ELSEVIER

Microbial genomes and vaccine design: refinements to the classical reverse vaccinology approach

Marirosa Mora, Claudio Donati, Duccio Medini, Antonello Covacci and Rino Rappuoli

The advent of whole-genome sequencing of bacteria and advances in bioinformatics have revolutionized the study of bacterial pathogenesis, enabling the targeting of possible vaccine candidates starting from genomic information. Nowadays, the availability of hundreds of bacterial genomes enables identification of the genetic differences across several genomes from the same species. The unexpected degree of intra-species diversity suggests that a single genome sequence is not entirely representative and does not offer a complete picture of the genetic variability of a species. The practical consequence is that, in many cases, a universal vaccine is possible only by including a combination of antigens and this combination must take into account the pathogen population structure.

Addresses
Novartis Vaccines, Via Fiorentina, 53100 Siena, Italy

Corresponding author: Rappuoli, Rino (rino_rappuoli@chiron.com)

Current Opinion in Microbiology 2006, **9**:532–536

This review comes from a themed issue on
Genomics
Edited by Dave Ussery and Tim Read

Available online 4th August 2006

1369-5274/$ – see front matter

DOI 10.1016/j.mib.2006.07.003

Introduction

The late 1990s marked the beginning of the era of genomics, with the publication of the first genome sequence of a free-living organism, that of the bacterium *Haemophilus influenzae* [1]. Since then, access to the entire genetic content of human pathogens and advances in bioinformatics analysis have opened innovative and efficient research avenues for the identification of a wide array of novel antigens for potential use as vaccines against infectious disease.

Genomic sequencing has since experienced an exponential growth; almost 300 bacterial genomes have been completed and more than 500 are currently being sequenced. All these sequences are available in public databases and represent hundreds of species, as well as multiple strains of the same species.

It is now possible to compare the sequences of related bacteria, compare pathogens with commensals of the same species or those of a related species, and even compare bacteria with different or with similar pathogenic profiles, enabling the putative identification of disease-related genes.

'Classical' reverse vaccinology

An example of the first application of reverse vaccinology that is a demonstration of the power of genomic approaches for target antigen identification is the attempt to develop a vaccine against serogroup B *Neisseria meningitidis* [2], the major cause of sepsis and meningitis in children and young adults.

For several decades, meningococcus posed insurmountable obstacles to conventional vaccinology approaches; these were overcome by mining the information from the sequenced *N. meningitidis* genome. While the genome sequencing project was still in progress — starting from the concept that secreted or extracellular proteins, which associate with bacterial membranes, are more easily accessible to antibodies and therefore represent ideal vaccine candidates — the incompletely assembled DNA fragments were screened using computer analysis to select proteins predicted to be present on the bacterial surface or proteins with homologies to known bacterial factors involved in pathogenesis and virulence.

More than 600 genes predicted to code for surface-exposed proteins were identified, half of which were expressed and tested for immunogenicity. From this, 91 antigens were identified, 29 of which were novel protective antigens. Some of these candidates are now in clinical trials.

The successful MenB (serogroup B *N. meningitides*) example has prompted the application of the reverse vaccinology to other pathogens, such as *Streptococcus pneumoniae* [3], *Porphyromonas gingivalis* [4], *Chlamydia pneumoniae* [5], *Bacillus anthracis* [6] and many others.

This genome-based approach is now used routinely in vaccine development, and is a major tool in the quest for vaccines, becoming therefore a 'classical' approach.

'Pan-genomic' reverse vaccinology

Following the pioneering effort on *N. meningitidis*, the classical reverse vaccinology approach has been applied to a growing number of microorganisms. In 2002, the

complete sequence of a virulent isolate of *Streptococcus agalactiae* (group B streptococcus or GBS), one of the leading causes of bacterial sepsis, pneumonia and meningitis in neonates in the USA and Europe, was determined [7]. Using the sequenced strain as a reference, comparative genomic hybridization (CGH) was applied to circumvent the need for sequencing multiple closely related genomes, enabling study of the genetic diversity of this species. Comparisons of this kind reveal regions of loss and/or retention with respect to the reference strain. It was found that approximately 18% of the genes encoded in the sequenced strain are absent in at least one of the other nineteen tested *S. agalactiae* strains. However, CGH experiments are only able to identify the portion of the sequenced genome that is shared with other test strains, and is not able to detect genes that are absent in the reference genome; thus, this leaves open the question of defining the panel of genes that pertain to the entire bacterial species. Following this work, the sequences of six additional strains of *S. agalactiae* were determined [8,9•]. Using this new information, mathematical extrapolation enabled researchers to estimate that 1806 genes are shared by all strains of *S. agalactiae*, and these genes form the species 'core genome'. This represents approximately 80% of the average number of genes encoded in each strain. Although the core genome includes genes that belong to many functional categories, the housekeeping genes are over-represented, suggesting that the core genome mainly encodes factors for functions that contribute to the major metabolic pathways, which are shared by all strains and define the species identity. The complementary set of genes that are absent in at least one strain form the 'dispensable genome', which is probably responsible for the adaptation of individual strains to environmental conditions specific to particular ecological niches. In the case of GBS, it was found that each new sequence contributed between thirteen and 61 genes previously not found in GBS. On the basis of these data, the size of the species pan-genome (i.e. the set of genes that would be found at least once if an infinite number of strains were to be sequenced) was estimated. Surprisingly, mathematical extrapolation of the existing data predicts that, no matter how many strains have been sequenced, each new sequence would contain genes that have not been encountered before, leading the counter-intuitive conclusion that this species pan-genome continues to grow without bounds as the number of sequenced strains grows.

These findings proved instrumental to the design of a universal vaccine against GBS [10•]. Use of computational algorithms enabled the prediction that GBS contains 589 surface-associated proteins, of which 396 were core genes and the remaining 193 were genes absent in at least one strain. Each of these proteins was tested for protection against GBS, and four antigens were able to elicit protective immunity in an animal model. The important novelty of this study is that none of these antigens could be classified as universal, because three of them were absent in a fraction of the tested strains, and the fourth core gene showed negligible surface accessibility in some strains. The use of multi-genome sequence information for vaccine design represented a major conceptual step from the common concept that a single genome sequence is sufficient to identify surface-associated proteins to be tested as potential vaccine candidates. Because a single genomic sequence is not sufficient to represent the variability of bacterial populations, multiple sequences might be needed to identify a vaccine formulation that is effective in the case of a highly differentiated species, and this situation is likely to be common to many important bacterial pathogens.

Recent evidence has shown that the genetic variability within single naturally occurring, seemingly homogenous populations of bacteria could be much higher than expected. For instance, an high-throughput study of the genomic diversity within a single population of a coastal bacterioplankton, showed that a single environmental niche could host greater than a thousand distinct genotypes, all sharing at least 99% 16S rRNA identity [11••]. Analyzing in detail twelve randomly chosen clones, the authors showed extensive allelic heterogeneity, with genomes varying in size by greater than 20%, and with no spatial or temporal substructure in the population. These results could be relevant in the design of vaccines against mainly commensal pathogens that occasionally become pathogenic, such as non-pathogenic and pathogenic types of *E. coli* [12].

For uropathogenic strains of *E. coli*, acquisition of a pathogenicity island resulted in conferring it the ability of to infect the urinary tract and bloodstream and evade host defences without compromising its ability to harmlessly colonize the intestine. If the genotypic variability of the colonizing population shows a similar degree of heterogeneity to that found in environmental samples, vaccine formulates against these pathogens should be designed to cover a wide panel of circulating strains.

From population genomics to population vaccinology

The accumulation of evidence concerning the degree of genomic diversity of pathogenic microbial species has resulted in increased attention to the characterization of internal structures of bacterial populations. Indeed, when a universal vaccine is only obtainable by using a combination of antigens chosen from different strains, their selection should take into account the population structure of the microorganism, weighting each representative strain with its relevance in the epidemiology of the disease.

The first epidemiological studies that observed a certain degree of association of particular serotypes with disease

[13], suggested an oversimplified, clonal view of bacterial populations. Subsequent multi-locus enzyme electrophoresis studies [14,15] enabled the application of theoretical and statistical methods of population biology to bacteria [16], suggesting the existence of at least three different kinds of bacterial populations: clonal species, like *Salmonella*, where mutation is the only active evolutionary force; panmictic (i.e. fully sexual) species, such as *Neisseria gonorrhoeae*, whose population structure is mainly determined by recombination; and epidemic species, such as *Neisseria meningitidis*, composed of a sexual background of relatively rare, highly recombining and unrelated genotypes, as well as single, highly adaptive genotypes that expand clonally giving rise to hypervirulent clusters.

The advent of multi-locus sequence typing (MLST) [17] — on the basis of the analysis of DNA sequences collected from seven, selectively neutral housekeeping loci — greatly enhanced the resolution of bacterial population studies. Collection of such sequences from several isolates enabled quantification of the contributions of mutation and recombination to the population structure of various species [18,19]. At present, MLST schemes are publicly available for 21 microbial species (see http://www.mlst.net), with thousands of isolates characterized for each of these.

However, the assumption that loci such as genes coding for antigenic proteins, under environmental selection share the same evolutionary history as neutral loci might lead to incorrect results.

Whereas MLST studies based on neutral loci provide accurate representations of the micro-evolutionary structures of a bacterial population, recent studies demonstrated that genes exposed to the selective pressure of the host immune system might substantially violate such predictions [20].

As a consequence, in a vaccine-focused perspective, it appears that the identification, characterization and mapping of non-neutral loci in order to detect adaptive genetic signatures and obtain a more comprehensive and epidemiologically-oriented picture of bacterial populations are increasingly important.

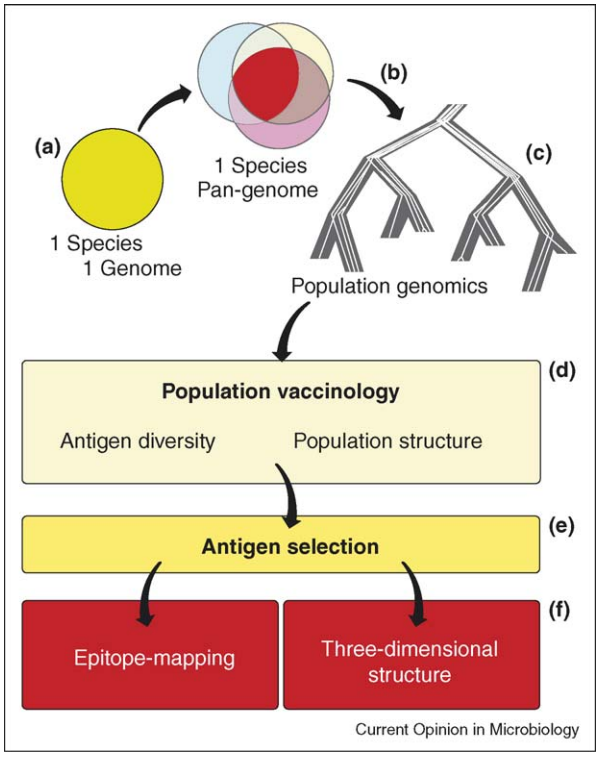
The recent application of genome-wide single nucleotide polymorphism analysis to *Mycobacterium tuberculosis* [21] provides a significant contribution is this direction, where the final goal remains the ability to exhaustively sample the genomic content of microbial populations, from a 'genome-wide population vaccinology' perspective.

Conclusions

The availability of an increasing number of bacterial genomes has prompted the application of the reverse vaccinology approach to pathogens, for which classical vaccinology has failed. The major drawback of using the genome sequence of a single strain is that it does not offer a complete picture of the genetic diversity of a species. The attempt to develop a universal vaccine against GBS [6] has demonstrated that the sequences of multiple genomes from each species are needed to cover the diversity of many bacterial pathogens, opening the era of the pan-genomic reverse vaccinology.

The natural next step to achieve a more comprehensive and epidemiologically related picture of bacterial populations will be population vaccinology, leading to the formulation of vaccines from a collection of proteins that, together, protect against the major circulating populations of a pathogen. In addition, the sequencing of human and pathogen genomes has provided vast amounts of data relevant to the study of human immune responses and

Figure 1



Flow chart for antigen discovery and refinement of the search. Three major genome-based approaches are involved in the identification of new potential vaccine candidates: **(a)** analysis of a single genome sequences in order to select secreted or extra-cellular proteins to identify potential vaccine candidates, **(b)** comparison of multiple genomes of the same species to assess intra-species diversity, **(c)** population genomics to achieve a more comprehensive coverage against the major circulating species. These three steps lead to **(d)** population vaccinology, which takes into account antigen variability and population structure, **(e)** allowing a more rational design of a new generation of vaccine targets. **(f)** Further *in silico* screening, such as epitope-mapping and structure-oriented bioinformatics, will refine the search.

complex host–pathogen interactions. The improvement of immuno-informatic tools, such as T-cell and B-cell epitope-mapping algorithms, and of structure-oriented bioinformatics [22,23] will enable the refinement of the search for (Figure 1) and design of a totally synthetic vaccine containing strings of the best epitopes encoded by the microorganism.

Complex vaccines containing T-cell and B-cell epitopes alongside cytotoxic T-lymphocyte epitopes derived from a variety of pathogens have already been constructed and tested [24]. A more recent study has shown that entirely synthetic epitope-driven vaccines elicit strong T-cell and B-cell responses, reaching the minimal requirements for an efficient vaccine in a single molecule [25].

Indeed, the application of epitope-driven vaccination to *Streptococcus pyogenes* resulted in protective immunity against a lethal challenge in an experimental animal model [26].

Over the next few years, integration of all these strategies will permit a more rational design of a new generation of vaccine targets.

Acknowledgements

We thank Giorgio Corsi for artwork and to Catherine Mallia for manuscript editing.

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb JF, Dougherty BA, Merrick JM *et al.*: **Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd**. *Science* 1995, **269**:496-512.

2. Pizza M, Scarlato V, Masignani V, Giuliani MM, Arico B, Comanducci M, Jennings GT, Baldi L, Bartolini E, Capecchi B *et al.*: **Identification of vaccine candidates against serogroup B meningococcus by whole-genome sequencing**. *Science* 2000, **287**:1816-1820.

3. Wizemann TM, Heinrichs JH, Adamou JE, Erwin AL, Kunsch C, Choi GH, Barash SC, Rosen CA, Masure HR, Tuomanen E *et al.*: **Use of a whole genome approach to identify vaccine molecules affording protection against *Streptococcus pneumoniae* infection**. *Infect Immun* 2001, **69**:1593-1598.

4. Ross BC, Czajkowski L, Hocking D, Margetts M, Webb E, Rothel L, Patterson M, Agius C, Camuglia S, Reynolds E *et al.*: **Identification of vaccine candidate antigens from a genomic analysis of *Porphyromonas gingivalis***. *Vaccine* 2001, **19**:4135-4142.

5. Montigiani S, Falugi F, Scarselli M, Finco O, Petracca R, Galli G, Mariani M, Manetti R, Agnusdei M, Cevenini R *et al.*: **Genomic approach for analysis of surface proteins in *Chlamydia pneumoniae***. *Infect Immun* 2002, **70**:368-379.

6. Ariel N, Zvi A, Grosfeld H, Gat O, Inbar Y, Velan B, Cohen S, Shafferman A: **Search for potential vaccine candidate open reading frames in the *Bacillus anthracis* virulence plasmid pXO1: *in silico* and *in vitro* screening**. *Infect Immun* 2002, **70**:6817-6827.

7. Tettelin H, Masignani V, Cieslewicz MJ, Eisen JA, Peterson S, Wessels MR, Paulsen IT, Nelson KE, Margarit I, Read TD *et al.*: **Complete genome sequence and comparative genomic analysis of an emerging human pathogen, serotype V *Streptococcus agalactiae***. *Proc Natl Acad Sci USA* 2002, **99**:12391-12396.

8. Glaser P, Rusniok C, Buchrieser C, Chevalier F, Frangeul L, Msadek T, Zouine M, Couve E, Lalioui L, Poyart C *et al.*: **Genome sequence of *Streptococcus agalactiae*, a pathogen causing invasive neonatal disease**. *Mol Microbiol* 2002, **45**:1499-1513.

9. • Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, Ward NL, Angiuoli SV, Crabtree J, Jones AL, Durkin AS *et al.*: **Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome"**. *Proc Natl Acad Sci USA* 2005, **102**:13950-13955.

A study that examines the diversity in eight isolates of GBS, suggesting that the number of genes associated with this species might be theoretically unlimited.

10. • Maione D, Margarit I, Rinaudo CD, Masignani V, Mora M, Scarselli M, Tettelin H, Brettoni C, Iacobini ET, Rosini R *et al.*: **Identification of a universal Group B streptococcus vaccine by multiple genome screen**. *Science* 2005, **309**:148-150.

The authors use a powerful strategy for identifying potential vaccine candidates against highly variable pathogens by the use of multistrain genomes.

11. •• Thompson JR, Pacocha S, Pharino C, Klepac-Ceraj V, Hunt DE, Benoit J, Sarma-Rupavtarm R, Distel DL, Polz MF: **Genotypic diversity within a natural coastal bacterioplankton population**. *Science* 2005, **307**:1311-1313.

The authors demonstrate the genetic variability within single naturally occurring, apparently homogenous populations of bacteria.

12. Welch RA, Burland V, Plunkett G III, Redford P, Roesch P, Rasko D, Buckles EL, Liou SR, Boutin A, Hackett J *et al.*: **Extensive mosaic structure revealed by the complete genome sequence of uropathogenic *Escherichia coli***. *Proc Natl Acad Sci USA* 2002, **99**:17020-17024.

13. Orskov F, Orskov I: **From the national institutes of health. Summary of a workshop on the clone concept in the epidemiology, taxonomy, and evolution of the *Enterobacteriaceae* and other bacteria**. *J Infect Dis* 1983, **148**:346-357.

14. Caugant DA, Levin BR, Selander RK: **Genetic diversity and temporal variation in the *E. coli* population of a human host**. *Genetics* 1981, **98**:467-490.

15. Whittam TS, Ochman H, Selander RK: **Multilocus genetic structure in natural populations of *Escherichia coli***. *Proc Natl Acad Sci USA* 1983, **80**:1751-1755.

16. Smith JM, Smith NH, O'Rourke M, Spratt BG: **How clonal are bacteria?** *Proc Natl Acad Sci USA* 1993, **90**:4384-4388.

17. Maiden MC, Bygraves JA, Feil E, Morelli G, Russell JE, Urwin R, Zhang Q, Zhou J, Zurth K, Caugant DA *et al.*: **Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms**. *Proc Natl Acad Sci USA* 1998, **95**:3140-3145.

18. Feil EJ, Maiden MC, Achtman M, Spratt BG: **The relative contributions of recombination and mutation to the divergence of clones of *Neisseria meningitides***. *Mol Biol Evol* 1999, **16**:1496-1502.

19. Feil EJ, Holmes EC, Bessen DE, Chan MS, Day NP, Enright MC, Goldstein R, Hood DW, Kalia A, Moore CE *et al.*: **Recombination within natural populations of pathogenic bacteria: short-term empirical estimates and long-term phylogenetic consequences**. *Proc Natl Acad Sci USA* 2001, **98**:182-187.

20. Urwin R, Russell JE, Thompson EA, Holmes EC, Feavers IM, Maiden MC: **Distribution of surface protein variants among hyperinvasive meningococci: implications for vaccine design**. *Infect Immun* 2004, **72**:5955-5962.

21. Filliol I, Motiwala AS, Cavatore M, Qi W, Hazbon MH, Bobadilla del Valle M, Fyfe J, Garcia-Garcia L, Rastogi N, Sola C *et al.*: **Global phylogeny of *Mycobacterium tuberculosis* based on single nucleotide polymorphism (SNP) analysis: insights into tuberculosis evolution, phylogenetic accuracy of other DNA**

fingerprinting systems, and recommendations for a minimal standard SNP set. *J Bacteriol* 2006, **188**:759-772.

22. De Groot AS: **Immunomics: discovering new targets for vaccines and therapeutics**. *Drug Discov Today* 2006, **11**:203-209.

23. Arcus VL, Lott JS, Johnston JM, Baker EN: **The potential impact of structural genomics on tuberculosis drug discovery**. *Drug Discov Today* 2006, **11**:28-34.

24. Falugi F, Petracca R, Mariani M, Luzzi E, Mancianti S, Carinci V, Melli ML, Finco O, Wack A, Di Tommaso A *et al.*: **Rationally designed strings of promiscuous CD4(+) T cell epitopes provide help to *Haemophilus influenzae* type b oligosaccharide: a model for new conjugate vaccines**. *Eur J Immunol* 2001, **31**:3816-3824.

25. Jackson DC, Lau YF, Le T, Suhrbier A, Deliyannis G, Cheers C, Smith C, Zeng W, Brown LE: **A totally synthetic vaccine of generic structure that targets Toll-like receptor 2 on dendritic cells and promotes antibody or cytotoxic T cell responses**. *Proc Natl Acad Sci USA* 2004, **101**:15440-15445.

26. Schulze K, Olive C, Ebensen T, Guzman CA: **Intranasal vaccination with SfbI or M protein-derived peptides conjugated to diphtheria toxoid confers protective immunity against a lethal challenge with Streptococcus pyogenes**. *Vaccine* 2006, DOI: 10.1016/j.vaccine.2006.05.024.