SESSION VI: BASIC RESEARCH AND PATHOGENESIS

CHAIRPERSON: DR M. J. COLSTON

Preliminary analysis of the genome sequence of *Mycobacterium leprae*

S. T. COLE, N. HONORE & K. EIGLMEIER
Unité de Génétique Moléculaire Bactérienne, Institut Pasteur, Paris, France

The genome sequence of a strain of *Mycobacterium leprae*, originally isolated in Tamil Nadu and designated 'TN', has been completed recently, in accord with one of the priorities defined for leprosy research programmes at the joint WHO/Sasakawa Memorial Health Foundation workshop held in Bangkok in 1995. The sequence was obtained by a combined approach, employing automated DNA sequence analysis of selected cosmids and whole-genome 'shotgun' clones.[1,2] After the finishing process, the genome sequence was found to contain 3,268,203 base-pairs (bp), and to have an average G + C content of 57·8%, values much lower than the corresponding values for *M. tuberculosis*, which are 4,441,529 bp and 65·6% G + C.[3] The genome of *M. tuberculosis* is estimated to contain 4,000 genes encoding proteins.

A battery of gene identification programmes was used to analyse the genome sequence, and detailed comparisons with the genome and proteome sequences of *M. tuberculosis* were carried out.[3,4] By these means, it was established that there are approximately 1500 genes which are common to both *M. leprae* and *M. tuberculosis*, and, therefore, likely to be functional in both organisms. From the combined results of BLASTX and BLASTN searches,[5,6] using filters to reduce noise, it was established that the genome of *M. leprae* contains at least 1000 pseudogenes, with two or more mutations which should prevent their expression, and that a further 1686 genes have been 'deleted' from the genome. This latter conclusion is based upon the not unreasonable assumption that both mycobacteria derived from a common ancestor and, at one stage, had gene pools of similar size, as suggested by comparative analysis.[7] Downsizing from a genome of 4·41 Mb, such as that of *M. tuberculosis*, to one of 3·27 Mb would account for the loss of some 1200 protein coding sequences.

There is clear evidence from inspection of the genomic context, and from the presence of extensively truncated coding sequences, that many of these genes were once present in the genome of *M. leprae*, and have truly been lost. It is also certain that *M. tuberculosis* harbours a substantial number of genes that were never present in *M. leprae*, and that *M. leprae* contains more than 100 genes that have no counterpart in the genome of *M. tuberculosis.* From the results of highly sensitive dot-matrix comparisons,[8] it is probable that the extensively decayed coding sequences of approximately 450 once-common genes remain within the genome of *M. leprae*, but that these have mutated so much that they are now below the threshold of the BLAST analysis. The corresponding genes may have been lost at an early stage of degeneration of the genome, and have been under lower selective pressure.

Of particular relevance to the epidemiology of leprosy is the finding that the chromosome *of M. leprae* contains approximately 65 segments, varying in length from five to more than 200 genes, that show synteny to the genome of *M. tuberculosis*, but differ in their relative order and distribution. The presence of all three members of the dispersed repeat families, RLEP,[9] REPLEP, and LEPREP, at the junctions of regions of discontinuity is especially noteworthy. Assuming that the genomes of *M. tuberculosis* and *M. leprae* were once topologically equivalent, it is most probable that the current mosaic arrangement of the genome of *M. leprae* reflects multiple recombination events between conserved repetitive sequences. In at least one instance, recombination probably resulted in deletion of a block of genes located between REPLEP sequences, whereas, in others that are more difficult to document, translocation of chromosomal regions occurred. Preliminary evidence suggests that these dispersed repeats are capable of transposition, because, in a few cases, they were found within sequences corresponding to known genes of *M. tuberculosis*. These observations are highly encouraging, in the context of studies of strain variation and molecular epidemiology; PCR assays that survey all sites for these repetitive elements might uncover ‘hotspots’ for genome rearrangement. If these elements are associated with polymorphisms, they could then form the basis of a molecular tool for epidemiological monitoring of leprosy, that may allow us to distinguish between relapse and reinfection, similar to those used for *M. tuberculosis.*[10]

There is hope that, by pursuing these comparisons with *M. tuberculosis*, we may be able to identify the missing genes for key metabolic steps that enable other mycobacteria to grow. This information could find practical application, allowing us to cultivate *M. leprae* by supplementing the growth medium with the missing or rate-limiting nutrients, or to introduce the corresponding genes from *M. tuberculosis* into *M. leprae* by genetic engineering. More rapidly growing derivatives of *M. leprae* and *M. leprae* cultivated *in vitro* would be extremely useful for production of a vaccine, and would represent a cheaper, more attractive alternative to the armadillo.

If we are to develop a specific immunological test for the early diagnosis of leprosy, in particular the tuberculoid form of the disease, it is essential to define the repertoire of proteins accurately by a combination of genomic and proteomic studies, and to identify those proteins that are confined to *M. leprae*, or that show extensive diversity from their counterparts in other mycobacteria. At the present time, particularly informative results of preliminary proteome comparisons *in silico* are available. Roughly 10% of the genome of *M. tuberculosis* encodes 167 proteins that belong to the novel, glycine-rich PE and PPE families.[3,11–13] Only about 10 of these proteins are present in *M. leprae*; this may account for the significant difference of the size of the genomes. Of interest is the finding that one of the PPE proteins, the serine-rich antigen, is recognised by sera from leprosy patients,[14] and is strikingly different from the equivalent protein of *M. tuberculosis*. Consistent with this downsizing trend, the class of proteins referred to as ‘conserved hypotheticals’, i.e. they are present in two or more bacteria, but of unknown function[15], is two-thirds smaller in the leprosy bacillus (approximately 300) than in *M. tuberculosis* (915). Similarly, of the 606 proteins, predicted to be present in the proteome of *M. tuberculosis*, that previously had no counterparts elsewhere, 130 were also found in *M. leprae*, and some of these show extensive sequence divergence. These polypeptides could be exploited for diagnostic purposes for the identification of intracellular mycobacteria.

Perhaps the best hope of developing a successful skin test reagent for leprosy lies within the class of approximately 100 proteins of *M. leprae* that have no homologue in *M. tuberculosis*, although some of them may be related to proteins found in other microorganisms. Having confirmed species specificity, the corresponding proteins may offer great potential as reagents for diagnostic skin tests for leprosy. It is also possible that they might confer novel biological properties on *M. leprae*, and be involved in such functions as neurotropism or nerve damage.[16–18] Finally, this comparative approach should enable us to identify novel drug targets, and will be invaluable in the rational design of new therapeutic agents and drugs to treat leprosy.

Acknowledgements

Special thanks are due to Bart Barrell, Julian Parkhill and the Pathogen Group at the Sanger Centre. Financial support from the New York Community Trust, Association Française Raoul Follereau, ILEP and the Institut Pasteur is gratefully acknowledged.

References

[1] Eiglmeier K, Honore N, Woods SA *et al.* Use of an ordered cosmid library to deduce the genomic organisation of *Mycobacterium leprae. Mol Microbiol*, 1993; **7**: 197–206.

[2] Honore N, Bergh S, Chanteau S *et al.* Nucleotide sequence of the first cosmid from the *Mycobacterium leprae* genome project: structure and function of the Rif-Str regions. *Mol Microbiol*, 1993; **7**: 207–214.

[3] Cole ST, Brosch R, Parkhill J *et al*. Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature*, 1998; **393**: 537–544.

[4] Tekaia F, Gordon SV, Garnier T *et al.* Analysis of the proteome of *Mycobacterium tuberculosis in vitro. Tubercle Lung Disease*, 1999; **79**: 329–342.

[5] Altschul SF, Boguski MS, Gish W, Wooton JC. Issues in searching molecular sequence databases. *Nature Genet*, 1994; **6:** 119–129.

[6] Altschul S, Gish W, Miller W *et al*. A basic local alignment search tool. *J Mol Biol*, 1990; **215**: 403–410.

[7] Philipp W, Schwartz DC, Telenti A, Cole ST. Mycobacterial genome structure. *Electrophoresis*, 1998; **19**: 573–576.

[8] Sonnhammer ELL, Durbin R. A dot-matrix program with dynamic threshold control suitable for genomic DNA and protein sequence analysis. *Gene*, 1995; **167**: GC1–10.

[9] Woods SA, Cole ST. A family of dispersed repeats in *Mycobacterium leprae. Mol Microbiol*, 1990; **4**: 1745–1751.

[10] van Embden JDA, Cave WM, Crawford JT *et al.* Strain identification of *Mycobacterium tuberculosis* by DNA fingerprinting: recommendations for a standardized methodology. *J Clin Microbiol*, 1993; **31**: 406–409.

[11] Cole ST, Barrell BG. Analysis of the genome of *Mycobacterium tuberculosis* H37Rv. In: Chadwick DJ, Cardew G (eds) *Analysis of the genome of* Mycobacterium tuberculosis *H37Rv*. Wiley, Chichester, 1998, pp 160–172.

[12] Espitia C, Laclette JP, Mondragon-Palomino M *et al.* The PE-PGRS glycine-rich proteins of *Mycobacterium tuberculosis*: a new family of fibronectin-binding proteins? *Microbiology*, 1999; **145**: 3487–3495.

[13] Ramakrishnan L, Federspiel NA, Falkow S. Granuloma-specific expression of mycobacterium virulence proteins from the glycine-rich PE-PGRS family. *Science*, 2000; **288**: 1436–1439.

[14] Vega-Lopez F, Brooks LA, Dockrell HM *et al.* Sequence and immunological characterization of a serine-rich antigen from *Mycobacterium leprae. Infect Immun*, 1993; **61**: 2145–2153.

[15] Arigoni F, Talabot F, Peitsch M *et al.* A genome based approach for the identification of essential bacterial genes. *Nature Biotechnol*, 1998; **16**: 851–856.

[16] Rambukkana A, Yamada H, Zanazzi G *et al.* Role of alpha-dystroglycan as a Schwann cell receptor for *Mycobacterium leprae. Science*, 1998; **282**: 2076–2079.

[17] Rambukkana A, Salzer JL, Yurchenco PD, Tuomanen EI. Neural targeting of *Mycobacterium leprae* mediated by the G domain of the laminin-$\alpha 2$ chain. *Cell*, 1997; **88**: 811–821.

[18] Shimoji Y, Ng V, Matsumura K *et al.* A 21-kDa surface protein of *Mycobacterium leprae* binds peripheral nerve laminin-2 and mediates Schwann cell invasion. *Proc Natl Acad Sci USA*, 1999; **96**: 9857–9862.

DISCUSSION

Dr van Brakel: What are drug efflux systems?

Dr Cole: These are pumps. Drugs must penetrate into bacteria in order to exert their toxic effects on the organisms. They can enter the bacterial cell by a number of different means. Once they are in the cytoplasm, the drugs seek out their targets and inactivate them. Many bacteria possess natural means of resisting drugs; these are efflux systems that are capable of taking the drugs from the cytoplasm and pumping them back outside the cell, before they have exerted their toxic effects. Efflux systems are particularly common among environmental bacteria, such as *Pseudomonas* spp.

Dr Gillis: When will the annotation of the *M. leprae* genome be completed?

Dr Cole: It should be completed by another month or so. The bulk of the analysis has been done, but we must go through it once more, to make certain there are no internal inconsistencies or contradictions. At this moment, we're at this stage.

Dr Colston: The repetitive sequences that you have demonstrated in the genome suggest that there are recombination events. Do you believe recombination to be effective in *M. leprae*?

Dr Cole: I think that the recombination occurred at some point in the evolution of *M. leprae*. Its quite possible that some of the recombination genes have since been lost, and that the organism is now 'stuck' in its current configuration. One could now design experiments to test this hypothesis, and these experiments would also offer a means of examining possible strain differences among isolates of *M. leprae*.

Dr Kaplan: Has someone tried to clone *M. leprae* genes into *M. tuberculosis*? Are the *M. leprae* genes then expressed? Is this possibly a useful means of trying to learn how the genes function?

Dr Cole: Most, but not all, *M. leprae* genes would be expressed in another mycobacterial host. It might be even easier to express the *M. leprae* genes in *E. coli*, because the G:C content of *M. leprae* is much more similar to that of *E. coli* than to that of *M. tuberculosis*.

Dr Sengupta: Can the pseudogenes by reactivated?

Dr Cole: A mechanism exists in eukaryotic cells know as RNA-editing, which permits some of the defects to be overcome. This is unlikely to be the case in mycobacteria; I believe that most of the pseudogenes are truly dead. Many of them have incurred small deletions; the function of bits of DNA that have lost numbers of codons is unlikely to be reacquired by such editing.

Professor Grosset: Among the genes that have disappeared from the *M. leprae* genome, are many of them to be found in eukaryotic cells in which the organisms multiply?

Dr Cole: There is no evidence for that. On the other hand, evidence exists that *M. leprae* has acquired genes from eukaryotic hosts.

Dr Colston: That *M. leprae* can't multiply *in vitro*, for which we now have a genetic explanation, and that it is capable of multiplying *in vivo* suggest that the organism gains something from the *in-vivo* situation.

Dr Cole: What is odd about *M. leprae* is that, like *M. tuberculosis*, it has retained most of the anabolic pathways, so that it is capable of synthesizing much of what it needs. This is a bit surprising. Pursuing the comparison with *M. tuberculosis*, it appears that *M. leprae* acquires substrates, perhaps lipids, from the host, which it then degrades.

Professor Ji: I find your description of potential targets for new drugs very encouraging. Are there any examples of drugs that have been developed for any infectious agent that have been based on leads from this kind of genetic information?

Dr Cole: This hasn't yet occurred for bacterial agents, but a good example is that of the protease inhibitors that have been developed for HIV. Investigators first identified the protease gene in the viral genome, produced the encoded protein, defined its structure, and designed inhibitors that inactivate the enzyme.

Dr Colston: In spite of the great success of programmes based upon MDT for the control of leprosy, many fundamental questions remain unanswered. Paradoxically, at the same time that laboratory-based research in leprosy has declined in recent years, unprecedented advances have occurred in basic biomedical research, as the result of which we now have unique opportunities to address some of these unanswered fundamental questions. The sequencing of the genome of *M. leprae*, and the availability of the sequences of the genomes of other mycobacteria and eubacteria means that we can now begin to understand the basic biology of the leprosy bacillus. In addition, the expanding information on the sequence of the human genome, and the ability to carry out targeted gene-disruption in mice present us with new opportunities to understand the immunological interaction between *M. leprae* and its host, and how immune recognition is translated into a protective immune response. Thus far in this Workshop, we have heard about some of the work aimed at understanding the regulation and dysregulation of the immune response during leprosy reactions, and the application of molecular biological techniques to the rapid detection of drug resistant *M. leprae*. In this session, we heard an update on the sequencing and sequence analysis of the *M. leprae* genome, a presentation on the application of modern cell biology techniques to understand the interaction between *M. leprae* and Schwann cells, and presentations in which murine models have been used to further our understanding of protective immunity against *M. leprae*. I wish to present some ideas about how genetic microarrays might be employed to maximize the use of genome-sequence information.

Gene microarrays provide a rapid means of determining changes in gene expression on a global scale. A single microscope slide can contain probes for tens of thousands of genes, making it possible to investigate entire genomes in a single experiment. Such an approach has recently been used to investigate genomic variation, for example between BCG and *M. tuberculosis*, and changes of bacterial gene expression under different environmental conditions. Three potential applications of microarray technology illustrate the possibilities available to the leprosy research community.

GENOMIC VARIATION AMONG STRAINS OF *M. LEPRAE*

By constructing a microarray based on the genome of *M. leprae*, it should be possible to test DNA from different isolates of *M. leprae* to investigate genomic differences. A PCR-derived probe for each of the open reading frames of the genome of *M. leprae* (one might also include pseudogenes, inactivated genes and insertion sequences) is prepared and gridded robotically onto a microscope slide. Hybridization with DNA prepared from different isolates of *M. leprae* is then carried out to identify genomic differences that ultimately might be used for strain typing.

CHANGES OF GENE EXPRESSION OF *M. LEPRAE* MAINTAINED *IN VITRO*

A number of research groups have described systems in which metabolic activity of *M. leprae* has been maintained for several weeks *in vitro*. In some instances, DNA transferred into *M. leprae* using bacteriophages has been shown to be transcribed and translated. It might be possible to use an *M. leprae* microarray to investigate transcription of genes, and how transcription changes during incubation *in vitro*. Such an approach could provide novel information on why it has proved impossible to grow *M. leprae in vitro*.

A MICROARRAY APPROACH TO INVESTIGATING IMMUNE RESPONSES IN LEPROSY

A microarray based on the human genome-sequence could be used to increase our understanding of the immune response and the immunopathology of leprosy. In the first instance, a 'mini-array' of 500–1000 genes putatively involved in immunoregulation could be constructed. Approximately 500 such genes have been identified; many of these are available as 'sequence-verified clones', and the relevant gene may be amplified using vector-specific primers. For those genes that are not available in the sequence-verified clone library, sequence-specific primers may be constructed that will enable one to obtain an appropriate probe by means of PCR. The genes thus far identified include apoptosis-related genes, apoptosis suppressor genes, cytokine genes, cytokine receptor genes, chemokine genes, chemokine receptor genes, integrin genes, and TGFβ superfamily genes. Such an array could be used to investigate in great detail the regulation of immune responses during the interaction between *M. leprae* and host cells.

I hope that it will prove possible to develop this technology as a general resource for leprosy researchers.