Functional genomics of pathogenic bacteria

E. R. Moxon^{1*}, D. W. Hood¹, N. J. Saunders², E. K. H. Schweda³
and J. C. Richards⁴

¹University of Oxford, Molecular Infectious Diseases Group and Department of Paediatrics, Institute for Molecular Medicine, John Radcliffe Hospital, Oxford, UK

²University of Oxford, Sir William Dunn School of Pathology, South Parks Road, Oxford OX1 3RE, UK

³Karolinska Institute, Clinical Research Centre, Novum, S-141 86 Huddinge, Sweden

⁴Immunochemistry Program, National Research Council, Canada Institute for Biological Sciences, 100 Sussex Drive, Ottawa, Ontario, Canada K1A 0R6

Microbial diseases remain the commonest cause of global mortality and morbidity. Automated-DNA sequencing has revolutionized the investigation of pathogenic microbes by making the immense fund of information contained in their genomes available at reasonable cost. The challenge is how this information can be used to increase current understanding of the biology of commensal and virulence behaviour of pathogens with particular emphasis on *in vivo* function and novel approaches to prevention. One example of the application of whole-genome-sequence information is afforded by investigations of the pathogenic role of *Haemophilus influenzae* lipopolysaccharide and its candidacy as a vaccine.

Keywords: *Haemophilus influenzae*; functional genomics; lipopolysaccharide; bacterial virulence; vaccine development

1. INTRODUCTION

Microbial diseases remain the commonest cause of global mortality and morbidity. Of the estimated millions of deaths from microbial infections per annum, more than half are the result of bacterial infections including tuberculosis, pneumococcal infections (and other bacterial pneumonias), pertussis, bacterial diarrhoea, neonatal tetanus and meningitis. In the more affluent, socio-economically endowed countries, the famous pronouncement by the US Surgeon General of the 1960s that infections could be consigned to the history books, has proved to be a false dawn. In particular, one of the most serious issues is the escalating prevalence of resistance to antimicrobials. Overall, the lack of adequate control of bacterial and other microbial infections continues to be a major global public health challenge and priority. What has changed radically is the means to meet this challenge through a range of scientific advances, especially those in molecular and cellular biology. Beginning in the 1970s, recombinant-DNA techniques provided incisive reductionist methods for analysing the genetic basis of virulence, a molecular version of Koch's postulates (Falkow 1988). In the 1980s, host-microbial interactions were increasingly studied through sophisticated experimental approaches that recruited the repertoire of new knowledge provided by cell biology; a body of science now formally recognized as cellular microbiology. In the 1990s automated-DNA sequencing brought about a revolution in microbiology through the provision of complete genome sequences of bacteria, fungi and parasites. The transition into a new

millennium coincided with the post-genomic age of functional biology. One particular challenge is how to exploit the raw data of whole-genome sequences to understand the *in vivo* behaviour of pathogenic microbes (Moxon & Tang 2000). In this paper, the aim is to provide a selective review of how whole-genome sequences can be used to advance current knowledge of bacterial pathogens. This review addresses some general points and then focuses specifically on selected aspects of our research on the commensal and pathogenic bacterium *Hi*.

2. WHOLE-GENOME ANALYSIS OF BACTERIA

Automated-DNA sequencing has revolutionized biology by making available the immense fund of information contained in the genomes of organisms. For an investment of less than £1 million, a typical bacterial genome sequence identifies several hundred predicted genes. Given the long-standing commitment to the human genome project, it is perhaps surprising in retrospect that it took so long to apply the rich potential of the availability of high-throughput sequencing technology to elucidate the whole-genome sequences of major pathogens. The turning point was the completion of the *Hi* (Rd strain) bacterial genome, 1.8 Mb of double standard, circular DNA, by TIGR (Fleischmann *et al.* 1995). This achievement took the scientific world by surprise and elicited great excitement as its potential contributions to public health and wealth were grasped. At the time of its completion (early in 1995) many had not known that the sequencing of the *Hi* genome (Rd strain) was in progress, and even microbiologists had underestimated the impact of this milestone. In subsequent years, the complete genome sequences of about

*Author for correspondence (richard.moxon@paediatrics.oxford.ac.uk).

40 bacteria have become publicly available, including more than a dozen bacterial pathogens.

One of the interesting aspects of the sequencing of the *Hi* genome was the innovative use of the shotgun sequencing strategy. In the accepted, classical approach, genome projects depended upon the prior construction of a physical map. Overlapping cloned fragments were then sequenced. Venter *et al.* (1996) argued that this assembly first approach was not necessary and that sequencing a library of randomly sheared fragments (shotgun approach), coupled with sophisticated computer programs to carry out the assembly might be preferable. This approach had originally been taken by Sanger to complete the sequence of bacteriophage, almost 50 kb in size. However, the use of this strategy to tackle a genome of almost two million base pairs was a bold step that posed a huge challenge in bio-informatics to achieve random-sequence assembly.

Hi was the first free-living organism to be completely sequenced and thus has provided a unique opportunity to probe the absolute requirements for the independent life of a cell. The utility of a complete microbial genome sequence is that it identifies the number of predicted genes as well as non-coding sequences, provides direct information on genome organization, and opens the door to studies of gene expression (stable proteins and RNA), evolution and population biology. With respect to pathogens, genomes provide a comprehensive inventory of virulence factors as well as candidate vaccines and diagnostics.

3. WHAT IS IN A GENOME?

Homology searches and annotation are the first analyses performed on a genome sequence. The *Hi* genome was estimated to contain 1743 predicted genes (ORFs). For these, homologous sequence matches in the database could be found for more than 90% and allowed some indication of function in more than 80%, although this is an overestimate of validated biological functions. A more realistic conclusion is that only 50–70% of these genes could be reasonably assigned a function.

An important point is that annotation based on homology data, even when the analysis is of the highest possible quality, will often include incomplete or erroneous indications of gene function. It may only take a few changes in an enzyme's active site to alter its substrate specificity and in the absence of experimental evidence, it is not possible to conclude that the best match for a gene, however good it is, represents a true orthologue. Concluding that a metabolic function is absent is even more problematic since it can be carried out by proteins that are completely unrelated to known enzymes or by molecules that are so divergent that they cannot be considered as homologues. Several examples illustrate the problem of sequence divergence where comparative analysis has suggested that orthologues can have a sequence identity as low as 18%. Furthermore, gene functions can be complemented between species by proteins with amino acid identities as low as 23%. Except for the limitations of homology-based annotation, this approach provides valuable information that reveals much about the biology of the organisms, and it is the starting point from which most investigators approach the sequence data. For example, it is apparent

that those organisms that have evolved to a host-adapted and restricted lifestyle have lost many metabolic functions compared to those present in organisms capable of independent growth in the environment, such as *Escherichia coli*. A good example of this is the three missing enzymes of the TCA cycle in *Hi* (Tatusov *et al.* 1996).

A valuable new method of searching for homologues has been developed. Tatusov and co-workers identified COGs between genomes (Tatusov *et al.* 1997). This method uses the identification of the most similar ORFs between groups of genomes to identify orthologues, rather than seeking strong homology between each individual ORF and the sequences deposited in the available databases. The value of the COG analysis increases with the number of available genome sequences, raising the possibility of not only automated annotation, but also recognition of gene families that are restricted to groups of bacteria whose shared biological properties provide insights into specialization. The analysis of the biosynthetic pathway of LPS in *Hi* (Hood *et al.* 1996a) is an example of the power of a homology-search-led analysis that made full use of the complete nature of the genome sequence.

A sequenced bacterial genome represents only the genetic complement of a single strain. The extent to which it is representative of the larger population, both in terms of the genes that are variable between different strains and those that are present in the sequence strain but absent in others and vice versa, is not known. Conventional taxonomy and microbiology have traditionally relied upon phenotypic characterization. Despite its utility, phenotypic characterization cannot capture and may even obscure the appreciation of genetic relationships and population structure. More than a decade ago, multilocus enzyme electrophoresis was recruited to investigate population structures of bacteria, focusing particularly on bacterial pathogens, by indexing variation based on the accumulated polymorphisms within neutral genes, such as essential metabolic enzymes (Selander *et al.* 1987). It allows large numbers of organisms to be characterized and, using appropriate statistical analyses, strong inferences can be drawn concerning the extent of horizontal gene transfer and inter-genomic recombination among strains of the species. With the advent of automated sequencing, multilocus enzyme electrophoresis has given way to MLST (Maiden *et al.* 1998). The phylogenetic trees derived from MLST provide a robust framework for analysing diversity and perhaps the evolutionary origins of an organism's genomic content. In a few instances, these data have been supplemented by more detailed analysis of the inter-strain variation in nucleotide sequence data of selected genes such as those involved in virulence. The sequenced strain may or may not be pathogenic, but even if it is, it will still not include all of the coding potential present in the larger population of disease-associated organisms. Thus, comparative analysis of genomes is important and will be an extremely valuable technique for comparing related organisms that cause different patterns of disease, especially where there are multiple pathogenic strains of the same bacterial species. An alternative approach to address this issue is to sequence multiple strains using only two to three times coverage as opposed to the sixfold or greater coverage needed to achieve assembly.

An organism's ability to adapt to changing environ-

ments to which it is exposed during its normal life cycle, including its interactions with natural and opportunistic hosts, is an important characteristic of pathogenic bacteria. Some of the genes involved in this process exhibit phase and/or antigenic variation. An important genetic mechanism underlying this variation is the presence in the genome of runs of tandemly repeated DNA. These repeats range from single nucleotides (homopolymeric tracts) to motifs of up to eight bases in size. These genes have been called 'contingency' genes to emphasize their potential role in facilitating the adaptation of a bacterial population to a variety of environments (Moxon *et al.* 1994). The key point is that the presence of these repeats is immediately revealed by complete genome sequences and provides a useful indication of genes whose functions will probably be relevant to pathogenesis and host adaptation (Hood *et al.* 1996b; Saunders *et al.* 1998, 2000).

A remarkably constant proportion (*ca.* 40%) of ORFs in each newly sequenced genome do not have identifiable homologues, at least based on conventional search methods. *Hi* is no exception, and these novel genes of unknown function represent an exciting challenge and an example of the richness of the treasure trove revealed by complete genome sequences. Many of these genes will probably have functions related to species specialization. Non-coding DNA in the *Hi* genome has also proved particularly interesting. There were 1400 copies of motifs (23 nucleotides) involved in DNA transformation. It is proposed that these sequences provide a mechanism through which *Hi* can recognize its own DNA, released following spontaneous lysis of sibling or closely related bacterial cells, and then incorporate it into the genome of recipient cells by homologous recombination (Smith *et al.* 1995).

4. GENE EXPRESSION DURING INFECTION

A variety of methods for the investigation of the role of genes during infection have been developed. Signature tagged mutagenesis in particular has been championed as a method for investigation of virulence genes (Hensel *et al.* 1995). This methodology identifies genes whose absence results in a reduced fitness and includes metabolic functions that cannot really be considered to be directly associated with the pathogenic potential of the organism as it is conventionally understood. It would be more correct to say that they are fitness genes, some of which are directly responsible for host interactions, including virulence. The utility of this and similar methods is greatly affected by the organism and the model systems that are used and the difference between *in vitro* and *in vivo* growth conditions. This method has been recently used to identify novel genes of *Neisseria meningitidis* (Sun *et al.* 2000).

Another approach to global analysis is to look at gene expression. This is an area where genome-sequence information makes a much more direct contribution to this type of study. This technology has been used to determine alterations in gene expression in yeast and bacteria, to detect polymorphisms between strains. Thus, genome sequences open the door to an understanding of the integration and coordination of genotype and phenotype under different environmental conditions. For example, the changes in expression associated with bacterial entry into host cells from an extracellular residence, or differ-

ences in the patterns of gene activity when the bacterium is in an insect vector compared to its mammalian host. Such analysis can be undertaken using a library of genes, representative of the whole genome. Each gene can be arrayed on a solid phase and probed with labelled RNA extracted from bacteria grown under different conditions (DeRisi *et al.* 1997). Translation products of expressed genes can be investigated by combining two-dimensional protein electrophoresis with mass spectroscopy. The resolving power of the latter, requiring only minute quantities of the relevant peptide, allows the detection of post-translational modifications. Although encouraging results have been obtained, there are currently many technical unresolved imperfections with nucleic acid arrays and proteomics, but these will undoubtedly be overcome. As technology becomes more reliable, economical and widely available, it will probably have a large impact in facilitating the study of microbial pathogens.

Whole-genome sequences provide another level of resolution, for virulence is multi-factorial and must be considered in the context of the products of many genes including the molecules that determine tropism, invasion, acquisition of nutrients *in vivo* replication and cytotoxic factors (Moxon & Tang 2000). As a specific example of how genomic sequences facilitate the analysis of virulence and how these data inform approaches to infection control, we will now describe in detail some of our investigations into *Hi* LPS.

5. APPLICATION OF COMPLETE BACTERIAL GENOME-SEQUENCE DATA TO THE BIOLOGICAL BASIS OF *HAEMOPHILUS INFLUENZAE* PATHOGENICITY AND VACCINE DEVELOPMENT

Haemophilus influenzae is a gram-negative bacterium that routinely colonizes the upper respiratory tract of humans, its only known host. It is potentially pathogenic, encapsulated type-b strains having the potential to cause serious invasive (bacteraemic) infections, such as meningitis. Unencapsulated or NTHi, strains, are less prone to cause invasive, bacteraemic systemic infections but are a common cause of otitis media, sinusitis and lower respiratory infections, including life-threatening pneumonia. LPS is a critical component of the outer membrane of *Hi* that influences the permeability of the cell and has an important role in both commensal and virulence behaviour (Moxon & Maskell 1992). LPS is involved in each stage of the pathogenesis of infections caused by this bacterium: colonization of the upper respiratory tract, systemic dissemination and the cytotoxic injury to target tissues, such as the central nervous system. LPS is inserted into the outer membrane through a membrane-anchor, lipid A. This component of the LPS molecule, designated endotoxin, is responsible for much of the pathological damage associated with disease (Beutler 2000). The fatty acyl groups and 4'-phosphate-containing substituents of lipid A are required for mediating the 'pyrogenic' effects that cause inflammation and the resulting pathological damage characteristic of *Hi* disease. The lipid A of *Hi* LPS exhibits heat-stable endotoxic properties. Intracisternal instillation of purified LPS from *Hib* in rabbits rapidly induces an inflammatory reaction in the CSF, evidenced

by inflammatory cells, increased serum proteins and a meningeal exudate (Syrogiannopoulos *et al.* 1988). Macrophages respond to LPS by producing cytokines and other host defence molecules and can aid the recruitment of PMNs to the pulmonary spaces. LPS binds to LBP protein and then principally to the receptor CD14 on the macrophage surface, triggering the inflammatory response and cytokine cascade essential to keep infection under control. More recently, a role for the family of Toll receptors has been proposed in signal transduction and the induction of responses to LPS. The role of LPS in the virulence of *Hi* has mainly been studied through experimental infection in infant rats. A deep rough mutant, I69, of a type-b strain, was avirulent in the infant rat model of infection (Zwahlen *et al.* 1985). Lipid A is covalently linked to an oligosaccharide core, the part of the LPS molecule that interacts specifically with molecules of host body fluids or host tissues. The core saccharides lack the polymerized side chains (O-antigens) characteristic of many other gram-negative bacteria, such as the enterobacteraceae. The molecule is nonetheless exceedingly complex as, in addition to the several sugars (e.g. keto-octulosonic acid, heptose, glucose and galactose), there are a variety of variably expressed substituents such as PO₄, PEA-PPEA (Masoud *et al.* 1997), O-acetyl groups, ChoP groups (Schweda *et al.* 2000) or NANA (Hood *et al.* 1999), that are characteristic of the mature glycoforms of the molecule. *Hi* has also evolved mechanisms for promoting antigenic variation of its LPS through genetic switching (phase variation) of LPS biosynthetic genes (Weiser *et al.* 1989).

To identify LPS-related genes from the *Hi* Rd strain genome sequence, the sequences of a wide range of LPS biosynthetic genes from gram-negative bacteria, deposited in GenBank-EMBL data banks, were used as 'database probes' to screen for homologous genes, by DNA or amino acid similarity (Hood *et al.* 1996a). Searches with amino acid sequences against the translated Rd-strain genome sequence proved the most useful for identification of putative LPS biosynthetic genes of *Hi*. DNA sequences identified from the Rd-strain genome sequence as potential homologues of LPS biosynthetic genes were translated then searched against the combined GenBank-EMBL data banks to confirm probable function(s). Due to a tendency in bacteria for clustering of genes with related functions, stretches of DNA sequence upstream and downstream of confirmed loci of interest were also isolated and treated in a similar fashion. The putative LPS associated genes that were identified from the Rd-strain genome sequence are shown in figure 1.

Our directed search for candidate LPS genes allowed us to identify sequences listed either as genes of unknown function or as incomplete ORFs, in the published version of the Rd-strain database. Several identified sequences were previously characterized as *Hi* LPS genes, indicating the validity of the approach. Following identification of candidate LPS genes, the relevant sequences were amplified by the PCR from *Hi* type-d strain RM118 (Rd); the strain used for the sequencing project. In addition, the equivalent DNA tracts from two epidemiologically distinct encapsulated type-b strains, RM153 (Eagan) and RM7004 used in our previous investigations of LPS biosynthesis, were amplified. In all cases, the oligonucleotide

Figure 1. Genes for LPS biosynthesis in the *Hi* Rd-strain are shown in their relative positions on a circular 1.83 Mbp map. The scattered distribution of the genes for the biosynthesis of LPS is in contrast to the clustering of genes for the biosynthesis of core and O-antigen chains of the enterobacteraceae LPS.

primers designed from the Rd-strain genome sequence allowed us to amplify the relevant loci from strain RM118 chromosomal DNA. Similarly, the primers enabled us to amplify most of the corresponding loci from the type-b strains. The failure rate (16%) reflected the divergence of the type-b strains in the DNA sequences selected for the primers, and possibly also differences in the genetic organization between the strains. For some loci the primers were designed to hybridize against sequences from adjacent genes or intergenic regions.

PCR-amplified products were cloned, mutated by insertional inactivation, then transformed into the appropriate *Hi* strains to obtain mutants by allelic replacement. LPS molecules were then analysed by colony immunoblotting, using a panel of mAbs, each specific for particular LPS epitopes. Differences in the patterns of mAb reactivity between wild-type and mutant strains provided strong indications of the role of mutated genes in LPS biosynthesis, findings which were then confirmed by fine structural analysis of the relevant LPS using NMR and electrospray MS. Some key LPS biosynthetic genes were not detected so easily. For example, our initial search of the Rd-strain genome sequence failed to identify a candidate transferase required for the addition of the third heptose to the growing LPS chain. The DNA sequence database was searched for less conserved homologues of known heptosyl-transferases known to add sugars to the second heptose of LPS in other organisms such as *rfaK* from *Neisseria* and *rfaG* from *E. coli*. This showed an *Hi* sequence with a very low match to *orfH*, an ORF that had been initially annotated as a gene of unknown function. Mutation of this homologue of *orfH* resulted in a strain with LPS of a size corresponding to that predicted for a molecule containing two heptose sugars. Further evidence

to support the role of this gene as a heptosyl transferase was obtained by fine structural analysis of the *orfH* mutant. Based on these mutants, a relatively complete assignment of the relevant transferases required for assembly of the inner and outer core sugars of the Rd strain was compiled. The location of the genes for LPS biosynthesis in *Hi* compared to the enterobacteraceae is rather striking. In the latter, the genes for LPS biosynthesis are located together on the chromosome as large regions whereas the LPS biosynthetic genes of *Hi* are located at multiple loci scattered around the chromosome.

The range of LPS mutants constructed in this study has allowed us for the first time to undertake a comprehensive study to correlate LPS structure with the virulence of *Hi* strains in the infant rat model of infection (Smith *et al.* 1973*a,b*). Strains that were attenuated and resulted in a level of bacteraemia greater than 2 log below that of the wild-type were those mutated in genes that produced very truncated LPS molecules. Mutant strains with more subtle changes to the LPS gel migration profile, namely those mutated in *lgtC*, *lpsA* and *rfbP*, gave rise to levels of bacteraemia 5–100fold less than that of the wild-type, a statistically significant difference in each case. Thus, there are some broad correlations between LPS structure and virulence. However, some puzzles remain and examples of genes that are exceptions to the rule include the mutations in *rfbB* and *orfZ*. Although there was only an apparently minor change to the LPS profile of these mutants when compared to the wild-type, there was a decrease by 1000-fold in the level of bacteraemia. Mutant *rfbB* has been constructed a second time to ensure that the observed attenuation was not related to a fortuitous secondary mutation, and similar results were obtained. No gross change in other virulence determinants such as capsule or outer membrane proteins has yet been found for any of the mutants, but minor changes have not yet been ruled out for all mutations in all strain backgrounds.

In summary, the complete genome sequence of the *Hi* Rd strain has allowed us to identify 25 candidate LPS genes. Detection of homologues of LPS biosynthetic genes in the *Hi* genome database using DNA sequences as probes was much more rapid and reliable than could have been achieved by hybridization experiments. However, the major advantage of the genome-sequence information was to identify candidate loci, found only, or most reliably, by amino acid homology. This is particularly important when studying biological systems such as LPS where it is not uncommon for proteins of related function, such as the sugar transferases, to be encoded by genes of divergent sequence. About 60% of the genes could not have been reliably identified by DNA sequence alone and therefore would not have been found by hybridization experiments using the relevant heterologous probe.

As indicated earlier, a major interest of our laboratory is in the phase variation of LPS structures exerted through nucleotide repeats situated within biosynthetic genes. These repeat regions, normally of tetranucleotides, are subject to loss or gain of one or more four base repeats through slipped strand mispairing during DNA replication, and result in translational frame shifts and non-functional gene products. This reversible on–off switching of LPS biosynthetic functions at high frequencies underlies the phase-variable phenotype characteristic of *Hi* LPS.

It is now becoming clear that these structural changes in LPS associated with phase variation have functional implications in pathogenesis. As examples, some details of the biological effects of *lic1*, *lic3* and *lgtC* will be reviewed.

Structural studies of *lic1* phase variants have indicated that they differ in the presence or absence of ChoP present on the cell surface-exposed outer core of the LPS. Weiser and colleagues have investigated these ChoP⁺ and ChoP[−] variants in an infant rat model and have found that, following intranasal inoculation, there was selection for ChoP⁺ variants. In this case the ChoP⁺ mutant showed a reduced incidence and magnitude of bacteraemia following intranasal inoculation and colonization. Incorporation of ChoP into *Hi* LPS increases the bacterium's susceptibility to the bactericidal activity of human serum owing to its interaction with CRP and activation of complement through a classical pathway. In summary, during nasopharyngeal carriage in infant rats there was a gradual selection for the ChoP⁺ phenotype indicating that this variant is optimal for organisms to persist on the nasopharynx mucosal surface but that a ChoP[−] phenotype may contribute to increased invasion and serum resistance. Independent studies (Tong *et al.* 2000) have confirmed the importance of phase variation of ChoP in the pathogenesis of otitis media in a chinchilla model. Consistent with these findings in experimental models studies of *Hi* from human respiratory secretions indicated that the genotype of these organisms was also predominantly ChoP⁺ (Weiser *et al.* 1998). Further studies have indicated that the greater fitness of the ChoP⁺ variants may reflect an increased resistance to innate immune clearance mechanisms (Lysenko *et al.* 2000). A further gene contributing to phase variation of LPS, *lgtC*, was identified from the Rd-strain genome sequence (Hood *et al.* 1996*b*). The loss of reactivity to a mAb 4C4, specific for a digalactoside structure indicated that LgtC influences the same region of LPS core structure as does a product of the *lic2* locus which includes another phase variable gene, *lic2A* (High *et al.* 1996). mAb 4C4 recognizes the gal α (1–4) β gal epitope that is expressed in the LPS of related bacteria and also on human cells (Virji *et al.* 1990). This epitope is expressed as the terminal extension from heptose II in the LPS from strain RM153 and loss of mAb 4C4 reactivity is consistent with the truncation of the LPS structure found by structural studies of the LPS from an RM153/*lgtC* mutant strain. A contribution of LgtC to the phase-variable expression of the digalactoside epitope has been shown (Hood *et al.* 1996*b*). In contrast to the epitope synthesized by genes in the *lic1* locus which diminishes clearance through interference with innate immunity, the importance of the gal α (1–4) β gal structure and its capacity for molecular mimicry appears to influence antibody-mediated clearance through acquired immunity (Weiser & Pan 1998). Interestingly, the *Hi lgtC* gene is so designated owing to its homology to the gene in *Neisseria* (Gotschlich 1994). The *lgtC* gene in *Neisseria* has a homopolymeric tract of guanosine nucleotides located within the reading frame that appears to be responsible for the phase-variable expression of a terminal digalactoside epitope of the meningococcal LPS (Jennings *et al.* 1995). Knowledge of the complete repertoire of repeated motifs in the Rd strain (Hood *et al.* 1996*b*) allowed us to propose that the CAAT and GACA associated loci are the only

structural genes contributing directly to a phase-variable LPS phenotype in that strain. The LPS biosynthetic locus (*lex2*) associated with a GCAA repeat motif in type-b strains (Jarosik & Hansen 1994) was not present in the Rd-strain genome sequence.

A further gene *lic3A* encodes a sialyl transferase and the associated 5'-CAAT-3' tetranucleotide repeats within this gene result in the phase variation of NANA (Hood *et al.* 2001). Interestingly, *lic3A* is located immediately upstream of *galE*, a UDP-galactose-4-epimerase. Mutation of the *galE* gene perturbs the balance of activated galactose for incorporation into LPS during synthesis and a *galEgalK* mutant of strain RM7004 elaborates LPS with no galactose (Maskell *et al.* 1991). The addition of NANA to terminal sugar residues has been described and was found to be a feature in the majority of, if not all, *Hi* strains. A change in the LPS profiles upon Nm treatment was observed for 7 out of 24 (29%) of the *Hi* typeable strains and 22 out of 31 (71%) of the NTHi strains tested (Hood *et al.* 1999). Some correlation between the serotype and the occurrence of the sialylated LPS phenotype was observed; a proportion of type a, c and d had NANA in their LPS, although the degree of sialylation varied. When NANA was added as a growth supplement to the medium it could enhance the degree of sialylation. The gene encoding the CMP-NANA synthetase (*siaB*), which catalyses the formation of the nucleotide sugar donor used by sialyl-transferases, is present in the Rd genome and when this gene is inactivated in *Hi* strains, the highest molecular weight sialylated LPS glycoforms are not detected. Structural analysis of the LPS from a NTHi strain, 375, has shown NANA is linked to a β -galactose of a (hex)3 glycoform. Structural analysis of the *Hib* strain A2 LPS has also confirmed the presence, but not detailed the linkage, of NANA to the rest of the molecule (Gibson *et al.* 1996). A minor component of NTHi strain 375 LPS contains a di-NANA moiety, a finding that suggests that *Hi* may have the potential to add multiple charged molecules to LPS thus modulating significantly the properties of the bacterial cell surface.

6. CANDIDACY OF INNER-CORE *Hi* LPS AS A VACCINE

Molecular modelling has been used to generate a three-dimensional model and to calculate conformational flexibility of the core oligosaccharide region of LPS from the RM153 strain (Hood *et al.* 1996a) (see figure 2). The LPS consists of a conserved heptose containing inner-core trisaccharide unit attached via a phosphorylated KDO residue to a lipid A component. In the major LPS population group, each heptose is substituted with further chain elongation from the central unit where molecules containing four, five and six residues are most prevalent. The inner-core element is also a common feature of LPS from the RM7004 strain. As noted for the LPS of *E. coli* (Kastowsky *et al.* 1992), calculations indicated that the heptose containing inner-core region forms a compact structural unit. The galabiose-containing side chain attached to heptose II adopts an elongated shape and shows considerable conformational flexibility. It was found that even though these outer-core regions of the molecule are flexible, the heptose III and the β -glucose in

Figure 2. Space-filling model of LPS from *Hi* Rd-strain, generated by MMC calculations. At the bottom of the model, the lipid A moiety (black) is substituted by KDO depicted in grey. KDO is in turn substituted by pyroPEA (mauve). The heptose-containing trisaccharide (red) is attached to KDO. HeptoseI is substituted by glucose (obscured in this representation) which is substituted by ChoP (ochre). The main oligosaccharide chain is attached to heptose III comprising glucose (green), galabiose (blue) and a terminal *N*-acetyl galactosamine (slate-grey).

the side chain were facing on the same side of the molecule. This modelling is being extended to investigate the conformation of the LPS from the mutants derived from the RM153, RM7004 and RM118 strains. Most of the strains found in the respiratory tract lack capsules and are therefore not a target for immune responses stimulated by the *Hib* conjugate vaccines.

A vaccine to control other *Hi* infections, responsible each year for millions of cases of otitis media and many hundreds of thousands of childhood deaths from acute lower respiratory tract infections around the world is now a public health priority. The comparatively conserved inner-core LPS of *Hi* can still fulfil many of the criteria for a good vaccine candidate (Moxon *et al.* 1998).

The surface-exposed oligosaccharides are accessible to host clearance mechanisms and antibodies to LPS are detectable in the acute phase sera of children with *Hi* infection. Even if native LPS structures are poorly immunogenic, this deficiency can often be enhanced if the specific carbohydrate antigens are conjugated to proteins. Structural analyses to date have confirmed that key inner-core epitopes are conserved across the range of *Hi* strains tested and may allow a rational approach to vaccine development, initiated through whole-genome sequencing and supported through the systematic application of classical genetic, structural and immunobiological analyses.

The authors acknowledge J. R. Brisson (Department of

Biology, National Research Council, Canada Institute for Biological Sciences, Ottawa, Ontario, Canada).

REFERENCES

- Beutler, B. 2000 Endotoxin, Toll-like receptor 4, and the afferent limb of innate immunity. *Curr. Opin. Microbiol.* **3**, 23–28.
- DeRisi, J. L., Iyer, V. R. & Brown, P. O. 1997 Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* **278**, 680–686.
- Falkow, S. 1988 Molecular Koch's postulates applied to microbial pathogenicity. *Rev. Infect. Dis.* **10**, S274–S276.
- Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Krikness, E. F. & Kerlavage, A. R. 1995 Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* **269**, 496–512.
- Gibson, B. W., Phillips, N. J., Melaugh, W. & Engstrom, J. J. 1996 In *Biochemical and biotechnological applications of electrospray ionization mass spectrometry*, pp. 166–184. Washington, DC: American Chemical Society.
- Gotschlich, E. C. 1994 Genetic locus for the biosynthesis of the variable portion of *Neisseria gonorrhoeae* lipooligosaccharide. *J. Exp. Med.* **180**, 2181–2190.
- Hensel, M., Shea, J. E., Gleeson, C., Jones, M. D., Dalton, E. & Holden, D. W. 1995 Simultaneous identification of bacterial virulence genes by negative selection. *Science* **269**, 400–403.
- High, N. J., Jennings, M. P. & Moxon, E. R. 1996 Tandem repeats of the tetramer 5'-CAAT-3' present in *lic2A* are required for phase variation but not lipopolysaccharide biosynthesis in *Haemophilus influenzae*. *Mol. Microbiol.* **2000**, 165–174.
- Hood, D. W., Deadman, M. E., Allen, T., Masoud, H., Martin, A., Brisson, J. R., Fleischmann, R., Venter, J. C., Richards, J. C. & Moxon, E. R. 1996a Use of the complete genome sequence information of *Haemophilus influenzae* strain Rd to investigate lipopolysaccharide biosynthesis. *Mol. Microbiol.* **22**, 951–965.
- Hood, D. W., Deadman, M. E., Jennings, M. P., Bisercic, M., Fleischmann, R. D., Venter, J. C. & Moxon, E. R. 1996b DNA repeats identify novel virulence genes in *Haemophilus influenzae*. *Proc. Natl Acad. Sci. USA* **93**, 11 121–11 125.
- Hood, D. W., Makepeace, K., Deadman, M. E., Rest, R. F., Thibault, P., Martin, A., Richards, J. C. & Moxon, E. R. 1999 Sialic acid in the lipopolysaccharide of *Haemophilus influenzae*: strain distribution, influence on serum resistance and structural characterization. *Mol. Microbiol.* **33**, 679–692.
- Hood, D. W. (and 13 others) 2001 Identification of a lipopolysaccharide α -2,3-sialyltransferase from *Haemophilus influenzae*. *Mol. Microbiol.* **39**, 341–351.
- Jarosik, G. P. & Hansen, E. J. 1994 Identification of a new locus involved in expression of *Haemophilus influenzae* type b lipooligosaccharide. *Infect. Immun.* **62**, 4861–4867.
- Jennings, M. P., Hood, D. W., Peak, I. R. A., Virji, M. & Moxon, E. R. 1995 Molecular analysis of a locus for the biosynthesis and phase variable expression of the lacto-N-neotetraose terminal LPS structure in *Neisseria meningitidis*. *Mol. Microbiol.* **18**, 729–740.
- Kastowsky, M., Gutberlet, T. & Bradaczek, H. 1992 Molecular modelling of three-dimensional structure and conformation flexibility of bacterial lipopolysaccharide. *J. Bacteriol.* **174**, 4798–4806.
- Lysenko, E. S., Gould, J., Bals, R., Wilson, J. M. & Weiser, J. N. 2000 Bacterial phosphorylcholine decreases susceptibility of the antimicrobial peptide LL-37/hCAP18 expressed in the upper respiratory tract. *Infect. Immun.* **68**, 1664–1671.
- Maiden, M. C. (and 13 others) 1998 Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Proc. Natl Acad. Sci. USA* **95**, 3140–3145.
- Maskell, D. J., Szabo, M. J., Butler, P. D., Williams, A. E. & Moxon, E. R. 1991 Molecular analysis of a complex locus from *Haemophilus influenzae* involved in phase-variable lipopolysaccharide biosynthesis. *Mol. Microbiol.* **5**, 1013–1022.
- Masoud, H., Moxon, E. R., Martin, A., Krajcarski, D. & Richards, J. C. 1997 Structure of the variable and conserved lipopolysaccharide oligosaccharide epitopes expressed by *Haemophilus influenzae* serotype b strain Eagan. *Biochemistry* **36**, 2091–2103.
- Moxon, E. R. & Maskell, D. 1992 *Haemophilus influenzae* lipopolysaccharide: the biochemistry and biology of a virulence factor. In *Molecular biology of bacterial infection: current status and future perspectives* (ed. C. Hormaeche, C. W. Penn & C. J. Smyth), pp. 75–96. SGM Symposium no. 49. Reading: Society for General Microbiology.
- Moxon, E. R. & Tang, C. 2000 Challenge of investigating biologically relevant functions of virulence factors in bacterial pathogens. *Phil. Trans. R. Soc. Lond. B* **355**, 643–656.
- Moxon, E. R., Rainey, P. B., Nowak, M. A. & Lenski, R. E. 1994 Adaptive evolution of highly mutable loci in pathogenic bacteria. *Curr. Biol.* **4**, 24–33.
- Moxon, E. R., Hood, D. & Richards, J. C. 1998 Bacterial lipopolysaccharides: candidate vaccines to prevent *Neisseria meningitidis* and *Haemophilus influenzae* infections. In *Advances in experimental medicine and biology*. 435. *Glycoimmunology*, vol. 2 (ed. J. S. Axford), pp. 237–243. New York: Plenum.
- Saunders, N. J., Peden, J. F., Hood, D. W. & Moxon, E. R. 1998 Simple sequence repeats in the *Helicobacter pylori* genome. *Mol. Microbiol.* **27**, 1091–1098.
- Saunders, N. J., Jeffries, A. C., Peden, J. F., Hood, D. W., Tettelin, H., Rappuoli, R. & Moxon, E. R. 2000 Repeat-associated phase variable genes in the complete genome sequence of *Neisseria meningitidis* strain MC58. *Mol. Microbiol.* **37**, 207–215.
- Schweda, E. K. H., Brisson, J. R., Alvelius, G., Martin, A., Weiser, J. N., Hood, D. W., Moxon, E. R. & Richards, J. C. 2000 Characterization of the phosphocholine substituted oligosaccharide in lipopolysaccharides of type b *Haemophilus influenzae*. *Eur. J. Biochem.* **267**, 1–12.
- Selander, R. K., Musser, J. M., Caugant, D. A., Gilmour, M. N. & Witam, T. S. 1987 Population genetics of pathogenic bacteria. *Microbiol. Pathogenesis* **3**, 1–7.
- Smith, A. L., Averill, D. R., Moxon, E. R., Weller, P. F., Marino, J. & Smith, D. H. 1973a *Haemophilus influenzae* type b meningitis in infant rats. In *Haemophilus influenzae: proceedings of a conference on antigen-antibody systems, epidemiology and immunophylaxis* (ed. S. H. W. Sell & D. T. Karzon), pp. 203–212. Nashville, TN: Vanderbilt University Press.
- Smith, A. L., Smith, D. H., Averill, D. R., Marino, J. & Moxon, E. R. 1973b Production of *Haemophilus influenzae* b meningitis in infant rats by intraperitoneal inoculation. *Infect. Immun.* **8**, 278–290.
- Smith, H. O., Tomb, J. F., Dougherty, B. A., Fleischmann, R. D. & Venter, J. C. 1995 Frequency and distribution of DNA uptake signal sequences in the *Haemophilus influenzae* Rd genome. *Science* **269**, 538–540.
- Sun, Y. H., Bakshi, S., Chalmers, R. & Tang, C. M. 2000 Functional genomics of *Neisseria meningitidis* pathogenesis. *Nature Med.* **6**, 1269–1273.
- Syrogianopoulos, G. A., Hansen, E. J., Erwin, A. L., Hunford, R. S., Rutledge, J., Reisch, J. S. & McCracken Jr, G. H. 1988 *Haemophilus influenzae* b lipooligosaccharide induces meningeal inflammation. *J. Infect. Dis.* **157**, 157–244.
- Tatusov, R. L., Mushegian, A. R., Bork, P., Brown, N. P.,

- Hayes, W. S., Borodovsky, M., Rudd, K. E. & Koonin, E. V. 1996 Metabolism and evolution of *Haemophilus influenzae* deduced from a whole genome comparison with *Escherichia coli*. *Curr. Biol.* **6**, 279–291.
- Tatusov, R. L., Koonin, E. V. & Lipman, D. J. 1997 A genomic perspective on protein families. *Nature* **278**, 631–637.
- Tong, H. H., Blue, L. E., James, M. A., Chen, Y. P. & DeMaria, T. F. 2000 Evaluation of phase variation of non-typeable *Haemophilus influenzae* lipooligosaccharide during nasopharyngeal colonization and development of otitis media in the chinchilla model. *Infect. Immun.* 4593–4597.
- Venter, J. C., Smith, H. O. & Hood, L. 1996 A new strategy for genome sequencing. *Nature* **381**, 364–366.
- Virji, M., Weiser, J. N., Lindberg, A. A. & Moxon, E. R. 1990 Antigenic similarities in lipopolysaccharides of *Haemophilus* and *Neisseria* and expression of digalactoside structure also present on human cells. *Microbiol. Pathog.* **89**, 441–450.
- Weiser, J. N. & Pan, N. 1998 Adaptation of *Haemophilus influenzae* to acquired and innate humoral immunity based on phase variation of lipopolysaccharide. *Mol. Microbiol.* **30**, 747–775.
- Weiser, J. N., Lindberg, A. A., Manning, E. J., Hansen, E. J. & Moxon, E. R. 1989 Identification of a chromosomal locus for expression of lipopolysaccharide epitopes in *Haemophilus influenzae*. *Infect. Immun.* **57**, 3045–3052.
- Weiser, J. N., Pan, N., McGowan, K. L., Musher, D., Martin, A. & Richards, J. 1998 The position of phosphorylcholine on the lipopolysaccharide of *Haemophilus influenzae* contributes to persistence in the respiratory tract and sensitivity to serum killing mediated by C-reactive protein. *J. Exp. Med.* **187**, 631–640.
- Zwahlen, A., Rubin, L. G., Connelly, C. J., Inzana, T. J. & Moxon, E. R. 1985 Alteration of the cell wall of *Haemophilus influenzae* type b by transformation with cloned DNA: association with attenuated virulence. *J. Infect. Dis.* **152**, 485–492.