Ann. N.Y. Acad. Sci. ISSN 0077-8923

ANNALS OF THE NEW YORK ACADEMY OF SCIENCES
Issue: *The Year in Immunology*

Reverse vaccinology in the 21st century: improvements over the original design

Claudio Donati and Rino Rappuoli

Novartis Vaccines and Diagnostics, Siena, Italy

Address for correspondence: Rino Rappuoli, Novartis Vaccines and Diagnostics, Via Fiorentina 1, 53100 Siena, Italy. rino.rappuoli@novartis.com

Reverse vaccinology (RV), the first application of genomic technologies in vaccine research, represented a major revolution in the process of discovering novel vaccines. By determining their entire antigenic repertoire, researchers could identify protective targets and design efficacious vaccines for pathogens where conventional approaches had failed. Bexsero, the first vaccine developed using RV, has recently received positive opinion from the European Medicines Agency. The use of RV initiated a cascade of changes that affected the entire vaccine development process, shifting the focus from the identification of a list of vaccine candidates to the definition of a set of high throughput screens to reduce the need for costly and labor intensive tests in animal models. It is now clear that a deep understanding of the epidemiology of vaccine candidates, and their regulation and role in host-pathogen interactions, must become an integral component of the screening workflow. Far from being outdated by technological advancements, RV still represents a paradigm of how high-throughput technologies and scientific insight can be integrated into biotechnology research.

Keywords: reverse vaccinology; vaccines; bacteria; genomics; proteomics

Introduction

In its original formulation, vaccination is the practice by which individuals injected with inactivated or attenuated forms of an infectious agent become immune to infection from the injected agent. The application of this practice has essentially remained unchanged for nearly two centuries. The first major revolution was the development of subunit vaccines, when it was realized that components of the microorganism could be sufficient to elicit immune response, decreasing the probability of unwanted side effects. Since then most efforts in vaccine research have been dedicated to identifying the component or mixture of components able to protect against the disease. The main characteristics of these molecules are their presence and conservation in the infectious agent, their visibility to the host immune system, and their ability to elicit a protective immune response. Given these assumptions, the preferred method for vaccine target identification has been the analysis of sera from infected individuals who are protected from reinfection. This procedure is usually able to identify a restricted set of candidates that dominate the host immune response; it fails to identify those components that are not highly immunogenic during infection, but are able to confer protective immunity. A typical example is tetanus toxin.

Since the late 1990s, the development of sequencing technologies has changed the landscape of the slowly evolving field of vaccinology. When the genome of the first living organism was sequenced in 1995,[1] it was realized that genomic technologies, by determining the whole proteomic potential of the infectious organism, could boost the chances of identifying the protein, or mixture of proteins, that could be used to develop an efficacious vaccine. The method, reverse vaccinology (RV), offered two main advantages. First, it allowed identification of a much broader spectrum of candidates, including proteins that had not been identified before because they were masked by other, immunodominant targets. Second, it allowed the identification of potential

doi: 10.1111/nyas.12046

vaccine targets in organisms that were difficult to cultivate in the laboratory.

The RV protocol was originally developed to overcome the hurdles that had hampered the development of an efficacious vaccine against serogroup B *Neisseria meningitidis* (MenB), a gram-negative bacterium responsible for about 50% of the bacterial meningitis worldwide.[2] *N. meningitidis* is a natural component of the commensal flora that colonizes the upper respiratory tract of healthy individuals. In a small proportion of cases, the bacterium can invade the host bloodstream and, after crossing the blood–brain barrier, cause meningitis. To escape reconnaissance by the host immune system and survive in blood, *N. meningitidis* is coated by a polysaccharide capsule that, based on its chemical properties, is classified into five major serogroups: A, B, C, Y, and W135. Given its exposure on the surface of the cell and role in pathogenicity, the capsular polysaccharide constitutes the antigen of choice for the meningococcus, and is an excellent target for bactericidal antibodies elicited by conjugate vaccines against serogroups A, C, Y, and W135. However, in the case of serogroup B, the conjugate vaccine was not feasible because the capsular polysaccharide is an $\alpha(2, 8)$ polysialic acid, identical to the polysialic acid present in human glycoproteins such as N-CAM. The capsule of MenB is thus a human self-antigen, and much effort has been directed toward the development of a protein-based vaccine specific for it. In the mid 1990s, all of these efforts were frustrated by the inconsistency of the protection data, in that there was extreme variability in the surface proteins tested as vaccine antigens.

The term *reverse vaccinology* originates from the change in perspective allowed by the advancements in sequencing technologies. The entire genome of the virulent MC58 strain was sequenced,[3] and from the genomic data, potential vaccine targets were selected.[2] The main idea behind the procedure was that all of the successful examples of protein-based vaccines include targets that are either exposed on the surface of the cell or secreted into the extracellular milieu. Starting from the 2,158 proteins encoded in the MC58 sequenced genome, bioinformatic analysis predicted that over 600 were either exposed on the surface or secreted. Of these, 350 were cloned in *Escherichia coli*, successfully expressed in soluble form, purified, and used to immunize mice. The sera of immunized animals were then screened in a serum bactericidal assay that is known to correlate with protection. At each of these steps, candidates not satisfying quality criteria were discarded; the process led to the identification of five previously unknown vaccine candidates[4] that subsequently have completed clinical trials in a vaccine combination known as 4CMenB, and received a positive opinion from the European Medicines Agency (and approved with the commercial name of Bexsero®). Since the pioneering MenB project, the RV approach has been applied to a variety of other important pathogens, including *Streptococcus pneumonia*,[5] *Porphyromonas gingivalis*,[6] *Chlamydia pneumoniae*,[7] *Streptococcus agalactiae*,[8] *E. coli*,[9] *Leishmania major*, and *L. infantum*,[10] and has earned a dedicated entry on Wikipedia (http://en.wikipedia.org/wiki/Reverse_vaccinology).

When first proposed, the idea behind the RV approach was for it to be the definitive solution to the problem of antigen discovery. Once all the genes encoded in the genome of a pathogenic species are known, the list of vaccine candidates is finite and, in principle, can be tested in animal models. Therefore, protective antigens would not be missed, although it might take time and effort to find them and to define the most effective vaccine formulation. However, in each of the vaccine projects based on RV, the original design had to be modified to adapt to the peculiarities of the target species. In turn, following the enormous research effort that was needed to overcome many unforeseen difficulties, there is a clearer understanding of many aspects of bacterial population biology and of the interactions between pathogens and the human host, as well as how they impact the development of a vaccine.

The experience accumulated in the last decade has demonstrated that the formulation of a vaccine able to guarantee broad coverage requires a deep understanding of the population structure of the pathogen. Theoretical and experimental work of sequence analysis has shown that a certain degree of strain-to-strain variability, both in sequence and expression level, is an unavoidable characteristic of the antigens identified using RV. Thanks to the advent of high-throughput sequencing technologies, it is now feasible to determine the complete genome sequences of hundreds of bacterial isolates and to use these, instead of a single genome, as a starting point for the vaccine target selection process.

In this way, epidemiological characterization of the pathogen has become an integral part of the RV approach.

The testing of vaccine candidates in animal models—a step at the end of which usually only a handful of the screened candidates proves to induce protective immunity—still produces the major bottleneck of the entire RV process. Therefore, a large effort has been devoted to defining a set of screening procedures that could integrate the original bioinformatic selection to significantly reduce the number of candidates that need to be tested in animal models. New technologies, including microarrays, RNA-Seq, and proteomics, are now providing data that, integrated into the RV process, can turn the original brute force approach into a much more efficient and streamlined process. In addition, there has been a general technological advancement in the software tools that are used throughout the selection process, and in many cases their predictions can now be tested using new experimental methods.

In the following pages we will review the major improvements of the original RV workflow that occurred in the last decade, with particular attention to those genome-wide experimental methods that constitute a valuable complement to bioinformatic screening. In doing so, we will also review the major discoveries that were made in the context of RV projects (Table 1).

Table 1. Major milestones in the evolution of reverse vaccinology

1995	First complete genome sequence of a living organism.[1]
2000	First application of whole genome sequencing in vaccine research: formulation of the reverse vaccinology approach.[2]
2002	First application of DNA microarray technology to antigen discovery.[96,97]
2005	First comparative genomic study of multiple isolates of the same bacterial species and formulation of the pan-genome concept.[17]
2005	First rational design of a multi-component protein vaccine for GBS based on the analysis and screening of multiple bacterial genomes.[8]
2006	First formulation of a broadly protective vaccine against MenB based on reverse vaccinology.[4]
2006	First use of proteomics to identify surface exposed proteins in the screening of vaccine candidates.[83]
2011	First rational, structure-based design of an antigen inducing broad protective immunity against heterologous strains for MenB.[136]

Genomic variability

Fifteen years ago the sequencing of the entire genome of a single bacterial isolate was a significant achievement; thanks to improvements in sequencing technologies within the last decade, the sequencing of hundreds of bacterial genomes is now routinely done at a fraction of the cost of that of a single genome at the beginning of the RV era. Comparative analysis of multiple genomes of the same bacterial species has shown that genomic variability in bacteria is much more extensive than initially anticipated, and is a mechanism by which bacterial species are able to adapt to many different environments and escape reconnaissance by the host immune system. There are two main ways in which this affects the selection of a vaccine candidate: different strains can have different antigenic repertoires, and the sequence of shared antigens can vary from strain to strain. This highlights the importance of characterizing the epidemiology of selected candidates that might require the development of an entirely new typing system.

Variability of genome content: core and pan-genome

When RV was first proposed, several aspects of bacterial population genomics were not known in detail. Although it was clear that at the genomic level bacteria are more variable than species in other realms of life, the extent of this variability was not fully understood, and it was thought to be confined to a few structures, known as pathogenicity islands (PI), strictly involved in interaction with the host.[11] These regions, usually including more than 10 kb of sequences, are present in pathogenic strains and absent from non-pathogenic strains of a given species, are frequently associated with mobile elements, and can often be identified from a distinct GC content from the hosting genome that is a relic of their recent acquisition from a foreign origin. Well-studied examples include the PI in uropathogenic *E. coli*,[12] the SPI-1 and SPI-2 islands in *Salmonella*,[13] the Yop virulon in *Yersinia pestis*,[14] and the Cag island in *Helicobacter pylori*.[15]

This anthropocentric vision was put into a wider perspective when the availability of multiple genomes of the same bacterial species showed that a certain degree of variability in gene content is not an exceptional phenomenon limited to PI, but is instead an essential component of bacterial population biology with profound implications for the way bacteria adapt to changing environments. Comparing the genomic sequences of eight strains of *S. agalactiae* (group B *Streptococcus*, GBS), it was shown that each strain had genes that were missing from one or more of the other strains. Mathematical extrapolation of this concept led to the definition of the *core genome* of a species (i.e., the portion of the genome that is shared by all strains). In the case of GBS, the core genome is composed of 1806 genes, representing approximately 80% of the genome of any given strain.[16] The remaining 20% includes genes either shared only by a subset of the strains or that are strain-specific. The surprising result of this analysis was the prediction that even after a large number of strains have been sequenced, each new sequence will contribute an average of thirty-three new genes, suggesting that the size of the pan-genome (i.e., the total set of distinct genes that can be found in at least one strain of a named species) can continue to grow as more strains are sequenced.

This model can be applied to any bacterial species.[17] The size of the core genome reflects the evolutionary history and lifestyle of each species, and can be as little as 42% of the genome in species, such as *E. coli,* that are able to colonize very different environments. In addition, although the core genome mostly encodes metabolic functions that are essential for cell viability, the rest of the genome that was, perhaps inappropriately, defined as dispensable is comparatively rich in poorly characterized genes and genes associated with mobile and extrachromosomal elements, supporting the hypothesis that the majority of strain-specific traits depend on lateral gene transfer events.[16,18]

Later analysis has shown that the size of a species pan-genome can be related to fundamental parameters of the population genetics of a species,[19] supporting a vision in which bacterial cells have access to a large pool of genes (the species pan-genome) from which they occasionally derive traits that are not essential for survival.[20] In the case of *E. coli*, the pool of genes present at least once in a panel of 100 genomic sequences of randomly chosen isolates was estimated to include 17,838 genes, almost four times bigger than the average size of *E. coli* genomes with 4,721 genes;[18] in contrast, for the same number of strains, the pan-genome of *S. pneumonia* was predicted to include 3,221 genes, compared with the 2,104 genes encoded, on average, in each *S. pneumoniae* genome.[19]

The probability of exchange of genetic material among intracellular pathogens that have a low chance of sharing the same environment with unrelated strains is low.[17] However, it has recently been shown that even in a strictly intracellular pathogen such as *Chlamydia trachomatis* horizontal gene transfer is an important phenomenon essentially shaping the phylogeny of the species.[21] The frequency at which gene exchange occurs can be significantly increased by natural processes such as inflammation, or by selective pressure induced by clinical intervention. An example of the former is the boost of the efficiency of plasmid exchange between *Salmonella enteric* serovar Typhimurium and *E. coli* following pathogen-driven inflammatory responses in the gut.[22] Frequent capsular exchange between pneumococcal strains following the introduction of a capsular polysaccharide vaccine, and spread of antibiotic resistance loci identified in *S. pneumoniae*, are clear examples of the latter.[23]

Recently, advancements in sequencing technologies, allowing characterization of the genetic composition of entire microbial communities (the microbiome), have shown that even the pan-genome concept is probably too restrictive, and that a network of gene exchange connects the bacteria forming the normal commensal flora of the human body, where the main limitation is the ecology of the microbial species.[24] This is especially relevant in vaccine research for those bacterial pathogens, such as *N. meningitidis*, that are essentially harmless components of the human microbial flora of the upper respiratory tract and only cause life-threatening disease in a small minority of cases for reasons that are still poorly understood.

From a vaccine discovery perspective, the distinction between core and pan-genome can be seen as both a limitation and an opportunity. If it is required that a single antigen protects against all strains, the antigen needs to be part of the core genome, thus limiting the panel of potential candidates. On the

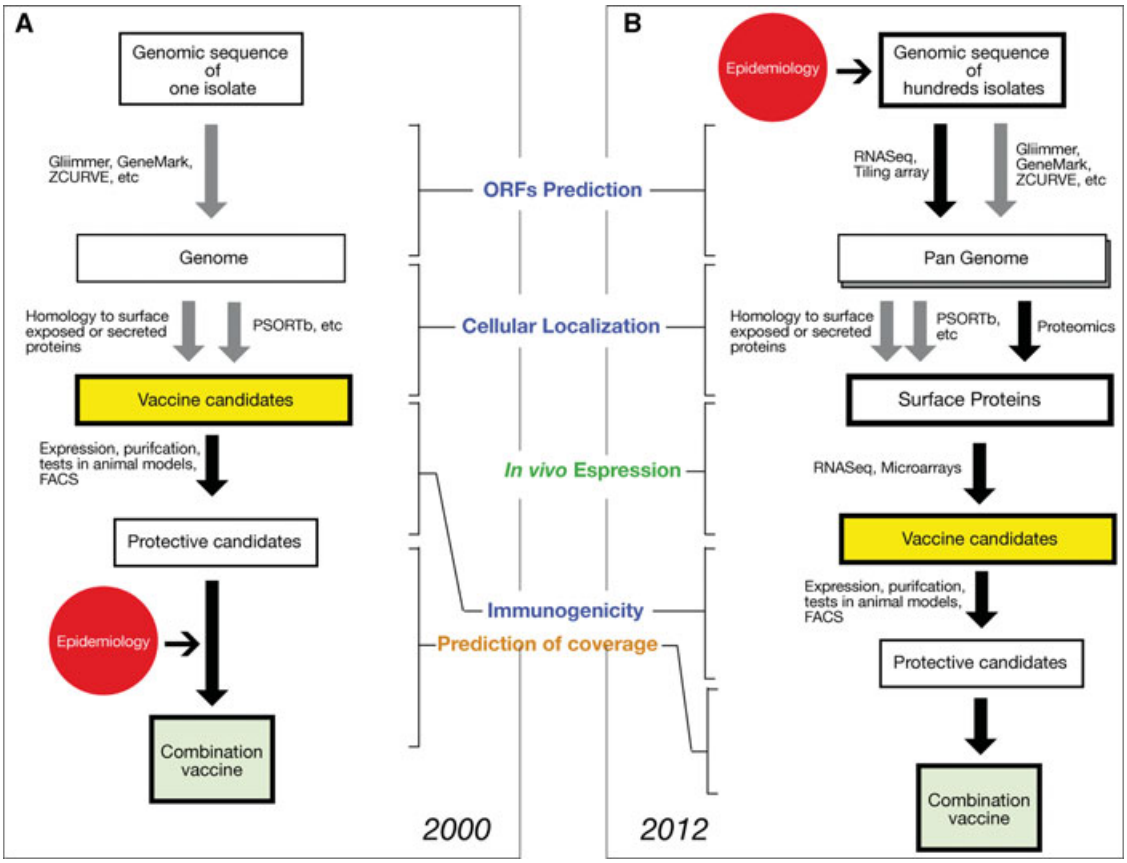

Figure 1. Reverse vaccinology (RV) pipeline. (A) In the original formulation, RV was designed to quickly lead from the genome of a single virulent isolate of a bacterial pathogen to the definition of a list of vaccine candidates through bioinformatics screening (grey arrows). The process involved the testing of a large panel of candidates in animal models, and only at the end, epidemiological data were generated for the protective antigens. (B) In the latest applications of RV, several steps have been modified and bioinformatic predictions have been complemented by experimental screenings (black arrows). The process starts with the sequencing of the genomes of a panel of representative isolates selected on the basis of epidemiological data. The complete set of genes encoded by these isolates (the pan-genome) is determined using bioinformatic predictions and transcriptome analysis. A list of surface-exposed proteins is extracted using bioinformatic predictions and refined using proteomic analysis, and additionally filtered using expression studies. This produces a focused list of high priority candidates to be tested in animal models. The ability of the selected antigens to protect against heterologous strains is tested, and a combination guaranteeing broad coverage is identified on the basis of epidemiology. Besides the introduction of several experimental steps to reduce the number of candidates to be tested in animal models, the main innovation is the use of next generation sequencing technologies to sequence large collections of isolates as a starting point for the selection pipeline. In this way, the conservation and epidemiology of a vaccine target is taken into account at the beginning of the process.

other hand, vaccine candidates able to guarantee partial protection can be derived from a gene pool that is larger than the genome of a single strain, and multicomponent vaccines including two or more of these antigens might guarantee broad protection. However, identification of the best combination requires a deep understanding of the epidemiology of the species, as an essential component of the initial screening of candidates (Fig. 1).

Variability of the antigens

The screening of large collections of isolates has shown that, even in the case of core genes, an aspect that must be taken into account is the sequence variability of the antigens and the effect that this has on their ability to elicit protection against heterologous strains. In an epidemiological study of a sample of 107 isolates, *fetA*, one of the major antigens in MenB, was found to have 56 distinct

amino acid sequences.[25] Five major variant families were identified, and polyclonal mouse sera raised against four of the variants were shown to react poorly with other variants. Similarly, one of the antigens that was selected in the first RV project, the factor H binding protein (fHbp), was found in three major variants that are not cross protective.[26] Subsequent studies showed that distinct subvariants differ in their level of surface accessibility and intrinsic reactivity to serum from a vaccinated individual.[27] Given its importance as an antigen, the epidemiology of fHbp has been extensively characterized.[28–32] Presently, more than 570 distinct peptide sequences have been deposited in a public database (http://pubmlst.org/neisseria/fHbp/).

Theoretical considerations and studies based on sequence analysis have shown that the high variability of the molecules having antigenic properties is related to the pressure exerted by the host immune response. In the case of PorB, another major meningococcal antigen, the variable regions of the protein coincide with the loops that are exposed on the surface of the bacterial cell and display a higher than expected rate of nonsynonymous mutation,[33] a phenomenon that is known as positive selection. Similar results have been shown for other antigens in MenB[34] and in other species, such as internalin A (inlA) in *Listeria monocytogenes,*[33] the antigenic membrane protein (Amp) in the plant pathogen *Phytoplasmas*,[35] the Hrp pilin HrpE of the plant pathogen *Xanthomonas,*[36] the intimin protein in *E. coli,*[37] the outer surface protein OspC in *Borrelia burgdoferi,*[38] and the major structural components RrgA and RrgB of the pilus in *S. pneumoniae*.[39]

Such results in many different species and for many different surface-exposed proteins suggest that a certain degree of variability of those proteins that are able to elicit a protective immune response is an unavoidable consequence of the evolutionary pressures exerted by their interaction with the host immune system and that select for the maintenance of polymorphisms within pathogen populations. Therefore, vaccine projects based on RV should take into account the possibility of having vaccine candidates with a low degree of cross protection against heterologous strains. To overcome these difficulties, strategies should be devised based on extensive characterization of the epidemiology of the candidate, or candidates of interest, in order to individuate the main variants and formulate a combination that is broadly cross protective. These variants can be expressed independently, or fused into a single construct.[40] Alternatively, a promising approach is the use of structural information of the target protein to rationally design a chimera that merges the major antigenic regions of the main variants into a single molecule, as recently shown in pioneering work on meningococcal fHbp.[41]

Population genomics and epidemiology of the bacterial species

Given the sequence variability often found in vaccine candidates, it is essential that sequences are characterized from a panel of isolates representative of the target bacterial population. This procedure requires an adequate description of the population structure of the pathogen. Considerable effort has been devoted to the definition of typing schema, such as multilocus sequence typing (MLST), based on the sequencing of small numbers of carefully chosen genomic loci (typically seven, but extended schema have also been proposed[42]); online databases for many pathogens are available.[43–49] Once a panel of strains reproducing the relevant characteristics of the population of circulating strains has been selected (in terms of, for instance, MLST typing), the antigens are sequenced and alignment pipelines, based on software tools like BLAST,[50] FastA,[51] ClustalW,[52] or MUSCLE,[53] are used to determine their variability and the distribution of distinct allelic variants. Since the advent of next generation sequencing (NGS),[10] it is often more practical to sequence the entire genome of representative strains and then extract the sequences of the MLST loci and the potential vaccine candidates from the genome. Two computational procedures are possible. The short sequencing reads can be directly aligned against one or more reference genomes to identify structural genomic variants, such as substitutions or short indels, using dedicated software suites.[54,55] Alternatively, the genomes can first be assembled and then the sequences of the potential antigens can be extracted and aligned. Although the first approach is usually quicker, given the short reads generated by NGS it can be applied only in the case of highly similar sequences, whereas preliminary assembly is required in the case of variable sequences, as is the case for many candidate antigens.

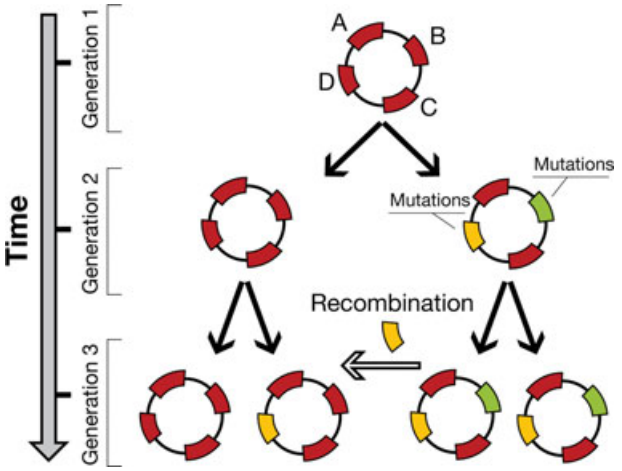

Figure 2. Clonal relationships and gene content. By transferring portions of the genome between unrelated strains, recombination events break the correlation between clonal relationships and gene content. Therefore, the presence, absence or allelic form of a given antigen cannot be predicted on the basis of the molecular typing of the strains. In the example sketched here, typing of isolates using loci A and B can predict the allelic state of locus D at generation 2, but not at generation 3, after an event of homologous recombination.

The work on MenB and GBS has shown that novel vaccine antigens do not necessarily follow the typing systems currently used and described in the textbooks, complicating the task of predicting which strains will be covered by the vaccine and, ultimately, the expected decrease in the number of disease cases following vaccination. Therefore, developing a vaccine based on RV may require developing a new typing system and educating the scientific community accordingly. A typical example is the 4CMenB vaccine against *N. meningitidis*. Pathogenic *N. meningitidis* is classified into serogroups A, B, C, Y, W-135, and X based on the chemical composition of the capsular polysaccharide, and on clonal complexes and sequence types based on the genetic make-up of the seven genes defined by the MLST schema.[43] It was found that neither of these typing systems was able to accurately describe the potential strain coverage of the 4CMenB vaccine. This lack of predictive value was attributed, on one hand, to the confounding effect of homologous recombination that breaks the linkage between the MLST loci and the antigens[28–32] and, on the other hand, to the lack of correlation between expression of the antigens and molecular typing (Fig. 2). As a generalization of MLST, sequencing of complete genomes is now increasingly used to gain a deeper understanding of the relation between bacterial strains,[21,23,56] and will soon become the tool of choice for the typing of bacterial species. The high resolution guaranteed by whole genome typing can readily be used to dissect finer relationships within known clades,[23,56] and these efforts are likely to result in improved phylogenetic characterization of bacterial species.[57] However, although in principle genomic data contain all of the information concerning the expression, surface exposure, and accessibility of each antigen, using them to predict coverage by a given vaccine formulation is still not feasible.

For this reason the meningococcal antigen typing system (MATS) was developed to predict the coverage of the 4CMenB vaccine (Fig. 3).[35] MATS combines a series of three antigen-specific sandwich ELISA assays for fHbp, NadA, and NHBA, with genotyping of the variable region of the PorA component. MATS ELISA simultaneously quantifies the amount of the antigens expressed by a given bacterial isolate and its immunological cross-reactivity with the antigens present in the 4CMenB vaccine. Given the correlation between MATS typing and susceptibility to killing in a serum bactericidal assay, MATS has the potential to predict vaccine coverage in large panels of strains, a result that even genome-based typing schema cannot presently guarantee.

Improvements to the antigen prediction pipeline

There has been a general evolution of the quality and accuracy of the bioinformatics tools that can be used to perform the selection of antigens, rivaling

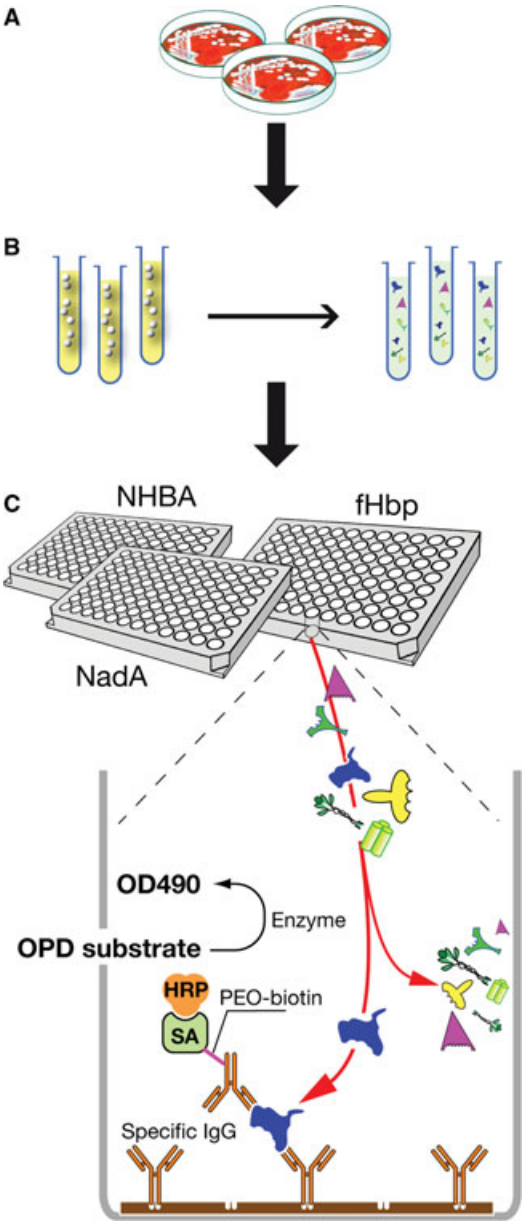

Figure 3. MATS typing system for the 4CMenB vaccine. (A) MenB bacteria are grown overnight on chocolate agar. (B) A suspension of bacteria taken from the plate is prepared to a specified OD600, and detergent is added to the suspension to extract the capsule and expose the antigens. (C) Serial dilutions of extract are tested in the MATS ELISA. A specific capture antibody binds one of the antigens from the extract, which is then detected with a specific biotin-labeled antibody and a streptavidin–enzyme conjugate. Plates are read in an ELISA reader.

in many cases the accuracy of experimental methods. The bioinformatics pipeline of RV involves several prediction steps, including (1) prediction of the protein open reading frames (ORFs), (2) annotation of the ORFs, and (3) prediction of the cellular localization of proteins. Because of the increasing popularity of the RV approach and the need for tools that allow its application, including by research groups that could not afford a large team of bioinformatics experts, dedicated software platforms were developed,[26–29] often using machine learning methods that leverage on accumulated knowledge from the experimental screens of large sets of potential vaccine candidates. Moreover, the increased availability of genome-scale experimental protocols, in many cases, now allows independent confirmation of bioinformatics predictions; in particular, it is now feasible to experimentally validate the gene prediction algorithms and the cellular localization algorithms at the genomic level. These experimental techniques are both able to correct errors and to identify new genes or surface-exposed proteins that were missed in the predictions.

Experimental validation of gene prediction algorithms

In the original RV efforts, identification of protein coding genes was based on gene prediction algorithms such a Glimmer,[58–61] GeneMark,[62] ZCURVE,[63] EasyGene,[64] or MED.[65] Despite continuous improvements, these algorithms suffer from systematic errors that are often not easy to quantify because of the relatively low prevalence of experimentally confirmed protein coding sequences in the genomic databases. Critical reannotation of 143 prokaryotic genomes has shown that the number of genes for which the wrong starting site has been predicted can be as high as 60%, especially for GC–rich genomes.[66] Given that, in many cases, the signal for cellular localization of the proteins is encoded near the start site, accuracy of the gene prediction is vital for the RV selection process of vaccine candidates.

Tiling arrays and RNA-Seq experiments can identify protein coding regions, transcriptional units, and regulatory elements independently from prediction algorithms.[67–69] Dedicated software suites for the assembly of RNA-Seq data are continuously improving our understanding of the structure of transcriptional units in both eukaryotes and prokaryotes.[70,71] Previously unpredicted coding

sequences, alternative transcripts, and regulatory elements have been identified in *Mycoplasma pneumoniae*, one of the smallest self-replicating organisms.[72] Alternatively, proteomic approaches using high resolution mass spectra obtained by tandem mass spectrometry can experimentally validate gene predictions by directly identifying peptides from expressed proteins. A proteome-derived ORF model for *M. pneumoniae* was able to confirm 81% of the predicted ORFs, additionally discovering 16 previously unknown ORFs and extending 19 ORFs at their N-termini.[73] In a recent genome-wide study on *S. typhimurium*, MS-based proteomics confirmed more than 40% of the predicted ORFs, correcting 47 start sites and identifying 12 novel genes not predicted by current gene-finding algorithms.[74]

Improved bioinformatics algorithms

Experimental determination of the cellular localization of proteins is a time consuming task that is difficult to perform on a genomic scale. Therefore, RV projects usually rely on teams of experts that complement protein cellular localization predictions provided by computational tools with literature-based searches of experimental evidence.[2] In the last ten years, the accuracy of computational predictions has greatly improved, limiting the need for expert supervision.

In bacteria, proteins are synthesized in the cytoplasm and then directed to specific sites or cell compartment by signals that are encoded in their amino-acid sequences. Although the nature of these signals is not completely known, their existence allows one to approach the protein localization problem through bioinformatics tools. The computational pipelines leading to a reliable prediction can be roughly classified into two main categories: homology-based inferences in which databases of proteins whose cellular localization has been experimentally determined are scanned to identify possible homologies with the unknown protein, and tools based on the identification of sequence-encoded signals known to influence protein localization. The homology-based method relies on the assumption that a certain degree of sequence conservation will lead to conservation of localization,[75] and its high accuracy is supported by the surprisingly sharp transition found in the relation between sequence similarity and identity in subcellular localization.[76] Other studies have shown that the level of homology needed for an accurate prediction of localization can be as low as 30%.[77]

In the last decade, many computational methods for predicting cellular localization based on machine learning algorithms have been introduced.[77–82] Because the first version was applicable only to Gram-negative bacteria, PSORTb[83] has been the most widely used localization prediction tool in RV projects. PSORTb has now been updated to include Gram-positive bacteria and to increase its coverage, while maintaining high precision.[84,85] Web-accessible databases of precompiled PSORTb predictions for most of the sequenced genomes also exist.[86] Other methods, combining more than one predictor, have been proposed.[87–89] Comprehensive databases that include the predictions of many computational tools for large samples of bacterial genomes are now freely available on the web.[90] Using these methods it is now possible to predict the cellular localization of large protein datasets with a precision that exceeds that of many experimental methods.[91] An approach that could substantially further improve the accuracy of computational methods is the integration of the available knowledge from the body of scientific literature using text mining and machine learning approaches,[92] supporting the role of human experts with modern tools of knowledge management. More generally, literature mining methods have been developed to identify proteins involved in host–pathogen interaction that could therefore constitute promising vaccine targets.[93]

Proteomics

The availability of bacterial genomes, together with the advancement of bioinformatic analysis tools, has greatly improved the ability of mass spectroscopy to identify bacterial proteins in uncharacterized samples. Using these technologies, known as *proteomics*, the large scale identification of proteins exposed on the surface of bacterial cells (the *surfome*) is now feasible.[94] In Gram-positive bacteria, the technology is based on "shaving" the surface of bacteria with proteases under conditions that preserve bacterial viability.[95] For Gram-negatives, the technology leverages on the fact that these organisms naturally release outer membrane vesicles (OMVs), organelles that bud out of the outer membranes without impairing cell viability. OMVs are then

purified and their protein content, prevalently composed by membrane-associated proteins, is characterized.[96,97]

Surfome analysis has allowed confirmation of surface localization predictions obtained by computational tools in many bacteria, and to identify previously unknown antigens. Novel antigens were identified by 2D gel electrophoresis of membrane extracts followed by MALDI-TOF identification in *Clostridium difficile*[98] and *Bordetella pertussis*.[99] Shaving of the bacterial cell surface followed by identification of the released peptides by mass spectroscopy have allowed the identification of surface-exposed proteins in group A *Streptococcus* (GAS)[95] and in GBS.[100] Furthermore, proteomic technologies have been used to identify biomarkers that distinguish virulent isolates of bacterial pathogens from harmless commensals.[101]

Differently from bioinformatic predictions, proteomic technologies also have the ability to identify posttranslational modifications that change the structure, and therefore potentially the antigenic properties, of the vaccine targets. Multiply charged isoforms that represent posttranslational modification such as phosphorylation, as well as smaller mass variants because of proteolytic cleavage, are commonly found in proteomic screens of outer membrane proteins by 2D-PAGE in *E. coli*.[102,103] On the basis of these data, it has been estimated that the fraction of proteins that undergo posttranslational modifications can represent up to 42% of the *E. coli* proteome.[102]

Prediction of immunogenicity

Of obvious interest for the selection of candidate antigens is the possibility to computationally predict the immunogenicity of the proteins that are on the surface of the bacterial cells. Most vaccines work by inducing serum antibodies, requiring the activation of B cells. Despite numerous efforts using a variety of computational techniques (for a review of available software and online resources, see Ref. 104), an exhaustive assessment of their predictive power has shown that their rate of success is only marginally better than random,[105] possibly because B cell epitopes are often conformational. Although recent approaches based on 3D protein structures have shown encouraging results,[106] their large-scale use is limited by the availability of crystallographic data.

In contrast to the case of B cell epitopes, considerable improvements have been made in T cell epitope prediction methods, thus allowing the inclusion of cellular mediated immunity in the RV framework. A large number of computational algorithms have been proposed,[104,107] and databases of predicted and experimentally determined epitopes are available.[108] In parallel, many technological advancements and experimental studies have allowed the validation of prediction algorithms and the elucidation of the role of cellular immunity in vaccine efficacy (for a recent comprehensive review, see Ref. 107).

Gene expression

It is increasingly appreciated that it is essential for potential vaccine targets to be expressed by pathogens during the infection process, and that in many cases antigens poorly expressed in *in vitro* conditions can be induced *in vivo* by specific signals. One example is the NadA protein, an adhesin present in approximately 50% of the MenB isolates. NadA was found to be protective in an infant rat model of bacteremia[109,110] and was selected as one of the components of the 4CMenB vaccine against MenB.[4] *In vitro*, the expression of NadA depends on the growth phase and is different in different strains of MenB.[33,109] However, the presence of the *nadA* gene associates significantly with hypervirulent strains; in contrast to most meningococcal surface-associated proteins, sera from children recovering from invasive meningococcal disease react specifically to recombinant forms of NadA.[25] Later studies have shown that NadA expression is tightly regulated by a complex combination of genetic and environmental factors involving the binding of a repressor, NadR, to regions flanking the phase variable promoter of the *nadA* gene (Fig. 4).[111] The presence of 4-hydroxyphenylacetic acid (4HPA), a metabolite of aromatic amino acids found in saliva, can induce NadA expression by alleviating the binding activity of NadR, suggesting that *in vivo* NadA expression can be induced by environmental factors present in the oropharynx.[111,112]

Although measuring the expression of large panels of proteins in a large number of strains or experimental conditions is not feasible, the identification of genes differentially transcribed can be a powerful tool to identify proteins that play a role in the infection process. These approaches

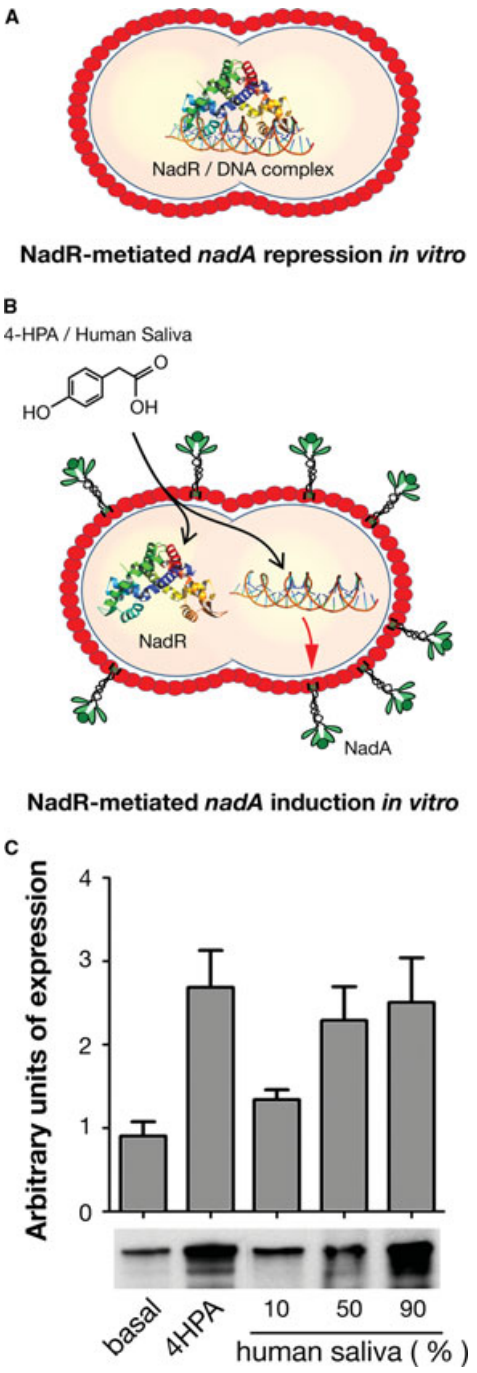

Figure 4. NadR mediated regulation of NadA expression. (A) The expression of NadA is repressed *in vivo* by the binding of NadR to regions flanking the promoter of NadA. (B) The binding activity of NadR is alleviated by the presence of 4HPA, a metabolite of aromatic amino acids found in saliva. (C) In human saliva NadA is expressed at levels similar to that obtained in the presence of 4HPA, suggesting that environmental factors can trigger the expression of NadA *in vivo*.

use RNA-based techniques—from the conventional expression microarray to the more recent tiling arrays and RNA-Seq technologies—to infer the amount of protein synthesized by bacterial cells at a given time. Soon after the first applications of RV, microarray studies of the differential expression between pathogenic *N. meningitidis* and nonpathogenic *Neisseria lactamica*, upon contact with epithelial cells, was used to identify previously unknown vaccine candidates.[113,114] Overexpression of the *fHbp* protein of *N. meningitidis* has been measured in human blood.[115] Selective enrichment techniques of transcribed sequences[116] have allowed measurement of the level of mRNA transcription *in vivo* in *S. typhi*.[117] An integrated systems biology approach applied to *S. pyogenes* has identified many ways in which this pathogen reacts in a niche-specific manner to the interaction with the human host, identifying new potential vaccine candidates.[118]

New technologies, like RNA-Seq and tiling arrays, that are able to determine the transcriptome independently from genome annotation are increasingly used to validate genome annotation and to elucidate the structure of transcriptional units. In one of the first genome-wide transcriptome determinations using tiling arrays, it was shown that the structure and regulation of the transcriptional units in *L. monocytogenes* are far more intricate than previously known.[119] Strand-specific RNA-Seq has been used to define both the coding and non-coding transcripts in *S. typhi*, identifying many small noncoding RNAs that are likely to be an important regulatory mechanism.[120] Trascriptome analysis through RNA-Seq of several pathogens, such as *H. pylori*,[121] *Campyobacter jejuni*,[122] and *C. trachomatis*,[123] has been performed. The application of these techniques to *in vivo* studies will help to identify genes that are differentially expressed during infection, allowing a more effective selection of target antigens for vaccine formulation.

Discovery of new mechanisms of pathogenesis

The large body of research on previously unknown or poorly characterized surface proteins, initiated by projects based on RV, has substantially increased our knowledge of the mechanisms of interaction between pathogens and the human host, and has led to the discovery of new mechanisms of pathogenesis.

One example is fHbp in *N. meningitidis*. Originally known as GNA1870, fHbp is a surface-exposed lipoprotein that was selected as a vaccine target because of its ability to elicit protective antibodies in an infant rat model of bacteremia. GNA1870 was shown to bind factor H, a component of the alternative complement pathway, and for this reason was renamed fHbp.[36] By inhibiting the activation of the complement cascade, fHbp enhances the ability of bacteria to survive in human serum and might be essential for survival during carriage, when bacteria are often unencapsulated.

Another example is the discovery of pili, long filamentous structures protruding from the surface of cells in Gram-positive organisms. Although it has long been known that Gram-negative bacteria can express pilus-like structures to mediate attachment to host tissues,[124] the presence, structure, and function of pili in Gram-positive bacteria has only recently been characterized, mainly because of their importance as vaccine antigens. Pilus-like structures in a Gram-positive bacterium were first observed in *Corynebacterium renale* using electron microscopy in the late 1960s,[34] and their role in attachment to cells was soon discovered thereafter.[37] However, only recently has the structure and mechanism of assembly of Gram-positive pili been studied in detail.[125] Since then pili have been identified in several species, including *Actinomyces*s pp,[126] GAS,[38] GBS,[39,41] and *S. pneumonia*;[127,128] and the way in which they were discovered represents an example of how applied research can produce results of general interest. Gram-positive pili are very different from pili found in Gram-negative bacteria because they are formed by covalent linkage between protein subunits by the action of specialized sortase enzymes and are then anchored to the cell wall. Sortase-mediated attachment to the cell wall is a general mechanism used by Gram-positive bacteria to expose proteins on their surface; and sortase enzymes specifically recognize a conserved carboxy terminal motif (LPXTG). For this reason, proteins containing the LPXTG motif are among those routinely selected in RV projects on Gram-positive bacteria. In most cases, Gram-positive pili are encoded in genomic islands that contain between one and five LPXTG proteins, and between one and three specific sortases. The presence of these genomic structures in many species of *Streptococcus* attracted the attention of researchers looking for vaccine candidates in a specialized variant of RV called *targeted reverse vaccinology*.[129] Components of these genomic structures often prove to be protective in animal models against homologous strains,[38,39,41,130] although their use as vaccine targets is challenging because of their extreme variability[131] and presence only in a subset of virulent strains.[128,132]

Future developments

Integrated approaches

One of the key points in an RV project is that both the bioinformatic analysis and the proteomic analysis identify a large number of candidate vaccines, all of which need to be tested in an animal model to measure their ability to elicit protective antibodies. In an attempt to narrow down the list of potential vaccine targets and to streamline the vaccine development process, the concomitant use of an array of technologies, including bionformatics, proteomics, screening of human sera using protein arrays, and FACS analysis, has been proposed.[133] By simultaneously predicting/testing the conservation, expression, surface exposure, and immunogenicity of the predicted antigens, this experimental strategy allows an accurate selection of bacterial vaccines without using *in vitro* and/or *in vivo* protection assays, thus dramatically decreasing the costs of vaccine projects and maximizing the chances of success.

Single cell expression

Most of the steps in the RV antigen selection procedure rely on the assumption that genetically homogeneous populations of bacterial cells will display homogeneous phenotype. However, it is well known that marked phenotypic heterogeneities can be identified even in clonal cultures,[134] where individual microbial cells can exhibit variable degrees of resistance to antibiotics, motility, or expression of virulence determinants.

Phenotypic heterogeneity of a population is thought to allow bacteria to adapt to variations in the environment, leveraging on their large population size. The main advantage of these mechanisms is that because they do not involve irreversible genetic modifications the population of bacteria can revert to the original state once the transient stimulus is removed, thus providing an ideal solution to the problem of adapting to short-time environmental fluctuations. Bacterial pathogens have adopted these mechanisms in cases in which

the infection process involves selective bottlenecks, and to provide a reservoir of escape mutants that survive the action of the adaptive host immune system. In other cases, structures with high fitness costs are expressed only transiently, as recently shown for the type III secretion system *ttss-1,* the main virulence factor of *S. enteric* serovar Typhimurium that, at any given time, is present only in a fraction of the bacterial population *in vitro*.[135] This phenomenon poses a significant challenge to the use of vaccine targets of antigens that are not consistently expressed by the entire bacterial population, as in the case of the Var antigen in *Plasmodium falciparum*.[136]

Several mechanisms have been identified as the cause of these heterogeneities, including the presence of variations at the genome or epigenetic levels. Hypermutable loci have been recognized as one of the mechanisms that allow the generation of a huge number of genomic variants via the combination of different variants at individual loci. Polymerase slippage on simple sequence repeats is one of the best understood mechanism that generate hypermutability at selected loci, often resulting in inactivation of genes through phase variation.[137] These and other subtle mechanisms have been suggested to be responsible for the observed virulence differences between carriage and disease strains of *N. meningitidis*.[138] The existence of minority sequence heterogeneities in a clonal population of bacteria is very difficult to identify with conventional sequencing technologies that usually determine only a consensus sequence in which low copy number variants are averaged out. NGS technologies, with the high level of coverage that they can attain, will provide the basis for a genome-wide identification of these loci.

Other mechanisms that are likely to be much more widespread than currently thought can produce phenotypic heterogeneities in the absence of genotypic differences, as suggested in the case of the pilus expression in pilus-positive strains of *S. pneumonia*.[139] These mechanisms rely on the intrinsically non-linear nature of many expression regulatory networks.[134] A certain degree of cell-to-cell and time-dependent fluctuations in protein expression is unavoidable, and the relevance of this grows for proteins present in low copy numbers.[140] These fluctuations, when occurring in a molecular circuitry that includes feedback loops, can allow single cells to reversibly switch between mutually exclusive states,[141] effectively turning metabolic pathways into decision-capable circuits.[142] One of the first well-studied examples of a fluctuation-induced switch between two alternative stable states is the lambda switch in bacteriophages;[143] the role of fluctuation-induced switches has also been studied in many other microbial biological processes, from chemotaxis in *E. coli*[144] to viral latency in HIV.[145] Although it is hard to imagine the fitness advantage of these mechanisms in a fixed environment, it is now well understood that these mechanisms can actually confer a selective advantage in fluctuating environments.[146] Moreover, it has been shown that these complex regulatory circuitries can, in turn, modulate host physiology in a nontrivial way.[147]

Technical and methodological advances in experimental techniques and data analysis are boosting interest in cell-to-cell variability.[148–152] Determination of copy number variations in a clonal population of *E. coli* has shown the extent to which such fluctuations are an intrinsic property of single cell measurements of both the transcriptome and the proteome.[149] Large-scale application of single cell expression measurement technologies to populations of bacterial cells grown in infection-mimicking cultures is likely to provide valuable information that can be used in screening libraries of potential vaccine targets, allowing concentration on those that are consistently expressed by all members of the bacterial population or whose expression is essential during the infectious process.

Conclusions

Despite the many difficulties and pitfalls, RV has been a revolution in the field of vaccine discovery for infectious diseases, and has been instrumental in advancing vaccines for bacteria that were thought intractable. At the same time, the introduction into vaccine research of high-throughput technologies, such as NGS and proteomic technologies, has brought new insights into many aspects of the biology of infectious diseases and allowed the use of experimental data in several steps of the RV process that a decade ago were based only on bioinformatic predictions. Since their introduction, these technological advancements have greatly improved the efficiency of the vaccine target identification, selection, and development process. Further improvements are still possible that will widen the scope of

RV. For instance, use of large-scale screening technologies based on sequencing of complete genomes could increase the chances of success and speed the development of some viral vaccines, particularly for complex viruses, such as human cytomegalovirus, that have a coding capacity, the complexity and sophisticated regulation mechanisms of which we are now starting to appreciate.[153]

In this review, we have presented an overview of the aspects of RV that have undergone major changes, with an emphasis on the possibilities, opened by new technologies, to focus the costly testing in animal models on those molecules that have a high chance of success.

Bringing a new vaccine from basic research to a product ready for the market through the many stages of development is a challenging task that sometimes requires formulating completely new scientific paradigms; RV is a good example of such a paradigm. Although many of its original premises have changed, the RV approach, and more generally the use of omics approaches, is, and will continue to be in the future, a fundamental part of vaccine research projects.

Acknowledgments

The authors wish to tank Antonello Covacci and Marirosa Mora for critical reading of the manuscript; Isabel Delany, Luca Fagnocchi, Sébastien Brier, and Nathalie Norais for giving access to unpublished data, and Giorgio Corsi for artwork.

Conflicts of interest

The authors declare no conflicts of interest.

References

1. Fleischmann, R.D., M.D. Adams, O. White, *et al*. 1995. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science* **269:** 496–512.
2. Pizza, M., V. Scarlato, V. Masignani, *et al*. 2000. Identification of vaccine candidates against serogroup B meningococcus by whole-genome sequencing. *Science* **287:** 1816–1820.
3. Tettelin, H., N.J. Saunders, J. Heidelberg, *et al*. 2000. Complete genome sequence of Neisseria meningitidis serogroup B strain MC58. *Science* **287:** 1809–1815.
4. Giuliani, M.M., J. Adu-Bobie, M. Comanducci, *et al*. 2006. A universal vaccine for serogroup B meningococcus. *Proc. Natl. Acad. Sci. U.S.A.* **103:** 10834–10839.
5. Git, A., H. Dvinge, M. Salmon-Divon, *et al*. 2010. Systematic comparison of microarray profiling, real-time PCR, and next-generation sequencing technologies for measuring differential microRNA expression. *RNA* **16:** 991–1006.
6. Boerno, S.T., C. Grimm, H. Lehrach & M.R. Schweiger. 2010. Next-generation sequencing technologies for DNA methylation analyses in cancer genomics. *Epigenomics* **2:** 199–207.
7. Nobuta, K., K. McCormick, M. Nakano & B.C. Meyers. 2010. Bioinformatics analysis of small RNAs in plants using next generation sequencing technologies. *Methods Mol. Biol.* **592:** 89–106.
8. Maione, D., I. Margarit, C.D. Rinaudo, *et al*. 2005. Identification of a universal Group B streptococcus vaccine by multiple genome screen. *Science* **309:** 148–150.
9. Cheng, L., W. Lu, B. Kulkarni, *et al*. 2010. Analysis of chemotherapy response programs in ovarian cancers by the next-generation sequencing technologies. *Gynecol Oncol.* **117:** 159–169.
10. Metzker, M.L. 2010. Sequencing technologies – the next generation. *Nat. Rev. Genet.* **11:** 31–46.
11. Hacker, J. & J.B. Kaper. 2000. Pathogenicity islands and the evolution of microbes. *Annu. Rev. Microbiol.* **54:** 641–679.
12. Nakano, M., S. Yamamoto, A. Terai, *et al*. 2001. Structural and sequence diversity of the pathogenicity island of uropathogenic Escherichia coli which encodes the USP protein. *FEMS Microbiol. Lett.* **205:** 71–76.
13. Galan, J.E. 2001. Salmonella interactions with host cells: type III secretion at work. *Annu. Rev. Cell Dev. Biol.* **17:** 53–86.
14. Cornelis, G.R. 2000. Molecular and cell biology aspects of plague. *Proc. Natl. Acad. Sci. U.S.A.* **97:** 8778–8783.
15. Odenbreit, S., J. Puls, B. Sedlmaier, *et al*. 2000. Translocation of Helicobacter pylori CagA into gastric epithelial cells by type IV secretion. *Science* **287:** 1497–1500.
16. Tettelin, H., V. Masignani, M.J. Cieslewicz, *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.
17. Tettelin, H., D. Riley, C. Cattuto & D. Medini. 2008. Comparative genomics: the bacterial pan-genome. *Curr. Opin. Microbiol.* **11:** 472–477.
18. Touchon, M., C. Hoede, O. Tenaillon, *et al*. 2009. Organised genome dynamics in the Escherichia coli species results in highly diverse adaptive paths. *PLoS Genet.* **5:** e1000344.
19. Donati, C., N.L. Hiller, H. Tettelin, *et al*. 2010. Structure and dynamics of the pan-genome of Streptococcus pneumoniae and closely related species. *Genome Biol.* **11:** R107.
20. Muzzi, A. & C. Donati. 2011. Population genetics and evolution of the pan-genome of Streptococcus pneumoniae. *Int. J. Med. Microbiol.* **301:** 619–622.
21. Harris, S.R., I.N. Clarke, H.M. Seth-Smith, *et al*. 2012. Whole-genome analysis of diverse Chlamydia trachomatis strains identifies phylogenetic relationships masked by current clinical typing. *Nat. Genet.* **44:** 413–419, S411.
22. Stecher, B., R. Denzler, L. Maier, *et al*. 2012. Gut inflammation can boost horizontal gene transfer between pathogenic and commensal Enterobacteriaceae. *Proc. Natl. Acad. Sci. U.S.A.* **109:** 1269–1274.
23. Croucher, N.J., S.R. Harris, C. Fraser, *et al*. 2011. Rapid pneumococcal evolution in response to clinical interventions. *Science* **331:** 430–434.

24. Smillie, C.S., M.B. Smith, J. Friedman, *et al*. 2011. Ecology drives a global network of gene exchange connecting the human microbiome. *Nature* **480:** 241–244.
25. Litt, D.J., S. Savino, A. Beddek, *et al*. 2004. Putative vaccine antigens from Neisseria meningitidis recognized by serum antibodies of young children convalescing after meningococcal disease. *J. Infect. Dis.* **190:** 1488–1497.
26. Masignani, V., M. Comanducci, M.M. Giuliani, *et al*. 2003. Vaccination against Neisseria meningitidis using three variants of the lipoprotein GNA1870. *J. Exp. Med.* **197:** 789–799.
27. Seib, K.L., B. Brunelli, B. Brogioni, *et al*. 2011. Characterization of diverse subvariants of the meningococcal factor H (fH) binding protein for their ability to bind fH, to mediate serum resistance, and to induce bactericidal antibodies. *Infect. Immun.* **79:** 970–981.
28. Brehony, C., D.J. Wilson & M.C. Maiden. 2009. Variation of the factor H-binding protein of Neisseria meningitidis. *Microbiology* **155:** 4155–4169.
29. Kodama, Y., E. Kaminuma, S. Saruhashi, *et al*. 2010. Biological databases at DNA Data Bank of Japan in the era of next-generation sequencing technologies. *Adv Exp Med. Biol.* **680:** 125–135.
30. Ravn, U., F. Gueneau, L. Baerlocher, *et al*. 2010. By-passing in vitro screening–next generation sequencing technologies applied to antibody display and in silico candidate selection. *Nucleic Acids Res.* **38:** e193.
31. Nagarajan, N. & M. Pop. 2010. Sequencing and genome assembly using next-generation technologies. *Methods Mol. Biol.* **673:** 1–17.
32. Cahill, M.J., C.U. Koser, N.E. Ross & J.A. Archer. 2010. Read length and repeat resolution: exploring prokaryote genomes using next-generation sequencing technologies. *PLoS One* **5:** e11518.
33. Martin, P., T. van de Ven, N. Mouchel, *et al*. 2003. Experimentally revised repertoire of putative contingency loci in Neisseria meningitidis strain MC58: evidence for a novel mechanism of phase variation. *Mol. Microbiol.* **50:** 245–257.
34. Yanagawa, R., K. Otsuki & T. Tokui. 1968. Electron microscopy of fine structure of Corynebacterium renale with special reference to pili. *Jpn. J. Vet. Res.* **16:** 31–37.
35. Donnelly, J., D. Medini, G. Boccadifuoco, *et al*. 2010. Qualitative and quantitative assessment of meningococcal antigens to evaluate the potential strain coverage of protein-based vaccines. *Proc. Natl. Acad. Sci. U.S.A.* **107:** 19490–19495.
36. Madico, G., J.A. Welsch, L.A. Lewis, *et al*. 2006. The meningococcal vaccine candidate GNA1870 binds the complement regulatory protein factor H and enhances serum resistance. *J. Immunol.* **177:** 501–510.
37. Honda, E. & R. Yanagawa. 1975. Attachment of Corynebacterium renale to tissue culture cells by the pili. *Am. J. Vet. Res.* **36:** 1663–1666.
38. Mora, M., G. Bensi, S. Capo, *et al*. 2005. Group A Streptococcus produce pilus-like structures containing protective antigens and Lancefield T antigens. *Proc. Natl. Acad. Sci. U.S.A.* **102:** 15641–15646.
39. Lauer, P., C.D. Rinaudo, M. Soriani, *et al*. 2005. Genome analysis reveals pili in Group B Streptococcus. *Science* **309:** 105.
40. Roh, S.W., G.C. Abell, K.H. Kim, *et al*. 2010. Comparing microarrays and next-generation sequencing technologies for microbial ecology research. *Trends Biotechnol.* **28:** 291–299.
41. Rosini, R., C.D. Rinaudo, M. Soriani, *et al*. 2006. Identification of novel genomic islands coding for antigenic pilus-like structures in Streptococcus agalactiae. *Mol. Microbiol.* **61:** 126–141.
42. Crisafulli, G., S. Guidotti, A. Muzzi, *et al*. 2012. An extended multi-locus molecular typing schema for Streptococcus pneumoniae demonstrates that a limited number of capsular switch events is responsible for serotype heterogeneity of closely related strains from different countries. *Infection Genet. Evol.*
43. Maiden, M.C., J.A. Bygraves, E. Feil, *et al*. 1998. Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Proc. Natl. Acad. Sci. U.S.A.* **95:** 3140–3145.
44. Enright, M.C. & B.G. Spratt. 1998. A multilocus sequence typing scheme for Streptococcus pneumoniae: identification of clones associated with serious invasive disease. *Microbiology* **144**(Pt 11)**:** 3049–3060.
45. Enright, M.C., N.P. Day, C.E. Davies, *et al*. 2000. Multilocus sequence typing for characterization of methicillin-resistant and methicillin-susceptible clones of Staphylococcus aureus. *J. Clin. Microbiol.* **38:** 1008–1015.
46. Enright, M.C. & B.G. Spratt. 1999. Multilocus sequence typing. *Trends Microbiol.* **7:** 482–487.
47. Enright, M.C., B.G. Spratt, A. Kalia, *et al*. 2001. Multilocus sequence typing of Streptococcus pyogenes and the relationships between emm type and clone. *Infect. Immun.* **69:** 2416–2427.
48. Grundmann, H., S. Hori, M.C. Enright, *et al*. 2002. Determining the genetic structure of the natural population of Staphylococcus aureus: a comparison of multilocus sequence typing with pulsed-field gel electrophoresis, randomly amplified polymorphic DNA analysis, and phage typing. *J. Clin. Microbiol.* **40:** 4544–4546.
49. Chan, M.S., M.C. Maiden & B.G. Spratt. 2001. Database-driven multi locus sequence typing (MLST) of bacterial pathogens. *Bioinformatics* **17:** 1077–1083.
50. Altschul, S.F., W. Gish, W. Miller, *et al*. 1990. Basic local alignment search tool. *J. Mol. Biol.* **215:** 403–410.
51. Pearson, W.R. & D.J. Lipman. 1988. Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. U.S.A.* **85:** 2444–2448.
52. Thompson, J.D., T.J. Gibson & D.G. Higgins. 2002. Multiple sequence alignment using ClustalW and ClustalX. In *Current Protocols in Bioinformatics*. Andreas D. Baxevanis, Chapter 2: Unit 2. 3. Wiley. New York.
53. Edgar, R.C. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32:** 1792–1797.
54. Li, H. & R. Durbin. 2010. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26:** 589–595.
55. Li, H., B. Handsaker, A. Wysoker, *et al*. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25:** 2078–2079.

56. Harris, S.R., E.J. Feil, M.T. Holden, *et al*. 2010. Evolution of MRSA during hospital transmission and intercontinental spread. *Science* **327:** 469–474.
57. Marttinen, P., W.P. Hanage, N.J. Croucher, *et al*. 2012. Detection of recombination events in bacterial genomes from large population samples. *Nucleic Acids Res.* **40:** e6.
58. Delcher, A.L., K.A. Bratke, E.C. Powers & S.L. Salzberg. 2007. Identifying bacterial genes and endosymbiont DNA with Glimmer. *Bioinformatics* **23:** 673–679.
59. Delcher, A.L., D. Harmon, S. Kasif, *et al*. 1999. Improved microbial gene identification with GLIMMER. *Nucleic Acids Res.* **27:** 4636–4641.
60. Salzberg, S.L., A.L. Delcher, S. Kasif & O. White. 1998. Microbial gene identification using interpolated Markov models. *Nucleic Acids Res.* **26:** 544–548.
61. Salzberg, S.L., M. Pertea, A.L. Delcher, *et al*. 1999. Interpolated Markov models for eukaryotic gene finding. *Genomics* **59:** 24–31.
62. Lukashin, A.V. & M. Borodovsky. 1998. GeneMark.hmm: new solutions for gene finding. *Nucleic Acids Res.* **26:** 1107–1115.
63. Guo, F.B., H.Y. Ou & C.T. Zhang. 2003. ZCURVE: a new system for recognizing protein-coding genes in bacterial and archaeal genomes. *Nucleic Acids Res.* **31:** 1780–1789.
64. Larsen, T.S. & A. Krogh. 2003. EasyGene–a prokaryotic gene finder that ranks ORFs by statistical significance. *BMC Bioinformatics* **4:** 21.
65. Zhu, H., G.Q. Hu, Y.F. Yang, *et al*. 2007. MED: a new non-supervised gene prediction algorithm for bacterial and archaeal genomes. *BMC Bioinformatics* **8:** 97.
66. Nielsen, P. & A. Krogh. 2005. Large-scale prokaryotic gene prediction and comparison to genome annotation. *Bioinformatics* **21:** 4322–4329.
67. Denoeud, F., J.M. Aury, C. Da Silva, *et al*. 2008. Annotating genomes with massive-scale RNA sequencing. *Genome Biol.* **9:** R175.
68. Martin, J., W. Zhu, K.D. Passalacqua, *et al*. 2010. Bacillus anthracis genome organization in light of whole transcriptome sequencing. *BMC Bioinformatics* **11**(Suppl 3)**:** S10.
69. Mader, U., P. Nicolas, H. Richard, *et al*. 2011. Comprehensive identification and quantification of microbial transcriptomes by genome-wide unbiased methods. *Curr. Opin. Biotechnol.* **22:** 32–41.
70. Li, W., J. Feng & T. Jiang. 2011. IsoLasso: a LASSO regression approach to RNA-Seq based transcriptome assembly. *J. Comput. Biol.* **18:** 1693–1707.
71. Trapnell, C., B.A. Williams, G. Pertea, *et al*. 2010. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28:** 511–515.
72. Guell, M., V. van Noort, E. Yus, *et al*. 2009. Transcriptome complexity in a genome-reduced bacterium. *Science* **326:** 1268–1271.
73. Jaffe, J.D., H.C. Berg & G.M. Church. 2004. Proteogenomic mapping as a complementary method to perform genome annotation. *Proteomics* **4:** 59–77.
74. Ansong, C., N. Tolic, S.O. Purvine, *et al*. 2011. Experimental annotation of post-translational features and translated coding regions in the pathogen Salmonella Typhimurium. *BMC Genomics* **12:** 433.
75. Gardy, J.L. & F.S. Brinkman. 2006. Methods for predicting bacterial protein subcellular localization. *Nat. Rev. Microbiol.* **4:** 741–751.
76. Nair, R. & B. Rost. 2002. Sequence conserved for subcellular localization. *Protein Sci.* **11:** 2836–2847.
77. Yu, C.S., Y.C. Chen, C.H. Lu & J.K. Hwang. 2006. Prediction of protein subcellular localization. *Proteins* **64:** 643–651.
78. Chou, K.C. & H.B. Shen. 2006. Large-scale predictions of gram-negative bacterial protein subcellular locations. *J. Proteome Res.* **5:** 3420–3428.
79. Shen, H.B. & K.C. Chou. 2007. Gpos-PLoc: an ensemble classifier for predicting subcellular localization of Gram-positive bacterial proteins. *Protein Eng. Des. Sel.* **20:** 39–46.
80. Chang, J.M., E.C. Su, A. Lo, *et al*. 2008. PSLDoc: protein subcellular localization prediction based on gapped-dipeptides and probabilistic latent semantic analysis. *Proteins* **72:** 693–710.
81. Su, E.C., H.S. Chiu, A. Lo, *et al*. 2007. Protein subcellular localization prediction based on compartment-specific features and structure conservation. *BMC Bioinformatics* **8:** 330.
82. Matsuda, S., J.P. Vert, H. Saigo, *et al*. 2005. A novel representation of protein sequences for prediction of subcellular location using support vector machines. *Protein Sci.* **14:** 2804–2813.
83. Gardy, J.L., C. Spencer, K. Wang, *et al*. 2003. PSORT-B: Improving protein subcellular localization prediction for Gram-negative bacteria. *Nucleic Acids Res.* **31:** 3613–3617.
84. Gardy, J.L., M.R. Laird, F. Chen, *et al*. 2005. PSORTb v.2.0: expanded prediction of bacterial protein subcellular localization and insights gained from comparative proteome analysis. *Bioinformatics* **21:** 617–623.
85. Yu, N.Y., J.R. Wagner, M.R. Laird, *et al*. 2010. PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. *Bioinformatics* **26:** 1608–1615.
86. Yu, N.Y., M.R. Laird, C. Spencer & F.S. Brinkman. 2011. PSORTdb–an expanded, auto-updated, user-friendly protein subcellular localization database for Bacteria and Archaea. *Nucleic Acids Res.* **39:** D241–244.
87. Bulashevska, A. & R. Eils. 2006. Predicting protein subcellular locations using hierarchical ensemble of Bayesian classifiers based on Markov chains. *BMC Bioinformatics* **7:** 298.
88. Niu, B., Y.H. Jin, K.Y. Feng, *et al*. 2008. Using AdaBoost for the prediction of subcellular location of prokaryotic and eukaryotic proteins. *Mol. Divers* **12:** 41–45.
89. E-komon, T., R.J. Burchmore, P. Herzyk & R.L. Davies. 2012. Predicting the outer membrane proteome of Pasteurella multocida based on consensus prediction enhanced by results integration and manual confirmation. *BMC Bioinformatics* **13:** 63.
90. Goudenege, D., S. Avner, C. Lucchetti-Miganeh & F. Barloy-Hubler. 2010. CoBaltDB: Complete bacterial and archaeal

orfeomes subcellular localization database and associated resources. *BMC Microbiol.* **10:** 88.
91. Rey, S., J.L. Gardy & F.S. Brinkman. 2005. Assessing the precision of high-throughput computational and laboratory approaches for the genome-wide identification of protein subcellular localization in bacteria. *BMC Genomics* **6:** 162.
92. Shatkay, H., A. Hoglund, S. Brady, *et al.* 2007. SherLoc: high-accuracy prediction of protein subcellular localization by integrating text and protein sequence data. *Bioinformatics* **23:** 1410–1417.
93. Thieu, T., S. Joshi, S. Warren & D. Korkin. 2012. Literature mining of host-pathogen interactions: comparing feature-based supervised learning and language-based approaches. *Bioinformatics* **28:** 867–875.
94. Walters, M.S. & H.L. Mobley. 2010. Bacterial proteomics and identification of potential vaccine targets. *Expert Rev. Proteomics* **7:** 181–184.
95. Rodriguez-Ortega, M.J., N. Norais, G. Bensi, *et al.* 2006. Characterization and identification of vaccine candidate proteins through analysis of the group A Streptococcus surface proteome. *Nat. Biotechnol.* **24:** 191–197.
96. Ferrari, G., I. Garaguso, J. Adu-Bobie, *et al.* 2006. Outer membrane vesicles from group B Neisseria meningitidis delta gna33 mutant: proteomic and immunological comparison with detergent-derived outer membrane vesicles. *Proteomics* **6:** 1856–1866.
97. Berlanda Scorza, F., F. Doro, M.J. Rodriguez-Ortega, *et al.* 2008. Proteomics characterization of outer membrane vesicles from the extraintestinal pathogenic Escherichia coli DeltatolR IHE3034 mutant. *Mol. Cell Proteomics* **7:** 473–485.
98. Wright, A., R. Wait, S. Begum, *et al.* 2005. Proteomic analysis of cell surface proteins from Clostridium difficile. *Proteomics* **5:** 2443–2452.
99. Tefon, B.E., S. Maass, E. Ozcengiz, *et al.* 2011. A comprehensive analysis of Bordetella pertussis surface proteome and identification of new immunogenic proteins. *Vaccine* **29:** 3583–3595.
100. Doro, F., S. Liberatori, M.J. Rodriguez-Ortega, *et al.* 2009. Surfome analysis as a fast track to vaccine discovery: identification of a novel protective antigen for Group B Streptococcus hypervirulent strain COH1. *Mol. Cell Proteomics* **8:** 1728–1737.
101. Cash, P. 2011. Investigating pathogen biology at the level of the proteome. *Proteomics* **11:** 3190–3202.
102. Lopez-Campistrous, A., P. Semchuk, L. Burke, *et al.* 2005. Localization, annotation, and comparison of the Escherichia coli K-12 proteome under two states of growth. *Mol. Cell Proteomics* **4:** 1205–1209.
103. Alteri, C.J. & H.L. Mobley. 2007. Quantitative profile of the uropathogenic Escherichia coli outer membrane proteome during growth in human urine. *Infect. Immun.* **75:** 2679–2688.
104. Korber, B., M. LaBute & K. Yusim. 2006. Immunoinformatics comes of age. *PLoS Computat. Biol.* **2:** e71.
105. Blythe, M.J. & D.R. Flower. 2005. Benchmarking B cell epitope prediction: underperformance of existing methods. *Protein Sci.* **14:** 246–248.
106. Haste Andersen, P., M. Nielsen & O. Lund. 2006. Prediction of residues in discontinuous B-cell epitopes using protein 3D structures. *Protein Sci.* **15:** 2558–2567.
107. Sette, A. & R. Rappuoli. Reverse vaccinology: developing vaccines in the era of genomics. *Immunity* **33:** 530–541.
108. Kim, Y., J. Ponomarenko, Z. Zhu, *et al.* 2012. Immune epitope database analysis resource. *Nucleic acids Res.* **40:** W525–530.
109. Comanducci, M., S. Bambini, B. Brunelli, *et al.* 2002. NadA, a novel vaccine candidate of Neisseria meningitidis. *J. Exp. Med.* **195:** 1445–1454.
110. Capecchi, B., J. Adu-Bobie, F. Di Marcello, *et al.* 2005. Neisseria meningitidis NadA is a new invasin which promotes bacterial adhesion to and penetration into human epithelial cells. *Mol. Microbiol.* **55:** 687–698.
111. Metruccio, M.M., E. Pigozzi, D. Roncarati, *et al.* 2009. A novel phase variation mechanism in the meningococcus driven by a ligand-responsive repressor and differential spacing of distal promoter elements. *PLoS Pathog.* **5:** e1000710.
112. Brier, S., L. Fagnocchi, D. Donnarumma, *et al.* 2012. Structural insight into the mechanism of DNA-binding attenuation of the Neisserial adhesin repressor NadR by the small natural ligand 4-hydroxyphenylacetic acid. *Biochemistry*. **51:** 6738–6752.
113. Grifantini, R., E. Bartolini, A. Muzzi, *et al.* 2002. Previously unrecognized vaccine candidates against group B meningococcus identified by DNA microarrays. *Nat. Biotechnol.* **20:** 914–921.
114. Grifantini, R., E. Bartolini, A. Muzzi, *et al.* 2002. Gene expression profile in Neisseria meningitidis and Neisseria lactamica upon host-cell contact: from basic research to vaccine development. *Ann. N. Y. Acad. Sci.* **975:** 202–216.
115. Echenique-Rivera, H., A. Muzzi, E. Del Tordello, *et al.* 2011. Transcriptome analysis of Neisseria meningitidis in human whole blood and mutagenesis studies identify virulence factors involved in blood survival. *PLoS Pathog.* **7:** e1002027.
116. Daigle, F., J.Y. Hou & J.E. Clark-Curtiss. 2002. Microbial gene expression elucidated by selective capture of transcribed sequences (SCOTS). *Methods Enzymol.* **358:** 108–122.
117. Sheikh, A., R.C. Charles, N. Sharmeen, *et al.* 2011. In vivo expression of Salmonella enterica serotype Typhi genes in the blood of patients with typhoid fever in Bangladesh. *PLoS Negl. Trop. Dis.* **5:** e1419.
118. Musser, J.M. & F.R. DeLeo. 2005. Toward a genome-wide systems biology analysis of host-pathogen interactions in group A Streptococcus. *Am. J. Pathol.* **167:** 1461–1472.
119. Toledo-Arana, A., O. Dussurget, G. Nikitas, *et al.* 2009. The Listeria transcriptional landscape from saprophytism to virulence. *Nature* **459:** 950–956.
120. Perkins, T.T., R.A. Kingsley, M.C. Fookes, *et al.* 2009. A strand-specific RNA-Seq analysis of the transcriptome of the typhoid bacillus Salmonella typhi. *PLoS Genet.* **5:** e1000569.
121. Sharma, C.M., S. Hoffmann, F. Darfeuille, *et al.* 2010. The primary transcriptome of the major human pathogen Helicobacter pylori. *Nature* **464:** 250–255.

122. Chaudhuri, R.R., L. Yu, A. Kanji, *et al*. 2011. Quantitative RNA-seq analysis of the Campylobacter jejuni transcriptome. *Microbiology* **157:** 2922–2932.
123. Albrecht, M., C.M. Sharma, R. Reinhardt, *et al*. 2010. Deep sequencing-based discovery of the Chlamydia trachomatis transcriptome. *Nucleic Acids Res.* **38:** 868–877.
124. Proft, T. & E.N. Baker. 2009. Pili in Gram-negative and Gram-positive bacteria–structure, assembly and their role in disease. *Cell Mol. Life Sci.* **66:** 613–635.
125. Ton-That, H. & O. Schneewind. 2003. Assembly of pili on the surface of Corynebacterium diphtheriae. *Mol. Microbiol.* **50:** 1429–1438.
126. Kelstrup, J., J. Theilade & O. Fejerskov. 1979. Surface ultrastructure of some oral bacteria. *Scand. J. Dent. Res.* **87:** 415–423.
127. Barocchi, M.A., J. Ries, X. Zogaj, *et al*. 2006. A pneumococcal pilus influences virulence and host inflammatory responses. *Proc. Natl. Acad. Sci. U.S.A.* **103:** 2857–2862.
128. F. Bagnoli, M. Moschioni, C. Donati, *et al*. 2008. A second pilus type in Streptococcus pneumoniae is prevalent in emerging serotypes and mediates adhesion to host cells. *J. Bacteriol.* **190:** 5480–5492.
129. Mora, M. & J.L. Telford. 2010. Genome-based approaches to vaccine development. *J. Mol. Med. (Berl).* **88:** 143–147.
130. Gianfaldoni, C., S. Censini, M. Hilleringmann, *et al*. 2007. Streptococcus pneumoniae pilus subunits protect mice against lethal challenge. *Infect. Immun.* **75:** 1059–1062.
131. Falugi, F., C. Zingaretti, V. Pinto, *et al*. 2008. Sequence variation in group A Streptococcus pili and association of pilus backbone types with lancefield T serotypes. *J. Infect. Dis.* **198:** 1834–1841.
132. Moschioni, M., C. Donati, A. Muzzi, *et al*. 2008. Streptococcus pneumoniae contains 3 rlrA pilus variants that are clonally related. *J. Infect. Dis.* **197:** 888–896.
133. Bensi, G., M. Mora, G. Tuscano, *et al*. 2012. Multi High-Throughput Approach for Highly Selective Identification of Vaccine Candidates: the Group A Streptococcus Case. *Mol. Cell Proteomics.* **11:** M111.015693.
134. Avery, S.V. 2006. Microbial cell individuality and the underlying sources of heterogeneity. *Nat. Rev. Microbiol.* **4:** 577–587.
135. Sturm, A., M. Heinemann, M. Arnoldini, *et al*. 2011. The cost of virulence: retarded growth of Salmonella Typhimurium cells expressing type III secretion system 1. *PLoS Pathog.* **7:** e1002143.
136. Ralph, S.A. & A. Scherf. 2005. The epigenetic control of antigenic variation in Plasmodium falciparum. *Curr. Opin. Microbiol.* **8:** 434–440.
137. Moxon, R., C. Bayliss & D. Hood. 2006. Bacterial contingency loci: the role of simple sequence DNA repeats in bacterial adaptation. *Annu. Rev. Genet.* **40:** 307–333.
138. Schoen, C., H. Tettelin, J. Parkhill & M. Frosch. 2009. Genome flexibility in Neisseria meningitidis. *Vaccine* **27**(Suppl 2)**:** B103–111.
139. De Angelis, G., M. Moschioni, A. Muzzi, *et al*. 2011. The Streptococcus pneumoniae pilus-1 displays a biphasic expression pattern. *PLoS One* **6:** e21269.
140. Elowitz, M.B., A.J. Levine, E.D. Siggia & P.S. Swain. 2002. Stochastic gene expression in a single cell. *Science* **297:** 1183–1186.
141. Smits, W.K., O.P. Kuipers & J.W. Veening. 2006. Phenotypic variation in bacteria: the role of feedback regulation. *Nat. Rev. Microbiol.* **4:** 259–271.
142. Balázsi, G., A. van Oudenaarden & J.J. Collins. 2011. Cellular decision making and biological noise: from microbes to mammals. *Cell* **144:** 910–925.
143. Shea, M.A. & G.K. Ackers. 1985. The OR control system of bacteriophage lambda. A physical-chemical model for gene regulation. *J. Mol. Biol.* **181:** 211–230.
144. Korobkova, E., T. Emonet, J.M. Vilar, *et al*. 2004. From molecular noise to behavioural variability in a single bacterium. *Nature* **428:** 574–578.
145. Weinberger, L.S., J.C. Burnett, J.E. Toettcher, *et al*. 2005. Stochastic gene expression in a lentiviral positive-feedback loop: HIV-1 Tat fluctuations drive phenotypic diversity. *Cell* **122:** 169–182.
146. Thattai, M. & A. van Oudenaarden. 2004. Stochastic gene expression in fluctuating environments. *Genetics* **167:** 523–530.
147. Tan, C., P. Marguet & L. You. 2009. Emergent bistability by a growth-modulating positive feedback circuit. *Nat. Chem. Biol.* **5:** 842–848.
148. Brehm-Stecher, B.F. & E.A. Johnson. 2004. Single-cell microbiology: tools, technologies, and applications. *Microbiol. Mol. Biol. Rev* **68:** 538–559.
149. Taniguchi, Y., P.J. Choi, G.W. Li, *et al*. 2010. Quantifying E. coli proteome and transcriptome with single-molecule sensitivity in single cells. *Science* **329:** 533–538.
150. Bandura, D.R., V.I. Baranov, O.I. Ornatsky, *et al*. 2009. Mass cytometry: technique for real time single cell multitarget immunoassay based on inductively coupled plasma time-of-flight mass spectrometry. *Anal. Chem.* **81:** 6813–6822.
151. Qiu, P., E.F. Simonds, S.C. Bendall, *et al*. 2011. Extracting a cellular hierarchy from high-dimensional cytometry data with SPADE. *Nat. Biotechnol.* **29:** 886–891.
152. Kalisky, T. & S.R. Quake. 2011. Single-cell genomics. *Nat. Methods* **8:** 311–314.
153. Stern-Ginossar, N., B. Weisburd, A. Michalski, *et al*. 2012. Decoding human cytomegalovirus. *Science* **338:** 1088–1093.