The integrated genome map of *Mycobacterium leprae*

KARIN EIGLMEIER*, SYLVIE SIMON, THIERRY GARNIER & STEWART T. COLE Unité de Génétique Moléculaire Bactérienne, Institut Pasteur, 28, rue du Dr Roux, 75724 Paris, Cedex 15, France

Summary The integrated map of the Mycobacterium leprae genome unveiled for the first time the genomic organization of this obligate intracellular parasite. Selected cosmid clones, isolated from a genomic library created in the cosmid vector Lorist6, were identified as representing nearly the complete genome and were subsequently used in the *M. leprae* genome sequencing project. Now a new version of the integrated map of *M. leprae* can be presented, combining the mapping results from the Lorist6 cosmids with data obtained from a second genomic library constructed in an *Escherichia coli*-mycobacterium shuttle cosmid, pYUB18. More than 98% of the *M. leprae* genome is now covered by overlapping large insert genomic clones representing a renewable source of well defined DNA segments and a powerful tool for functional genomics.

Introduction

When Gerhard Henrik Armauer Hansen established in 1873 the contagious nature of leprosy, and thereby associated for the first time bacteria with a disease in man, he probably did not envisage that over 100 years later leprosy, caused by *Mycobacterium leprae*, would still be an important health problem in several countries. Only in the last 2 decades of the 20th century, due to the leprosy elimination strategy recommended by the WHO^{1,2} and the widespread implementation of multidrug therapy (MDT), there was a substantial reduction of leprosy prevalence achieved, resulting in an estimated 1–2 million registered cases. Effective MDT kills the bacteria; however, it does not reverse or cure the nerve damage in the afflicted patients. For this reason, some of the current major leprosy research priorities are the development of new diagnostic tools for the detection of early leprosy cases, new methods for detecting drug resistance or persistence, and the design of new drugs with increased effectiveness, capable of shortening the duration of antileprosy chemotherapy and which are operationally less demanding. To prevent the onset of infection by developing a leprosy vaccine could also contribute towards leprosy elimination efforts. The wide-spread administration of BCG, *Mycobacterium bovis* Bacille Calmette Guérin, is probably already

^{*} Corresponding author: Karin Eiglmeier. Mailing address: Unité de Génétique Moléculaire Bactérienne, Institut Pasteur, 28, rue du Dr Roux, 75724 Paris, Cedex 15, France. Phone: 33-1-45 68 84 49, Fax: 33-1-40 61 35 83, E-mail: kei@pasteur.fr

contributing to the reduction of leprosy prevalence, offering a significant degree of protection from clinical leprosy.^{3,4} The characterization of protective antigens from M. *leprae* to be included in a prophylactic vaccine and the understanding of its unusual life style, particularly its ability to enter and live inside Schwann cells, are research areas where many questions remain unanswered.

Research on *M. leprae*—the challenge

M. leprae, is an obligate intracellular parasite which has, so far, evaded cultivation *in vitro* outside of living animals. The development of the mouse model for leprosy research⁵⁻⁸ was the first step towards the study of *M. leprae* in the laboratory, and is still in use for testing drug susceptibility and monitoring of treatment. The introduction of the armadillo (*Dasypus noveminctus* Linnaeus) as an animal model⁹ made it possible for the first time to prepare a sufficient quantity of bacteria to envisage biochemical or genetic studies of *M. leprae*. Together with the rapidly developing recombinant DNA technologies, this represented a watershed for leprosy research. Preparation of chromosomal *M. leprae* DNA and first estimations of its G+C content and size became possible with the availability of armadillo-derived bacteria.¹⁰ The values obtained (G+C 56%, genome size of about 3.38 Mb) were ultimately shown to be very close to the data generated by the *M. leprae* sequencing project (57.8%, 3,268 Mb) several years later.¹¹

The λ gt11 period in *M. leprae* research and the generation of other genomic libraries

The advent of recombinant DNA technologies opened new avenues for leprosy research by analysis of gene products expressed in a cultivable bacterium, *Escherichia coli*. In the bacteriophage λ gt11 system, a relatively small (2–8 kb) foreign insert DNA is expressed under the control of the *E. coli* β -galactosidase promoter, therefore ensuring that the cloned sequence will be efficiently transcribed and translated in *E. coli*, generating a β -galactosidase fusion protein. Armadillo-derived genomic DNA was utilized in λ gt11 expression libraries and many investigators successfully identified recombinant clones producing antigens recognized by a panel of monoclonal antibodies or patient's sera^{12–19}; see references ²⁰ and ²¹ for a compilation. These antigenic determinants were subsequently characterized; however, it became evident that despite using different murine and human sera to screen the libraries, a limited number of immunodominant protein antigens was identified,²² and the same antigen was sometimes independently re-isolated in different laboratories,²⁰ suggesting an element of bias in the system used. In addition, the size limitation in the λ gt11 cloning system, while ideal for the detection of small antigenic epitopes, was not suitable for the analysis of full length genes.

To date, all attempts to obtain high molecular weight chromosomal DNA isolated from armadillo-derived *M. leprae* have been unsuccessful. The strong physical barrier of the cell envelope is one explanation, necessitating a relatively rough treatment to obtain permeabilization and preventing *in situ* preparations of *M. leprae* DNA suitable for pulsed-field gel electrophoresis (PFGE). The second reason is that *M. leprae* isolated from an animal host is a heterogenous population of bacilli with many cells probably harbouring already degraded DNA. This limits the maximum insert length to 40–60 kb. Chromosomal DNA fragments

464 K. Eiglmeier et al.

of this size can be prepared in sufficient quantities from armadillo-derived bacteria. The technique used to manipulate DNA of this size is the cosmid cloning strategy. However, because of the relatively small size of the starting material, the conventional partial digest methods employed for library construction are not always appropriate due to the risk of obtaining chimeric inserts. Nevertheless, investigators subjected *M. leprae* chromosomal DNA to partial restriction enzyme digestion and subsequently utilized size fractionated DNA fragments for the construction of different types of large insert libraries.¹⁰

The integrated map of M. leprae

With the aim of centralizing and unifying the *M. leprae* data and creating a new starting point for leprosy research, an integrated genome map was initiated, with the first step consisting of the construction of an ordered cosmid library. No M. leprae strain has ever been cloned, strictly speaking, so working with a clinical isolate seemed to be an acceptable solution. Thus M. leprae, originally derived from a patient from Tamil Nadu, was purified from an experimentally infected armadillo and the chromosomal DNA was prepared as described by Clark-Curtiss et al.¹⁰ In order to circumvent the problem of chimeric clones, the DNA was gel-fractionated, end-repaired and blunt-end fragments in the 35-50kb size range were cloned directly into the cosmid vector Lorist6.²³ About 1000 independent clones harbouring M. leprae DNA were obtained, corresponding to about 10 genome equivalents, and the cosmids were subjected to a modified fingerprint analysis.²⁴ Briefly, clones are 'fingerprinted' by treatment with restriction enzymes, measurement of the sizes of all the resulting fragments, followed by employing a computer-assisted matching technique²⁵ to detect regions common to several clones. The fingerprint data permit overlapping cosmids to be identified, enabling a contiguous physical map of the *M. leprae* chromosome to be built up of overlapping clones.²⁶

This contig map was refined and completed by hybridization, either using complete cosmids or suitable end-fragments of cosmids as probes. The understanding of the basic organization of the M. *leprae* chromosome was further improved by the successful positioning of all the available cloned M. *leprae* genes and loci, along with well-conserved, heterologous genes from other bacteria (such as housekeeping genes or stable RNAs). The ultimate comparison of fingerprinting data and hybridization results with genetic markers finally resulted in the integrated map of the M. *leprae* chromosome.

Based on this integrated map, the chromosome was estimated to be ~ 2.8 Mb in size. Seven years later the *M*. *leprae* sequencing project revealed a genome of ≈ 3.2 Mb for the leprosy bacillus,¹¹ indicating that the data from the fingerprint analysis had led to the underestimation of the contig sizes, due to paucity of bands in the fingerprint or different restriction fragments having similar lengths and hence appearing to be the same fragment. Nevertheless, the overall clone and gene distribution on the integrated map was left nearly unchanged and the hypothesis of a circular chromosome was confirmed.

First genome comparisons of *M. leprae* with *M. tuberculosis*²⁷ based on the genetic maps were within reach, revealing existing regions of synteny. However, it rapidly became evident that the order of these limited genome sections was different, observations which were later substantiated by the corresponding genome sequencing projects of *M. leprae* and *M. tuberculosis*.

Shuttle vector libraries

A prerequisite for studying proteins produced by mycobacterial pathways or those whose expression is catalysed by the products of other mycobacterial genes is the introduction of their genes in an appropriate surrogate host such as Mycobacterium smegmatis or M. bovis BCG where faithful gene expression can be obtained. The replicative shuttle cosmid pYUB18,²⁸ which contains an origin of replication functional in *E. coli* and a second, mycobacterial origin of replication for stable maintenance in mycobacteria, proved to be very useful for this purpose. As a selective marker pYUB18 carries a kanamycin resistance gene, functional in both E. coli and mycobacteria; however, due to the size of the vector, slightly smaller chromosomal DNA inserts can be cloned (between 30 and 35 kb). In order to pave the way for new functional genomic experiments, we wanted to determine a minimum tiling path (i.e. the combination of cosmids which gave the minimum overlap) of M. leprae shuttle cosmid clones representing the entire genome. Having the complete M. leprae genome sequence at our disposal, the positioning of M. leprae shuttle clones by cosmid end-sequencing became feasible as a direct mapping approach. About 800 shuttle cosmid clones from a pYUB18-based library (obtained from W. Jacobs Jr), generated from a different source of chromosomal M. leprae DNA, were characterized by end-sequencing and positioned on the *M. leprae* genome ; however, only incomplete coverage with shuttle cosmids was obtained. A high number of chimeric clones were identified and eliminated, due to an apparent discrepancy with the genome sequence. Additional proof for the chimeric nature of these shuttle cosmids was generated by comparison of their fingerprint data with the complete cosmid fingerprint dataset.

These comparisons also reinforced the initial assumption, that *M. leprae* strains from different origins exhibit no obvious, important genome diversity,^{10,29-31} a hypothesis that was also strengthened by sequence comparisons of the generated end-sequences with the genome sequence. However, the previously reported polymorphisms, the RLEP distribution in the genome³² and the variations in the TTC repeats,³³ have not yet been exploited in the pYUB18 based library.

The integration of the Lorist6-based cosmids and the pYUB18 shuttle cosmid clones yielded a new integrated map, which, as in the first version, still exhibits the four gaps in the library coverage creating contig breaks. Comparisons with the annotated *M. leprae* genome reveal gap sizes of approximately 2.9, 7.5, 5.1 and 42.3 kb, corresponding to about 1.7% of the genome. Table 1 summarizes the open reading frames which are absent from the two cosmid libraries; in total 34 possible ORFs are concerned, comprising only 19 genes and 10 conserved hypotheticals. Why these regions are not present in the two different libraries is unclear. There could be, as in other similar genome projects,^{34–37} *M. leprae* sequences that cannot be cloned in *E. coli* using these vectors. Another explanation could be the quality of the DNA preparation used as starter material for the library construction in that these regions might be underrepresented in the cloned, size fractionated region and more clones should be specifically screened for the presence of these missing sequences.

Conclusions and future studies

These ordered libraries are valuable tools and will facilitate genetic research considerably as they represent renewable sources of well-defined segments of the *M. leprae* genome, easy to distribute and convenient to handle in a non-pathogenic, cultivable host bacterium. They

Coordinates (nts)		Genes	Conserved hypotheticals	Hypotheticals	Pseudogenes
174 526-177 499	ML0131 fadD29	- oxidoreductase - acyl-CoA synthetase	1	T	_
254 015-261 597	ML0192	- membrane protein	2	1	4
655 781-660 912	PPE mihF gmk dfp metK	 PPE-family protein integration host factor guanylate kinase flavoprotein S-adenosylmethionine synthase 	1	_	-
911 512–953 873	sahH tmk mtrA mtrB lpqB secA ML0782 aroA ML0793 ML0803 whiBl	 S-adenosyl-L-homocysteine hydrolase thymidylate kinase two-component response regulator two-component system sensor kinase lipoprotein preprotein translocase subunit transmembrane transport protein 3-phosphoshikimate 1-carboxyvinyl transferase bacteriophage protein two-component system sensor kinase putoting transpirate page. 	6	4	17
Total	whiD1	19	10	5	21

Table 1. Regions of the *M. leprae* chromosome absent from the Lorist6 and also from the pYUB 18 genomic libraries. The first column indicates the limits, corresponding to the sequenced *M. leprae* genome, of the regions absent in the two mapped libraries. In column two the putative genes present in these regions are given. Columns three and four indicate the number of conserved hypotheticals and hypothetical genes present in these uncloned regions and the fifth column contains the number of identified pseudogenes

remove the tedious step of preparation of mycobacterial chromosomal DNA and, in addition, the insert size permits investigations of longer operons or genes with related functions, often linked on the chromosome.

Nowadays, complete bacterial genome sequences are relatively easy to produce without prior mapping of clones with large inserts or the establishment of a complete integrated map. However, sequencing projects concerning species necessitating laborious culture conditions, biohazard facilities or with complex genomes which make the sequencing approach by a whole genome shotgun difficult due to their special features (for example, extensive, numerous repetitive regions or DNA segments which are difficult to clone) benefit from clones carrying defined DNA fragments and a detailed cosmid map, serving as a framework for the assembly. Selected clones represented ideal starting material for the systematic genome sequencing, giving early access to thoroughly analysed and annotated regions of the chromosome, long before the genome was completely sequenced, as for example, the first completely sequenced cosmid, B1790.³⁸ Sequence analysis of this cosmid revealed the presence of the genes for four potential drug targets: rpoB (rifampicin), tuf (tetracyclin derivatives), rpsL (streptomycin) and efg (fusidic acid). These results permitted the early development of molecular diagnostic tests for determining rifampicin resistance in M. $leprae^{39}$ which, refined further, culminated in a relatively simple assay for resistance detection⁴⁰ (Honoré et al., this issue). In addition, benefiting from the sequence data obtained



Figure 1. Latest version of the integrated map of *M. leprae*. A minimal set of Lorist6-based/PYUB18 shuttle vector clones is shown, covering in 4 contigs more than 98% of the *M. leprae* genome. The central black line corresponds to the genome sequence, with the positions indicated in nucleotides. Below the sequence line, the lower bars represent the sections of the chromosome present in completely sequenced clones of the Lorist6-based library. The upper bar indicates the genome sections represented in cosmids of the pYUB18-based library. The exact coordinates of the selected clones can be obtained at the following web site: http://genolist.pasteur.fr/Leproma.

from the evolutionary well-conserved rpsL gene of M. leprae, the molecular basis of streptomycin resistance in *M. tuberculosis*⁴¹ could be determined.

The resources described here also provide a new foundation for functional genomics. Comparisons of the genomes of *M. leprae* and *M. tuberculosis* reveal a number of genes common to both species which could be characteristic for mycobacteria, plus another group of genes which are specific to *M. leprae*. Capitalizing on the contig map plus the genome sequence and its annotation, defined shuttle vector cosmids could be used to introduce, in a straightforward way, these M. leprae specific genes into fast-growing mycobacteria allowing M. leprae proteins to be expressed and studied. Clearly comparative mycobacterial genomics in combination with the well-defined *M. leprae* cosmids can open new and systematic ways of searching for vaccine candidates, immunodiagnostic reagents and molecules involved in the pathogenesis of *M. leprae*.

Acknowledgements

We wish to acknowledge the financial support of the Institut Pasteur, the Association Française Raoul Follereau, and the National Institutes of Health, National Institute of Allergy and Infectious Diseases (grant no. 1 RO1 AI47197-01A1).

References

- ¹ WHO Study group. Chemotherapy of leprosy for control programmes. World Health Organization, Geneva, Switzerland, 1982, Tech. Rep. Ser. 675.
- ² WHO. Action programme for the elimination of leprosy. World Health Organization, Geneva, Switzerland, 1998, Status Rep. 98.2.
- ³ Fine PEM. BCG vaccination against tuberculosis and leprosy. Br Med Bull, 1988; 44: 691–703.
- ⁴ Fine PEM, Clayton D, Ponnighaus JM, Warndorf DK. Randomized controlled trial of single BCG, repeated BCG, or combined BCG and killed Mycobacterium leprae vaccine for prevention of leprosy and tuberculosis in Malawi. Lancet, 1996; 348: 17-24.
- ⁵ Shepard CC. The experimental disease that follows the injection of human leprosy bacilli into foot pads of mice. J Exp Med, 1960; 112: 445-454.
- ⁶ Rees RJ, Waters MF, Weddell AG, Palmer E. Experimental lepromatous leprosy. *Nature*, 1967; **215:** 599–602.
- ⁷ Rees RJ, Weddell AG. Biology of the mycobacterioses: experimental models for studying leprosy. Ann NY Acad Sci, 1968; 154: 214-236.
- ⁸ Lancaster RD, Hilson GR, McDougall AC, Colston MJ. Mycobacterium leprae infection in nude mice: bacteriological and histological responses to primary infection and large inocule. Infect Immun, 1983; 39: 865-872.
- ⁹ Kirchheimer WF, 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.
- ¹⁰ Clark-Curtiss JE, Jacobs WR, Docherty MA et al. Molecular analysis of DNA and construction of genomic libraries of Mycobacterium leprae. J Bacteriol, 1985; 161: 1093-1102.
- ¹¹ Cole S, Eiglmeier K, Parkhill J et al. Massive gene decay in the leprosy bacillus. Nature, 2001; 409: 1007–1011.
- ¹² Young RA, Mehra V, Sweetser D et al. Genes for the major protein antigens of the leprosy parasite Mycobacterium leprae. Nature, 1985; **316:** 450–452. ¹³ Young RA, Bloom BR, Grosskinsky CM *et al.* Dissection of Mycobacterium tuberculosis antigens using
- recombinant DNA. Proc Natl Acad Sci USA, 1985; 82: 2583-2587.
- ¹⁴ Cherayil BJ, Young RA. A 28-kDa protein from *Mycobacterium leprae* is a target of the human antibody response in lepromatous leprosy. J Immunol, 1988; 141: 4370-4375.
- ¹⁵ Young D, Lathigra R, Hendrix R et al. Stress proteins are immune targets in leprosy and tuberculosis. Proc Natl Acad Sci USA, 1988; 85: 4267-4270.
- ¹⁶ Sathish M, Esser RE, Thole JER, Clark-Curtiss JE. Identification and characterization of antigenic determinants

of *Mycobacterium leprae* that react with antibodies in sera from leprosy patients. *Infect Immun*, 1990; 58: 1327-1336.

- ¹⁷ Thole JE, Schoningh R, Janson AA et al. Molecular and immunological analysis of a fibronectin-binding protein antigen secreted by Mycobacterium leprae. Mol Microbiol, 1992; 6: 153–163.
- ¹⁸ Mehra V, Sweetser D, Young RA. Efficient mapping of protein antigenic determinants. *Proc Natl Acad Sci USA*, 1986; 83: 7013–7017.
- ¹⁹ Laal S, Sharma YD, Prasad HK *et al.* Recombinant fusion protein identified by lepromatous sera mimics native *Mycobacterium leprae* in T-cell responses across the leprosy spectrum. *Proc Natl Acad Sci USA*, 1991; 88: 1054–1058.
 ²⁰ V. D. D. M. C. D. D. M. C. D. D. M. C. D. M. M. C. D. M. C. D.
- ²⁰ Young DB, Kaufmann SHE, Hermans PWM, Thole JER. Mycobacterial protein antigens: a compilation. *Mol Microbiol*, 1992; 6: 133-145.
- ²¹ Thole JE, Wieles B, Clark-Curtiss JE et al. Immunological and functional characterization of Mycobacterium leprae protein antigens: an overview. Mol Microbiol, 1995; 18: 791-800.
- ²² Khanolkar-Young S, Kolk AHJ, Andersen AB et al. Results of the Third Immunology of Leprosy/Immunology of Tuberculosis Antimycobacterial Monoclonal Antibody Workshop. Infect Immun, 1992; 60: 3925–3927.
- ²³ Gibson TJ, Rosenthal A, Waterson RW. Lorist 6, a cosmid vector with BamHI, Notl, Scal and HindIII cloning sites and altered neomycin phosphotransferase gene expression. Gene, 1987; 53: 283-286.
- ²⁴ Coulson A, Sulston J, Brenner S, Karn J. Toward a physical map of the genome of the nematode *Caenorhabditis elegans*. Proc Natl Acad Sci USA, 1986; 83: 7821-7825.
- ²⁵ Sulston J, Mallett F, Staden R *et al.* Software for genome mapping by fingerprinting techniques. *CABIOS*, 1988; 4: 125–132.
- ²⁶ Eiglmeier K, Honoré N, Woods SA et al. Use of an ordered cosmid library to deduce the genomic organization of Mycobacterium leprae. Mol Microbiol, 1993; 7: 197-206.
- ²⁷ Philipp W, Poulet S, Eiglmeier K et al. An integrated map of the genome of the tubercle bacillus Mycobacterium tuberculosis H37Rv, and comparison with Mycobacterium leprae. Proc Natl Acad Sci USA, 1996; 93: 3132– 3137.
- ²⁸ Jacobs Jr WR, Kalpana GV, Cirillo JD et al. Genetic systems for mycobacteria. Methods Enzymol, 1991; 204: 537-555.
- ²⁹ Clark-Curtiss JE, Walsh GP. Conservation of genomic sequences among isolates of *Mycobacterium leprae*. J Bacteriol, 1989; 171: 4844-4851.
- ³⁰ Sela S, Clark-Curtiss JE, Bercovier H. Characterization and taxonomic implications of the rRNA genes of Mycobacterium leprae. J Bacteriol, 1989; 171: 70-73.
- ³¹ 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.
- ³² Fsihi H, Cole ST. The Mycobacterium leprae genome: systematic sequence analysis identifies key catabolic enzymes, ATP-dependent transport systems and a novel polA locus associated with genomic variability. Mol Microbiol, 1995; 16: 909-919.
- ³³ Shin Y-C, Lee H, Lee H et al. Variable numbers of TTC repeats in Mycobacterium leprae DNA from leprosy patients and use in strain differentiation. J Clin Microbiol, 2000; **38**: 4535–4538.
- ³⁴ Kohara Y, Akiyama K, Isono K. The physical map of the whole *E. coli* chromosome: application of a new strategy for rapid analysis and sorting of a large genomic library. *Cell*, 1987; **50**: 495–508.
- ³⁵ Knott V, Blake DJ, Brownlee G. Completion of the detailed restriction map of the *E. coli* genome by the isolation of overlapping cosmid clones. *Nucl Acids Res*, 1989; **17:** 5901–5912.
- ³⁶ Birkenbihl RP, Vielmetter W. Cosmid derived map of *E. coli* strain BHB2600 in comparison to the map of strain W3110. *Nucl Acids Res*, 1989; **17**: 5057–5069.
- ³⁷ Charlebois RL. Detailed physical map and set of overlapping clones covering the genome of the archebacterium Halofax volcanii DS2. J Mol Biol, 1991; 222: 509–524.
- ³⁸ 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.
- ³⁹ Honoré N, Perrani E, Telenti A et al. A simple and rapid technique for the detection of rifampin resistance in Mycobacterium leprae. Int J Lepr Other Mycobact Dis, 1993; 61: 600-604.
- ⁴⁰ Roche PW, Shrestha N, Thomas A et al. Rapid detection of resistance to rifampicin in Mycobacterium leprae. Lepr Rev, 2000; 71: S96-97.
- ⁴¹ Honoré N, Cole ST. Streptomycin resistance in mycobacteria. Antimicrob Agents Chemother, 1994; 38: 238-242.