The utility of complete genome sequences in the study of pathogenic bacteria

D. W. HOOD

Molecular Infectious Diseases Group, University of Oxford Department of Paediatrics, Institute of Molecular Medicine, John Radcliffe Hospital, Headington, Oxford, OX3 9DU, UK

SUMMARY

The availability of complete genome sequences is a revolution in the study of microorganisms. A fully annotated genome sequence provides an interactive tool for scientists and influences the approach and focus of research. In this article I discuss the impact of genome sequencing projects of bacteria. Much useful data have been obtained but the experimental methods needed to fully exploit the information continue to develop. Some of the approaches and particular applications relevant to bacteria of clinical importance are discussed.

Key words: genome sequence, pathogenic bacteria, infection control.

INTRODUCTION

July 1995 was the advent of a new era in microbiology with the publication of the complete 1·83 Mb genome sequence of the bacterium *Haemophilus influenzae* strain Rd (Fleischmann *et al*. 1995). The significance of this achievement was not just that it presented for the first time a complete genome from a free-living organism but that it proved that complete genomes could be sequenced more rapidly, and at a lower cost, than had been previously thought. Indeed, the *Escherichia coli* sequencing project, initiated in the late 1980s, was widely expected to be the first bacterial genome sequencing project concluded. The novel approach taken by Venter and his colleagues at The Institute for Genomic Research (TIGR) for *H. influenzae* was that random shotgun-cloned small fragments of chromosomal DNA were sequenced then the vast amount of information was assembled into large contiguous sequences by computers and specifically developed software. Gaps were closed using homology comparisons, the polymerase chain reaction and inserts from lambda libraries. This method circumvented the construction of a physical map with overlapping cosmids which was a prerequisite of the ordered approach of sequencing the genome of *E. coli* (Blattner *et al*. 1997).

The interest following the publication of the genome sequence of *H. influenzae* was enormous and dramatically changed the somewhat sceptical attitude of many scientists and researchers. Here for the first time was the complete genetic information displayed for a free living organism. This gave the impetus for the extensive list of bacterial and eukaryotic sequencing projects subsequently undertaken. Within a short time bacterial genome sequencing became a highly resourced area and only 3 years later there are 16 completed and published, and 51 current, genome sequencing projects. It is estimated that by the year 2000 there could be around 100 completed bacterial genome sequences encompassing the majority of significant bacterial pathogens of humans.

The linear DNA sequence of a bacterial chromosome gives information not obtainable by other means, identifying all potential open reading frames (orfs), intergenic sequences and their organization within the genome. Gene functions can be ascertained by homology comparisons with known data bank sequences from the same or other organisms, and gene products can be grouped by function. The catalogue of genes represents all aspects of cell growth, division and metabolism, and for pathogenic bacteria contains information on the genetic determinants for each virulence factor, vaccine candidate and potential drug and diagnostic targets.

It is somewhat surprising that the majority of published studies based upon genome sequence data have been restricted to comparative computer analyses. Fewer than 1% of the more than 1000 papers citing the *H. influenzae* genome sequence paper have used the genome sequence to perform experimental work (Saunders & Moxon, 1998). Partly this reflects the exponential rate at which the sequences are becoming available and the problems in comprehending and dealing with the significant amount of information. The establishment of funding in laboratories to maximize the use of the data has somewhat lagged behind the funding of the genome sequence projects themselves. The completion of a genome sequence signals the end of many traditional strategies in microbiology, in particular for the analysis of gene function. In the remainder of this article I discuss bacterial genome sequencing and the redirection of existing technologies and some new

https://doi.org/10.1017/S0031182099004072 Published online by Cambridge University Press

technologies that can be applied to general investigations of microbial cells, and some particular investigations relevant to bacterial pathogens.

IDENTIFICATION AND THE STUDY OF GENES

A genome sequence lists the unique order of all nucleotides and their arrangement into coding and non-coding regions and is essentially a starting point for scientific research. The order of the genes can give information about co-regulated transcriptional units and the direction of transcription relative to the origin of replication can given information on chromosome organisation and fluidity. For example in *Mycoplasma genitalium* the polarity of transcription is predominantly away from the origin of replication (Fraser *et al*. 1995). Fluctuations in the DNA base composition can identify regions of DNA acquired by horizontal transfer, for example a mu-like prophage was identified in the *H. influenzae* genome sequence (Fleischmann *et al*. 1995). The gene pool contains all of the genes required for the existence of the free living bacterium. Riley (1993) has proposed a system to catalogue genes into broad functional groups. These 12 groups have proved useful as a framework for classifying genes that have been identified by sequencing projects. Gene lists comprise a basic ingredient for the publication of all new genomes sequences and in particular, they allow researchers to access the DNA sequence information in a directed and informative manner relevant to their particular research interest. To take, for example, *H. influenzae*; 1743 putative genes were identified in the 1·83 Mb genome sequence, of which 60 % have been assigned to these broad functional categories. Gene functions are assigned after homology comparisons by automated data bank searching by methods based on the BLAST and FASTA search programmes. Importantly, as with the majority of bacterial genome sequences, about 40 % of the *H. influenzae* orfs remained unassigned with respect to function by homology comparisons alone. These are often referred to as 'orphan' or 'FUN' genes. The number of genes which remain with no data-base match has decreased as subsequent sequencing projects identified homologous orphan genes in other organisms.

A limitation to a genome annotation procedure based solely upon homology comparisons is that it can be erroneous and does not always provide rigorous evidence for the biological functions of genes and their products. Existing data bank sequences may have putative, probable or confirmed gene functions and each will be given equal weighting by annotation teams. Good quality partial and lower quality overall matches may not always be distinguished but could be crucial in deciding gene function. Examples are known where gene functions can be complemented between species with amino acid identity of just over 20 % (Tatusov *et al*. 1996). Also, examples are known where only a very few changes to an amino acid sequence results in altered substrate or functional specificity of a protein. When analyzing genomes, useful terms for considering genes of comparable function are paralogues, homologous genes in the same organism whose products perform related but not identical functions and orthologues, homologous gene functions from different organisms (Koonin, Tatusov & Rudd, 1996). An estimate is that half of the genes in *E. coli* compared to only 35 % of the genes in *H. influenzae* and 17 % of the genes in *H. pylori* are paralogues (Hinton, 1997). The higher level of paralogy in *E. coli* correlates with its larger genome size and its highly regulated ability to adapt to novel environments. Homology-based annotation of genome sequences continues to improve and as more genome sequences become available the comparison of functional groups, such as clusters of orthologues (COGs), will become more significant in gene assignment (Tatusov, Koonin & Lipman, 1997).

Gene lists alone can give the first indication of novel findings from a genome sequence. For example, *E. coli* is the best studied bacterium and the model system that has led to many of the fundamental advances in biology. It is the workhorse for molecular biologists and serves as a model system in which to study gene regulation. Of the 4288 genes in the 4·64 Mb genome there remains a staggering 40 % for which we have no real clues as to their function. Thus in the most studied organism there still remain significant gaps in our knowledge and perhaps some aspects of fundamental processes are undiscovered. Of the classified genes, new biosynthetic functions have been described which were previously unknown (Tyrell *et al*. 1998) and new tRNA genes and cryptic phage have been found. Gene lists allow observations to be made regarding the predicted metabolic pathways present in an organism and how these influence the niches in which it is found. For example, *H. influenzae*, found only in the human host, has a metabolism biased towards an anaerobic, nitrogen-rich environment and key enzymes in the TCA cycle and electron transport chain are apparently absent. *E. coli* allocates more than 4 % of its genome to regulatory genes, including 45 or more two-component regulatory systems, whereas *H. influenzae* has only 4 such systems. This reflects the metabolic versatility and diverse habitats of *E. coli*, being able to respond to environmental signals, and the host dependence and restricted niche of *H. influenzae*.

For certain key bacteria there is a good argument for undertaking a systematic experimental approach to mutate all reading frames to maximize the information on gene function. Such an approach has been established for *Saccharomyces cerevisiae* by the European Union of Yeast Genome Sequencing

https://doi.org/10.1017/S0031182099004072 Published online by Cambridge University Press

Network using collaborating laboratories around the world (Goffeau *et al*. 1996) and could prove especially useful to help understand the high proportion of genes of unknown function present in most organisms. The role of orphan genes may be of particular interest in pathogenic bacteria such as *Helicobacter pylori*, *Mycobacterium tuberculosis* and *H. influenzae* where their restriction to particular host compartments may not allow the function to be defined in laboratory studies. A systematic, or even random, knockout of many of these genes, or analysis of *in vivo* expression, would be invaluable in determining their importance and potential function.

MOTIF SEARCHING

The availability of complete genome sequences permits new approaches to the identification of genes of particular interest. Motif searching can allow rapid screening for target DNA features or amino acid consensus sequences. For example, a finding in the *H. influenzae* genome sequence was the prevalence of a DNA uptake signal sequence (USS) which mediates the preferred uptake of homologous DNA in an organism which is considered to be naturally competent. A total of 1465 copies of the USS, comprising a conserved 9 bp core with larger consensus 29 bp sequence were identified dispersed, mainly in integenic regions, throughout the chromosome (Smith *et al*. 1995). Only 9 copies of this sequence would be expected by chance in a genome of this size.

A further application has been to identify simple DNA repeats which have been shown to be associated with some virulence related genes. The length of a tract of repeats of a DNA motif of between 1 and 6 nucleotides in length is unstable and can change by polymerase slippage during nucleic acid replication, altering gene expression. If located within the orf, alteration in the number of repeats leads to translational frame shifting and if in the promoter region can lead to altered transcription by affecting RNA polymerase binding. If the gene product is a cell surface-associated structure or is required for virulence then random on-off switching of expression (phase variation) could enhance survival of the bacterium in different host compartments or microenvironments. These genes are often termed 'contingency genes'. A search for simple repeats in the *H. influenzae* genome sequence revealed 12 loci containing tetranucleotides, repeated 6 or more times, within the 5′ end of orfs (Hood *et al*. 1996*a*). Three were previously characterized LPS biosynthetic genes and of the 9 novel genes, 1 was involved in LPS biosynthesis, one was related to an adhesin of *Yersinia pseudotuberculosis* and *Y. pestis*, 4 were homologues of genes involved in iron acquisition and 1 was a DNA methyl transferase. Roles for each of these loci in the virulence of *H. influenzae* can be postulated. A similar but more complete survey of the *H. pylori* genome sequence revealed 27 genes containing simple repeats (mostly mono- or di-nucleotides), the genes fitting into the broad categories described for *H. influenzae* above (Saunders *et al*. 1998). This general principal can be extended to screening genome sequences with any conserved motif representing gene or protein functions of interest.

BEYOND THE INDEX GENOME

A recurring problem for every genome project is the choice of the strain of an organism as the source of DNA to be sequenced. A sequenced genome of the index strain represents only the genetic complement from an isolate of a single strain of a single species. How this represents sequence drift and the presence or absence of unique DNA sequences in the population is an important question. Often the sequenced strain is non-pathogenic and/or related strains have altered spectrums of disease. Comparative genome analysis of related bacteria can tackle these issues if the coding potential is directly related to the capacity to cause disease. Current examples of sequencing related bacteria are of a pathogenic *E. coli* strain 0157 to compare with the completed non-pathogenic K-12 sequence. For *Mycobacterium tuberculosis*, two clinical isolates distinct in geography and time, have been sequenced (Cole *et al*. 1998) and for *Neisseria meningitidis* a group A and group B strain have been selected for sequencing projects and when complete can be compared with a strain of the related pathogen, *N. gonorrhoeae*. Each comparison could help answer different questions such as the particular and overall drift of DNA sequence between organisms and the specific requirements for infection by related organisms. The currently available genome sequences include those of *E. coli* (Blattner *et al*. 1997) and *Bacillus subtilus* (Kunst *et al*. 1997), the most completely studied Gram-negative and Gram-positive organisms, respectively, and these model strains are used as reference for the analysis of other bacterial species. An alternative to complete genome sequencing is to partially sequence (2–3 × coverage) related strains, negating the time-consuming and costly assembly process, but still giving a theoretical 95 % or greater coverage of a genome. This would be useful for studying gene organization but would give information on the unique sequences present only in the partially sequenced strain. Beyond DNA sequencing, a solution for comparative analysis is to use subtractive hybridization. This technique has been used to compare *E. coli* and *Salmonella typhimurium* (Groisman *et al*. 1993) and can identify, if of suitable size, sequences that are unique to each bacterium. Further to this, DNA fragments or oligonucleotides representing all genes from the

https://doi.org/10.1017/S0031182099004072 Published online by Cambridge University Press

index organism can be assembled as a microarray on a DNA chip (Marshall & Hodgson, 1998) and then hybridized with DNA from a test organism (Gingeras *et al*. 1998). A limitation is that this technique will not identify sequences that are unique to the test isolate.

Complete genome sequencing has allowed validation of the taxonomic classification of organisms proposed by Woese & Fox (1977) based upon ribosomal RNA sequence data. Comparison of available genome sequences supercedes current assessments of evolutionary relatedness based largely upon limited sequence comparisons, structural features and multi-locus enzyme electrophoresis and has confirmed the three lines of descent of the eubacteria, archaebacteria and eukarya.

GENE EXPRESSION ANALYSIS

The information available from sequencing projects can be used to examine either individual or global gene function. One comprehensive approach to the analysis of gene function is to investigate gene expression through examination of total messenger RNA or protein. An extension to the microarray technnique described above is used to analyse expression of genes by RT-PCR of mRNA then to hybridize the products to the DNA chip (DeRisi, Iyer & Brown, 1997; Saizieu *et al*. 1998). In this way the specific expression of gene products under different growth conditions or at different stages of the disease process can be determined and compared to commensal or *in vitro* grown organisms. The power of this system is that there is no reliance on identification of gene function and it may indicate polymorphisms between strains that are essential for particular aspects of pathogenesis.

An alternative strategy for analysing gene expression is to investigate total cellular proteins, so-called proteomics. Reference 2-dimensional gel electrophoresis mapping of bacteria integrated to the genome sequence (Cash *et al*. 1997; Humphery-Smith, Cordwell & Blackstock, 1997) provides a powerful technique with which to investigate bacterial environmental responses and transitions encountered during infection. This has the potential to be integrated with mass spectrometry and be used directly to monitor expressed protein levels within cells. A recent development is the protein chip, constructed with expressed gene products which interact differently with the chip surface. This can be used for expression or immunological studies of bacterial growth or pathogenesis.

VIRULENCE AND PATHOGENICITY

An obvious basis for the selection of bacteria for genome sequencing projects is driven by our need and desire to understand and control bacteria of clinical significance. The linear DNA sequence from a microbial pathogen catalogues every gene product and therefore the basis for every determinant influencing the host-parasite interaction, every vaccine candidate and potential therapeutic and diagnostic target for the organism. As an example of the information gained from the gene catalogue alone, the genome sequence of *H. influenzae* strain Rd revealed that genes involved in capsule and pilin formation, both of which are virulence determinants, were absent in the relatively non-pathogenic strain sequenced (Fleischmann *et al*. 1995). Several aspects of the virulence of the gastric pathogen *H. pylori* were evident from interpretation of the genome sequence alone (Berg *et al*. 1997). The conundrum to sequence non-virulent strains of pathogenic bacteria has been repeated with *E. coli* where the sequenced strain, K-12, is a laboratory-adapted commensal lacking some important virulence determinants found in other strains and in related bacteria of the Enterobacteriaceae. However, *E. coli* K-12 contains many genes and features important for the virulence of other strains and species and remains important because it is genetically amenable and facilitates the study of heterologous genes from other organisms. The release of the genome sequences from related bacteria such as *E. coli* O157, *Salmonella*, *Shigella* and *Yersinia* will permit comparative analysis to identify features influencing the different spectra of disease caused by the relevant bacterium. Many of the 'orphan' genes may have functions related to the species specialization. A particular emphasis is placed upon identifying so-called pathogenicity islands, regions of DNA containing clusters of virulence-related genes. These may be mobile and divergent in DNA base composition and when acquired are sometimes all that is apparently required to convert a bacterium to a virulent phenotype (Groisman & Ochman, 1996). Assuming a basic degree of conservation of genome sequence and arrangement it should be possible to detect such islands by relatively simple comparison techniques such as subtractive hybridization, genome walking or limited coverage genome sequencing.

One particular experimental application of a genome sequence to pathogenicity studies has been to analyse the biosynthesis and biology of lipopolysaccharide (LPS), an important cell surface molecule and a characterized virulence determinant in *H. influenzae*. LPS is a complex glycolipid involved in host cell interaction and the lipid A portion of the molecule initiates the cytokine cascade in response to the endotoxin. Little information was available on the biosynthetic pathway, and the relevant genes, prior to the elucidation of the genome sequence of *H. influenzae*. 25 genes potentially involved in LPS biosynthesis were identified from the genome sequence by homology comparisons

https://doi.org/10.1017/S0031182099004072 Published online by Cambridge University Press

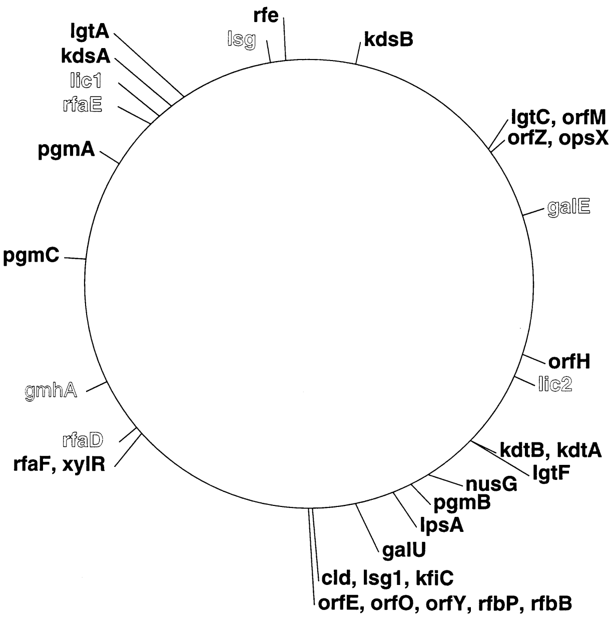

Fig. 1. Identification of genes for lipopolysaccharide biosynthesis in *Haemophilus influenzae*. The genes indicated in outline are those which had been identified and characterized as *H. influenzae* LPS-related genes prior to the availability of, and are present in, the strain Rd genome data base. The remaining genes were identified from the genome sequence and are shown at their relative positions on a circular 1·83 Mbp strain Rd genome map.

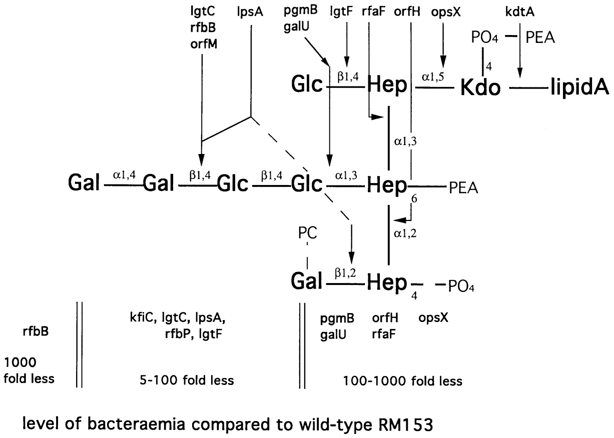

Fig. 2. A schematic representation of the structure of LPS from *H. influenzae* strain RM153 (Eagan) based on the results of the analysis of Masoud *et al.* (1996). The proposed site of action in LPS biosynthesis of loci identified from the complete genome sequence of strain Rd are shown above the structure, based on the combined results form T-SDS-PAGE and comparative ES-MS analysis of purified LPS. The grouping of mutant strains shown below the structure represents the level of bacteraemia after intraperitoneal inoculation of infant rats, compared to the wild-type. Represented in the LPS structure: Kdo, 2-keto-3-deoxyoctulosonic acid; Hep, L-glycero-D-manno-heptose; Glc, D-glucose; Gal, D-galactose; PEA, phosphoethanolamine; PO_4, phosphate; PC, phosphorylcholine.

with characterized genes from other organisms (Hood *et al.* 1996*b*). Fig. 1 illustrates the contribution of the genome sequence to determination of the genetic basis of *H. influenzae* LPS biosynthesis. Once isolated and mutated, these 25 genes were re-introduced into the sequenced strain and a pathogenic strain (Eagan) and their effect on LPS structure and biology was assessed through analysis of isolated LPS and the behaviour of mutant strains in infant rats, a valid model for the infection of *H. influenzae*. It was possible to identify by these methods most of the major steps in LPS biosynthesis and for the first time to estimate the minimal LPS structure required for the bacterium to survive in the host vascular compartment (Fig. 2). The significance of this work was the speed and completeness of a task that would have been virtually impossible without the genome sequence. The LPS molecule is immunogenic and such analyses can form the rational basis for studying the molecule as a candidate for vaccine development (Moxon, 1997; Moxon, Hood & Richards, 1998), as discussed below. The availability of complete genome sequences enhances other methodologies such as sequence-tagged mutagenesis (STM) (Hensel *et al.* 1995) and *in vivo* expression technology (IVET) (Slauch, Mahan & Mekalanos, 1994); both techniques have proven to identify genes required for virulence of pathogenic organisms. Candidate virulence-related genes implied by these methods can be much more rapidly identified when the chromosomal sequence of the target organism is known.

VACCINES AND THERAPEUTIC TARGETS

The complete genome sequence of an organism of clinical interest offers information on every gene product responsible for the synthesis of each vaccine candidate and potential therapeutic target. The identification of suitable targets is somewhat involved, some are cell surface-exposed features involving multiple gene products, so targeting a single protein may not always be optimally effective. Genome searches can readily identify numerous novel or particular classes of enzymes or structures but the lengthy investigation and screening involved with a single organism can easily stretch the research resources of even the largest pharmaceutical companies. Some pharmaceutical companies have undertaken in-house bacterial sequencing projects and many others rely on sequence supply companies such as Incyte to supply data to a number of pharmaceutical partners. Often less emphasis is given to completion of a DNA sequence and full annotation.

Homology searches and structural predictions can identify all surface-accessible proteins and other molecules as immunological targets and potential

https://doi.org/10.1017/S0031182099004072 Published online by Cambridge University Press

vaccine candidates. This search can be fine-tuned with information on known immunogenic structures in the same or related bacteria but screening of candidates is still a significant undertaking. Gene products can be assayed directly by injecting relevant DNA directly into eukaryotic cells (Johnston & Barry, 1997) but the more labour-intensive cloning, expression and immunisation procedures may prove to have greater utility. As the number of genome sequences continues to grow, it is likely that a detailed comparison and systematic informatics approach will help focus upon the more suitable candidates.

It is now over 30 years since the discovery of the last significant new antibacterial agent. Undoubtedly, genome sequences offer a means to combat the emergence of antibiotic resistance in common bacterial pathogens. Structurally novel classes of antibiotics can be tested against selected bacterial targets, chosen from the genome sequence. Such development is still an enormous undertaking and target investigation must include studies on its novelty, spectrum of action, selectivity of target, *in vivo* expression, essentiality of the gene product, cellular location and targets must be amenable to high-throughput screening against banks of bioactive compounds. The first three criteria listed for study can be investigated by comparative genome analyses as described above.

CONCLUSION

Complete genome sequences are a revolution in biology. The sequences of the major bacterial pathogens will be available within the next two years and with advances in bioinformatics will provide the starting point for significant research on bacterial metabolism, host adaptation and virulence. Developments in expression analysis will allow us to answer many general and particular questions on how bacteria function and how they become pathogens, and as new technologies appear should allow us to better understand the organisms and develop successful measures for their control.

REFERENCES

BERG, D. E., HOFFMAN, P. S., APPELMELK, B. J. & KUSTERS, J. G. (1997). The *Helicobacter pylori* genome sequence: genetic factors for long life in the gastric mucosa. *Trends in Microbiology* **5**, 468–474.

BLATTNER, F. R., PLUNKETT, G., BLOCH, C. A., PERNA, N. T., BURLAND, V., RILEY, M., COLLADO-VIDES, J., GLASNER, J. D., RODE, C. K., MAYHEW, G. F., GREGOR, J., DAVIS, N. W., KIRKPATRICK, H. A., GOEDEN, M. A., ROSE, D. J., MAU, B. & SHAU, Y. (1997). The complete genome sequence of *Escherichia coli* K12. *Science* **277**, 1453–1462.

CASH, P., ARGO, E., LANGFORD, P. R. & KROLL, J. S. (1997). Development of a *Haemophilus* two-dimensional protein database. *Electrophoresis* **18**, 1472–1482.

COLE, S. T., BROSCH, R., PARKHILL, J., GARNIER, T., CHURCHER, C., HARRIS, D., GORDON, S. V., EIGLMEIER, K., GAS, S., BARRY, C. E., TEKAIA, F., BADCOCK, K., BASHAM, D., BROWN, D., CHILLINGWORTH, T., CONNOR, R., DAVIES, R., DEVLIN, K., FELTWELL, T., GENTLES, S., HAMLIN, N., HOLROYD, S., HORNSBY, T., JAGELS, K., KROGH, K., McLEAN, J., MOULE, S., MURPHY, L., OLIVER, J., OSBORNE, J., QUIAL, M. A., RAJANDREAM, M., ROGERS, J., RUTTER, S., SEEGER, K., SKELTON, J., SQUARES, R., SQUARES, S., SULSTON, J. E., TAYLOR, K., WHITEHEAD, S. & BARRELL, B. G. (1998). Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature* **393**, 537–544.

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.

FLEISCHMANN, R. D., ADAMS, M. D., WHITE, O., CLAYTON, R. A., KIRKNESS, E. F., KERLAVAGE, A. R., BUTT, C. J., TOMB, J.-F., DOUGHERTY, B. A., MERRICK, J. M., McKENNEY, K., SUTTON, G., FITZHUGH, W., FIELDS, C., GOCAYNE, J. D., SCOTT, J., SHIRLEY, R., LIU, L.-I., GLODEK, A., KELLEY, J. M., WIEDMAN, J. F., PHILLIPS, C. A., SPRIGGS, T., HEDBLOM, E., COTTON, M. D., UTTERBACK, T. R., HANNA, M. C., NGUYEN, D. T., SAUDEK, D. M., BRANDON, R. C., FINE, L. D., FRITCHMAN, J. L., FUHRMANN, J. L., GEOGHAGEN, N. S. M., GNEHM, C. L., McDONALD, L. A., SMALL, K. V., FRASER, C. M., SMITH, H. O. & VENTER, J. C. (1995). The genome of *Haemophilus influenzae* Rd. *Science* **269**, 496–512.

FRASER, C. M., GOCAYNE, J. D., WHITE, O., ADAMS, M. D., CLAYTON, R. A., FLEISCHMANN, R. D., BULT, C. J., KERLAVAGE, A. R., SUTTON, G., KELLEY, J. M., FRITCHMAN, J. L., WEIDMAN, J. F., SMALL, K. V., SANDUSKY, M., FUHRMANN, J., NGUYEN, D., UTTERBACK, T. R., SAUDEK, D. M., PHILLIPS, C. A., MERRICK, J. M., TOMB, J.-F., DOUGHERTY, B. A., BOTT, K. F., HU, P.-C., LUCIER, T. S., PETERSON, S. N., SMITH, H. O., HUTCHISON, C. A. & VENTER. J. C. (1995). The minimal gene complement of *Mycoplasma genitalium*. *Science* **270**, 397–403.

GINGERAS, T. R., GHANDOUR, G., WANG, E., BERNO, A., SMALL, P. M., DROBNIEWSKI, F., ALLAND, D., DESMOND, D., HOLODNIY, M. & DRENKOW, J. (1998). Simultaneous genotyping and species identification using hybridisation pattern recognition analysis of generic *Mycobacterium* DNA arrays. *Genome Research* **8**, 435–448.

GOFFEAU, A., BARRELL, B., BUSSEY, H., DAVIS, R., DUJON, B., FELDMANN, H., GALIBERT, F., HOHEISEL, J., JACQ, C., JOHNSTON, M., LOUIS, E. J., MEWES, H. W., MURAKAMI, Y., PHILLIPSEN, P., TETTELIN, H. & OLIVER, S. G. (1996). Life with 6000 genes. *Science* **274**, 546.

GROISMAN, E. & OCHMAN, H. (1996). Pathogenicity islands: bacterial evolution in quantum leaps. *Cell* **87**, 791–794.

GROISMAN, E., STURMOSKI, M., SOLOMON, F., LIN, R. & OCHMAN, H. (1993). Molecular, functional and evolutionary aspects of sequences specific to *Salmonella*. *Proceedings of the National Academy of Sciences, USA* **90**, 1033–1037.

https://doi.org/10.1017/S0031182099004072 Published online by Cambridge University Press

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.

HINTON, J. (1997). The *Escherichia coli* genome sequence: the end of an era or the start of the FUN? *Molecular Microbiology* **26**, 417–422.

HOOD, D. W., DEADMAN, M. E., JENNINGS, M. P., BISCERIC, M., FLEISCHMANN, R. D., VENTER, J. C. & MOXON, E. R. (1996*a*). DNA repeats identify novel virulence genes in *Haemophilus influenzae*. *Proceedings of the National Academy of Science*, *USA* **93**, 11121–11125.

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. (1996*b*). Use of the complete genome sequence information of *Haemophilus influenzae* strain Rd to investigate lipopolysaccharide biosynthesis. *Molecular Microbiology* **22**, 951–965.

HUMPHERY-SMITH, I., CORDWELL, S. J. & BLACKSTOCK, W. P. (1997). Proteome research: Complementarity and limitations with respect to RNA and DNA worlds. *Electrophoresis* **18**, 1217–1242.

JOHNSTON, S. A. & BARRY, M. A. (1997). Genetic to genomic vaccination. *Vaccine* **15**, 808–809.

KOONIN, E. V., TATUSOV, R. L. & RUDD, K. E. (1996). Protein sequence comparison at genome scale. *Methods in Enzymology* **266**, 295–322.

KUNST, F., OGASAWARA, N., MOSZER, I., ALBERTINI, A. M., ALLONI, G., AZEVEDO, V., BERTERO, M. G., BESSIERES, P., BOLOTIN, A., BORCHERT, S., BORRISS, R., BOURSIER, L., BRANS, A., BRAUN, M., BRIGNELL, S. C., BRON, S., BROUILLET, S., BRUSCHI, C. V., CALDWELL, B., ACPUANO, V., CARTER, N. M., CHOI, S. K., CODANI, J. J., CONNERTON, I. F., DANCHIN, A. *et al*. (1997). The complete genome sequence of the Gram-positive bacterium *Bacillus subtilis*. *Nature* **390**, 249–256.

MARSHALL, A. & HODGSON, J. (1998). DNA chips: an array of possibilities. *Nature Biotechnology* **16**, 27–31.

MASOUD, H., MOXON, E. R., MARTIN, A., KRAJCARSKI, D. & RICHARDS, J. C. (1996). Structure of the variable and conserved lipopolysaccharide oligosaccharide epitopes expressed by *Haemophilus influenzae* serotype b strain Eagan. *Biochemistry* **36**, 2091–2103.

MOXON, E. R. (1997). Applications of molecular microbiology to vaccinology. *Lancet* **278**, 631–637.

MOXON, E. R., HOOD, D. & RICHARDS, J. (1998). Bacterial lipopolysaccharides: Candidate vaccines to prevent *Neisseria meningitidis* and *Haemophilus influenzae* infections. *Advances in Experimental Medicine and Biology* **435**, 237–243.

RILEY, M. (1993). Functions of the gene products of *Escherichia coli*. *Microbiology Reviews* **57**, 862–952.

SAIZIEU, A., CERTA, U., WARRINGTON, J., GRAY, C., KECK, W., & MOUS, J. (1998). Bacterial transcript imaging by hybridisation of total RNA to oligonucleotide arrays. *Nature Biotechnology* **16**, 45–48.

SAUNDERS, N. J. & MOXON, E. R. (1998). Implications of sequencing bacterial genomes for pathogenesis and vaccine development. *Current Opinions in Microbiology* **9**, 618–623.

SAUNDERS, N. J., PEDEN, J. F., HOOD, D. W. & MOXON, E. R. (1998). Simple sequence repeats in the *Helicobacter pylori* genome. *Molecular Microbiology* **27**, 1091–1098.

SLAUCH, J. M., MAHAN, M. J. & MEKALANOS, J. J. (1994). *In vivo* expression technology for selection of bacterial genes specifically induced in host tissues. *Methods in Enzymology* **235**, 481–492.

SMITH, H. O., TOMB, J.-F., DOUGHERTY, B. A., FLEISCHMANN, R. D. & VENTER, J. C. (1995). Frequency and distribution of DNA uptake sequences in the *Haemophilus influenzae* Rd genome. *Science* **269**, 538–540.

TATUSOV, R. L., KOONIN, E. V. & LIPMAN, D. J. (1997). A genomic perspective on protein families. *Nature* **278**, 631–637.

TATUSOV, R. L., MUSHEGIAN, A. R., BORK, P., BROWN, N. P., HAYES, W. S., BORODOVSKY, M., RUDD, K. E. & KOONON, E. V. (1996). Metabolism and evolution of *Haemophilus influenzae* deduced from a whole-genome comparison with *Escherichia coli*. *Current Biology* **6**, 279–291.

TYRELL, C., BAUSCH, C., PEEKHAUS, N. & UTZ, C. (1998). Post-genomic elucidation of a novel pathway for L-iodinate catabolism in *Escherichia coli*. *Microbial Comparative Genomics* **3**, 54–55.

WOESE, C. R. & FOX, G. E. (1977). Phylogenetic structure of the prokaryotic domain: the primary kingdom. *Proceedings of the National Academy of Sciences*, *USA* **74**, 5088–5090.

https://doi.org/10.1017/S0031182099004072 Published online by Cambridge University Press