Leading article

From genes to genome biology: a new era in *Helicobacter pylori* research

In August 1997 the complete annotated genome sequence of *Helicobacter pylori* was published, just 15 years after the organism was first cultured. This is an important milestone in gastroenterology research as *H. pylori* is the first enteropathogen to be fully sequenced. The availability of a complete genetic data set heralds a new era in *H. pylori* research as it will provide a framework for global studies of virulence and other aspects of the organism’s biology. In this article we outline the genomes approaches that can now be applied and highlight the potential practical benefits of this research in terms of eradication therapy and disease prevention.

The *H. pylori* genome sequence: new opportunities

Scientists at The Institute for Genomic Research (TIGR) (http://www.tigr.org) have determined the order of the 1,667,867 nucleotides that constitute the circular chromosomal content of *H. pylori* strain ACTC 26695. Initial computer analysis suggests the presence of 1,590 open reading frames (ORFs or genes), of which nearly 70% can be matched to genes encoding proteins of known function. This immense achievement gives us unequalled opportunities to relate gene sequence to biology and will provide insight into many aspects of the organism’s biology, such as replication, DNA repair, metabolism, ion and protein transport, cell wall structure, and virulence. The most intriguing findings are as follows:

- **Almost one third of the predicted coding sequences are ORFans**—that is, they have no known homologues or any clues as to function. These ORFans may encode proteins unique to *H. pylori* and provide selective targets for antibiotic therapy.
- **Very few regulatory proteins were found**—by contrast to the *Escherichia coli* sequence, 10-fold less regulatory sequences have been identified. Tight regulation of gene expression is imperative for enteropathogenic bacteria whether they are continually responding to the harsh acidic environment of the stomach or bile salts in the intestine. The remarkable economy of regulatory elements in *H. pylori* may reflect the very limited range of environments in which it survives, suggesting a highly evolved inter-relationship between bacteria and microbe.
- **Outer membrane proteins (OMPs) and DNA repeats**—the OMPs identified provide a tractable subset of the total genome, comprising most of the proteins known to be involved in virulence (for example, those required for adherence to gastric epithelial cells and evasion of the immune system). Of particular interest is the presence of tandem repeat sequences upstream of some OMPs. In other mucosal pathogens increasing or decreasing the numbers of repeats by slipped-strand mispairing and recombination affects transcription of the downstream genes. In this way minor reversible mutations rapidly change the antigen profile of the pathogen, leading to evasion of the host immune system. Repeat DNA sequences can therefore be used as markers for the rapid identification of potential virulence genes.

Global approaches to the analysis of gene function (functional genomics)

The acquisition and analysis of sequence data is not an end in itself; instead it is a starting point for generating hypotheses that can be tested in the laboratory. Homology provides clues, but does not prove the gene function. In the past geneticists assigned gene function by specifically disabling or knocking out a single gene and comparing the phenotypes of the mutant and parent strains. The release of the *H. pylori* genome sequence has coincided with important technological advances in four areas: bioinformatics, hybridisation technology, mutagenesis, and protein chemistry. This allows an integrated approach to the functional genomics of *H. pylori* at the mutational, transcriptional and protein expression levels (fig 1). These advances will liberate scientific understanding from the piecemeal study of individual genes or operons towards a comprehensive analysis of the entire gene and protein complement of the bacterial cell. The following provides an outline of what is now possible.

![Figure 1 An integrated approach to the functional genomics of H pylori.](http://gut.bmj.com/)

Global deletion mutagenesis analysis

Bioinformatics

Complete nucleotide sequence

Expression profile of all transcripts in a genome

"Transcriptome"

Bioinformatics

The total expressed protein complement of a genome

"Proteome"

Bioinformatics

"Mass murder" Global deletion mutagenesis analysis

Leading articles express the views of the author and not those of the editor and editorial board.
Bioinformatics (in silico analysis of sequence data)

Bioinformatics is poised to change the way in which we tackle biomedical research forever. As applied to genome sequences, it is essentially the evolution of computer-based technology dedicated to the mining of genome sequences. The past few years have seen vast improvements in the algorithms used to analyse sequence data. Furthermore, an increasing range of bioinformatics software has been developed and released into the public domain via the Internet. Careful and intelligent use of this software can afford important new insights into protein structure and function and allow the generation of testable hypotheses.

The published annotated form of the genome sequence of *H pylori* by scientists at TIGR falls short of being definitive. The authors used a narrow set of programs in their analysis and doubtless several reanalyses of the data will be undertaken. Already such analysis is available on the Internet at the PEDANT web site: (http://pedant.mips.biochem.mpg.de/frishman/pedant.html). In addition to the benefits of such “static” analysis, an on-going dynamic analysis is needed, constantly re-evaluating the *H pylori* sequence data in the light of newly published sequences.

Differential gene expression (DGE) and high density hybridisation arrays

In terms of genome analysis, DGE refers to the studies of the transcriptional activity of all an organism’s genes. This is possible by extracting the mRNA expressed under a range of environmental conditions and hybridising these sequences to a high-density gridded array of the DNA content of an organism. The availability of ever-cheaper oligonucleotides, 96-well PCR technology, and complete genome sequence data, makes possible the highly attractive option of using gridded libraries of PCR products, constituting a defined and complete set of ORFs and intergenic regions. The Affymetrix Biochip is an example of the emerging technology of nucleic acid arrays which extends this principle. Biochip technology makes it possible to perform thousands of hybridisations in parallel, so that—for example, the effect of a given stimulus on transcription, such as low pH or the interaction with gastric epithelial cells, can be assayed simultaneously for all genes in a genome. A *H pylori* biochip could also be used for differential genomics by comparing the gene complements of various clinical isolates, or even testing genome plasticity of the same strain.

Global gene deletion analysis (the “mass murder” approach)

Although some traditional genetic techniques have not been possible with *H pylori* (for example, direct transposon mutagenesis or phage mediated transduction), the construction of defined mutants by allele replacement has proved to be reliable and relatively simple. The availability of the entire genomic sequence of *H pylori* means that the large-scale systematic construction of defined mutants is now possible. In conjunction with global gene deletion analysis is the potential to label each mutant with a unique DNA signature tag. The original use of DNA signature tags used randomly tagged transposons. More recently, we have developed this technology by coupling the incorporation of tags with allele replacement, referred to as STAR (signature tagged allele replacement). Thus the systemic construction and tagging of different *H pylori* mutants means that hundreds of mutants can be analysed simultaneously for phenotypic features, such as the ability to survive acid shock. Although in vitro screens are useful, they cannot be expected to identify all virulence genes of a pathogen because they do not reflect the complex environment that a pathogen encounters within the host. The greatest potential for the wholesale tagging of mutants is for in vivo studies. This will allow a drastic reduction in the number of animal experiments required for the assessment of virulence. Furthermore, the need for onerous and repetitive filter based radioactive hybridisations can be avoided by quantitating the survival of mutants using a specifically designed Affymetrix Biochip containing all the different sequence tags used.

Proteome analysis (the entire complement of proteins expressed by a cell)

The protein products of many of the newly identified *H pylori* genes are unknown. Recent improvements in high sensitivity biological mass spectrometry have provided a powerful adjunct to traditional two dimensional gel electrophoresis. Proteins cut out of a two dimensional gel can now be peptide-mass-fingerprinted, and constituent peptides can be sequenced by mass spectrometry. New software takes data from mass spectrometry and uses them to find the best match in a sequence database, allowing one to go from a spot on a PAGE gel to protein identification in a matter of hours. Thus the entire complement of proteins expressed by a cell (the proteome) can be defined. This kind of approach has already been used to provide insights into the function of an anatomical subset of the proteome such as the cell envelope. Proteome studies are made even more powerful when applied to an organism whose genome has been sequenced—synergistic interactions between the two approaches maximise information return, so that—for example, ORFan gene products can be shown to be real proteins.

Practical implications for genomic studies on *H pylori*

A multidimensional analysis of *H pylori*, looking at sequences, mutants, transcripts, and proteins, will result in a quantum leap in our understanding of the biology and pathogenesis of this bacterium. In determining the activity of large sets of genes and proteins and the interactions between them, an important step towards constructing a functional model of the entire organism will be taken. This basic information will provide the framework for future research and for highlighting potential drug targets or vaccine components. Rather than one or two drug targets or vaccine candidates, this approach should yield dozens.

Drug resistant *H pylori* is a real threat to current therapies and new targets need to be identified. Ideally, new drug design should be specific for the organism, alleviating aggressive disruption of the bowel microflora. One rationale is to identify novel ORFsequences specific to *H pylori* whose disruption is lethal to the organism. Thus several *H pylori* specific gene products which are essential for the organism can be identified. If such a lethal knockout can be shown to encode a protein that is important for survival under acidic conditions then further selectivity can be assured increasing the potential for developing monotherapeutic agents. A comprehensive understanding of the mechanisms involved in the adhesion of *H pylori* to gastric cells offers further opportunities for drug design to inhibit colonisation the stomach.

The availability of the genome sequence offers unprecedented opportunities for vaccine design as the complete inventory of genes encoding every virulence factor and potential immunogen are available for selection as potential vaccine candidates. The combination of two dimensional SDS-PAGE and immunoblot analysis of the whole organism, or a subset such as the cell envelope, should identify all immunodominant proteins. The discovery that naked DNA is immunogenic has opened up new
Concluding remarks

All sufferers from *H pylori* associated disease and researchers in the field should be indebted to the scientists at TIGR for releasing the complete *H pylori* sequence into the public domain. This has not been the trend among some companies who have hoarded these important data. Closer collaboration between academia and industry in the postgenomic era can only be mutually beneficial.

Given that we have known of the existence of many bacterial pathogens for over a century, it is ironic that *H pylori*, one of the most recently identified pathogens, promises to be one of the most well characterised. It will improve our understanding of other bacteria, especially other enteropathogens such as *Campylobacter jejuni*. Indeed, *H pylori* may become the model organism for a new approach to research in the postgenomic era. *H pylori* and humans have been partners for a very long time. The challenge is how to translate this “Aladdin’s cave” of information into explaining the multiple facets of *H pylori* associated disease. Undoubtedly, environmental and host susceptibility play a major role in colonisation and disease. The likely completion of the human genome project early in the new millennium promises to increase further our understanding of this complex disease. It would be fitting if the “*H pylori*” model of genomic research were the prototype for functional analysis of the human genome.

N DORRELL
B W WREN

Microbial Pathogenicity Research Group,
Department of Medical Microbiology,

St Bartholomew’s and the Royal London School of Medicine & Dentistry,
West Smithfield, London EC1A 7BE, UK

Correspondence to: Dr Wren (email b.w.wren@mds.qmw.ac.uk).

Nick Dorrell is funded by the Joint Research Board, St Bartholomew’s Hospital.