Proc. Natl. Acad. Sci. USA
Vol. 93, pp. 3132–3137, April 1996
Microbiology

An integrated map of the genome of the tubercle bacillus, *Mycobacterium tuberculosis* H37Rv, and comparison with *Mycobacterium leprae*

(genome mapping/contig mapping/ordered libraries/bacterial genomics/tuberculosis)

WOLFGANG J. PHILIPP*, SYLVIE POULET*, KARIN EIGLMEIER*, LISA PASCOPELLA†, V. BALASUBRAMANIAN†, BEATE HEYM*, STAFFAN BERGH*‡, BARRY R. BLOOM†, WILLIAM R. JACOBS, JR.†, AND STEWART T. COLE*

*Unité de Génétique Moléculaire Bactérienne, Institut Pasteur, 28 rue du Docteur Roux, 75724 Paris Cedex 15, France; †Howard Hughes Medical Institute, Department of Microbiology and Immunology, Albert Einstein College of Medicine, New York, NY 10461; and ‡Department of Biochemistry and Biotechnology, Royal Institute of Technology, 100 44 Stockholm, Sweden

Contributed by Barry R. Bloom, December 12, 1995

ABSTRACT **An integrated map of the genome of the tubercle bacillus, *Mycobacterium tuberculosis*, was constructed by using a twin-pronged approach. Pulsed-field gel electrophoretic analysis enabled cleavage sites for *Asn* I and *Dra* I to be positioned on the 4.4-Mb circular chromosome, while, in parallel, clones from two cosmid libraries were ordered into contigs by means of fingerprinting and hybridization mapping. The resultant contig map was readily correlated with the physical map of the genome via the landmarked restriction sites. Over 165 genes and markers were localized on the integrated map, thus enabling comparisons with the leprosy bacillus, *Mycobacterium leprae*, to be undertaken. Mycobacterial genomes appear to have evolved as mosaic structures since extended segments with conserved gene order and organization are interspersed with different flanking regions. Repetitive sequences and insertion elements are highly abundant in *M. tuberculosis*, but the distribution of IS*6110* is apparently nonrandom.**

In spite of the availability of effective short-course chemotherapy, and the bacillus Calmette–Guérin (BCG) vaccine, *Mycobacterium tuberculosis* still accounts for more deaths worldwide than any other single infectious agent (1). Recent increases in tuberculosis in both developing and industrialized countries, together with the emergence of drug-resistant strains and synergy with the human immunodeficiency virus (HIV) pandemic, have combined to raise considerable public concern and to highlight the need for radical improvements in control strategies (2). The development, improvement, and use of genetic tools for mycobacteria lie at the center of current research programs (3, 4). Paradoxically, in the last few years, there has been a quantum jump in our understanding of the related leprosy bacillus, *Mycobacterium leprae*, as a result of the application of genome research and systematic DNA sequence analysis (5), and this has opened new avenues for research in immunology, therapeutics, and drug development. A similar approach for *M. tuberculosis* was thus urgently required.

The specific objectives of this project were to construct, characterize, and maintain ordered clone libraries corresponding to the chromosome of *M. tuberculosis* and to establish a contig map on which the positions of all known genes and markers were established. In parallel, a physical map of the genome was determined by means of pulsed-field gel electrophoresis (PFGE) of macro-restriction fragments, and this was correlated with the contig map to produce an integrated map (6) suitable for dissemination through the dedicated mycobacterial database, MycDB (7).

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

MATERIALS AND METHODS

Analysis of *M. tuberculosis* DNA by PFGE. To prepare chromosomal DNA suitable for PFGE, *M. tuberculosis* strain H37Rv was inoculated into Dubos medium (Pasteur Diagnostics) supplemented with OADC (oleic acid, albumin, dextrose, catalase) (Difco), and incubated for 10 d at 37°C. In some cases D-cycloserine was then added (1 mg/ml) and incubation continued for an additional 24 h prior to cell harvesting, enclosure in low melting point agarose (GIBCO/BRL), and further processing to release intact genomic DNA, as described previously (8).

Samples were digested with restriction endonucleases and analyzed on a clamped homogeneous electric field (CHEF) apparatus (LKB 2015 Pulsaphor Plus) using 1.2% agarose gels (in 0.5× TBE; 1× TBE = 89 mM Tris/89 mM boric acid/2 mM EDTA) as described (8, 9). To separate fragments <100 kb in size, the run time was 18 h with a pulse time of 3 s and a voltage of 270 V; for fragments >100 kb in size, the run time was extended to 21 h with a 10-s pulse time; whereas for very large fragments (>1 Mb) the run time was 48 h with a pulse time of 30 s. Gels were calibrated by using concatemerized λ or yeast chromosomes (*Saccharomyces cerevisiae* YPH80; 225–1,900 kb; New England Biolabs; or *Schizosaccharomyces pombe*, 3–6 Mb) then processed for Southern blotting onto Hybond C-extra membranes (Amersham Plc), as outlined (8).

To isolate linking clones carrying rare restriction sites, 400 cosmids from the pYUB18 library were screened by restriction digestion for the presence of *Asn* I or *Dra* I sites. Two-dimensional PFGE-analysis of reciprocal *Asn* I and *Dra* I digests was performed exactly as described in ref. 10.

Cosmid Library Construction. DNA was extracted from *M. tuberculosis* strain H37Rv, then subjected to partial digestion with *Sau*3AI, as described previously (3). Fragments of 30–45 kb were obtained after fractionation on a 0.4% agarose gel, then ligated to either dephosphorylated, *Bam*HI-cleaved pYUB18 (3) or later to pYUB328 (11). After *in vitro* packaging with Gigapack Gold II extracts (Stratagene), the pYUB18 recombinant cosmids were introduced into the *Escherichia coli* K-12 strain χ 2819 (12) and packaged *in vivo* to produce a high titer cosmid lysate. Aliquots were subsequently used to infect *E. coli* strain NM554 (13) and maintained either as gridded arrays of colonies or stored frozen at −80°C in microtiter plates containing 15% (vol/vol) glycerol. The pYUB328 cosmid library was constructed as described, except that no *in vivo* packaging was performed prior to infection and clone preparation.

Fingerprinting. Cosmid miniprep DNA (≈3 μg) was obtained from 2 ml overnight cultures and subjected to fingerprint analysis with *Eco*RI and *Alu* I as described (14). Samples were then lyophilized, denatured by heating, and loaded onto a 6% dena-

Abbreviation: PFGE, pulsed-field gel electrophoresis.

turing polyacrylamide gel calibrated with end-labeled *Sau*3AI fragments of λ DNA (15). Fingerprints were analyzed, and overlapping cosmids were organized into contigs with the program CONTIG9 (16).

Hybridization Mapping, Data Handling and Analysis. To facilitate gene mapping and the construction of contigs by hybridization mapping (17), arrays of cosmids were dot blotted onto Hybond N membranes and processed as recommended by Amersham, Plc. The probes were macrorestriction fragments isolated from pulsed-field gels and labeled in the gel slice, whole cosmids, repetitive sequences, or single-copy probes produced either from cloned genes or by PCR amplification of genomic DNA (18).

Hybridization was performed with ^{32}P-labeled probes, produced by nick translation or random priming, under stringent conditions [16 h at 37°C in 50% (vol/vol) formamide/5× SSC (1× SSC = 0.15 M NaCl/0.015 M sodium citrate) (19)], except when using heterologous probes (30°C in the same hybridization solution but with washes in 2× SSC instead of 0.1× SSC). Oligonucleotide probes were labeled by using terminal deoxynucleotidyl transferase and [α-^{32}P]dCTP and hybridized as described (20).

After autoradiography, the hybridization signals were scored manually then entered into input files (clone.file and probe.file) prior to analysis with the programs of Mott *et al.* (21) on a Sun Sparc II workstation. All hybridizations were run through PROBEORDER, which orders probes and clones. The resultant hybridization contigs were subsequently reanalyzed using REORDER, and in some cases additional joins and verifications were done by referring to the fingerprinting dataset constructed by CONTIG9 (16). The consensus contigs were then correlated with the PFGE map by using the *Dra* I and *Asn* I linking clones as landmarks.

RESULTS

PFGE Analysis of Genomic DNA from *M. tuberculosis*. Generally, the genomes of microorganisms with dG+dC-rich chromosomes like the tubercle bacillus are readily cleaved into a limited number of fragments by restriction endonucleases with dA+dT-rich recognition sequences, as these sites are underrepresented. On screening a battery of restriction enzymes, it was found that none of the enzymes with 8-bp recognition sites cut the chromosome and that the majority of the enzymes with 6-bp recognition sites (*Cla* I, *Hpa* I, *Nde* I, *Ssp* I, *Xba* I, etc.) yielded more than 50 fragments and were thus of limited use for map construction. By contrast, digestion of DNA from strain H37Rv with *Dra* I (TTTAAA) or *Asn* I (ATTAAT) gave a manageable number of fragments, 35 and 47, respectively (Fig. 1). Sixteen of the *Dra* I sites were associated with IS*6110*, as this insertion element contains a unique *Dra* I site (22, 23).

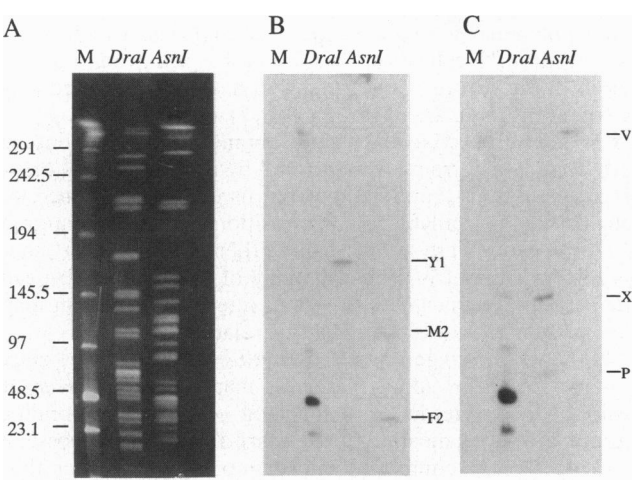

FIG. 1. PFGE analysis of *M. tuberculosis* H37Rv. (*A*) Ethidium bromide stained gel of DNA from H37Rv digested with *Asn* I or *Dra* I. (*B*) Southern blot of gel shown in *A* hybridized with *Asn* I linking clone T227. (*C*) Same blot as in *B* hybridized with *Dra* I linking clone T551. The size markers (in kb) and relevant fragments (see Table 1) are indicated. M, marker.

Table 1. Restriction fragments of the chromosome of *M. tuberculosis* H37Rv

Dra I		*Asn* I	
Fragment	Size, kb	Fragment	Size, kb
Z7	580	V	700
Z6	475	U	340
Z5	300	T	240
Z4	260	S	235
Z3	230	R	180
Z2	230	Q	170
Z1	220	P	155
Y2	190	O3	137
Y1	190	O2	135
X	165	O1	133
W	160	N	130
V	140	M1/2	125 (×2)
U	125	L	115
T	120	K	105
S	95	J1/2	100 (×2)
R	87	I	78
Q1	87	H	73
Q2	87	G8	65
P	82	G5/6/7	63 (×3)
O	78	G4	61
N	72	G2/3	59 (×2)
M	65	G1	57
L	60	F1/2	50 (×2)
K	47	E4	47
J	47	E3	46
I	40	E2	45
H	33	E1	44
G	30	D	37
F	26	C4	33
E	23	C3	31
D	22	C1/2	30 (×2)
C	13	B6	18
B	7	B4/5	15 (×2)
A	6	B3	13
A1	2.2	B2	11
	4394.2	B1	9
		A4	4.5
		A3	4
		A2	3.5
		A1	3
			4405

To determine the exact number of fragments and obtain an estimate of the chromosome size, samples were analyzed under a variety of electorphoretic conditions and pulse times. *Dra* I digestion gave rise to 35 fragments ranging in size from 2.2 to 580 kb (Fig. 1 and Table 1), whereas *Asn* I generated 47 fragments (3–700 kb; Fig. 1 and Table 1), the majority of which were in the lower range (3–100 kb). The size of the genome was estimated by summing the fragment sizes, and, in both cases, this was about 4.4 Mb (Table 1).

Construction of a Physical Map. To establish the order and contiguity of the *Dra* I or *Asn* I restriction fragments on the chromosome, a pYUB18 cosmid library of *M. tuberculosis* H37Rv was screened for clones carrying *Dra* I or *Asn* I restriction sites. The 26 independent *Dra* I linking clones (3 of which carried two sites) that resulted from this screen were

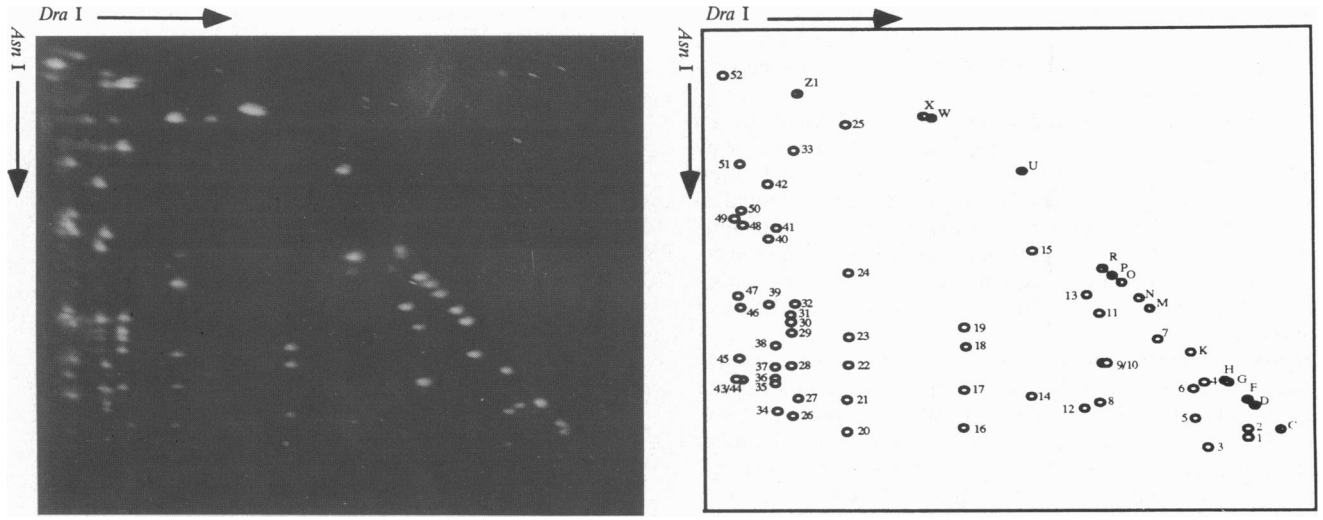

FIG. 2. Two-dimensional PFGE analysis of a representative *Dra* I (first dimension) and *Asn* I (second dimension) double digest of genomic DNA from *M. tuberculosis* H37Rv. (*Left*) Ethidium bromide stained gel. (*Right*) Schematic interpretation of gel shown in *Left*. Fragments were identified and labeled in accordance with Table 1 and correspond as follows: 1, dE/aC4; 2, dE/aM1; 3, dI/aB5; 4, dI/aM2; 5, dK/aC2; 6, dK/aL; 7, dL/aI; 8, dQ1/aT; 9/10, dQ2/aT/aG2; 11, dQ1/aQ; 12, dS/aG2; 13, dS/aR; 14, dT/aO1; 15, dT/aI; 16/17/18/19, dV/aB6/aH/aE3/aG7; 20, dY1/aB2; 21, dY1/aG6; 22, dY2/aD; 23, dY1/aF2; 24, dY1/aM2; 25, dY2/aP; 26, dZ1/aC4; 27, dZ2/aH; 28/29/30/31, dZ3/aO3/aF1/aG3/aG5; 32, dZ2/aG8; 33, dZ2/aO2; 34/35/36/37/38, dZ4/aB3/aC3/aJ1/aG6/aE2; 39, dZ5/aG1; 40, dZ5/aS; 41, ddZ4/aJ2; 42, dZ5/aS; 43, dZ7/aM1; 44, dZ6/aC1; 45, dZ6/aE4; 46, dZ6/aG4; 47, dZ7/aJ1; 48, dZ6/aV; 49, dZ7/aK; 50, dZ6/aS; 51, dZ6/aN; and 52, dZ7/aU. *Dra* I fragments running on the outermost diagonal lane (C, D, F, G, H, M, N, O, P, R, U, W, X) have no internal *Asn* I restriction sites. Fragments of <10 kb were detected independently.

used as probes on Southern blots of DNA digested with *Dra* I, and all revealed unambiguous linkage of adjacent *Dra* I fragments (Fig. 1*C*), thereby accounting for 29 of the 35 fragments. Similarly, linkage of many *Asn* I fragments was obtained by using some of the 34 *Asn* I linking clones as probes (Fig. 1*B*). To complete the map, various *Asn* I or *Dra* I restriction fragments were isolated and used as probes in reciprocal cross-hybridization experiments. Final confirmation of topology and fragment order was obtained by performing two-dimensional PFGE analysis (Fig. 2) of H37Rv DNA reciprocally digested with *Asn* I and *Dra* I (10).

This enabled all of the sites to be accounted for and revealed the presence of a single circular chromosome as found in most bacteria (24–26). To exclude the possibility that linear and circular forms might coexist, as in the related genus *Streptomyces* (27), undigested *M. tuberculosis* DNA was subjected to PFGE under conditions which resolved intact *Sch. pombe* chromosomes (data not shown). No bands in the 4-Mb range were observed, thus indicating that the chromosomal topology was predominantly circular.

Construction of Cosmid Libraries and a Contig Map. Initially, ≈970 of the pYUB18 clones were subjected to computer-assisted fingerprint analysis (16, 28), but this approach was later replaced by hybridization mapping (17) in which gridded arrays of cosmid DNAs were hybridized with suitable probes. The arrays consisted of 970 pYUB18 clones and 500 pYUB328 clones, and the probes were ≈150 genetic markers, 140 pYUB18 clones, and 300 pYUB328 clones. In some experiments, *Asn* I, *Dra* I, and *Xba* I restriction fragments, isolated by PFGE, were used as probes. Hybridization signals were recorded and analyzed with a suite of programs (21) which constructs contigs on the basis of linkage of two probes by means of a common clone.

This gave 16 "hybridization contigs," which showed generally good agreement with those obtained by fingerprinting, although a few of the contigs constructed by fingerprinting were found to be spurious due to chimeric clones or a paucity of bands in the fingerprint. As systematic gap closure has not yet been undertaken, it is possible that short overlaps between some contigs were not detected. To obtain map integration, the consensus contigs were then compared with the complete PFGE map by using the *Asn* I and *Dra* I linking clones as landmarks, and the current integrated map is depicted in Fig. 3.

A Gene Map. The genetic markers which have been precisely located on the integrated map fall into three groups: known genes or probes obtained by using PCR to amplify mycobacterial genes of conserved sequence as targets (18), sequences encoding protein antigens isolated from expression libraries by using antibodies, and repetitive DNA such as insertion elements (29, 30). There are >95 known genes (Table 2) and 20 loci encoding protein antigens recognized by monoclonal antibodies or sera from patients (data not shown). The insertion sequences which have been mapped are IS*6110* (22, 23), which is present in 16 copies, IS*1081* (31), present in six loci, and two copies of a recently identified IS-like element (32). At least 26 loci contain copies of the polymorphic G+C-rich repetitive sequence (PGRS; refs. 20 and 30), whereas the major polymorphic tandem repeat, MPTR, was so abundant that accurate localization could not be achieved. Further details of the nature of the genes that have been mapped may be found in Table 2 and in MycDB (7).

Comparison of Mycobacterial Genome Maps. The genome of *M. leprae* is currently represented by four contigs of organized cosmids (14), and ≈1.6 Mb of its genome sequence, or roughly half, is available. As the positions of several hundred *M. leprae* genes were known from either sequencing or mapping experiments (5–7) and as many of these had also been mapped in *M. tuberculosis*, it was of interest to perform map comparisons to assess the extent of relatedness.

Of the 95 known genes on the *M. tuberculosis* genome map, 61 have also been either physically mapped in *M. leprae* or positioned as a result of sequencing of the corresponding cosmid. However, meaningful comparisons were only possible for gene clusters containing two or more genes. It was thus established that the largest conserved stretch was a chromosomal region of ≈240 kb delimited by *dnaB* and *fbpA* that spans *oriC*, the chromosomal origin of replication. Likewise, a second cluster of over eight genes between *hemE* and *lexA* encoding two sigma factors and two members of the SOS regulon (RecA and LexA), is found in both cases (Fig. 3). At

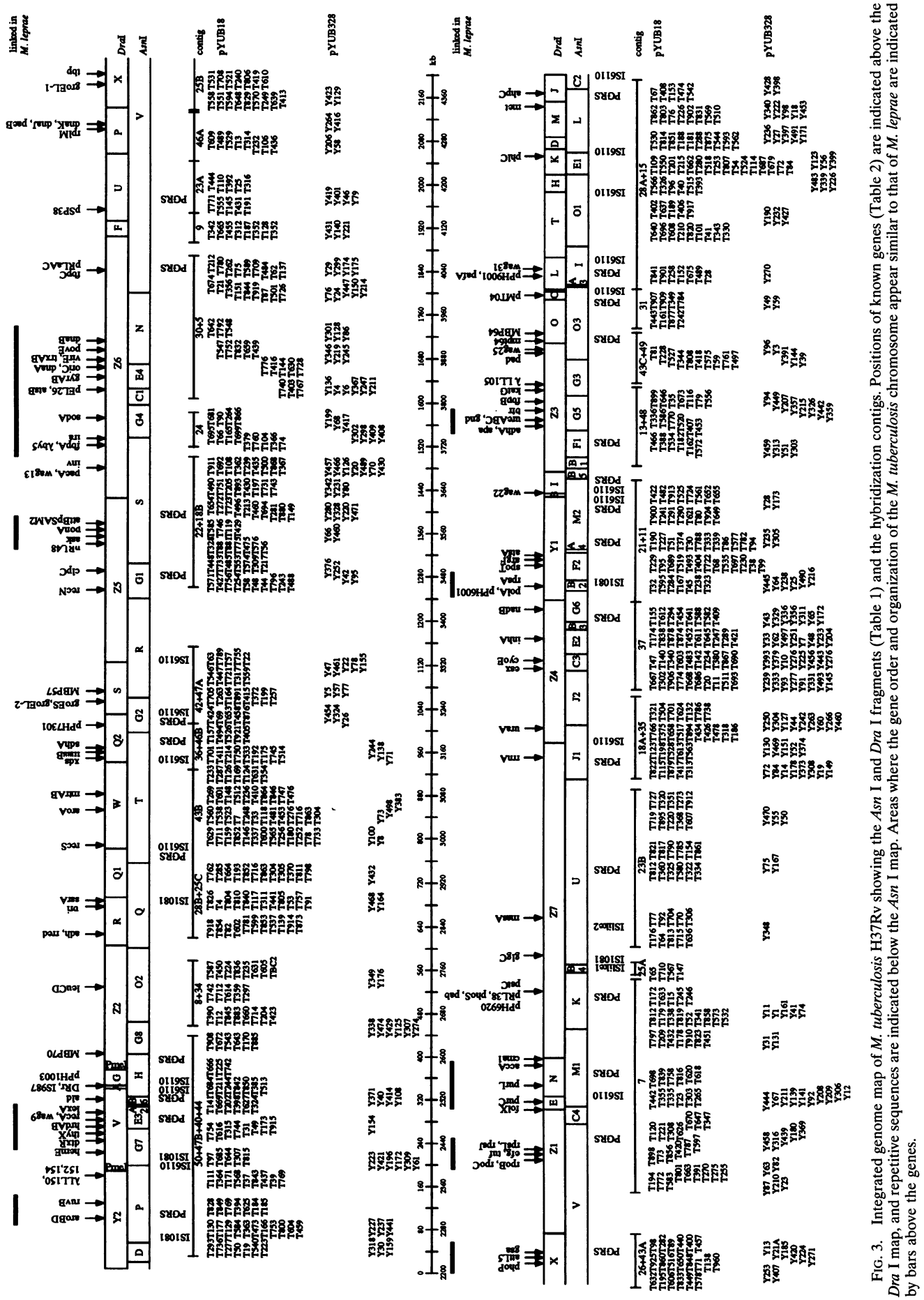

FIG. 3. Integrated genome map of *M. tuberculosis* H37Rv showing the *Asn* I and *Dra* I fragments (Table 1) and the hybridization contigs. Positions of known genes (Table 2) are indicated above the *Dra* I map, and repetitive sequences are indicated below the *Asn* I map. Areas where the gene order and organization of the *M. tuberculosis* chromosome appear similar to that of *M. leprae* are indicated by bars above the genes.

Table 2. Identity and source of known genetic markers mapped in *M. tuberculosis*

Locus	Canonical shuttle cosmid	Description of probe	Origin	Source/GenBank accession no.	Mapped in *M. leprae*
accA	T830	biotin binding protein	Mt	J. Dale/Z19549	+
adh	T854	alcohol dehydrogenase	Mb	J. Content/X63450	
adhA	T453	alcohol dehydrogenase	Mt	GMB	+
ahpC	T183	alkylhydroperoxide reductase	Mt	GMB/U18264	+
ald	T292	L-alanine dehydrogenase	Mt	Å. Andersen/X63069	+
apa	T407	secreted alanine-proline-rich antigen	Mt	A. Laqueyrerie	+
argF	T333	ornithine carbamoyl transferase	Mt	GMB	
aroA	T146	5-enolpyruvylshikimate-3-P synthase	Mt	D.B. Young/M62708	+
aroBD	T130	3-dehydroquinate synthase, 3-dehydroquinase	Mt	D.B. Young/X59509	+
ask	T149	aspartokinase	Mt	GMB	+
atsA	T535	AT10S gene	Mt	GMB/D17369	
atsB	T403	AT9S gene	Mt	GMB/D14355	
attB-L5	T606	tRNApro; tRNAgly; attachment site for L5	Ms	M65195	+
attB-pSAM2	T449	putative tRNApro; attachment site for pSAM2	Ml	GMB/X60720	+
bfr	T407	bacterioferritin	Mt	GMB	+
clpC	T485	proteolysis regulator; chaperone	Ml	GMB/M67510	+
cma1	T-/-	cycloprane mycolic acid synthase	Mt	GMB/U27357	
cyoE	T483	cytochrome O ubiquinol oxidase	Mt	GMB	+
dnaA	T776	initiation of DNA biosynthesis	Ml	H. E. Takiff/L39923	+
dnaB	T832	replication	Ml	G. Riccardi/L39923	+
dnaJ	T369	chaperone	Mt	D.B. Young/X58406	+
dnaK	T267	HSP-70, chaperone	Mt	D.B. Young/X58406	+
dtxR	T144	similar to diphtheria toxin repressor	Mt	I. Smith	+
efg	T663	elongation factor G	Ml	GMB/Z14314	+
esx	T174	ESAT6 T-cell antigen	Mt	B. Gicquel/X79562	+
fbpA	T366	fibronectin binding protein - 85A	Mt	J. Thole/M27016	+
fbpB	T336	" " " - 85B	Mt	J. Thole/X62398	+
fbpC	T464	" " " - 85C	Mt	J. Thole/X57229	+
folX	T597	putative dihydrofolate reductase	Mt	J. Dale/X59271	
glgC	T588	glucose-1-phosphate adenylytransferase	Mt	GMB	+
gnd	T407	6-phosphogluconate dehydrogenase	Mt	B. Gicquel	
groEL-1	T521	HSP-60; chaperone	Mt	D.B. Young/M15467	+
groEL-2	T721	HSP-60 homologue	Mt	GMB/X60350	+
groES	T721	HSP-12; chaperone	Mt	T. Shinnick/X60350	+
gsa	T440	glutamate-1-semialdehyde 2,1-aminot-ase	Mt	GMB	
gyrBA	T728	DNA gyrase, A and B subunits	Mt	H. E. Takiff/L27512	+
hemE	T685	uroporphyrinogen decarboxylase	Mt	GMB	+
hrdAB	T144	RNA polymerase sigma factors	Mt	GMB	+
inhA	T20	enoyl reductase	Mt	W.R. Jacobs/U02492	+
inv	T500	putative invasin homologue	Mt	S. Porter	
ira	T366	putative iron regulated antigen (28kD)	Mt	B. Gicquel	+
katG	T116	catalase-peroxidase	Mt	B. Heym/X68081	+
leuABCD	T224	isopropylmalate synthase, dehyrogenase, isomerase	Mt	W.R. Jacobs	
lexA	T616	SOS regulator	Mt	GMB	+
masA	T316	mycocerosic acid synthetase	Mt	GMB	+
masB	T192	putative mycocerosic acid synthetase isoenzyme	Mt	GMB	+
met	T902	?	Mt	W.R. Jacobs	
mtrAB	T148	response regulator	Mt	V. Deretic/U01971+U14909	
nadB	T174	quinolinate synthetase	Ml	GMB/U00010	+
nrdA	T918	ribonucleotide reductase	Mt	GMB/L34407	
oriC	T776	origin of replication	Mt	H. Schrempf	+
pab	T172	anonymous antigen AA59 probably PhoS	Mt	Å. Andersen/M30046	
pad	T808	anonymous antigen AA68	Mt	Å. Andersen	
paeA	T692	anonymous antigen AA62	Mt	Å. Andersen	
paeB	T369	anonymous antigen AA61	Mt	Å. Andersen	
pafA	T258	encodes 33 kD antigen AA63	Mt	Å. Andersen	
phlC	T54	putative phospholipase C	Mt	P. del Portillo/L11868	
phoP	T93	putative phosphate sensor	Mt	?	+
phoS	T172	phosphate binding protein	Mt	D.B. Young	+
polA	T167	DNA polymerase I	Mt	V. Mizrahi/L11920	+
ponA	T460	penicillin binding protein	Mt	GMB	+
povB	T832	?	Ml	E. de Rossi	+
pri	T441	putative 38 kD virulence regulator	Mt	GMB/X68281	
purC	T465	purine synthesis	Mt	B. Gicquel	
purL	T355	purine synthesis	Mt	B. Gicquel	+
recA	T49	homologous recombination	Mt	M.J. Colston/X58485	+
recN	./.	recombinase	Mt	GMB	+
recS	T252	putative recombinase	Mt	S. Nair	
regX	T449	response regulator	Mt	M. Pallen/X66591	
rplM	T13	50S ribosomal protein L13	Mt	GMB	+
rpoBC	T311	beta, beta' subunits of RNA polymerase	Ml	GMB	+
rpsA	T167	30S ribosomal protein S1	Ml	GMB/Z46257	+
rpsJ	T73	ribosomal protein S10	Ml	GMB/Z14314	+
rpsL	T663	ribosomal protein S12	Ml	GMB/Z14314	+
rrnA	T198	rRNA operon (16S/23S/5S)	Ml	X56657	+
ruvB	T184	DNA helicase	Mt	GMB	+
sdhA	T685	succinate dehydrogenase	Mt	GMB	+
sodA	T264	superoxide dismutase	Ml	X16453	+
spoT	T333	pyrophosphohydrolase	Mt	GMB	+
ssrA	T441	10Sa RNA; small stable RNA	Mt	GMB/X60301	
tbp	T240	tuberculin related peptide	Mt	D00815	
thyX	T31	putative thymidylate synthetase	Mt	J. Dale/X59273	+
trxAB	T776	thioredoxin	Mt	B. Wieles	+
tuf	T663	elongation factor Tu	Ml	GMB/Z14314	+
uraA	T112	orotidine 5'-P decarboxylase	Mb	R.A. Young/U01072	+
ureABC	T407	urease subunits ABC	Mt	B. Gicquel/L41141	
virE	T776	?	Ml	E. de Rossi	+
wag12	T156	antigen MPT64	Mt	K. Matsuo/X75361	
wag13	T500	19 kD lipoprotein antigen	Mt	D.B Young/J03838	
wag15	T170	MPB70, secreted antigen	Mt	K. Matsuo/D37968	
wag25	T527	47 kD antigen	Mt	M.J. Colston	
wag31	T258	antigen 84	Mt	P. Hermans/X77129	+
wag36	T825	14 kD antigen	Mt	D.B Young/M76712	
wag9	T173	35 kD antigen	Mt	GMB/M69187	
xds	T750	unique sequence for detection of *M. tuberculosis* complex	Mt	GMB/M75726	

*Mt, *M. tuberculosis*; Mb, *Mycobacterium bovis*; Ms, *Mycobacterium smegmatis*; Ml, *M. leprae*; GMB, Génétique Moléculaire Bactérienne.

least seven other regions of locally conserved linkage appear to exist (Fig. 3) although the neighboring genes on the *M. tuberculosis* map do not have counterparts at the same locus in *M. leprae* and, in some cases, additional genes are present within the cluster.

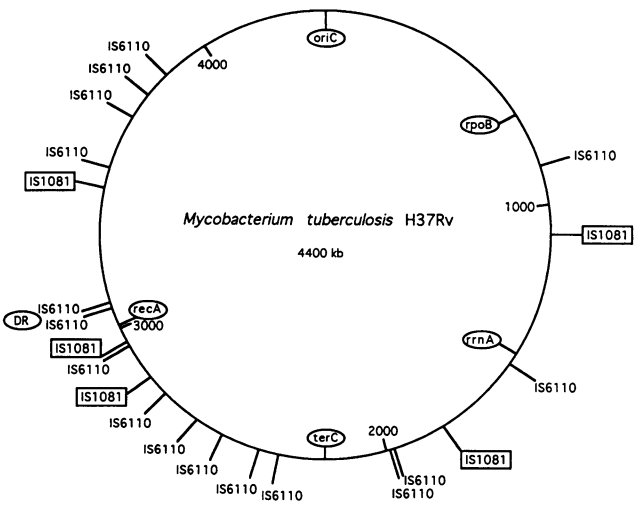

FIG. 4. Distribution of IS*6110* elements on the genome of H37Rv with respect to *oriC* and the putative terminus *terC*. The positions of the *rrn* operon, *recA*, and the DR locus are indicated.

DISCUSSION

The goal of this study was to elucidate the genomic organization of *M. tuberculosis* and to establish a set of ordered DNA fragments, a valuable genetic resource. The clones based on the shuttle vector pYUB18 should facilitate the dissection of the pathogenicity of the tubercle bacillus, as they can be introduced into easily manipulated surrogate hosts, such as *Mycobacterium smegmatis*, where faithful gene expression can be obtained (3). Several recent examples (11, 33, 34) leading to the identification of genes involved in drug resistance or encoding new therapeutic targets testify to the power of this approach. Clones from the pYUB328 library will undoubtedly find application in allelic exchange (35, 36) and systematic DNA sequence analysis (6). Although a whole genome shotgun sequencing approach appears to be the most efficient and cost-effective means of achieving this objective (37), the clones described here will greatly simplify analysis as they provide instant access to difficult regions, such as those harboring IS*6110* or other extended repetitive elements.

Analysis of the distribution of copies of IS*6110* around the genome was instructive as both clustering and bias in their location was observed (Figs. 3 and 4). In BCG, the sole IS*6110* element is located in the DR region, consisting of a series of short interspersed repeats (38). From the map of the *M. tuberculosis* genome it is clear that DR is situated roughly midway between *oriC* and the putative replication terminus *terC*. As the DR site is invariably occupied by IS*6110* in *M. tuberculosis*, it is probable that the original tubercle bacillus also had a single copy of IS*6110* in DR and that this subsequently migrated outwards by transposition in a stepwise manner, as suggested by the clustering seen in H37Rv (Fig. 4).

Comparison of the genome maps of *M. leprae* and *M. tuberculosis* revealed the absence of global similarity, although localized regions of conserved organization appear to exist. This suggests that, although mycobacterial genomes may have had a common origin, they have subsequently undergone extensive diversification. The current mosaic structure reflects groups of related genes which were primordially linked now being surrounded by novel chromosomal segments. It is conceivable that, in contrast to *M. leprae*, the large number of insertion sequences and the extensive amount of repetitive DNA present in *M. tuberculosis* may have contributed to genomic rearrangements (29, 30). Further analysis of the genomes of other pathogenic mycobacteria is required to see whether the mosaic arrangement is a general trend, for it is possible that only limited constraints are imposed on chromosome organization and gene order in these exceptionally slow-growing microbes.

We wish to thank the mycobacterial research community and people listed in Table 2 for kind gifts of probes, and Doug Smith for access to *M. leprae* genomic sequences prior to publication. We give special thanks to Thierry Garnier for cheerfully helping with the computer analysis, Burkhard Tümmler and Üte Römling for valuable advice, and Gérard Gugliemini and Brigitte Gicquel for useful discussions. This investigation received financial support from the World Health Organization, the Association Française Raoul Follereau, the Institut Pasteur, and from National Institutes of Health Grants AI26170 and AI23545.

1. Kochi, A. (1991) *Tubercle* **72,** 1–6.
2. Bloom, B. R. & Murray, C. J. L. (1992) *Science* **257,** 1055–1064.
3. Jacobs, W. R., Jr., Kalpana, G. V., Cirillo, J. D., Pascopella, L., Snapper, S. B., Udani, R. A., Jones, W., Barletta, R. G. & Bloom, B. R. (1991) *Methods Enzymol.* **204,** 537–555.
4. Young, D. B. & Cole, S. T. (1993) *J. Bacteriol.* **175,** 1–6.
5. Cole, S. T. (1994) *Int. J. Lepr.* **62,** 122–125.
6. Cole, S. T. & Smith, D. R. (1994) in *Tuberculosis: Pathogenesis, Protection, and Control*, ed. Bloom, B. R. (Am. Soc. for Microbiol., Washington, DC), pp. 227–238.
7. Bergh, S. & Cole, S. T. (1994) *Mol. Microbiol.* **12,** 517–534.
8. Canard, B. & Cole, S. T. (1989) *Proc. Natl. Acad. Sci. USA* **86,** 6676–6680.
9. Canard, B., Saint-Joanis, B. & Cole, S. T. (1992) *Mol. Microbiol.* **6,** 1421–1429.
10. Römling, U. & Tümmler, B. (1991) *Nucleic Acids Res.* **19,** 3199–3206.
11. Banerjee, A., Dubnau, E., Quémard, A., Balasubramanian, V., Um, K. S., Wilson, T., Collins, D., de Lisle, G. & Jacobs, W. R., Jr. (1994) *Science* **263,** 227–230.
12. Clark-Curtiss, J. E., Jacobs, W. R., Docherty, M. A., Ritchie, L. R. & Curtiss, R., III (1985) *J. Bacteriol.* **161,** 1093–1102.
13. Raleigh, E. A., Murray, N. E., Revel, H., Blumenthal, R. M., Westaway, D., Reith, A. D., Rigby, P. W. J., Elhai, J. & Hanahan, D. (1988) *Nucleic Acids Res.* **16,** 1563–1575.
14. Eiglmeier, K., Honoré, N., Woods, S. A., Caudron, B. & Cole, S. T. (1993) *Mol. Microbiol.* **7,** 197–206.
15. Coulson, A. & Sulston, J. (1988) in *Genome Analysis: A Practical Approach*, ed. Davies, K. E. (IRL, Oxford), pp. 19–39.
16. Sulston, J., Mallett, F., Staden, R., Durbin, R., Horsnell, T. & Coulson, A. (1988) *Comput. Appl. Biosci.* **4,** 125–132.
17. Hoheisel, J., Maier, E., Mott, R., McCarthy, L., Grigoriev, A. V., Schalkwyck, L. C., Nizetic, D., Francis, F. & Lehrach, H. (1993) *Cell* **73,** 109–120.
18. Philipp, W. J. & Cole, S. T. (1995) *FEMS Microbiol. Lett.*, in press.
19. Sambrook, J., Fritsch, E. F. & Maniatis, T. (1989) *Molecular Cloning: A Laboratory Manual* (Cold Spring Harbor Lab. Press, Plainview, NY), 2nd Ed.
20. Poulet, S. & Cole, S. T. (1995) *Arch. Microbiol.* **163,** 87–95.
21. Mott, R., Grigoriev, A., Maier, E., Hoheisel, J. & Lehrach, H. (1993) *Nucleic Acids Res.* **21,** 1965–1974.
22. Thierry, D., Brisson-Noël, A., Vincent-Lévy-Frébault, V., Nguyen, S., Guesdon, J. & Gicquel, B. (1990) *J. Clin. Microbiol.* **28,** 2668–2673.
23. Thierry, D., Cave, M. D., Eisenach, K. D., Crawford, J. T., Bates, J. H., Gicquel, B. & Guesdon, J. L. (1990) *Nucleic Acids Res.* **18,** 188.
24. Cole, S. T. & Saint-Girons, I. (1994) *FEMS Microbiol. Rev.* **14,** 139–160.
25. Fonstein, M. & Haselkorn, R. (1995) *J. Bacteriol.* **177,** 3361–3369.
26. Krawiec, S. & Riley, M. (1990) *Microbiol. Rev.* **54,** 502–539.
27. Leblond, P., Redenbach, M. & Cullum, J. (1993) *J. Bacteriol.* **175,** 3422–3429.
28. Coulson, A., Sulston, J., Brenner, S. & Karn, J. (1986) *Proc. Natl. Acad. Sci. USA* **83,** 7821–7825.
29. McAdam, R. A., Guilhot, C. & Gicquel, B. (1994) in *Tuberculosis: Pathogenesis, Protection, and Control*, ed. Bloom, B. R. (Am. Soc. for Microbiol., Washington, DC), pp. 199–216.
30. Poulet, S. & Cole, S. T. (1995) *Arch. Microbiol.* **163,** 79–86.
31. Collins, D. M. & Stephens, D. M. (1991) *FEMS Microbiol. Lett.* **83,** 11–16.
32. Mariani, F., Piccolella, E., Colizzi, V., Rapuolli, R. & Gross, R. (1993) *J. Gen. Microbiol.* **139,** 1767–1772.
33. Yuan, Y., Lee, R. E., Besra, G., Belisle, J. T. & Barry, C. E., III (1995) *Proc. Natl. Acad. Sci. USA* **92,** 6630–6634.
34. Zhang, Y., Heym, B., Allen, B., Young, D. & Cole, S. (1992) *Nature (London)* **358,** 591–593.
35. Reyrat, J.-M., Berthet, F.-X. & Gicquel, B. (1995) *Proc. Natl. Acad. Sci. USA* **92,** 8768–8772.
36. Balasubramanian, V., Pavelka, M. S., Bardarov, S. S., Martin, J., Weisbrod, T., McAdam, R. A., Bloom, B. R. & Jacobs, W. R. (1996) *J. Bacteriol.* **178,** 273–279.
37. Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., *et al.* (1995) *Science* **269,** 496–512.
38. Hermans, P. W. M., van Soolingen, D., Bik, E. M., de Haas, P. E. W., Dale, J. W. & van Embden, J. D. A. (1991) *Infect. Immun.* **59,** 2695–2705.