Lepr Rev (2001) **72,** 387–398

The decaying genome of *Mycobacterium leprae*

KARIN EIGLMEIER*, JULIAN PARKHILL**, NADINE HONORÉ*, THIERRY GARNIER*, FREDJ TEKAIA***, AMALIO TELENTI[+], PAUL KLATSER[++], KEITH D. JAMES**, NICOLAS R. THOMSON**, PAUL R. WHEELER[+++], CAROL CHURCHER**, DAVID HARRIS**, KAREN MUNGALL**, BART G. BARRELL** & STEWART T. COLE*

**Unité de Génétique Moléculaire Bactérienne, Institut Pasteur, 28 rue du Docteur Roux, 75724 Paris Cedex 15, France*

***Sanger Centre, Wellcome Trust Genome Campus, Hinxton, CB10 1SA, United Kingdom*

****Unité de Génétique Moléculaire des Levures, Institut Pasteur, 25 rue du Docteur Roux, 75724 Paris Cedex 15, France*

[+]Microbiologie, Centre Hospitalier Universitaire Vaudois, Rue du Bugnon, 46, 1011 Lausanne, Switzerland

[++]Royal Tropical Institute (KIT) N.H. Swellengrebel Laboratorium voor Tropische Hygiene, Meibergdreef 39, 1105 AZ Amsterdam, The Netherlands

[+++]Veterinary Laboratories Agency- Weybridge, Woodham Lane, New Haw, Addlestone, Surrey KT15 3NB, United Kingdom

Summary Everything that we need to know about *Mycobacterium leprae,* a close relative of the tubercle bacillus, is encrypted in its genome. Inspection of the 3·27 Mb genome sequence of an armadillo-derived Indian isolate of the leprosy bacillus identified 1,605 genes encoding proteins and 50 genes for stable RNA species. Comparison with the genome sequence of *Mycobacterium tuberculosis* revealed an extreme case of reductive evolution, since less than half of the genome contains functional genes while inactivated or pseudogenes are highly abundant. The level of gene duplication was ~34% and, on classification of the proteins into families, the largest functional groups were found to be involved in the metabolism and modification of fatty acids and polyketides, transport of metabolites, cell envelope synthesis and gene regulation. Reductive evolution, gene decay and genome down-sizing have eliminated entire metabolic pathways, together with their regulatory circuits and accessory functions, particularly those involved in catabolism. This may

* Correspondence to: S. T. Cole, Unité de Génétique Moléculaire Bactérienne, Institut Pasteur, 28 Rue du Docteur Roux, 75724 Paris Cedex 15, France; Phone: 33-1-45 68 84 46. Fax: 33-1-40 61 35 83. E-mail: stcole@pasteur.fr

0305-7518/01/072387+12 $1.00

explain the unusually long generation time and account for our inability to culture the leprosy bacillus.

Introduction

Determining the complete genome sequence of a strain of *Mycobacterium leprae* was one of the highest priorities defined for leprosy research and control programmes at the joint WHO/Sasakawa Memorial Health Fund meeting held in Bangkok in 1995. It was clear to all participants that genomics would not only provide understanding of the unusual biology of *M. leprae* but that the information thus obtained would underpin leprosy research in what are hopefully the final years of the elimination campaign. The choice of the strain to be sequenced was influenced by the disease burden and as a result we chose to work with a patient isolate from Tamil Nadu, India, one of the worst affected countries.

In order to produce sufficient bacilli to extract DNA for library construction this strain, referred to as TN, was passaged in the armadillo.[1] An immortalized source of DNA was then obtained in the form of a cosmid library and this will facilitate future functional studies.[2] The DNA sequence of the TN strain of *M. leprae* was obtained by a combined approach employing automated DNA sequence analysis of selected cosmids and whole genome shotgun clones.[2,3] After assembling the subsequences, the complete genome sequence was regenerated and this was subjected to bioinformatic analysis so that the genes, their control signals, repetitive elements and other genomic features could be identified. Comparison of the genome sequence and gene set of the TN strain with those of other organisms was then undertaken. Comparative genomic analysis with the tubercle bacillus, *Mycobacterium tuberculosis* H37Rv, was found to be a particularly powerful approach that enabled much of the biochemistry, physiology and genetics of *M. leprae* to be predicted and unraveled. The principal findings of this analysis were presented in the original landmark publication which the reader should consult for further details.[4] Here we present a brief summary and an update of some of the findings.

Results and discussion

ORGANIZATION AND SEQUENCE OF THE GENOME

The complete genome sequence of the TN strain of *M. leprae* was found to contain 3,268,203 bp, and to have an average G + C content of 57·8%. There is a single circular chromosome and no plasmids. Bioinformatics predicted the existence of 1605 genes encoding proteins and a further 50 that code for stable RNA molecules. These values are much lower than those reported for the genome of the H37Rv strain of *M. tuberculosis*, comprising ~4000 genes, at 4,411,532 bp and 65·6% G + C.[5]

On analysis of the first complete *M. leprae* cosmid sequence in 1993, only 50% of the sequence was found to contain functional genes.[3] At that time, meaningful comparisons were very limited as few other sequences were available and no clues could be found to the possible function of the remaining 50% of the DNA. However, the availability of the *M. tuberculosis* genome sequence in 1998 changed things radically and enabled detailed pairwise comparisons of the genome and proteome sequences of both pathogens to be undertaken.[5,6] These revealed that only 49·5% of the genome of *M. leprae* was occupied by protein-coding genes while at least 27% of the sequence contained recognizable

Figure 1. See enclosed poster. Organization of the genome of the TN strain of *M. leprae*. Only those genes predicted to be active are shown and their functions are colour coded as follows: lipid metabolism, black; intermediary metabolism and respiration, yellow; information pathways, pink; regulatory proteins, sky blue; conserved hypothetical proteins, orange; unknown, light green; insertion sequences and phage-related functions, blue; stable RNAs, dark blue; cell wall and cell wall processes, dark green; virulence, detoxification and adaptation, white; PE and PPE protein families, magenta. For the sake of clarity the pseudogenes are not indicated but these occur in all empty regions. The scale is shown with 10 or 100 kb intervals. Genes shown above the line are transcribed from left to right, those below from right to left.

pseudogenes, inactive reading frames with functional counterparts in the tubercle bacillus. The remaining 23·5% of the genome did not appear to be coding at all, and probably contains gene remnants mutated beyond recognition together with the regulatory sequences that usually occur in intergenic regions.

The distribution of the 1,114 recognizable pseudogenes was essentially random throughout the genome (Figure 1; see attached poster of *M. leprae* genome). However, the 1605 potentially active genes tend to occur in clusters often flanked on both sides by long stretches of seemingly non-coding DNA.

REDUCTIVE EVOLUTION

The process by which large-scale loss of gene function arises has been termed reductive evolution. It has been observed in a number of important human pathogens such as the obligate intracellular parasites *Rickettsia* and *Chlamydia* spp.[7] and this suggests that genes become inactivated once their functions are no longer required in these highly specialized niches. In some endosymbionts such as the *Buchnera* spp., which are related to the enteric bacteria and found in aphids, reductive evolution has proceeded so extensively that the genome size is thought to have been reduced from ~4·5 to 0·64 Mb.[8] There are few pseudogenes in this case, and deletion appears to have been the dominant means of genome downsizing. One hypothesis, known as Muller's ratchet, has been proposed to explain reductive evolution. This involves the stochastic loss of genetic material and results in decreased fitness and little genetic variability. In part, this is due to the inability of organisms with no sex-cycle to acquire DNA and hence to repair genetic lesions through acquisition of new genes or by recombination. Obviously, as a consequence of its highly specialized niche, the only organism with which *M. leprae* can exchange DNA is the human host.

Until the sequence of *M. leprae* became available, the most extensive genome degradation reported in a pathogen was in *Rickettsia prowazekii*, the typhus agent, where only 76% of the potential coding capacity was used.[9] In comparison with *M. leprae*, the level of gene loss detected in *R. prowazekii* was modest, and it is notable that elimination of pseudogenes by deletion lags far behind gene inactivation in both pathogens, in contrast to *Buchnera*.[8] Interestingly, the G + C content of the functional genes in the leprosy bacillus (60·1%) is higher than that of the recognisable pseudogenes (56·5%), which is in turn greater than that of the remainder of the genome (54·5%), which may have undergone the most extensive decay. This suggests that the relatively high G + C content of *M. leprae*, and by extension, the other mycobacteria, is driven by the codon preference of the active genes, while random mutation within the non-coding regions causes them to drift towards a more neutral G + C content that is closer to that of the host. Deamination of cytosine residues in the DNA is a possible mechanism to account for this trend. This process would account for the leprosy bacillus having the lowest G + C content of all mycobacteria and it is noteworthy that the genomes of organisms that have undergone reductive evolution are generally richer in A + T.[10]

When one examines closely the genes that have been lost or inactivated during the reductive evolutionary process, clear trends are observed that conform to Darwinian theory and testify to the importance of selective pressure. For instance, *M. tuberculosis* is capable of anaerobic respiration using nitrate as terminal electron acceptor in a reaction catalysed by nitrate reductase and using electrons from the quinone pool of the respiratory chain. Nitrate reductase comprises four subunits, encoded by *narGHIJ,* and uses a complex cofactor, molybdopterin, which is synthesized by at least nine *moe/moa* genes and requires molybdate to be taken up from the extracellular medium by the ABC-transporter encoded by *modABC*.[5] In the tubercle bacillus, the only other enzyme predicted to use molybdopterin as cofactor is formate dehydrogenase and many microbes are capable of growing on a defined medium containing formate and nitrate as sole carbon and energy sources. *M. leprae* has pseudogenes corresponding to both of these enzymes and for almost all of the proteins required to transport molybdate and to insert it into the pterin ring. Apparently, once the need to use the formate-nitrate pathway was lost, the genes for the entire system acquired mutations and decayed as none of their functions was required.

GENOME DOWNSIZING

Reductive evolution involves gene loss through mutational inactivation and deletion. If one makes the reasonable assumption that the genomes of *M. leprae* and *M. tuberculosis* were once topologically equivalent and roughly 4·4 Mb in size, as is the case for many other slow-growing mycobacteria,[11–13] then extensive downsizing must have occurred during evolution of the leprosy bacillus since its genome is <75% of the size of that of *M. tuberculosis*.[4] It seems likely that recombination events involving repetitive DNA were responsible for both a reduction in genome size and the rearrangement of chromosomal segments that led to the loss of global synteny between the genomes of the tubercle and leprosy bacilli.[4] This is discussed further in this issue.[14]

A 1·1 Mb reduction in the size of the genome would have eliminated ~1100 protein-coding sequences, and *M. leprae* should, therefore, produce 3000 proteins compared to the 4000 predicted in *M. tuberculosis*. On proteomic analysis of *M. leprae*, only 391 soluble protein species were detected,[15] compared to nearly 1800 in *M. tuberculosis*.[16] This is entirely consistent with the large number of pseudogenes in *M. leprae* and excludes the possibility of gene expression by a novel mechanism such as RNA editing. Since diverging from the last common mycobacterial ancestor, the leprosy bacillus may have lost over 2000 genes, and reductive evolution has probably defined naturally the minimal gene set for a pathogenic mycobacterium.

At present, little is known about the genomic diversity of different isolates of *M. leprae* as very few studies have been undertaken. No differences in genomic organization were uncovered by restriction fragment length polymorphism analysis[17] and the systematic sequence analysis of two regions of the genome of several different isolates of *M. leprae* revealed no single nucleotide polymorphisms.[3] These regions correspond to the pseudogenes ML1873, ML1874, ML1884 and ML1885, which are functionally inactive and should, therefore, be more prone to divergence as they are under less selective pressure. This sequence conservation suggests that the immediate ancestor of the present leprosy bacillus had already undergone reductive evolution and that a single clone then expanded and was disseminated globally. Sequence and micro-array analysis of more *M. leprae* isolates will shed further light on possible strain divergence and the emergence of the disease.

GENES ENCODING PROTEINS

Of the 1605 genes predicted in *M. leprae*, there are 1440 which are also found in tubercle bacilli and 165 genes that have no orthologue in *M. tuberculosis*. Of the latter, some functional information could be predicted for 29 genes by bioinformatics, while the remaining 136 show no similarity to known genes elsewhere. It is highly likely that many of these will also be pseudogenes as they are generally shorter than average and occur in regions of low gene density.[4] In consequence, *M. leprae* may produce as few as 1500 proteins. In most cases, these proteins have been assigned roles in housekeeping functions, information pathways, various metabolic pathways, and so on,[4] and they will not be discussed further here. A complete functional list may be found at http://www.pasteur.fr/recherche/unites/Lgmb/NATURE_DATA/ML_gene_list. From Figure 2, which presents the classification into the 11 functional categories used previously,[5] it can be seen that the major groups contain genes for central and intermediary metabolism, cell wall processes and conserved hypothetical proteins. Issues such as why *M. leprae* grows so slowly and resists attempts at in vitro culture are discussed elsewhere in this issue.[18] Instead, we will now concentrate on those proteins, and the principal protein families, that have not been described previously in detail.

MULTIGENE FAMILIES

The majority of the genes (52%) present in *M. tuberculosis* arose from gene duplication events and this process may have conferred extensive functional redundancy.[6] Many of these genes are involved in lipid metabolism or encode the novel glycine-rich proteins of the PE and PPE families. Slightly more than 34% of the proteins now found in *M. leprae* are the products of gene duplication events, or share common domains, as defined by the bioinformatic routines used here.[6,19] There are far fewer families of duplicated proteins than in the tubercle bacillus and they are much smaller in size as a result of reductive evolution. For instance, whereas *M. leprae* contains 87 families containing two members and 37 with three members, *M. tuberculosis* has 213 and 72, respectively.

The duplicated genes have been classified into partitions on the basis of their similarity and part of this information is presented in Table 1. As in the tubercle bacilli, the largest partition, P46.1, contains enzymes involved in polyketide synthesis and fatty acid metabolism (Table 1), and this again underlines the importance of these activities to slow growing mycobacterial pathogens. The second and third largest protein families in *M. tuberculosis* are those for the 167 PE and PPE proteins; their counterparts in *M. leprae* are considerably smaller accounting for only 12 members (P5.8, P7.2, Table 1) collectively, although about 30 pseudogenes also exist. None of the PE or PPE proteins predicted to be produced by *M. leprae* contain the multiple C-terminal repetitions that are suspected of being involved in antigenic variation.[5] Retraction of these gene families, which are exceptionally GC-rich, partly contributes to the difference in the respective genome sizes of the tubercle and leprosy bacilli, and to the much lower GC content of *M. leprae*.

From Table 1, it can be seen that the major protein families are involved in lipid or polyketide metabolism, modification and synthesis of cell envelope components (methyltransferases, glycosyltransferases), transport processes (ABC transporters, MmpL proteins), or in gene regulation (TetR, WhiB, two-component system response regulators).

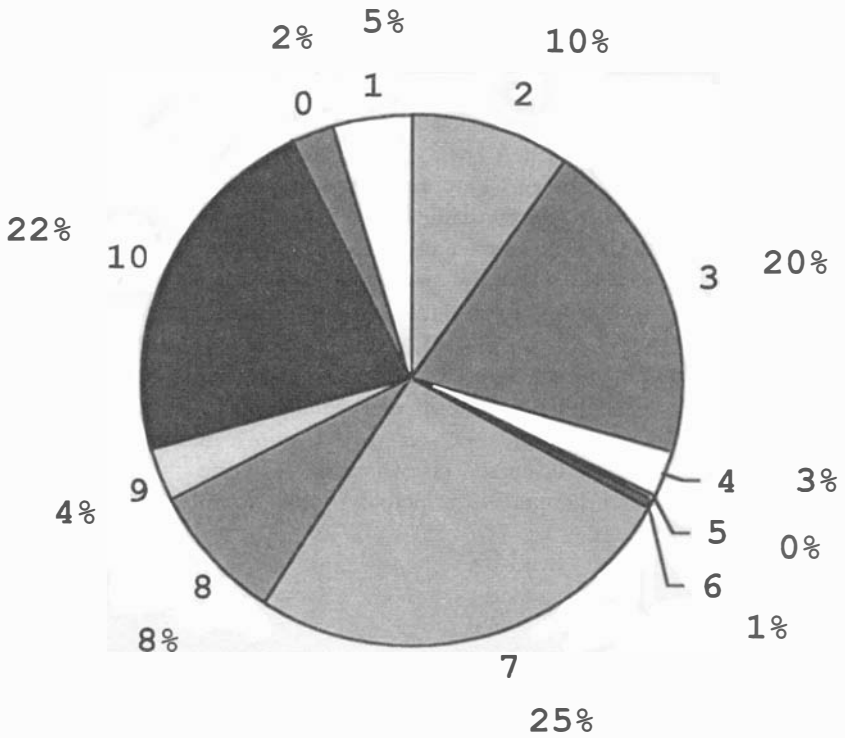

Figure 2. Broad functional classification of genes found in *M. leprae*. The functional groups and the percentage of total gene complement are shown, and correspond to: 1, lipid metabolism; 2, information pathways; 3, cell wall and cell wall processes; 4, stable RNAs; 5, insertion sequences and phage-related functions; 6, PE and PPE protein families; 7, intermediary metabolism and respiration; 8, unknown function; 9, regulatory proteins; 10, conserved hypothetical proteins; 0, virulence, detoxification and adaptation.

REGULATION AND SIGNAL TRANSDUCTION

Intracellular pathogens such as *Chlamydia* and *Rickettsia*,[10] as well as endosymbionts like *Buchnera*,[8] tend to have lost most of the genes involved in regulation, and *M. leprae* also displays this trend. The number of regulatory proteins predicted is roughly one-third of that recorded for *M. tuberculosis*, as a result of reductive evolution, and details of those remaining are presented in Table 2.

At the level of promoter recognition and transcription initiation controlled by the sigma factors of RNA polymerase, major differences exist between the tubercle and leprosy bacilli and these may explain some of the differences in their physiology. There are 13 sigma factors in *M. tuberculosis*, 10 of which belong to the extra-cytoplasmic function (ECF) family.[5] Only two of these ECF genes are still functional in *M. leprae, sigC* and *sigE,* while the eight others are present as pseudogenes. The sigma-70 family sigma factors SigA and SigB are predicted to be functional, however, whereas the 13th sigma factor gene, *sigF,* is inactive. It has been found recently in *M. tuberculosis* that the ability to survive at temperatures of 37°C and above is partly controlled by the ECF sigmas E and H, acting through the sigma 70 factor SigB.[20,21] In addition to the heat shock response, loss of *sigE* also affects resistance to SDS and oxidative stress, and survival in macrophages.[20,21] The

Table 1. Prominent protein families in *M. leprae*

Family	Description	Metabolic class
P5.1	Various methyl/cyclopropane mycolic acid synthases	Fatty acid metabolism
P5.10	WhiB transcriptional regulators	Gene regulation
P5.2	GTP-binding proteins	Regulation
P5.3	sugar-phosphate nucleotidyl transferases	Cell wall functions
P5.4	two-component system response regulators	Signal transduction
P5.5	acetohydroxyacid/acetolactate synthases	Central metabolism
P5.6	glycosyl transferases	Cell wall functions
P5.7	membrane/cell division proteins	Transport
P5.8	PE family	Unknown
P5.9	celldivision/anion transporting ATPase	Transport
P6.1	MmpL, conserved large membrane proteins	Transport
P6.2	phosphoserine phosphatase/acyl transferases	Fatty acid metabolism
P6.3	Mce proteins	Pathogenesis
P6.4	ABC-transport protein, inner membrane component	Transport
P7.1	methyltransferases	Fatty acid metabolism
P7.2	PPE-family (6)	Unknown
P8.1	conserved (membrane) protein	Transport
P8.2	phosphoglycerate mutases/mutT1	Central metabolism
P9.1	enoyl-CoA hydratase/isomerase	Fatty acid metabolism
P10.1	TetR-family transcriptional regulators	Regulation
P13.1	conserved hypothetical proteins	Unknown
P18.1	ABC-transport protein. ATP-binding component	Transport
P46.1	Polyketide synthesis/fatty acid metabolism	Fatty acid metabolism

optimal growth temperature of *M. leprae* is 32°C and this probably explains why the bacterium, which contains a full complement of heat shock proteins, multiplies principally in the extremities of the human body. It is conceivable that mutational inactivation of the sigma factor gene *sigH*, and possibly others, leads to lowered production of SigB at higher temperatures, and this may have led to *M. leprae* colonizing cooler regions of the body such as the skin and ears.

The interaction of RNA polymerase with the promoter regions of genes is often influenced by repressors or transcriptional activators that ensure expression under defined physiological conditions or in response to availability of a given substrate. In *M. tuberculosis* there are >110 proteins that have broad regulatory potential and in the leprosy bacillus this number has dwindled to 46, suggesting that the organism resides within a more stable niche than *M. tuberculosis (*Table 2). These regulatory proteins can be classed in the corresponding families on the basis of the characteristic motifs that they contain and, with one exception, all of these families are substantially smaller than those of the tubercle bacillus. The exception is the WhiB family[22] as there are five of these proteins in *M. leprae* compared to seven in *M. tuberculosis* (Table 2). The cysteine-rich WhiB proteins, also known as WhmA-G,[23] are confined to the *Actinomycetes* but in *Streptomyces* spp. they regulate developmental processes such as sporulation. Although mycobacteria do not sporulate, the change from exponential growth to persistancy or dormancy can be considered as a different state of development and the *whiB2* ortholog of *M. smegmatis, whmD*, has been found to be essential for septum formation and cell division.[24] The fact that so many WhiB proteins have been preserved in the face of reductive evolution strongly argues for their playing a major biological role in *M. leprae*.

Table 2. The regulatory repertoire of *M. leprae*

Sigma factors		
sigA (*rpoT*)	ML1022	RNA polymerase sigma-70 factor
sigB	ML1014	RNA polymerase sigma-70 factor
sigC	ML1448	ECF subfamily sigma factor
sigE	ML1076	ECF subfamily sigma subunit
Two component regulatory systems		
—	ML0174	response regulator
—	ML0175	sensor kinase
mtrA	ML0773	response regulator
mtrB	ML0774	sensor kinase
—	ML2123	response regulator
—	ML2124	sensor kinase
regX3	ML2439	response regulator
senX3	ML2440	sensor kinase
—	ML0803	sensor kinase
—	ML1286	response regulator
Serine-Threonine protein kinases and phosphoprotein phosphatases		
pknA	ML0017	serine-threonine protein kinase
pknB	ML0016	serine-threonine protein kinase
pknG	ML0304	serine-threonine protein kinase
pknL	ML0897	serine-threonine protein kinase
ppp	ML0020	probable phosphoprotein phosphatase
Repressors/activators		
TetR/AcrR-family		ML0064, ML0316, ML0717, ML0815, ML0949, ML1070, ML1733, ML2457, ML2568, ML2677
WhiB-family		ML0382 (*whiB3*), ML0639, (*whiB7*), ML0760 (*whiB2*), ML0804 (*whiB1*), ML2307 (*whiB4*)
MarR-family		ML0550, ML2696, ML2140
ArsR-family		ML0825
LysR-family		ML2041 (*oxyR*), ML2663 (*oxyS*)
NifR3-family		ML2186
Crp/Fnr-family		ML2302
Other families		ML0565 (*whiA*), ML0824 (*furB*), ML0988 (*recX*), ML1003 (*lexA*), ML1013 (*ideR*), ML1411 (*argR*), ML2188 (*phoY1*)
Possible others		ML0320, ML0592, ML0898, ML0919, ML1320, ML1328, ML1330, ML1367, ML1419, ML1652, ML1753, ML1783, ML2063, ML2156, ML2429, ML2530

Bacteria respond to changes in environmental conditions by means of diverse signal transduction systems that control gene expression through phosphorylation of regulatory proteins. These are of two types in mycobacteria, the classical two-component systems, comprising membrane-bound histidine protein kinases and phospho-aspartyl response regulators,[25] and the eukaryotic-like serine-threonine protein kinase phosphorelay system (STPK).[26,27] Of the 11 complete two-component systems present in the tubercle bacillus, only four have been retained by *M. leprae* together with isolated genes coding for a single histidine protein kinase and a response regulator. One of these systems, *mtrAB,* is essential for *M. tuberculosis* and this is almost certainly true of *M. leprae* as well. Indeed, one can

speculate with confidence that orthologs of the remaining two-component system genes of *M. leprae* will also prove to be essential in the tubercle bacillus.

Although formal evidence is still lacking, it is highly likely that the STPK, are also involved in signal transduction given their similarity to many other enzymes of the STPK superfamily that play this role. Only four of the 11 STPK genes of *M. tuberculosis* and the putative phosphoprotein phosphatase (Table 2), have functional orthologues in *M. leprae*. Indeed, it has been suggested that three of these genes *pknA*, *pknB* and *ppp* may control the timing of cell division or septation as they occur in an operon with other division genes.[26]

PATHOGENICITY

In some microbes, the combination of genomics and bioinformatics has been of great value in identifying genes for potential virulence factors that augment the degree of pathogenicity, and an excellent example is provided by the genome of the plague bacillus, *Yersinia pestis*.[28] When bioinformatic tools such as the GC skew[29] or dinucleotide bias[30] were applied to this genome, several new pathogenicity or adaptation islands of atypical base composition were uncovered. Using similar approaches to investigate the *M. leprae* genome, no such islands were detected, although they did provide evidence for recent chromosomal rearrangements.[4] Likewise, when database searches were performed few hits to genes for known virulence factors were obtained. Similar observations have been made previously for the tubercle bacillus,[31] and it now seems unlikely that any virulence genes were acquired. Instead, it seems more probable that the ability to survive in the macrophage or Schwann cell and hence to persist in the body represent the major determinants of pathogenicity. Although our understanding of the initial steps in infection of Schwann cells by *M. leprae* has improved considerably, thanks to the definition of the roles of laminin-binding protein and phenolic glycolipid 1 in this process,[32–36] we know little about the ensuing events or mycobacterial persistance in either Schwann cells or macrophages. Nevertheless, although genomics has not pinpointed a handful of potential candidate virulence genes, the sample size has been reduced to a tangible level by comparative genomics. A set of ~120 genes of unknown function has been defined that are common to both the leprosy and tubercle bacilli but no other sequenced pathogens[4] and this includes the *mce* genes encoding the cell entry factors[37,38] (Table 1). Testing their role in pathogenesis can now be undertaken by means of surrogate functional genomics.

NOVEL FUNCTIONS, HORIZONTAL GENE TRANSFER AND IMMUNODIAGNOSTICS

As outlined above, while most of the *M. leprae* genes have orthologues in *M. tuberculosis*, there are several that appear to be unique and may have novel activities. These include hypothetical proteins of unknown function and a number of potential enzymes such as the inorganic pyrophosphatase encoded by *ppa*, prolyl-tRNA synthetase, a eukaryotic-like uridine phosphorylase, phospho-*enol*-pyruvate carboxylase, adenylate cyclase, cytochrome P450 and enoyl-CoA hydratase. Furthermore, there are two transport systems that may play significant physiological roles: an ABC-transporter for sugars, and a second Nramp1-like protein, possibly involved in divalent metal ion uptake[39] that may offset the apparent absence of a siderophore system. It is probable that the phospho-*enol*-pyruvate (PEP) carboxylase replaces the pyruvate carboxylase of *M. tuberculosis*, as this enzyme is missing from *M. leprae*, and intervenes in the anaplerotic pathways. There is only one cytochrome

P450 (ML2088) present in the leprosy bacillus, compared to 20 in *M. tuberculosis* and, as this enzyme has no counterpart in *M. tuberculosis,* its function might be specific.

There is evidence that some of these enzymes have been acquired as a result of horizontal gene transfer and this is best illustrated by the prolyl-tRNA synthetase, ProS, which is the sole aminoacyl-tRNA synthetase of *M. leprae* with no counterpart in *M. tuberculosis.* Surprisingly, ProS is more similar to the enzymes of *Borrelia burgdorferi* and to eukaryotes such as Drosophila, humans and yeast. It has been proposed that horizontal transfer of tRNA synthetase genes occurs frequently, and that the pathogen *B. burgdorferi* may have acquired *proS* from its host.[40] Comparison of the genetic neighbourhood provides further support for this hypothesis as the *M. leprae proS* is both displaced and inverted with respect to the *M. tuberculosis* genome,[4] consistent with recent acquisition. In this case, the domain structure of the enzyme is indicative of a eukaryotic origin, and the human host appears the most likely candidate. Another example is found in the case of uridine phosphorylase, an enzyme that is not common in bacteria, as the closest relative of the *M. leprae* protein is that of the mouse.

Ensuring the elimination of leprosy as a public health problem will require both continued implementation of multidrug therapy and improved detection of infected individuals. Diagnosis is difficult in patients with few lesions and accurate information about subclinical infection is often rare. The identification of proteins that may be specific for the leprosy bacillus opens up new avenues for the development of immunodiagnostic tests possibly of the transdermic kind. In the post-genomic era it is important that the immunogenicity of these polypeptides be appraised and, if the initial findings are promising, attempts should be made to produce batches suitable for field testing. Determining the genome sequence of *M. leprae* has taught us much about the biology of the pathogen but it is the application of this new knowledge to disease control that should be prioritized now.

Acknowledgements

We are grateful to B. R. Bloom, P. J. Brennan, M. J. Colston, J. Grosset, and B. Ji for their advice, reagents and encouragement. This work was supported by the New York Community Trust, ILEP, the Association Française Raoul Follereau, the Wellcome Trust, the Institut Pasteur, and the UNDP/World Bank/WHO Special Programme for Research and Training in Tropical Diseases.

References

[1] Kirchheimer WK, Storrs EE. Attempts to establish the armadillo (*Dasypus novemcinctus Linn.*) as a model for the study of leprosy. I. Report of lepromatoid leprosy in an experimentally infected armadillo. *Int J Lepr*, 1971: **39:** 693–702.

[2] Eiglmeier K, Honoré 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.

[3] Honoré 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.

[4] Cole ST, Eiglmeier K, Parkhill J *et al.* Massive gene decay in the leprosy bacillus. *Nature*, 2001: **409:** 1007–1011.

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

[6] Tekaia F, Gordon SV, Garnier T *et al*. Analysis of the proteome of *Mycobacterium tuberculosis* in silico. *Tubercle Lung Dis*, 1999: **79:** 329–342.
[7] Andersson JO, Andersson SGE. Insights into the evolutionary process of genome degradation. *Curr Opin Genet Dev*, 1999: **9:** 664–671.
[8] Shigenobu S, Watanabe H, Hattori M *et al*. Genome sequence of the endocellular bacterial symbiont of aphids *Buchnera* sp. APS. *Nature*, 2000: **407:** 81–86.
[9] Anderssen SGE, Zomorodipour A, Andersson JO *et al*. The complete genome sequence of the obligate intracellular parasite *Rickettsia prowazekii*. *Nature*, 1998: **396:** 133–140.
[10] Tamas I, Klasson LM, Sandstrom JP, Andersson SG. Mutualists and parasites: how to paint yourself into a (metabolic) corner. *FEBS Lett*, 2001: **498:** 135–139.
[11] Brosch R, Gordon SV, Eiglmeier K *et al*. Comparative genomics of the leprosy and tubercle bacilli. *Res Microbiol*, 2000: **151:** 135–142.
[12] Philipp W, Schwartz DC, Telenti A, Cole ST. Mycobacterial genome structure. *Electrophoresis*, 1998: **19:** 573–576.
[13] Stinear TP, Jenkin GA, Johnson PDR, Davies JK. Comparative genetic analysis of *Mycobacterium ulcerans* and *Mycobacterium marinum* reveals evidence of recent divergence. *J Bacteriol*, 2000: **182:** 6322–6330.
[14] Cole ST, Supply P, Honoré N. Repetitive sequences in *Mycobacterium leprae* and their impact on genome plasticity. *Lepr Rev*, 2001, **72:** 387–398.
[15] Marques MAM, Chitale S, Brennan PJ, Pessolani MCV. Mapping and identification of the major cell-wall associated components of *Mycobacterium leprae*. *Infect Immun*, 1998: **66:** 2625–2631.
[16] Jungblut PR, Schaible UE, Mollenkopf H-J *et al*. Comparative proteome analysis of *Mycobacterium tuberculosis* and *Mycobacterium bovis* BCG strains: towards functional genomics of microbial pathogens. *Mol Microbiol*, 1999: **33:** 1103–1117.
[17] Williams DL, Gillis TP, Portaels F. Geographically distinct isolates of *Mycobacterium leprae* exhibit no genotypic diversity by restriction fragment-length polymorphism analysis. *Mol Microbiol*, 1990: **4:** 1653–1659.
[18] Wheeler P. The microbial physiologists guide to the leprosy genome. *Lepr Rev*, 2001, **72:** 399–407.
[19] Tekaia F, Dujon B. Pervasiveness of gene conservation and persistence of duplicates in cellular genomes. *J Mol Evol*, 1999: **49:** 591–600.
[20] Manganelli R, Dubnau E, Tyagi S *et al*. Differential expression of 10 sigma factor genes in *Mycobacterium tuberculosis*. *Mol Microbiol*, 1999: **31:** 715–724.
[21] Manganelli R, Voskuil MI, Schoolnik GK, Smith I. The *Mycobacterium tuberculosis* ECF sigma factor sigmaE: role in global gene expression and survival in macrophages. *Mol Microbiol*, 2001: **41:** 423–437.
[22] Hutter B, Dick T. Molecular genetic characterisation of whiB3, a mycobacterial homologue of a *Streptomyces* sporulation factor. *Res Microbiol*, 1999: **150:** 295–301.
[23] Soliveri JA, Gomez J, Bishai WR, Chater KF. Multiple paralogous genes related to the *Streptomyces coelicolor* developmental regulatory gene whiB are present in *Streptomyces* and other actinomycetes. *Microbiology*, 2000: **146:** 333–343.
[24] Gomez JE, Bishai WR. whmD is an essential mycobacterial gene required for proper septation and cell division. *Proc Natl Acad Sci USA*, 2000: **97:** 8554–8559.
[25] Grebe TW, Stock JB. The histidine protein kinase superfamily. *Adv Microb Physiol*, 1999: **41:** 139–227.
[26] Fsihi H, De Rossi E, Salazar L *et al*. Gene arrangement and organisation in a ~76 kilobase fragment encompassing the oriC region of the chromosome of *Mycobacterium leprae*. *Microbiology*, 1996: **142:** 3147–3161.
[27] Av-Gay Y, Davies J. Components of eukaryotic-like protein signaling pathways in *Mycobacterium tuberculosis*. *Microb Comp Genom*, 1997: **2:** 63–73.
[28] Parkhill J *et al*. Genome sequence of *Yersinia pestis*, the causative agent of plague. *Nature*, 2001: **412:** 777–999.
[29] Lobry JR. Asymmetric substitution patterns in the two DNA strands of bacteria. *Mol Biol Evol*, 1996: **13:** 660–665.
[30] Karlin S. Global dinucleotide signatures and analysis of genomic heterogeneity. *Curr Opin Microbiol*, 1998: **1:** 598–610.
[31] Cole ST. Learning from the genome sequence of *Mycobacterium tuberculosis* H37Rv. *FEBS Lett*, 1999: **452:** 7–10.
[32] Ng V, Zanazzi G, Timpl R *et al*. Role of the cell wall phenolic glycolipid-1 in the peripheral nerve predilection of *Mycobacterium leprae*. *Cell*, 2000: **103:** 511–524.
[33] Rambukkana A, Salzer JL, Yurchenco PD, Tuomanen EI. Neural targeting of *Mycobacterium leprae* mediated by the G domain of the laminin-α2 chain. *Cell*, 1997: **88:** 811–821.
[34] 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.
[35] Rambukkana A. Molecular basis for the peripheral nerve predilection of *Mycobacterium leprae*. *Curr Opin Microbiol*, 2001: **4:** 21–27.
[36] 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.

[37] Arruda S, Bomfim G, Knights R *et al.* Cloning of an *M. tuberculosis* DNA fragment associated with entry and survival inside cells. *Science*, 1993: **261:** 1454–1457.
[38] Chitale S, Ehrt S, Kawamura I *et al.* Recombinant *Mycobacterium tuberculosis* protein associated with mammalian cell entry. *Cell Microbiol*, 2001: **3:** 247–254.
[39] Makui H, Roig E, Cole ST *et al.* Identification of the *Escherichia coli* K-12 Nramp orthologue (MntH) as a selective divalent metal ion transporter. *Mol Microbiol*, 2000: **35:** 1065–1078.
[40] Wolf YI, Aravind L, Grishin NV, Koonin EV. Evolution of amino-acyl-tRNA synthetases – analysis of unique domain architectures and phylogenetic trees reveals a complex history of horizontal gene transfer events. *Genome Res*, 1999: **9:** 689–710.