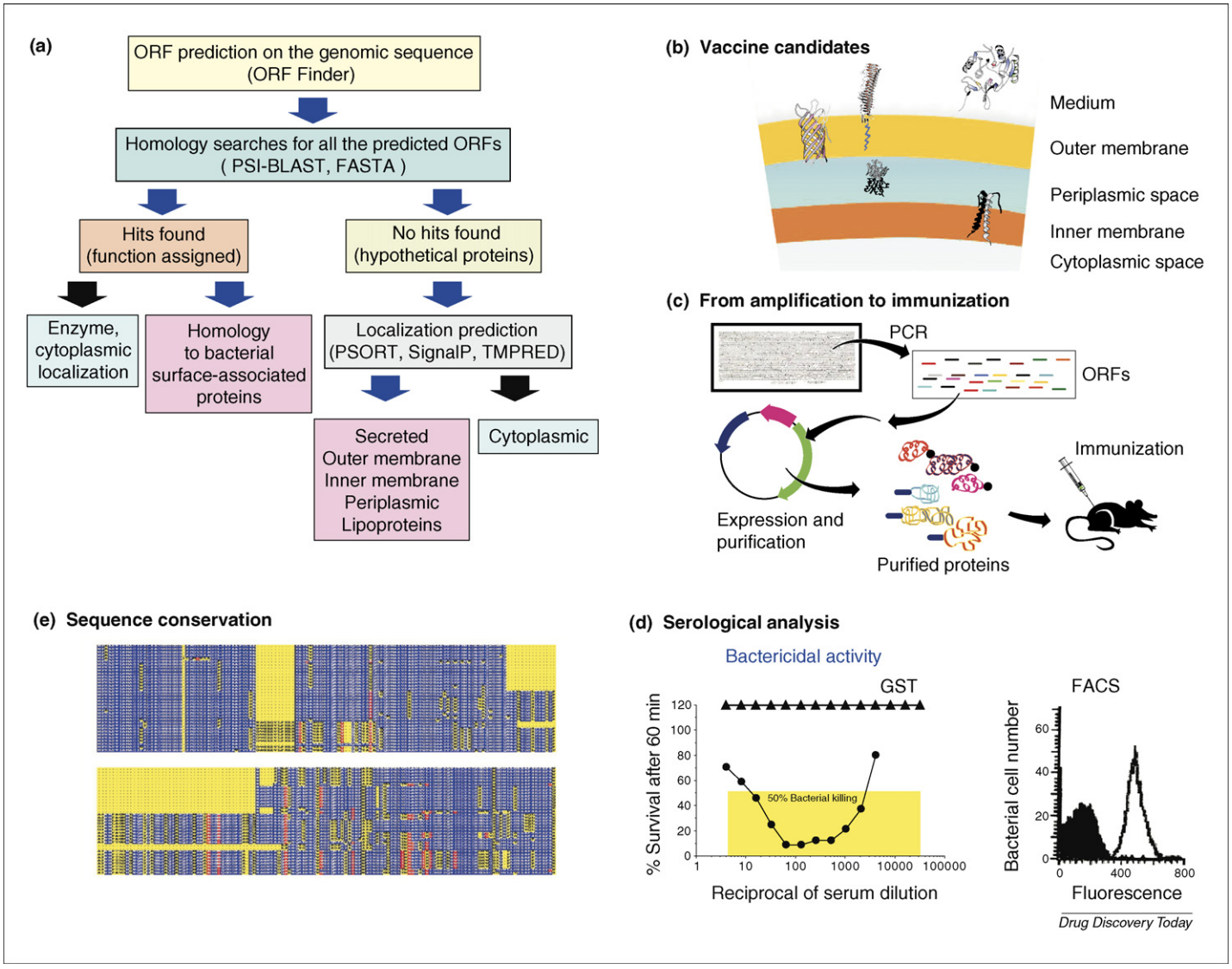

FIGURE 1

The reverse vaccinology approach, arbitrarily applied to a Gram negative organism. (a) Flow chart of the computer analysis strategy generally used for the identification of potential vaccine candidates. The program used for ORF prediction is ORFFinder (http://www.ncbi.nlm.nih.gov/projects/gorf/). The programs used for homology searches are: PSI-Blast (http://www.ncbi.nlm.nih.gov/BLAST/) and FASTA (http://fasta.bioch.virginia.edu/fasta_www2/fasta_list2.shtml). The programs used for cellular localization prediction are: i. PSORT (http://www.psort.org); ii. SignalP (http://www.cbs.dtu.dk/services/SignalP/); iii. TMPRED (http://www.ch.embnet.org/software/TMPRED_form.html). **(b)** The selected gene products belong to either of the following groups: exported proteins, outer membrane and porin-like proteins, periplasmic proteins or integral membrane proteins. **(c)** The candidate genes are PCR amplified, expressed in *Escherichia coli*, purified and used to immunize mice. **(d)** The sera can be used in *in vitro* assays, such as fluorescence-activated cell sorting analysis (FACS), to determine expression of the antigen on the cell surface of the bacterium and to measure the ability of specific antibodies to kill the pathogen (as in the bactericidal assay, which is used for *Neisseria meningitidis*). **(e)** Finally, all the selected vaccine candidates are screened for sequence conservation on a panel of strains.

reassociation values [12], making clear that the sequence of one genome was not sufficient to represent the genetic diversity of a microorganism. To overcome this limitation, the pan-genome concept was introduced [13,14] to define the global gene repertoire possibly pertaining to a given species.

Although providing a more comprehensive definition of a bacterial species, the pan-genome has also opened up new possibilities for vaccine design, and will hopefully prove useful to better select novel targets for antimicrobials. Here, we discuss how genomics has influenced the antimicrobial and vaccine research field, with a particular focus on the potential of the pan-genome concept.

Reverse vaccinology: a successful application

Biochemical, immunological and microbiological methods have been used for decades to identify microbial components that are important for understanding the pathogenesis and that are useful for vaccine development. Although several antigens have been identified using this conventional approach, in most cases, decades are required to characterize and propose them as vaccine components. Furthermore, this approach has been unsuccessful for those pathogens whose protective antigens are difficult to identify, or that are expressed only during infection or in particular growth conditions. By contrast, the complete sequence of a bacterial genome provides the opportunity to tackle vaccine development from a new perspective, showing simultaneously all the protein antigens, independent of their expression level and abundance [15].

On the same line of the 'reverse immunology' approach, where potential HLA-binding peptides are predicted on the basis of their sequence features, reverse vaccinology makes use of bioinformatics methods to exploit the information derived from the complete genomic sequence of a pathogen and to predict potential vaccine candidates 'in silico' (Figure 1). This approach demonstrated, for the first time, that sequencing a bacterial genome was not only a major scientific achievement, but could also be of practical use.

Reverse vaccinology was applied for the first time to serogroup B *Neisseria meningitidis* [16,17]. An invasive isolate of *N. meningitidis* (MC58) was fully sequenced and analyzed to identify suitable vaccine candidates by an *in silico* approach using few basic assumptions. The primary criterion for a bacterial protein to be considered an antigenic molecule to use for vaccination is its cellular localization. Cytosolic proteins are unlikely to be immunological targets, whereas surface-associated and secreted structures are more easily accessible to antibodies, the primary immune effector molecules against this bacterial pathogen.

Proteins were predicted to be surface exposed based on the combination of several pieces of information, including the presence of amino acid motifs that are responsible for targeting the mature protein to the outer membrane (signal peptides), to the lipid bilayer (lipoproteins), to the integral membrane (hydrophobic transmembrane domains), or for recognition and interaction with host proteins or structures (e.g. adhesins having integrin-binding domains) (Figure 1).

Within 18 months from the start of sequencing, 600 potential vaccine candidates were identified on the basis of these criteria. The selected proteins were expressed in *E. coli*, purified and used for immunization of mice. Antisera raised against the injected proteins were assayed for specificity (by western blot), accessibility on the surface of the pathogen (by flow cytometry), and for their ability to kill bacteria when combined *in vitro* with human complement (bactericidal assay). Each experimental step reduced the number of potential vaccine candidates to a defined set of proteins that satisfied all the criteria and warranted high probability of success for the development of a vaccine. At the end of the screening, the initial set of 600 potential candidates was reduced to 91 surface-exposed molecules, of which 29 could induce a complement-mediated bactericidal antibody response.

As one of the major problems with meningococcal antigens is sequence variability, the selected candidates were checked for conservation across a panel of diverse strains of *N. meningitidis* representing all serotypes and spanning the phylogeny of the species [16]. This analysis yielded a handful of antigens, which were both conserved in sequence and able to elicit a cross-bactericidal response against all the strains in the panel, demonstrating that they could confer general protection against meningococcus. These promising vaccine candidates are currently being tested in phase I clinical trials [18]. To strengthen the protective activity of the single protein antigens and to increase strain coverage, the final vaccine formulation comprises a 'cocktail' of the selected antigens. The success of reverse vaccinology for Meningococcus led to the application of this approach to a variety of other pathogens, such as *Streptococcus pneumoniae* [19], *Porphyromonas*

TABLE 1

Examples of human pathogens that have been studied using reverse vaccinology and pan-genomic approaches, and the status of corresponding vaccine development

Pathogen	Disease	Approaches for vaccine development	Status of vaccine development	Refs
Bacillus anthracis	Causative organism of potentially fatal anthrax	Reverse vaccinology	Preclinical testing	[22]
Chlamydia pneumoniae	Pneumonia; also associated with atherosclerotic and cardiovascular disease	Reverse vaccinology; proteomics	Preclinical testing	[21]
Neisseria meningitidis (Serogroup B)	Bacterial septicaemia and meningitis	Reverse vaccinology	Phase I clinical trials	[16]
Porphyromonas gingivalis	Periodontitis	Reverse vaccinology	Preclinical testing	[20,73]
Streptococcus pneumoniae	Pneumonia, middle ear infections and meningitis	Reverse vaccinology	Preclinical testing	[19]
Streptococcus agalactiae (Group B streptococcus)	Bacterial sepsis, pneumonia and meningitis in neonates	Reverse vaccinology; pan-genome	Phase I clinical trials	[4,38]

gingivalis [20], *Chlamydia pneumoniae* [21], *B. anthracis* [22], and others (Table 1).

Thus, the reverse vaccinology approach appears to be applicable to a range of pathogens and, in principle, also to eukaryotic parasites, with the only limits being the availability of the genome sequence and suitable *in vitro* or *in vivo* models to test the capacity of the antigens to induce a protective immune response. Furthermore, one advantage of reverse vaccinology is that all the antigens are produced as soluble recombinant proteins in *E. coli*, thus supporting the straightforward development of a suitable manufacturing process for large-scale production.

What about viral genomes?

Although viruses generally have a more simple structural organization compared with that of bacterial pathogens, the development of effective vaccine formulations against some of them is still complicated. For example, the virus responsible for hepatitis C (HCV) had never been cultivated *in vitro* and, therefore, any conventional approach to vaccine development was not applicable. In 1990, the availability of the genome sequence of a hepatitis C virus [23] finally enabled the identification of envelope proteins E1 and E2, which were then expressed in different hosts and shown to induce the production of antibodies that interfered with the binding of E2 to the host receptor [24]. More importantly, the recombinant proteins protected chimpanzees from infection with the homologous HCV virus [25]. This success made HCV vaccine the first case of a vaccine entirely developed by the use of reverse vaccinology [26].

For other viruses such as HIV and dengue, the development of effective universal vaccines has so far been impaired by the high antigenic diversity and the extremely rapid evolution associated with these microbial pathogens. To overcome these problems, bioinformatics technologies have been recently applied to the study of these viruses.

In the case of dengue, the virus exists in four serotypes, which show 30–40% amino acid variability. Strains within each serotype are more conserved, but different intra-serotype variants can be defined. In recent work, Khan and colleagues [27] applied a systematic computer-based approach to explore the antigenic diversity of dengue by analyzing >9000 dengue virus protein sequences to identify the minimal set of sequences encoding the complete genetic diversity of short peptides representing possible T-cell epitopes from all known sequences of dengue virus serotypes. The authors found that a limited number of short peptides of the viral proteins are sufficient to represent the whole antigenic diversity of T-cell epitopes, thus paving the way for similar analyses of other highly variable viral species (e.g. HIV and influenza virus).

Similar computer-based methods have also been applied to the study of HIV-1, to identify potential T-cell epitopes for the construction of 'mosaic' proteins to use as polyvalent vaccine antigens, which can cover the diversity of the HIV viral species [28]. The mosaics are generated from natural sequences, but systematically include only common potential epitopes, which are then computationally optimized to achieve the highest coverage rate.

These examples illustrate the value of the bioinformatics approach to treat different levels of biological problems, which were difficult or impossible to address with conventional microbiological strategies. Such examples suggest that genomics represents the future of biological research.

An evolution of genomics-based strategies: the bacterial pan-genome

Even before the pan-genome concept was developed, several studies analyzing multiple isolates of the same species by subtractive hybridization and comparative genome hybridization (CGH) had shown that bacterial species such as *H. pylori*, *S. aureus* and *E. coli* have extensive genetic diversity, with an average of 20–35% of

TABLE 2

Species with three or more sequenced genomes[a]

Species	No of species	No of genomes sequenced
Bacillus cereus group	1	21
Listeria monocytogenes	1	19
Escherichia coli	1	17
Streptococcus pyogenes	1	12
Burkholderia pseudomallei group, *Staphylococcus aureus*	2	11
Burkholderia cepacia complex, *Vibrio cholerae*	2	10
Burkholderia mallei, *Campylobacter jejuni*	2	9
Streptococcus agalactiae,*Yersinia pestis*	2	8
Francisella tularensis, *Pseudomonas aeruginosa*, *Rickettsia spotted fever* group	3	7
Mycobacterium tuberculosis complex, *Prochlorococcus marinus*	2	6
Ehrlichia canis group, *Rhodopseudomonas palustris*	2	5
Buchnera aphidicola, *Chlamydophila pneumoniae*, *Coxiella burnetii*, *Haemophilus influenzae*, *Xylella fastidiosa*	5	4
Borrelia burgdorferi group, *Clostridium perfringens*, *Fusobacterium nucleatum*, *Helicobacter pylori*, *Legionella pneumophila*, *Mycoplasma hyopneumoniae*, *Pseudomonas putida* group, *Pseudomonas syringae*, *Rhodobacter sphaeroides*, *Rickettsia canis* group, *Salmonella enterica*, *Shigella flexneri*, *Streptococcus pneumoniae*, *Streptococcus thermophilus*, *Xanthomonas campestris*, *Xanthomonas oryzae*	16	3

[a] Source: http://www.ncbi.nlm.nih.gov/genomes/MICROBES/microbial_taxtree.html.

genes that are not conserved in all strains of the same species [3,5,29].

This observation drew attention to the use of genomics for the investigation of the variability within a single species or genus and for the study of microevolution in bacteria. 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, the most represented being *Bacillus cereus* with 21 strains, *Listeria monocytogenes* with 19 strains, and *E. coli* with 17 strains (Table 2). The availability of genome sequences for different isolates of a single species enables quantitative analyses of their genomic diversity through comparative genomic analyses. The higher the number of isolates and the broader the selection of strains, the better the estimate of the whole species heterogeneity.

The definition of the bacterial pan-genome was proposed by Tettelin *et al.* [13] in a study in which eight different isolates of *S. agalactiae* (group B *Streptococcus* or GBS) representative of the genetic diversity of the species were completely sequenced and compared. The pan-genome can be defined as the global gene repertoire pertaining to a species. In general, it can be divided in three parts: the core-genome, which includes the set of genes invariably present and conserved in all the isolates; the 'dispensable genome', comprising genes present in some but not all the strains, and the strain-specific genes, which are present only in one single isolate (Figure 2).

The analysis performed on GBS genomes indicated that the different isolates have an estimated core-genome containing 1806 genes, whereas each single genome contained between 2000 and 2400 genes. In other words, each strain contains a

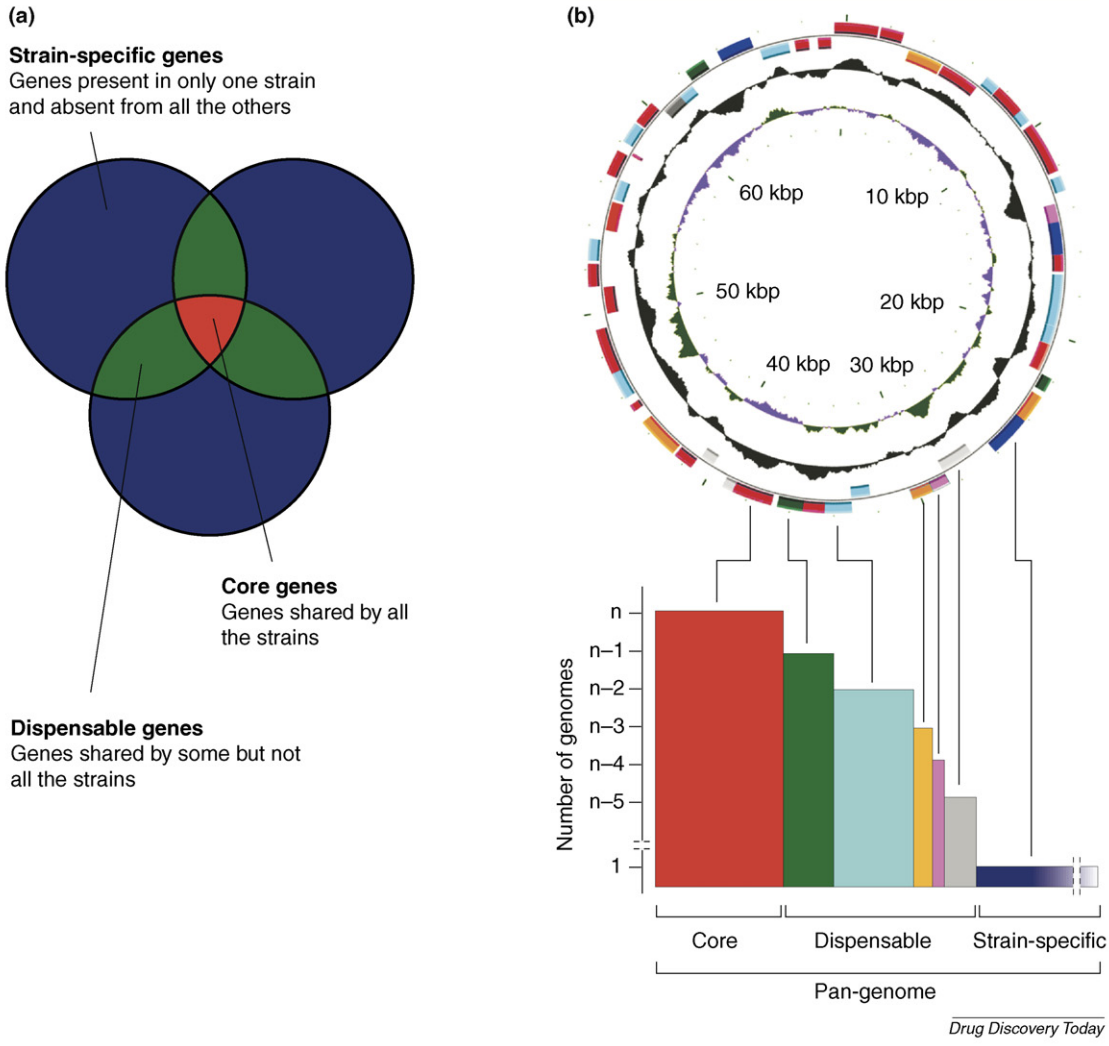

FIGURE 2

The species pan-genome. (a) Schematic image of the pan-genome structure. Core, dispensable and strain-specific genes are indicated in red, green and violet, respectively. Circles represent different strains. For simplicity, a comparative analysis performed on a set of three genomes is shown. **(b)** Gene classification of a new sequenced genome according to the pan-genome structure (core, dispensable and strain-specific genes). Each bar represents the group of genes conserved in n, n − 1, n − 2,. . ., 1 genomes. In species with an open pan-genome (e.g. *S. agalactiae and S. pyogenes*), each new sequenced strain contributes a finite number of novel genes; therefore, the bar corresponding to strain-specific genes increases continuously in size. By contrast, species having a closed pan-genome (e.g. *B. anthracis*) are more conserved, and a limited number of strains is sufficient to explore their genetic diversity. As a result, the bar corresponding to strain-specific genes does not increase; it either stays constant or drops to zero.

BOX 1

Mathematical definition of the pan-genome: methods to identify core, dispensable and strain-specific genes

To calculate the number of genes comprising the species pan-genome, an all-versus-all comparison of the strains is performed using three different methods: (i) a protein versus protein search; (ii) a DNA search of all the predicted ORFs of a strain against the genomic sequence of the other strain; and (iii) a translated protein search of all the predicted proteins of a strain against the complete DNA sequence of the other strain. A gene is considered conserved if at least one of these three methods produced an alignment with a minimum of 50% sequence conservation over 50% of the protein and/or gene length.

When each new genome is compared with the others, the number of new genes that it contains is measured. In the case of GBS, the curve extrapolated from the comparative data of eight genomes reaches a non-zero asymptotic value. In other words, after eight genomes, for every new GBS genome sequenced, an average of 33 new strain-specific genes will always be identified (Figure Ia, blue line). This implies that the GBS pan-genome is open, meaning that its size grows indefinitely with the number of independent strains sequenced.

Conversely, in the case of a clonal species such as *Bacillus anthracis*, the number of specific genes added to the species pan-genome drops to zero after the addition of a fourth genome, indicating that *B. anthracis* has a closed pan-genome (Figure Ib, red line).

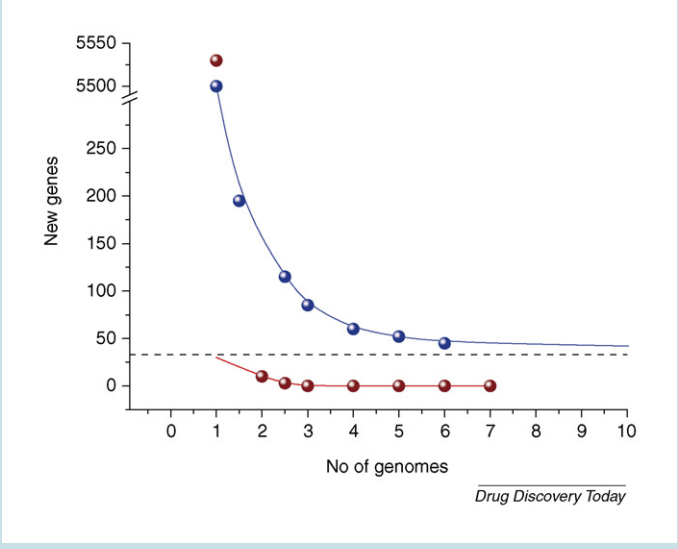

FIGURE I

The pan-genome. The number of specific genes is plotted as a function of the number *n* of strains sequentially added. For GBS (blue line), the extrapolated average number of strain-specific genes (33) is shown as a dashed line. For *Bacillus anthracis* (red line), the curve reaches zero after addition of the fourth genome. No new genes will be discovered after this threshold.

relatively large number of dispensable genes (200–600) that are missing in at least one of the other strain genomes. The most interesting finding is that when a new genome sequence is added to the pool of the others, an estimated number of 33 new strain-specific genes, which are exclusively present in that genome, are added. Consequently, the pool of genes comprising the core-genome, dispensable genome and all the strain-specific genes, globally defined as the pan-genome, represents an open entity (open pan-genome) that is continuously increasing in size (Box 1).

A similar analysis carried out on five strains of *S. pyogenes* revealed the same kind of genomic diversity, with an average of 27 specific genes for each new genome added, leading again to an open pan-genome. The *E. coli* pan-genome has been analyzed in a recent study [30] and has shown an extreme flexibility, with each new sequenced strain contributing 441 new genes to the core-genome comprising 2865 genes.

Interestingly, a different behaviour was observed in the study of eight independent *B. anthracis* isolates. In this case, the number of strain-specific genes added to the pan-genome was found to reduce rapidly to zero after the addition of a fourth genome (Box 1). This indicates that *B. anthracis* has a poorly variable structure and that four genomic sequences are enough to complete its pan-genome (close pan-genome). These results reflect the fact that *B. anthracis* is a recently evolved highly clonal species, where genome variability is only associated with the presence or absence of the virulence plasmids, and no further strain-specific genes have yet contributed to genome diversification [31].

In general, the core genome contains a variety of genes mostly encoding for housekeeping functions. Dispensable and strain-specific genes might confer selective advantages, especially for species with an open pan-genome, including adaptation to particular niches, antibiotic resistance and colonization of different hosts.

Investigation and functional annotation of the dispensable genome reveal that this subgroup mainly comprises hypothetical, phage- and transposon-related genes [32], thus showing that mobile elements, which are generally present in a limited number of strains, contribute poorly to the overall fitness and differentiation of the species, and only rarely harbour important genes [33,34]. As these genes are not necessary for the survival or maintenance of the species, they can be deleted from the genome. However, in pathogenic species, this loss is often accompanied by a parallel reduction in virulence, as in the case of the spontaneous loss of fimbriae, hair-like projections that are thought to have an important role in colonization, as it has been observed after laboratory manipulations of *H. influenzae* and *E. coli* [35].

In conclusion, species can have an open or a closed pan-genome. An open pan-genome is typical of those 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. *Streptococci*, *Meningococci*, *H. pylori*, *Salmonellae* and *E. coli* have these properties and are likely to have an open pan-genome. By contrast, other species such as *B. anthracis*, *Mycobacterium tuberculosis* and *Chlamydia 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. An extreme example is *Buchnera aphidicola*, which is an endosymbiont of aphids and whose genome has undergone no chromosome rearrangements, duplications or horizontal gene transfer during the past 50 million years, thus showing the most extreme genome stability to date [36].

Application of the pan-genome to vaccine design

From the determination of a species pan-genome, we can infer important practical information for vaccine design. Although the core genes represent the most desirable source for the selection of

conserved and, therefore, potentially universally applicable vaccine candidates, they are also more likely to be immunologically silent in any successful pathogen. The group of dispensable genes, by contrast, might be an invaluable source of novel antigens that, although only present in a subgroup of strains, might encode important virulence-associated functions [37] and might be exploited in appropriate combinations to elicit a broad immune response.

The first application of a pan-genomic approach to vaccine design was performed by Maione and colleagues on *S. agalactiae* (GBS) [38]. Following the classification of the GBS global gene pool into core and dispensable genome, classical bioinformatic algorithms were used to select genes from the two sub-genomes that encode putative surface-associated and secreted proteins. Among the identified proteins, 396 were part of the core and 193 were part of the dispensable genome. Selected proteins (potential antigens) were expressed as recombinant proteins, purified and tested for protection using an animal mouse model, in which mothers are immunized with the selected antigen, and the offspring are then challenged with the infecting strain. Four antigens were capable of significantly increasing the survival rate among challenged infant mice. Unexpectedly, only one of these antigens was part of the core genome, the remaining three were encoded by genes present in ~75% of strains. However, the conserved antigen was not able to confer global protection, as its level of expression was shown to be highly variable among different strains [38]. The final vaccine formulation comprises a combination of the four antigens, which provide overall almost universal strain coverage, with levels of protection similar to those seen when using capsular carbohydrate-based vaccines [38]. This example demonstrates the importance of having access to the genome sequence of multiple strains and performing up-front genome comparisons prior to developing vaccines based on genome data.

Although this concept should, in principle, be applied to all pathogenic organisms, it is especially important for bacteria with an open pan-genome, where the selection of targets on the basis of a limited number of sequenced genomes could lead to antigens able to protect only against a limited number of strains. Mobile genetic elements, such as transposons, plasmids and phage related genes, are frequently responsible for the pathogenic activity of bacteria or for drug resistance mechanisms. Generally, these genes do not belong to the core-genome; thus, their characterization will become more accurate as more genomes are sequenced.

The past and future of antimicrobials: help from the pan-genome?

Classic drug discovery strategies

Historically, we can distinguish two main strategies in antimicrobial drug discovery: 'whole-cell screening' and 'target-based screening' [39]. Whole-cell screening is based on the identification of compounds that can kill the pathogen directly. This approach uses intact bacterial cells to identify chemical or natural compounds that can inhibit bacterial growth *in vitro*. For drug discovery, the advantage of this approach is that drug penetration and microbial susceptibility are tested immediately, thus selecting only molecules with established antibacterial properties. Whole-cell screening is the classic approach for identifying antimicrobials and has led to the discovery of penicillin and sulfonamides. However, the increase in drug resistant bacterial strains and the small number of newly discovered compounds during the past three decades is a sign of the need for new approaches. An additional disadvantage of whole-cell screening is the complete lack of knowledge concerning the target and the mechanism of action. Consequently, drug optimization can be difficult.

In 'target-based screening', the drug is identified or modelled to hit a predefined metabolic process. In this case, rational drug design supported by specific molecular models can be envisaged. The first step is the identification of the genes that are essential for pathogen survival (*in vitro* or *in vivo*). Once identified, the essential genes are tested as targets of specific compounds derived from large chemical libraries. Having an essential biological role is the most important requirement for a drug target, although other characteristics are also desirable, such as molecular and functional conservation in multiple bacterial species (for broad spectrum antibiotics); selectivity against the pathogen to limit toxicity in humans; ability to test the enzymatic activity *in vitro*; a low rate of resistance; in addition, bactericidal compounds are preferable to bacteriostatics (i.e. compounds that inhibit growth).

Genomic-based strategies

With the advent of genomic technologies, comparative genomics was used to recognize sets of bacterial essential genes, assuming that genes conserved in phylogenetically distant species should encode products whose functions are indispensable to sustain cellular life. Following the publication of the first two genomic sequences of *H. influenzae* and *Mycoplasma genitalium*, a comparative study was performed to identify the set of genes shared between these two organisms [40,41]. This group, comprising 240 genes, was shown to represent an 'ancestral' core-genome of indispensable functions. Similar comparative studies have confirmed that genes conserved in different genomes are often essential [42–45].

The recently completed sequence of the human genome represents a major step in drug discovery [46,47]. One of the most recently adopted strategies for target identification is based on a comparative genomic approach between human host and pathogen genomes to identify the set of genes that are likely to be specific and essential to the pathogen but absent in the host [48,49]. Together with comparative genomics, several technologies based on whole-genome random or targeted mutagenesis were applied to test gene essentiality experimentally (for a comprehensive description of these methods, see Refs [39,50,51]). The first examples were the random conditional lethal mutants under non-permissive growth conditions, such as high temperatures, which enabled essential targets in a small number of Gram-positive and Gram-negative organisms to be validated. Another random mutagenesis approach uses either transposon insertions or plasmid or cassette integration, which disrupt the gene located at the insertion site. Typically, these techniques detect the disruption of non-essential genes and essential genes can be determined by subtraction. Using this approach, 265–350 genes were determined to be essential in *M. genitalium* compared with the total of 484 genes in the genome [52]. Similar studies performed on other pathogens led to the identification of essential genes for *H. influenzae* (478–670 essential genes out of 1272 screened) [53], *S. pneumoniae* (113 essential genes) [54], *E. coli* (620 essential genes out of 4291 in

TABLE 3

Databases of essential genes

Species	Databases of essential genes					Refs
	PEC[a]	DEG[b]	WISC[c]	NMPDR[d]	SEED[e]	
Bacillus subtilis		•		•	•	[45]
Escherichia coli				•	•	[77]
		•	•	•	•	[55]
	•					[74]
			•			[78]
			•			[79]
Haemophilus influenzae		•		•	•	[53]
Helicobacter pylori		•		•	•	[56]
Mycoplasma genitalium		•		•	•	[80]
		•				[52]
Mycobacterium tuberculosis		•		•	•	[81]
Pseudomonas aeruginosa				•	•	[82]
				•	•	[83]
Staphylococcus aureus		•	•	•	•	[57]
		•		•	•	[84]
Salmonella enterica		•		•	•	[85]
Streptococcus pneumoniae				•	•	[86]
		•		•	•	[54]
Vibrio cholerae		•				[87]

[a] Profiling of *E. coli* chromosome [74] (http://www.shigen.nig.ac.jp/ecoli/pec/).
[b] Database of essential genes [58] (http://tubic.tju.edu.cn/deg/).
[c] http://www.genome.wisc.edu/.
[d] NMPDR [75] (http://www.nmpdr.org/content/essential.php).
[e] SEED [76] (http://theseed.uchicago.edu/FIG/eggs.cgi).

total) [55] and *H. pylori* (344 out of 1590) [56]. Other approaches have exploited genomic information to help target identification. In an interesting study by Forsyth and colleagues, a rapid shotgun antisense RNA method was applied to identify a large number of essential genes for *S. aureus* [57].

On the basis of these and other studies, several websites that collect and categorize essential genes from various species have been created and made available to the scientific community (Table 3). In particular, Zhang *et al.* [58] compiled a list of all currently available essential genes into the Database of Essential Genes (DEG), which includes the essential genes identified in the genomes of *M. genitalium*, *H. influenzae*, *Vibrio cholerae*, *S. aureus*, *E. coli* and *Saccharomyces cerevisiae*.

Genomic studies comparing some of the results described above showed two phenomena: first, analogous metabolic pathways sometimes comprise structurally unrelated enzymes, which catalyze the same reactions; second, highly conserved genes shown to be essential in some organisms are not essential in others [52,59]. Salama *et al.* [56] compared the set of essential genes detected in *H. pylori* with those identified in different bacteria by other studies. The authors found that essential core-genomes of different organisms had a limited overlap, and only 11% of the *H. pylori* essential genes overlapped with those of other organisms. More than 50% of genes were found to be essential in some species but not in others.

As a conclusion, all these studies demonstrated that essentiality should be verified experimentally and not simply inferred from homology. Moreover, because no new antibiotics have been discovered, in spite of the availability of drug targets, strategies for finding novel high-quality targets still need to be improved.

Drug targets and pan-genome

The newly developed pan-genome concept can be instrumental in providing a novel selection criterion for the identification of potential targets. Interspecies comparative analyses should be the starting point for the selection of the species core genome. Once identified, this subset of genes should be compared to the core genomes of other species to select the limited number of functions that are not only conserved among the strains of a given organism, but are also shared among a variety of related species. Thus, it would be important to compare species with common features, such as causing the same type of infection, or colonizing the same niche. This would enable the identification of proteins that might be involved in the same mechanism of infection and that, therefore, represent potential common virulence factors.

In an extensive study, Gerdes *et al.* [55] applied a mutagenesis-based technique for a genome-wide assessment of genes that are essential for aerobic growth of the *E. coli* strain MG1655. The 620 identified essential genes were compared with the corresponding putative orthologues of 32 complete genomes of diverse bacterial species. Conserved genes were mapped according to their chromosomal position and grouped into functional classes. Nucleic acid metabolism, protein metabolism and secretion are the most

conserved functional categories and contain the highest number of essential genes. Interestingly, all functional classes contained a certain percentage of essential genes; almost 8% of the genes in each class are essential. In addition, the authors observed that, among the variable genes, there is always a fraction of essential genes (>12%). This means, from a pan-genomic view point, that essentiality is not strictly a property of the core genome but that there is a given level of essentiality also in the accessory genes; this finding is important for the discovery of novel potential drug targets.

Virulence factors as drug targets

Several bacterial species are asymptomatically carried by the host, but only some of them can replace commensal flora, traverse different epithelia, resist attack by the host immune system and reach a specific niche where they can reproduce and cause disease. Typically, bacteria have developed specific virulence factors to overcome these barriers. As these factors are essential for the pervasiveness and ability of the bacteria to cause disease, new lines of research suggest that they should be considered as drug targets. Thus, several studies have exploited genomics technologies to identify targets that are essential for the formation and persistence of an *in vivo* infection or for the expression of specific virulence factors. Functional genomics approaches such as signature-tagged mutagenesis (STM, [60]), *in vivo* expression technologies (IVET, [61]), differential fluorescence induction (DFI, [62]) and selective capture of transcribed sequences (SCOTS, [63]) were developed to identify bacterial genes required for disease. In STM, a pool of mutants (specifically targeted) is used to infect an animal model. Tagged mutant pools are compared before and after the infection and the tagged genes that are important in pathogenesis are determined by subtraction. IVET and DFI are based on promoter-trapping techniques and identify microbial promoters that are active under particular circumstances, for example in a specific niche or during interaction with the host. SCOTS is based on the identification of extracted RNA with different hybridization approaches.

Recently, the availability of complete genomic sequences also enabled the development of technologies, such as whole-genome expression microarrays, that can simultaneously measure the expression of thousands of genes. Several studies based on these technologies [64,65] showed that pathogens use various strategies to regulate the expression of virulence genes during infection and, in particular, that expression levels crucially depend on growth phases and on host interactions.

An interesting example of how targeting virulence factors can be successful in antibiotic discovery is the recent work by Mekalanos' group [66], in which the authors screened a chemical library and identified a small molecule (virstatin) that, after orogastric administration, was shown to be capable of protecting infant mice from intestinal colonization by *V. cholerae*. Interestingly, this compound inhibits virulence by blocking the activity of the transcriptional regulator ToxT, thereby preventing the expression of two crucial virulence factors, cholera toxin and the toxin-co-regulated pilus [66].

A further example of how virulence factors can be explored with a whole-genome approach can be found in recently published work on *M. tuberculosis*. Sassetti *et al.* [67] examined the complexity of the life cycle of the pathogen by mutating every nonessential gene. In particular, the authors generated a library of transposon mutants of *M. tuberculosis* and compared the growth behaviour of the whole library pool during different phases of infection. The genomic DNA of the whole pools recovered from a reference *in vitro* growth and from several *in vivo* growths at different disease stages were compared by using competitive hybridization on a microarray chip. This revealed the differences in the gene content of each pool, thus enabling the authors to infer which group of genes was necessary during each *in vivo* phase. By analyzing these groups of genes, the authors obtained a detailed view of the changing environment that the bacterium encounters during infection and established which the bacterial genetic mechanisms of adaptation are. Interestingly, most of the genes that are essential *in vivo* but not *in vitro*, are conserved only in *M. tuberculosis* and closely related species, suggesting that they could represent valid drug targets specific to this class of pathogens.

Recent comparative pan-genomic studies demonstrate how virulence factors can be identified when different strains of a bacterial species are analyzed or when more distant species are compared. For example, Takeuchi *et al.* [68] compared three different human-colonizing staphylococcal species (in particular *S. haemolyticus*, *S. aureus* and *S. epidermidis*) and showed that they have a large fraction of genes and the corresponding chromosomal positions in common. However, the authors also identified a region near the origin of replication that has little homology between the species but which is highly conserved in different strains of each species. This region probably contributes to the evolution and specialization of staphylococcal species and might contain genes whose products are specialized for their specific mechanisms of pathogenicity.

An alternative approach for the characterization of genes responsible for strain specialization has been proposed by Chen *et al.* [30]. The authors suggest that factors differentiating uropathogenic (UPEC) from non-uropathogenic *E. coli* strains can be determined not only by searching UPEC-specific genes, but also by studying the set of genes that are under positive selection only in the UPEC subgroup of isolates, as derived from a phylogenetic model. In fact, genes showed to be under positive selection are likely to be involved in pathogenesis and could, therefore, represent valuable drug targets.

An alternative application of comparative analyses can be also aimed at the targeting of a specific biological problem. For example, several pathogenic Gram-negative bacteria, including *Salmonella*, *Shigella*, *Yersinia*, *Chlamydia*, *Pseudomonas aeruginosa* and enteropathogenic *E. coli*, harbour a type III secretion system (T3SS), which is specialized for the injection of specific toxins into the cytoplasm of host cells. This system comprises several components, some of which are highly conserved among the different bacteria and, therefore, could be evaluated as potential targets for novel antibacterial agents [69]. Using this approach, a small chemical compound was found to specifically block the secretion of the *Yersinia* outer protein virulence factor Yop by targeting the T3SS of *Y. pseudotuberculosis* and, thus, attenuating the pathogen [70]. Interestingly, the same compound was also able to perform the same attenuating activity by blocking the T3SS of *C. trachomatis* [71]. These results show proof of concept that small compounds targeting the T3SS systems can be identified and that T3SS

inhibitors might constitute a novel class of antibacterial agents that could be able to attenuate different pathogenic species.

As a further example, the recent discovery that most Gram-positive pathogens express pili suggests that these structures are related to colonization of the host. The observation that pili-negative strains of *S. pneumoniae* are less efficient in colonizing lung epithelial cells [72] indicates that pili represent a conserved class of virulence factors, which could be exploited both as vaccine candidates and as possible targets for the production of broad spectrum antimicrobials.

Conclusions

For more than a century, infectious diseases have been controlled by vaccination and the administration of antibiotics. In spite of the technical progress of the past century, innovation in both fields came exclusively from traditional approaches. All vaccines licensed to date consist of killed or live-attenuated microorganisms or proteins purified from the microorganism (subunit vaccines), which have been discovered following the century-old principles of Pasteur to isolate, inactivate and inject the disease-causing agent. Similarly, antibiotics have been identified by screening natural compounds for their ability to kill bacteria grown *in vitro*.

Recently, the availability of the genome sequence of microorganisms and the possibility of determining the genome of entire species (pan-genome) has made it possible to collect and store detailed information in databases about the whole genetic repertoire of a microbial species. This has led to a shift in the discovery of novel vaccines and antimicrobials from the traditional empirical approach to a novel knowledge-based approach.

In the field of vaccines, this novel approach, named reverse vaccinology, has already been successful in bringing several novel vaccines to clinical development. Furthermore, the attempt to develop a universal vaccine against GBS has demonstrated that the sequences of multiple genomes from each species are needed to cover the diversity of many bacterial pathogens, thus leading to the new era of pan-genomic reverse vaccinology.

In the field of antibacterial research, 'target-based' approaches aimed at the identification of targets essential for *in vitro* growth did not improve the identification of novel antibiotics. This suggests that we need to enhance our knowledge of bacterial pathogenic strategies and host–pathogen interactions, because these mechanisms are likely to harbour successful targets. The increasing knowledge of bacterial diversity based on genomics and pan-genomics suggests that the way forward should be to focus discovery strategies on the identification of targets that are essential for the formation and persistence of an *in vivo* infection or in the expression of virulence factors.

References

1 Fleischmann, R.D. *et al.* (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* 269, 496–512

2 Galperin, M.Y. (2006) The Molecular Biology Database Collection: 2006 update. *Nucleic Acids Res.* 34, D3–D5

3 Dorrell, N. *et al.* (2001) Whole genome comparison of *Campylobacter jejuni* human isolates using a low-cost microarray reveals extensive genetic diversity. *Genome Res.* 11, 1706–1715

4 Tettelin, H. *et al.* (2002) Complete genome sequence and comparative genomic analysis of an emerging human pathogen, serotype V *Streptococcus agalactiae*. *Proc. Natl. Acad. Sci. U. S. A.* 99, 12391–12396

5 Fukiya, S. *et al.* (2004) Extensive genomic diversity in pathogenic *Escherichia coli* and *Shigella* strains revealed by comparative genomic hybridization microarray. *J. Bacteriol.* 186, 3911–3921

6 Fraser-Liggett, C.M. (2005) Insights on biology and evolution from microbial genome sequencing. *Genome Res.* 15, 1603–1610

7 Bjorkholm, B. *et al.* (2001) Comparison of genetic divergence and fitness between two subclones of *Helicobacter pylori*. *Infect. Immun.* 69, 7832–7838

8 Israel, D.A. *et al.* (2001) *Helicobacter pylori* genetic diversity within the gastric niche of a single human host. *Proc. Natl. Acad. Sci. U. S. A.* 98, 14625–14630

9 Perna, N.T. *et al.* (2001) Genome sequence of enterohaemorrhagic *Escherichia coli* O157:H7. *Nature* 409, 529–533

10 Field, D. *et al.* (2006) How do we compare hundreds of bacterial genomes? *Curr. Opin. Microbiol.* 9, 499–504

11 Raskin, D.M. *et al.* (2006) Bacterial genomics and pathogen evolution. *Cell* 124, 703–714

12 Wayne, L.G. *et al.* (1987) International Committee on Systematic Bacteriology. Report of the ad hoc committee on reconciliation of approaches to bacterial systematics. *Int. J. Syst. Bacteriol.* 37, 463–464

13 Tettelin, H. *et al.* (2005) Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial 'pan-genome'. *Proc. Natl. Acad. Sci. U. S. A.* 102, 13950–13955

14 Medini, D. *et al.* (2005) The microbial pan-genome. *Curr. Opin. Genet. Dev.* 15, 589–594

15 Rappuoli, R. (2000) Reverse vaccinology. *Curr. Opin. Microbiol.* 3, 445–450

16 Pizza, M. *et al.* (2000) Identification of vaccine candidates against serogroup B meningococcus by whole-genome sequencing. *Science* 287, 1816–1820

17 Tettelin, H. *et al.* (2000) Complete genome sequence of *Neisseria meningitidis* serogroup B strain MC58. *Science* 287, 1809–1815

18 Giuliani, M.M. *et al.* (2006) A universal vaccine for serogroup B meningococcus. *Proc. Natl. Acad. Sci. U. S. A.* 103, 10834–10839

19 Wizemann, T.M. *et al.* (2001) Use of a whole genome approach to identify vaccine molecules affording protection against *Streptococcus pneumoniae* infection. *Infect. Immun.* 69, 1593–1598

20 Ross, B.C. *et al.* (2001) Identification of vaccine candidate antigens from a genomic analysis of *Porphyromonas gingivalis*. *Vaccine* 19, 4135–4142

21 Montigiani, S. *et al.* (2002) Genomic approach for analysis of surface proteins in *Chlamydia pneumoniae*. *Infect. Immun.* 70, 368–379

22 Ariel, N. *et al.* (2002) Search for potential vaccine candidate open reading frames in the *Bacillus anthracis* virulence plasmid pXO1: *in silico* and *in vitro* screening. *Infect. Immun.* 70, 6817–6827

23 Kato, N. *et al.* (1990) Molecular cloning of the human hepatitis C virus genome from Japanese patients with non-A, non-B hepatitis. *Proc. Natl. Acad. Sci., U. S. A.* 87, 9524–9528

24 Rosa, D. *et al.* (1996) A quantitative test to estimate neutralizing antibodies to the hepatitis C virus: cytofluorimetric assessment of envelope glycoprotein 2 binding to target cells. *Proc. Natl. Acad. Sci. U. S. A.* 93, 1759–1763

25 Choo, Q.L. *et al.* (1994) Vaccination of chimpanzees against infection by the hepatitis C virus. *Proc. Natl. Acad. Sci. U. S. A.* 91, 1294–1298

26 Sarbah, S.A. and Younossi, Z.M. (2000) Hepatitis C: an update on the silent epidemic. *J. Clin. Gastroenterol.* 30, 125–143

27 Khan, A.M. *et al.* (2006) Large-scale analysis of antigenic diversity of T–cell epitopes in dengue virus. *BMC Bioinform.* 7 (Suppl 5), S4

28 Fischer, W. *et al.* (2007) Polyvalent vaccines for optimal coverage of potential T-cell epitopes in global HIV-1 variants. *Nat. Med.* 13, 100–106

29 Fitzgerald, J.R. *et al.* (2001) Evolutionary genomics of *Staphylococcus aureus*: insights into the origin of methicillin-resistant strains and the toxic shock syndrome epidemic. *Proc. Natl. Acad. Sci. U. S. A.* 98, 8821–8826

30 Chen, S.L. *et al.* (2006) Identification of genes subject to positive selection in uropathogenic strains of *Escherichia coli*: a comparative genomics approach. *Proc. Natl. Acad. Sci. U. S. A.* 103, 5977–5982

31 Read, T.D. *et al.* (2002) Comparative genome sequencing for discovery of novel polymorphisms in *Bacillus anthracis*. *Science* 296, 2028–2033

32 Daubin, V. and Ochman, H. (2004) Bacterial genomes as new gene homes: the genealogy of ORFans in *E. coli*. *Genome Res.* 14, 1036–1042

33 Brussow, H. *et al.* (2004) Phages and the evolution of bacterial pathogens: from genomic rearrangements to lysogenic conversion. *Microbiol. Mol. Biol. Rev.* 68, 560–602

34 Feil, E.J. (2004) Small change: keeping pace with microevolution. *Nat. Rev. Microbiol.* 2, 483–495
35 Geme, J.W., 3rd and Cutter, D. (1995) Evidence that surface fibrils expressed by *Haemophilus influenzae* type b promote attachment to human epithelial cells. *Mol. Microbiol.* 15, 77–85
36 Tamas, I. *et al.* (2002) 50 million years of genomic stasis in endosymbiotic bacteria. *Science* 296, 2376–2379
37 Lauer, P. *et al.* (2005) Genome analysis reveals pili in Group B *Streptococcus*. *Science* 309, 105
38 Maione, D. *et al.* (2005) Identification of a universal Group B streptococcus vaccine by multiple genome screen. *Science* 309, 148–150
39 Miesel, L. *et al.* (2003) Genetic strategies for antibacterial drug discovery. *Nat. Rev. Genet.* 4, 442–456
40 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. U. S. A.* 93, 10268–10273
41 Koonin, E.V. (2000) How many genes can make a cell: the minimal-gene-set concept. *Annu. Rev. Genomics Hum. Genet.* 1, 99–116
42 Itaya, M. (1995) An estimation of minimal genome size required for life. *FEBS Lett.* 362, 257–260
43 Tatusov, R.L. *et al.* (1997) A genomic perspective on protein families. *Science* 278, 631–637
44 Galperin, M.Y. and Koonin, E.V. (1999) Searching for drug targets in microbial genomes. *Curr. Opin. Biotechnol.* 10, 571–578
45 Kobayashi, K. *et al.* (2003) Essential *Bacillus subtilis* genes. *Proc. Natl. Acad. Sci. U. S. A.* 100, 4678–4683
46 Lander, E.S. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature* 409, 860–921
47 Venter, J.C. *et al.* (2001) The sequence of the human genome. *Science* 291, 1304–1351
48 Sakharkar, K.R. *et al.* (2004) A novel genomics approach for the identification of drug targets in pathogens, with special reference to *Pseudomonas aeruginosa*. *In Silico Biol.* 4, 355–360
49 Dutta, A. *et al.* (2006) In silico identification of potential therapeutic targets in the human pathogen *Helicobacter pylori*. *In Silico Biol.* 6, 43–47
50 Chalker, A.F. and Lunsford, R.D. (2002) Rational identification of new antibacterial drug targets that are essential for viability using a genomics-based approach. *Pharmacol. Ther.* 95, 1–20
51 Pucci, M.J. (2006) Use of genomics to select antibacterial targets. *Biochem. Pharmacol.* 71, 1066–1072
52 Hutchison, C.A. *et al.* (1999) Global transposon mutagenesis and a minimal *Mycoplasma* genome. *Science* 286, 2165–2169
53 Akerley, B.J. *et al.* (2002) A genome-scale analysis for identification of genes required for growth or survival of *Haemophilus influenzae*. *Proc. Natl. Acad. Sci. U. S. A.* 99, 966–971
54 Thanassi, J.A. *et al.* (2002) Identification of 113 conserved essential genes using a high-throughput gene disruption system in *Streptococcus pneumoniae*. *Nucleic Acids Re.* 30, 3152–3162
55 Gerdes, S.Y. *et al.* (2003) Experimental determination and system level analysis of essential genes in *Escherichia coli* MG1655. *J. Bacteriol.* 185, 5673–5684
56 Salama, N.R. *et al.* (2004) Global transposon mutagenesis and essential gene analysis of Helicobacter pylori. *J. Bacteriol.* 186, 7926–7935
57 Forsyth, R.A. *et al.* (2002) A genome-wide strategy for the identification of essential genes in *Staphylococcus aureus*. *Mol. Microbiol.* 43, 1387–1400
58 Zhang, R. *et al.* (2004) DEG: a database of essential genes. *Nucleic Acids Res.* 32, D271–D272
59 Arigoni, F. *et al.* (1998) A genome-based approach for the identification of essential bacterial genes. *Nat. Biotechnol.* 16, 851–856
60 Hensel, M. *et al.* (1995) Simultaneous identification of bacterial virulence genes by negative selection. *Science* 269, 400–403
61 Mahan, M.J. *et al.* (1993) Selection of bacterial virulence genes that are specifically induced in host tissues. *Science* 259, 686–688
62 Valdivia, R.H. and Falkow, S. (1997) Fluorescence-based isolation of bacterial genes expressed within host cells. *Science* 277, 2007–2011
63 Graham, J.E. and Clark-Curtiss, J.E. (1999) Identification of *Mycobacterium tuberculosis* RNAs synthesized in response to phagocytosis by human macrophages by selective capture of transcribed sequences (SCOTS). *Proc. Natl. Acad. Sci. U. S. A.* 96, 11554–11559
64 Mahan, M.J. *et al.* (2000) Assessment of bacterial pathogenesis by analysis of gene expression in the host. *Annu. Rev. Genet.* 34, 139–164
65 Rediers, H. *et al.* (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
66 Hung, D.T. *et al.* (2005) Small-molecule inhibitor of *Vibrio cholerae* virulence and intestinal colonization. *Science* 310, 670–674
67 Sassetti, C.M. and Rubin, E.J. (2003) Genetic requirements for mycobacterial survival during infection. *Proc. Natl. Acad. Sci. U. S. A.* 100, 12989–12994
68 Takeuchi, F. *et al.* (2005) Whole-genome sequencing of *Staphylococcus haemolyticus* uncovers the extreme plasticity of its genome and the evolution of human-colonizing staphylococcal species. *J. Bacteriol.* 187, 7292–7308
69 Muller, S. *et al.* (2001) The Type III secretion system of Gram-negative bacteria: a potential therapeutic target? *Expert. Opin. Ther. Targets* 5, 327–339
70 Nordfelth, R. *et al.* (2005) Small-molecule inhibitors specifically targeting type III secretion. *Infect. Immun.* 73, 3104–3114
71 Muschiol, S. *et al.* (2006) A small-molecule inhibitor of type III secretion inhibits different stages of the infectious cycle of *Chlamydia trachomatis*. *Proc. Natl. Acad. Sci. U. S. A.* 103, 14566–14571
72 Barocchi, M.A. *et al.* (2006) A pneumococcal pilus influences virulence and host inflammatory responses. *Proc. Natl. Acad. Sci. U. S. A.* 103, 2857–2862
73 Ross, B.C. *et al.* (2004) Characterization of two outer membrane protein antigens of *Porphyromonas gingivalis* that are protective in a murine lesion model. *Oral Microbiol. Immunol.* 19, 6–15
74 Hashimoto, M. *et al.* (2005) Cell size and nucleoid organization of engineered *Escherichia coli* cells with a reduced genome. *Mol. Microbiol.* 55, 137–149
75 Gerdes, S. *et al.* (2006) Essential genes on metabolic maps. *Curr. Opin. Biotechnol.* 17, 448–456
76 Overbeek, R. *et al.* (2005) The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Res.* 33, 5691–5702
77 Baba, T. *et al.* (2006) Construction of *Escherichia coli* K-12 in-frame, single–gene knockout mutants: the Keio collection. *Mol. Syst. Biol.* 2, 1–11
78 Herring, C.D. and Blattner, F.R. (2004) Conditional lethal amber mutations in essential *Escherichia coli* genes. *J. Bacteriol.* 186, 2673–2681
79 Kang, Y. *et al.* (2004) Systematic mutagenesis of the *Escherichia coli* genome. *J. Bacteriol.* 186, 4921–4930
80 Glass, J.I. *et al.* (2006) Essential genes of a minimal bacterium. *Proc. Natl. Acad. Sci. U. S. A.* 103, 425–430
81 Sassetti, C.M. *et al.* (2003) Genes required for mycobacterial growth defined by high density mutagenesis. *Mol. Microbiol.* 48, 77–84
82 Jacobs, M.A. *et al.* (2003) Comprehensive transposon mutant library of *Pseudomonas aeruginosa*. *Proc. Natl. Acad. Sci. U. S. A.* 100, 14339–14344
83 Liberati, N.T. *et al.* (2006) An ordered, nonredundant library of *Pseudomonas aeruginosa* strain PA14 transposon insertion mutants. *Proc. Natl. Acad. Sci. U. S. A.* 103, 2833–2838
84 Ji, Y. *et al.* (2001) Identification of critical staphylococcal genes using conditional phenotypes generated by antisense RNA. *Science* 293, 2266–2269
85 Knuth, K. *et al.* (2004) Large-scale identification of essential *Salmonella* genes by trapping lethal insertions. *Mol. Microbiol.* 51, 1729–1744
86 Song, J.H. *et al.* (2005) Identification of essential genes in *Streptococcus pneumoniae* by allelic replacement mutagenesis. *Mol. Cells* 19, 365–374
87 Judson, N. and Mekalanos, J.J. (2000) TnAraOut, a transposon-based approach to identify and characterize essential bacterial genes. *Nat. Biotechnol.* 18, 740–745