RESEARCH

Multiplex Sequencing of 1.5 Mb of the *Mycobacterium leprae* Genome

Douglas R. Smith,[1,2] Peter Richterich,[2] Marc Rubenfield,[2] Philip W. Rice,[2] Carol Butler,[2] Hong-Mei Lee,[2] Susan Kirst,[2] Kristin Gundersen,[2] Kari Abendschan,[2] Qinxue Xu,[2] Maria Chung,[2] Craig Deloughery,[2] Tyler Aldredge,[2] James Maher,[2] Ronald Lundstrom,[2] Craig Tulig,[2] Kathleen Falls,[2] Joan Imrich,[2] Dana Torrey,[2] Marcy Engelstein,[2] Gary Breton,[2] Deepika Madan,[2] Raymond Nietupski,[2] Bruce Seitz,[2] Steven Connelly,[2] Steven McDougall,[2] Hershel Safer,[2] Rene Gibson,[2] Lynn Doucette-Stamm,[2] Karin Eiglmeier,[5] Staffan Bergh,[5] Stewart T. Cole,[5] Keith Robison,[4] Laura Richterich,[4] Jason Johnson,[4] George M. Church,[1,3,4] and Jen-i Mao[2]

[2]Genome Therapeutics Corporation, Collaborative Research Division, Waltham, Massachusetts 02154; [3]Howard Hughes Medical Institute and [4]Department of Genetics, Harvard Medical School, Boston, Massachusetts 02115; [5]Unite de Genetique Moleculaire Bacterienne, Institut Pasteur, 75724 Paris CEDEX 15, France

The nucleotide sequence of 1.5 Mb of genomic DNA from *Mycobacterium leprae* was determined using computer-assisted multiplex sequencing technology. This brings the 2.8-Mb *M. leprae* genome sequence to ~66% completion. The sequences, derived from 43 recombinant cosmids, contain 1046 putative protein-coding genes, 44 repetitive regions, 3 rRNAs, and 15 tRNAs. The gene density of one per 1.4 kb is slightly lower than that of *Mycoplasma* (1.2 kb). Of the protein coding genes, 44% have significant matches to genes with well-defined functions. Comparison of 1157 *M. leprae* and 1564 *Mycobacterium tuberculosis* proteins shows a complex mosaic of homologous genomic blocks with up to 22 adjacent proteins in conserved map order. Matches to known enzymatic, antigenic, membrane, cell wall, cell division, multidrug resistance, and virulence proteins suggest therapeutic and vaccine targets. Unusual features of the *M. leprae* genome include large polyketide synthase (pks) operons, inteins, and highly fragmented pseudogenes.

[The sequence data described in this paper have been submitted to GenBank under accession nos. L78811–L78829, U00010–U00023, U15180–U15184, U15186, U15187, L01095, L01536, L04666, and L01263. On-line supplementary information for Table 1 is available at http://www.cshl.org/gr.]

Despite improved medical care and large vaccination programs, infectious organisms are still the leading cause of death, worldwide, and the pathogenic mycobacteria are among the worst offenders. There are estimated to be ~5 million cases of leprosy, globally, while tuberculosis kills ~3 million persons per year. The frequent occurrence of multidrug resistant *Mycobacterium tuberculosis* and the documented appearance of dapsone resistant *Mycobacterium leprae* are reminders that current therapies may not always be effective and that we should continue to search for and develop new antiinfective agents.

M. leprae is one of the few bacterial pathogens that infects humans and cannot be cultivated outside of animals. The organism is an intracellular parasite that grows extremely slowly (generation

[1]Corresponding authors.
E-MAIL church@salt2.med.harvard.edu; smith@cric.com; FAX (617) 432-7663.

7:802–819 ©1997 by Cold Spring Harbor Laboratory Press ISSN 1054-9803/97 $5.00

time, 14 days). A number of immunodominant protein antigens have been identified and characterized in *M. leprae* (Murray and Young 1992), but few metabolic enzymes have been studied. This combination of urgent problems and difficulties with classical biological approaches have made the mycobacteria prime candidates for comparative genome sequencing. This approach promises to aid in the identification of targets for vaccine and therapeutics development, possible regulatory elements and mechanisms, and will help us to understand the unique biochemistry of microbial intracellular parasites. The recent construction of a cosmid-based genome map for *M. leprae* has facilitated study of the genome by molecular biological techniques. This report summarizes DNA sequencing results on 43 cosmids selected from this set.

Advances in large-scale sequencing driven by the Human Genome Project have stimulated sequencing projects on a variety of small genomes. For example, at least six microbial genomes and one fungal genome have now been sequenced, ranging in size from 0.58 to 12 Mbp and representing all major phylogenetic kingdoms. These include *Haemophilus influenzae* (Fleischmann et al. 1995), *Mycoplasma genitalium* (Fraser et al. 1995), *Saccharomyces cerevisiae* (Dujon 1996), *Methanococcus jannaschii* (Bult et al. 1996), *Methanobacterium thermoautotrophicum* (Smith et al. 1996), *Synechocystis* sp. 6803 (Kaneko et al. 1996), and *Mycoplasma pneumoniae* (Himmelreich et al. 1996). Thirty-seven other small genome sequencing projects are now reportedly under way (Gaasterland and Sensen 1996). Thus, there is considerable biological and economic motivation for the development of more rapid DNA sequencing technologies that offer high accuracy and lower cost than current methods.

Multiplex sequencing is a rapid sequencing approach based on sample tagging, mixing, and molecular decoding by oligonucleotide hybridization (Church and Kieffer-Higgins 1988). The approach is compatible with a variety of sequencing strategies, including transposon-ordered and whole genome shotgun sequencing (Church and Kieffer-Higgins 1988). The potential throughput is very high, because all of the ‘‘front-end’’ steps, from DNA amplification and isolation through gel electrophoresis, are performed on mixtures of plasmid clones. Using pools of 20 plasmid clones (each clone provides two sequences), these front-end steps are facilitated by a factor of 40 compared to M13-based methods. Sequencing patterns are generated by ^{32}P-labeled film-based detection, by chemiluminescence (Richterich and Church 1993), by direct fluorescence, or by enzyme-linked fluorescence detection (Cherry et al. 1994). Digitized images of the sequencing patterns are then processed on computer workstations using automated image analysis and sequence reading software. These techniques have allowed the generation of significant volumes of sequencing data over the past few years of development on *Escherichia coli* (Church and Kieffer-Higgins 1988), *Salmonella typhimurium* (Roth et al. 1993), *Helicobacter pylori, M. tuberculosis, Staphylococcus aureus, Streptococcus pneumoniae, Clostridium acetobutylicum, M. thermoautotrophicum* (Smith et al. 1996), *Arabidopsis thaliana, Pyrococcus furiosus,* and *Homo sapiens* (Cawthon et al. 1990). Nevertheless, this is the first publication describing the application of the technology on a megabase scale. The sequences described here were generated over a 3.5-year period as the technology was developed and optimized.

Sequencing Strategy and Accuracy

The cosmids used in this study (Fig. 1) were constructed from *M. leprae* DNA isolated from armadillo liver infected with the dapsone-resistant Tamil Nadu strain of a clinical *M. leprae* isolate (Eiglmeier et al. 1993). The DNA sample has been shown to be heterogeneous, at least with respect to one putative transposon (Fsihi and Cole 1995). Cosmids were sequenced by a shotgun strategy at 5- to 10-fold redundancy followed by fragment assembly and primer-directed finishing to bridge contigs and eliminate single-stranded regions. The individual fragment sequences were proofread to correct obvious errors as the data were entered, and the contigs were proofread after assembly to correct errors detectable as discrepancies between individual fragments. The data were analyzed to identify errors resulting in frameshifts, and these were also corrected, wherever possible. The shotgun data were derived almost exclusively by chemical sequencing, which produced satisfactory data with very even band intensities although it suffered somewhat from a lack of reproducibility.

The average G + C content of the cosmids sequenced was 58%. This resulted in a significant electrophoretic gel compression every 200 bp or so, on average. Difficult compressions were resolved by careful analysis of reads from both strands, and by electrophoresis of the products of primer-directed cycled sequencing reactions on formamide gels, which were capable of resolving all compressions encountered. This worked well enough that in some of the later sets of cosmids, formamide gels were routinely used to generate ~30% of the shotgun cov-

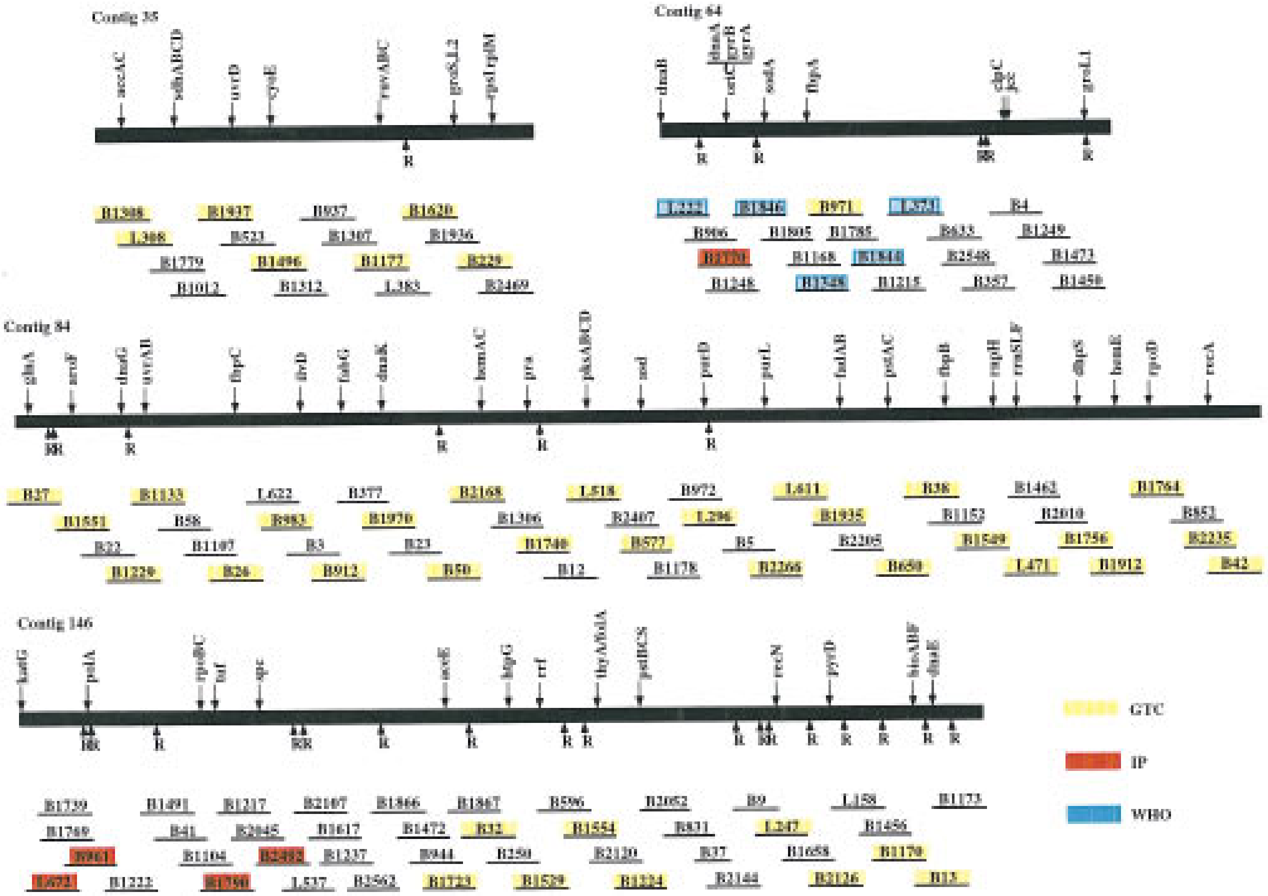

Figure 1 *M. leprae* genome map indicating regions discussed in the text. Cosmid clone names (starting with B or L) follow the *M. leprae* map (Eiglmeier et al. 1993). Cosmid sequences described here are indicated by yellow boxes (see Table 3, below). Red and blue boxes indicate cosmids sequenced by the Institut Pasteur (IP) and other members of *M. leprae* World Health Organization (WHO) genome consortium, respectively. Unboxed cosmids are mapped but not sequenced. Eighteen of the cosmids sequenced in this study overlapped to form contigs (eight contigs, averaging 67 kb in size). Most of the gaps remaining between adjacent sequenced cosmids were small, and many could be bridged by long-range PCR.

erage. This up-front measure significantly reduced the need for special-case compression resolution, although it often led to a reduction in gel resolution and a reduction in read length of ~10%. The insertion/deletion (indel) error rate after contig proofreading was estimated to average 1.8×10^{-4}, based on ~67 kb of overlapping sequence between pairs of cosmids that were finished independently. The frequency of missense errors was similar. Gel traces from all genes with potential frameshift errors were carefully examined for errors, and additional sequences were generated in ambiguous regions. After such homology-based frameshift editing, the indel error rate was reduced to 1.0×10^{-4} for sequences with database homologies (53 likely frameshift errors remaining in 31 genes out of a total of 562 with database homologies spanning a total of ~542 kb). Overall, the raw data indel error rate was considerably higher than that associated with ABI dye-terminator chemistry. The lack of an equivalent chemistry is one of the current limitations of multiplexing sequencing. Other limitations in comparison to ABI technology are the shorter read lengths and lower overall data quality.

Identification of Potential Gene Sequences

The sequences were analyzed for open reading frames (ORFs) using a set of computer programs unified through a single platform, GenomeBrowser (Robinson et al. 1994; Robinson and Church 1995). The programs identify all possible ORFs larger than a specified size (60 codons) and parse them to the National Center for Biotechnology Information

Table 1. List of 1064 Putative ***M. leprae*** Genes Identified in This Study

Intermediary metabolism	
acdC	Acyl CoA dehydrogenase
aceE	Pyruvate dehydrogenase E1 component
adi	Biodegradative arginine decarboxylase
ahh	6-aminohexanoate-cyclic-dimer hydrolase
ansB	L-asparaginase
atoB	Acetyl-CoA acyltransferase
dgt	dGTP triphosphohydrolase
fadA	3-ketoacyl-CoA-thiolase I
fadB	Multifunctional fatty acid degradation
fadE	Long-chain-fatty-acid--CoA ligase
gapA	Glyceraldehyde 3-P dehydrogenase B
gcvT	Aminomethyltransferase
gpm	Phosphoglyceromutase
pfkA	Weak match to fructokinase
pgk	Phosphoglycerate kinase
tdcB	Threonine dehydratase
thdF	Thiophene and furan oxidation
ureC	ureC / phosphomannomutase homolog
ycoD	Probable oxidase
ahpC	Alkyl hydroperoxide reductase
gabD	Succinate-semialdehyde dehydrogenase
gabT	4-aminobutyrate aminotransferase
glmS	Glucosamine--fructose-6-phosphate aminotransferase
gpsA	Glycerol 3-phosphate dehydrogenase
lpdA	Lipoamide dehydrogenase
mdh	Malate dehydrogenase
ppc	Phosphoenolpyruvate carboxylase
sfcA	Malate oxidoreductase
thiX	Thioredoxin
tlpA	Thioredoxin-like protein
tpiA	Triosephosphate isomerase
yceL	Weak match to glucose-6-phosphate 1-dehydrogenase
Respiration (aerobic and anaerobic)	
coxB	Putative cytochrome c oxidase polypeptide II
cyoC	Cytochrome o ubiquinol oxidase
cyoE	Ubiquinol oxidase operon protein
glpD	Aerobicglycerol-3-P dehydrogenase
narJ	Nitrate reductase beta subunit
narY	Respiratory nitrate reductase 2 beta
narZ	Respiratory nitrate reductase 2 alpha
Fermentation	
adhA	Alcohol dehydrogenase
adhB	Alcohol dehydrogenase
aldB	Aldehyde dehydrogenase
aldH	Putative aldehyde dehydrogenase
dhaP	Aldehyde dehydrogenase
mdlB	S-mandelate dehydrogenase
ybfF	Ketoacyl reductase hetn
ATP-proton motive force interconversion	
atpA	ATP synthase alpha chain
atpB	ATP synthase beta chain
atpD	ATP synthase delta chain
atpE	ATP synthase epsilon chain
atpF	Weak match to ATP synthase B chain
atpG	ATP synthase gamma chain
atpI	ATP synthase A chain
atpL	ATP synthase C chain
Broad regulatory functions	
ada	6-methylguanine-DNA methyltransferase
era	Ras-like GTP binding protein
hflX	E. coli GTP-binding protein hflx
hzfA	Probable zinc finger protein
icc	icc protein
lexA	Repressor for SOS regulon
oxyR	Activator H2O2-inducible genes
oxyS	Hydrogen peroxide-inducible protein
phoR	Pho regulatory gene sensor
pki	Protein kinase c inhibitor homolog
rpoE	RNA polymerase sigma-E factor
rpsA	RNA polymerase sigma factor
rpsB	Principle sigma factor
rpsC	Putative sigma factor (minor)
xtr	Probable trascriptional regulator
ybeY	Probable repressor
znfA	Zinc finger protein
Amino acids-Glutamate family/nitrogen assimilation	
argA	Glutamate N-acetyltransferase
argB	Acetylglutamate kinase
argC	N-acetyl-gamma-glutamyl-phosphate reductase
argD	Acetylornithine aminotransferase
argF	Ornithine carbamoyltransferase
argG	Argininosuccinate synthase
argH	Argininosuccinate lyase
argI	Arginine regulatory protein ahrc
argR	Repressor of arginine biosynthesis
glnA	Glutamine synthetase
glnB	Nitrogen regulatory protein p-II
glnE	Glutamate-ammonia-ligase adenylyltransferase
glnS	Glutamine synthetase
murI	Glutamate racemase
nifS	Nitrogen fixation gene nifS
nifU	Nitrogen fixation gene nifU
proC	Pyrroline-5-carboxylate reductase
Amino acids-Aspartate family, pyruvate family	
aar	alpha-Aminoadipate reductase
alr	Alanine racemase
apk	Aspartokinase III
apt	Similarity to P. putida OAPT
asd	Aspartate-semialdehyde dehydrogenase synthase
asnB	Asparagine synthetase B
aspC	Aspartate aminotransferase
dapA	Dihydrodipicolinate synthase
dapB	Dihydrodipicolinate reductase
dapE	Succinyl-diaminopimelate desuccinylase
dapF	Diaminopimelate epimerase
dha	Alanine dehydrogenase
ilvB	Acetolactate synthase
ilvD	Dihydroxy-acid dehydratase
ilvE	Weak match to branched-chain aa aminotransferase
leuA	Alpha-isopropylmalate synthase
lysA	Diaminopimelate decarboxylase
metB	Cystathionine gamma-synthase
metH	Homocysteine-THF-transmethylase
metL	Homoserine dehydrogenase
metY	Weak match to metH; possible isozyme
thrc	Threonine synthase
Amino acids-Glycine-serine family/sulfur metabolism	
cysE	Serine acetyltransferase
cysK	Cysteine synthase A
cysM	Cysteine synthase
serB	Phosphoserine phosphatase
thrB	Homoserine kinase
thtR	Thiosulfate sulfotransferase
Amino acids-Aromatic amino acid family	
aroB	3-dehydroquinate synthase
aroC	Chorismate synthase
aroD	3-dehydroquinate dehydratase
aroF	P-2-dehydro-3-deoxyheptonate A
aroK	Shikimate kinase I
dciA	Indolepyruvate decarboxylase
pccA	Propionyl-CoA carboxylase
trpG	Anthranilate synt. component II
yclF	Yeast aro1 gene for arom multifunctional enzyme
Amino acids-Histidine	
hisG	ATP-phosphoribosyl transferase
hisH	Imidazole acetol phosphate aminotransferase
hisI	Phosphoribosyl-AMP cyclohydrolase
Purine ribonucleotides	
guaA	GMP synthetase
guaB	IMP dehydrogenase
purD	Phosphoribosylamine-glycine ligase
purF	Amidophosphoribosyl transferase precursor
purI	Phosphoribosylaminoimidazole carboxylase catalytic subunit
purK	Phosphoribosylaminoimidazole carboxylase (ATPase)
purL	Phosphoribosylformyl-glycinamidine
purM	Phosphoribosylformulglycinamidine cyclo-ligase
purX	Phosphoribosylamine-glycine ligase
Pyrimidine ribonucleotides	
carA	Carbamoyl-phosphate synthase small chain
carB	Carbamoyl-phosphate synthase large chain
pyrB	Aspartate carbamoyltransferase
pyrC	Dihydroorotase
pyrD	Dihydro-orotate oxidase
pyrG	CTP synthetase
pyrH	Uridylate kinase

(*See p. 808 for Table 1 footnote.*)

Table 1. (*Continued*)

udp	Uridine phosphorylase
uraA	Orotidine-5'-monophosphate decarboxylase
2'-Deoxyribonucleotides	
dcd	Deoxycytidine triphosphate deaminase
dut	dUTP pyrophosphatase
thyA	Thymidylate synthase
Nucleotides-Salvage and interconversions	
add	Adenosine deaminase
deoD	Purine nucleoside phosphorylase
spoT	ppGpp 3'-pyrophosphohydrolase
upp	Uracil phosphoribosyltransferase
Sugars and sugar nucleotides	
cpsG	Phosphomannomutase
rfbA	Glucose-1-phosphate thymidylyltransferase
sucA	Sucrose synthase 1
tktA	Transketolase
ybkO	Putative sucrose-phosphate synthase
Biotin	
bioA	7,8-diaminopelargonic acid synthase
bioB	Biotin synthetase
bioD	Dethiobiotin synthetase
bioF	7-keto-8-aminopelargonic acid synthase
Folic acid	
folA	Dihydrofolate reductase type I
folP	Dihydropteroate synthase
Pantothenate	
panB	3-methyl-2-oxobutanoate hydroxymethytransferase
Pyridine nucleotides	
nadA	Quinolinate synthetase a protein
nadB	Quinolinate synthetase b protein
nadC	Quinolinate phosphoribosyltransferase
Thioredoxin, glutaredoxin, and glutathione	
trxA	Thioredoxin
Menaquinone and ubiquinones	
menE	O-succinylbenzoic acid--CoA ligase
Heme and porphyrins	
cysG	Uroporphyrinogen-III methyltransferase
hemA	Glutamyl-tRNA reductase
hemB	5-aminolevulinate dehydratase
hemC	Porphobilinogen deaminase
hemE	Uroporphyrinogen decarboxylase
hemK	Protoporphyrinogen oxidase
hemL	Glutamate-semialdehyde aminomutase
hemY	Coproporphyrinogen III oxidase
ybiR	Magnesium-chelatase 38 kDa subunit
Fatty acids and lipids	
accA	Acetyl-CoA carboxylase
acdA	Stearoyl- ACP-desaturase
acdB	Acyl-CoA dehydrogenase
acp	Acyl carrier protein (C)
atoB	Acetyl-CoA acetyltransferase (A)
caiC	Long-chain-fatty-acid--CoA ligase
cdh	CDP-diacylglycerol pyrophosphatase
cdsA	Phosphatidate cytidylyltransferase
cfa	Cyclopropane-fatty-acyl-phospholipase
choD	Cholesterol oxidase
fabD	Malonyl CoA-acyl carrier protein (A)
fabE	Beta ketoacyl-acyl carrier protein
fabF	Beta ketoacyl-acyl carrier protein (S)
fabG	3-oxoacyl-acyl-carrier protein (R)
falA	Long-chain fatty acid-CoA ligase
falB	Long-chain-fatty-acid-CoA ligase
falD	Long-chain-fatty-acid--CoA ligase
fasA	Fatty-acid synthase
fcb	Enoyl-CoA hydratase
glpK	Glycerol kinase
masA	Mycocerosic acid synthase (mas)
masB	Mas-associated gene (bcg orf-l)
pccB	Propionyl-CoA carboxylase beta chain
pgsA	Phosphotidylglycerophosphate synthase
pksA	Polyketide synthase (A)
pksB	Polyketide synthase (RC)
pksC	Polyketide synthase (SA)
pksD	Polyketide synth. (SADERC)
pksE	Polyketide synth. (SADC)
pksF	Polyketide synthase (SAC)
pksG	Polyketide synthase
pksX	Polyketide synthase (A)
ybbL	Weak match to ketoacylsynthase (S)
ycjW	4-chlorobenzoyl-CoA-dehalogenase (D)
yctL	Enoyl-CoA hydratase-like protein (D)
Ribosomal and "stable" RNAs	
rrnF	5S ribosomal RNA
rrnL	23S ribosomal RNA (truncated)
rrnS	16S ribosomal RNA
tsr	Rrna methylase
Ribosomal proteins and their modification	
rplM	Ribosomal protein L13
rpsI	Ribosomal protein S9
rplT	Ribosomal protein L20
rpmE	Ribosomal protein L31
rpmI	Ribosomal protein L35
tRNAs and aminoacyl-tRNA synthetases	
alaS	Alanyl-tRNA synthetase
argS	Arginyl-tRNA synthetase
argT	tRNA arginine (anti: ACG)
argU	tRNA arginine (anti: CCU)
asnT	tRNA asparagine (anti: GUU)
aspT	tRNA-aspartic acid (anti:GTC)
cysS	Cysteinyl-tRNA synthetase
gluT	tRNA-glutamic acid (anti: TTC)
glyS	Glycyl-tRNA synthetase
glyT	tRNA glycine (anti: CCC)
hisS	Histidyl-tRNA synthetase
ileS	Isoleucyl-tRNA synthetase
leuT	Leucine tRNA (anti: TAG)
lysT	tRNA-lysine (anti: TTT)
lysU	Lysyl-tRNA synthetase
miaA	tRNA delta(2)-isopentenyl-PPi transferase
pheS	Phenylalanyl-tRNA synthetase alpha chain
pheT	tRNA phenylalanine (anti GAA)
pheU	Phenylalanyl-tRNA synthetase beta chain
proT	tRNA-pro (anti: GGG)
proU	tRNA proline (anti: CGG)
serT	tRNA serine (anti: UGA)
serU	tRNA serine (anti: GCU)
thrT	tRNA threonine (anti: UGU)
trpS	Tryptophanyl tRNA synthetase
tyrS	Tyrosine tRNA synthetase
valT	tRNA valine (anti: UAC)
RNA synthesis, modification, and DNA transcription	
hsrA	Heat shock / transcription regulator
nusB	nusB Transcription terminator
rho	Transcription termination factor R
yccB	Regulatory protein whib
Proteins (translation and modification)	
ampM	Methionine aminopeptidase
efp	Elongation factor P (ef-P)
fusA	Elongation factor G
greA	Transcription elongation factor greA I
infC	Protein chain initiation factor IF-3
pepN	Aminopeptidase I homolog
pknA	Protein kinase
prfA	Peptide chain release factor 1
rrf	Ribosome releasing factor (rrf)
tsf	Elongation factor ef-TS
ybcC	Weak match to peptidyl-prolyl cis-transisomerase
Polysaccharide biosynthesis (cytoplasmic)	
glgC	Glucose-1-phosphate adenylyltransferase
Degradation of RNA	
rnc	Ribonuclease III
rnpH	Ribonuclease H
DNA replication, restriction/modification, recomb., repair	
dnaE	DNA polymerase III alpha chain
dnaG	DNA primase
dnaL	dnaJ protein isozyme
fpg	Formamidopyrimidine-DNA glycosylase
polI	DNA polymerase I
recA	Recombination and DNA repair
recG	Probable ATP-dependent DNA-helicase
recH	Repair DNA helicase (similar to recG)
recN	Recombination and DNA repair
recR	RecB/RecC-independent recombination repair protein
recX	Reca regulatory gene
ruvA	Holliday junction DNA helicase ruvA
ruvB	Holliday junction DNA helicase ruvB

Table 1. (*Continued*)

ruvC	Crossover junction endo-DNAse I
tag	DNA-3-methyladenine glycosidase I
traB	Putative transposase
ung	Probable uracil-DNA glycosylase
uvrA	Excinuclease ABC subunit A (N-terminal)
uvrB	Excinuclease ABC subunit B (N-terminal)
uvrD	DNA helicase II; E. coli uvrD
xerC	Probable integrase / recombinase
xheA	ycaA URF with intein homing endonuclease
yciM	12 kDa protein (E. coli ybaB)
ycoU	Conserved protein
Degradation of Proteins	
gcp	O-sialoglycoprotein endopeptidase
pepE	Lactococcal endopeptidase
pepP	Prolidase
prcB	Proteasome multicatalytic endopeptidase
pre	Prolyl endopeptidase
ycdR	Putative serine protease
Membrane components	
ybtS	Match to motility related promoter orf
Murein sacculus, cell wall and surface polysaccharides	
ddlA	D-alanine-D-alanine ligase
embA	ethambutol resistance protein
embB	ethambutol resistance protein
embC	ethambutol resistance protein
embD	ethambutol resistance protein
kreA	Weak homolog of S. cerevisiae KRE1
mraY	Phospho-N-acetylmuramoyl-pentapeptide-transferase
murB	Udp-N-acetylpyruvoylglucosamine reductase
murC	UDP-N-acetylmuramate--alanine ligase
murD	UDP-N-acetylmuramoylalanyl-d-glutamate synthase
murG	UDP-N-acetylglucosamine-N-acetylmuramyl-(pentapept.)
murZ	UDP-N-acetyglucos.-carboxyvinyltransferase
pbpA	Penicillin binding protein-1A
pbpB	Penicillin-binding protein 6
yceA	Stage V sporulation protein E
Antigens	
cpsA	Cpsa gene product
fbpA	Fibronectin binding protein (ag)
magA	Antigen ag84-M. tuberculosis
magB	Antigen ag84 (cie)
magC	Antigen 85-C precursor
pra	Proline-rich antigen
rfe	rfe protein
sraA	45 kDa serine-rich antigen
sraB	Serine rich antigen - M. leprae
yclD	Putative CMP-2-keto-3-deoxyoctulosonic acid synthetase
Transport/binding proteins	
aatA	aromatic aa transporter
aatB	aromatic aa transporter
abcA	Yeast mitochondrial protein ABC1
abcB	ABC transporter
abcD	ABC transporter (ATPase operon)
ach	Acetyl-hydrolase
actA	Transport protein - S. Coelic. ActII-3
actB	Transport protein
actC	Transport protein
aph	Acid phosphatase
aroP	Aromatic aa transport protein
aroQ	Aromatic aa permease
bfr	Bacterioferritin
caiD	Carnitine racemase
cobT	Nicotinate-nucleotide-dimebenzimidazole PRTase
corA	Mg, cobalt transport protein
ctrA	Calcium-transporting ATPase
cysA	Thiosulfate sulfur transferase
cysT	Putative sulfate permease T
cysU	Sulfate permease T protein
cysW	Sulfate permease W protein
dkgR	2,5-diketo-d-gluconic acid reductase
drrA	ABC family transporter
drrB	ABC-2 family transporter
drrC	ABC-2 family transporter
entB	Enterobactin synthetase component E
hup	Homolog of human P protein
lepA	Ras-like GTP binding protein
lysP	Lysine-specific permease ecoli
malF	Weak match to maltose transport permease
malG	Maltose transport inner membrane protein
malK	Maltose transport cytoplasmic membrane protein
nakA	N+, K+ ATPase
pheP	Aromatic aa permease
phoH	Phos. Starvation-inducible protein
phoK	Phosphate transport (pstA)
phoL	Phosphate transport (pstC)
phoM	Phosphate-binding periplasmic protein (pstS)
phoP	Transcriptional response regulator
phoS	Phosphate-repressible phosphate-binding protein
phoT	Phosphate transport protein pstb
phoU	Weak match to phosphate regulatory protein
phoW	Phosphate transport protein
phoX	Phosphate transport protein psta
rbsB	D-ribose periplasmic binding protein
rbsC	D-ribose high-affinity transporter
rfbD	Capsular polysaccharide export protein
rfbI	LPS /O antigen transport
secD	Membrane protein secD
secF	Membrane protein secF
subI	Sulfate-binding protein precursor
ybqS	68 kDa TPP-requiring protein (transketolase)
Cell division	
cdcH	Cell cycle control ATPase (cdc48)
ftsY	Cell division protein ftsY
ftsZ	Cell division protein ftsZ
Protein secretion	
ffh	Signal recognition particle protein
ybfI	Weak match to lipoprotein signal peptidase
ybqR	Possible suppressor protein
Osmotic adaptation	
otsA	Hypothetical protein
otsA	Alpha-trehalose-phosphate synthase
Virulence and toxin-related functions	
hlyX	S. hyodysenteriae hemolysin gene
invA	L. Ivanovi invasion precursor p60
mce	Mce gene product
ybkM	Protein p60 precursor
Plasmid / Transposon-related functions	
nreA	Nrea gene product
ybmC	Weak match to transposase of E. coli IS3411
ybmD	Weak match to transposase of M. bovis IS1081
Drug/analog sensitivity	
aga	Aminoglycoside 2'-N-acetyltransferase
arsA	Arsenical pump-driving ATPase
bacA	Putative bacitracin resistance protein
entE	Putative enterobactin synthetase
lmbE	Lmbe gene product
mdmC	O-methyltransferase
mmr	Methylenomycin a resistance protein
pit	Glvr-1 protein
tnrB	tnrB2 Protein
ybbA	Regulatory or antibiotic resistance protein
ybiV	Putative cephamycin export protein
ycjC	Smf2 protein
Adaptations to atypical conditions	
clpB	Heat shock clpB protein
cspA	Cold shock protein, activator
dnaJ	Dnaj protein
groE	Groe1 protein
groS	10 kDa chaperonin
hspA	Heat shock 70 kDa protein
hspB	Heat shock grpE protein
hspC	Probable heat shock protein
hspD	Heat shock protein C62.5
htrA	Heat shock protein htra
Unassigned function	
cfx	cfxQ protein - xanthobacter flavus
codH	Molybdenum-containing iron-sulfur flavoprotein
dim	Dimerase - streptomyces coelicolor
gcpE	gcpE protein (protein E)
kdtB	KDTB protein
mrp	MRP protein
msrA	Peptide methionine sulfoxide reductase

(*See following page for Table 1 footnote.*)

(NCBI) network BLAST server to identify database homologies. They also search for tRNAs (Fichant and Burks 1991), perform codon usage analysis, and perform a nucleotide BLAST search. The results were displayed using the GCG Figure program, or the Belmont Tool Kit, an interpretive object-oriented graphical environment. This provided a graphical representation of each cosmid displaying the locations of putative reading frames with corresponding BLAST homologies displayed above each frame. Below each frame was displayed a series of dots (which may merge into a solid line) if the dicodon usage matched an *M. leprae* gene-specific dicodon usage table.

Reading frames with dicodon usage similar to previously identified *M. leprae* genes were analyzed further for the presence of translation initiation sites. Acceptable sites were selected from a comprehensive list for each reading frame and contained an ATG or GTG initiation codon preceded by an optional spacer (0–8 nucleotides) and a sequence complementary to at least 4 out of 11 nucleotides from the 3′ terminus of *M. leprae* 16S rRNA (Shine and Dalgarno 1975; Liesack et al. 1990). Alignments with the amino termini of homologous proteins were also used to select translational start sites, in some cases. Possible coupled translation signals (an initiation codon within 20 nucleotides of a stop codon, characteristic of many bacterial operons) were also accepted as putative start sites. The positions of all putative genes meeting one or more of these criteria were recorded, together with the nature of the initiation site or operon linkage.

A list of putative *M. leprae* genes identified in this study and sorted by function is provided in Table 1 [a more comprehensive list is available on the Genome Research Web site (http://www.cshl.org/gr)]. Functional designations and gene names were assigned to genes with homologs having BLAST scores over 100; otherwise a name beginning with the letter ‘‘y’’ was assigned. We stress that the functional assignments must be viewed as provisional because of the inherent uncertainties in assigning gene function by sequence similarity. Gene names are based on existing mycobacterial names, where acceptable. Otherwise, names are based on *E. coli* nomenclature rules corresponding to the closest bacterial homologs in the following order of priority: *E. coli, S. typhimurium, Bacillus subtilis, Streptomyces* species, and other bacteria. In many cases, new names were assigned. A more extensive table of interpretations, including accession numbers, is available from http://www.cric.com/ and from MycDB, a database of mycobacterial mapping and sequence information (Bergh and Cole 1994) based on the acedb (Durbin and Thierry-Meig 1991–1995). The following sections describe some of the more striking findings from the data. We stress that owing to the limitations imposed by an incomplete geneome sequence, it is not possible to make definitive conclusions concerning the unique nature of mycobacterial metabolism relative to other organisms.

Repetitive Sequences and DNA Duplications

The *M. leprae* genome was found to contain several types of repetitive sequences by cross-searching for homology between different cosmids (precise location and size of repeats are given in Table 2). The most common repeats were a large family of 70- to 80-bp sequences, which we have called REP1 elements. The functional significance of these elements is unknown, but some of them were found to be located near the beginnings of genes. We found several RLEP elements (originally described as near-perfect 700-base repeats (Woods and Cole 1990). These occurred in cosmids where they had been previously located by physical mapping techniques (Eiglmeier et al. 1993). However, the RLEPs do not appear to encode any proteins. Of particular interest is the DNA polymerase I gene in cosmid L247 (Table 3), which is closely flanked by two inverted RLEP elements. This arrangement is reminiscent of certain composite transposons and provides a possible explanation for the origin of RLEPs as ‘‘IS-like’’ elements (Fsihi and Cole 1995). The only clearly identifiable IS element, the 1051-bp REP13 element, was found in cosmid B1620, and this shows 65% identity at the DNA level with IS*1081* from the *M. tuberculosis* complex (Poulet and Cole 1995).

Some smaller direct repeats were also seen. For instance, two identical copies of the 309-bp REP9

Table 1 (*Continued*) Here, 419 genes are sorted by name into 46 functional categories similar to M. Riley's *E. coli* categories (Belfort et al. 1995). Additional data are detailed in the *Genome Research* Web site, http://www.cshl.org/gr including database matches, scores, cosmid name, and gene start/stop positions within the cosmid (only one cosmid is designated in the case of genes that reside in overlapping regions on two or more cosmids), as well as 645 genes with only weak database matches or matches to only genes of unknown function.

Table 2. Strong DNA Sequence Matches and Repetitive Elements

Repeat	Cosmid	Position	Size (bp)
REP1	B1549	12421–12481	62
REP1	B1549	5568–5646	78
REP1	B1549	8856–8928	72
REP1	B1549	8940–8987	47
REP1	B1790	6026–6104	78
REP1	B1912	34462–34532	70
REP1	B1912	6864–6919	55
REP1	B1937	25910–25989	79
REP1	B2126	6878–6957	79
REP1	B2168	39401–39482	81
REP1	B2168	41486–41565	79
REP1	B2168	5506–5586	80
REP1	B2235	1229–1307	78
REP1	B2235	19105–19175	70
REP1	L247	22012–22049	37
REP1	B2266	14210–14283	73
REP1	B2266	18663–18721	58
REP1	B2266	21210–21266	56
REP1	B1764	10653–10740	87
REP1	B1764	22929–22984	55
REP1	B1756	14974–15044	70
REP1	B1740	33648–33723	75
REP1	B1496	22092–22131	39
REP1	L518	2738–2822	84
REP1	L471	17300–17367	67
REP9	B2168	8855–9164	309
REP9	B1790	15824–16133	309
REP13	B1620	7178–8228	1050
REP14	B1790	21759–21810	51
REP14	B1790	23970–24021	51
RLEP	L247	3399–4380	981
RLEP	B1177	25929–26861	932
RLEP	L247	1–641	641
RLEP	B1170	26569–27218	649
RLEP	B2126	11059–11612	553
pksA	L518	43–1677	1637
pksC	L518	5439–7091	1652
pksD	L518	10138–11763	1625
pksE	L518	16715–18172	1457
pksX	B1170	21311–22938	1627
aroP1	B2126	28854–30398	1544
aroP2	B2126	30520–32065	1545
CAC_6	B1935	12928–12944	18
TTC_{21}	L518	9592–9603	63

Table 3. ***M. leprae*** Cosmid Sequence Accession Numbers

B1133	L78811	B2235	U00019
B1170	U00010	B2266	U15182
B1177	U00011	B229	U00020
B1229	L78812	B26	L78816
B13	L78823	B27	L78817
B1308	U00012	B32	L78818
B1496	U00013	B38	L01095
B1529	L78824	B42	L78826
B1549	U00014	B50	L78827
B1551	L78813	B577	L01263
B1554	L78814	B650	U15184
B1620	U00015	B912	L78819
B1723	L78825	B937	L78820
B1740	U15183	B961	Z46257
B1756	U15180	B971	L78821
B1764	U15181	B983	L78828
B1770	Z70722	B998	L78829
B1790	Z14314	L222	L39923
B1912	L01536	L247	U00021
B1935	L04666	L296	U15187
B1937	U00016	L308	U00022
B1970	L78815	L471	U15186
B2126	U00017	L518	U00023
B2168	U00018	L611	L78822

Numbers in boldface type indicate sequences first described here.

element were contained in cosmids B2168 and B1790 (Honoré et al. 1993). The 52-bp REP14 element with 69% identity between two copies in B1790 also detected 12- to 18-bp stretches in several other cosmids (data not shown). Simple sequence repeats, including 6-copy CAC and 21-copy TTC tandem trinucleotide repeats, are longer than those in *E. coli*.

Several apparent gene duplication events were evident. One of these is a 1.6-kb sequence that recurs in several members of a family of polyketide synthase (pks) genes (including four within a single large operon in cosmid L518). The 1.6-kb repeat is composed of two segments, 120 bp and 1385 bp (separated by a 120-bp spacer), which are virtually identical between repeats *pksA* and *pksC* (Table 2). These two repeats, separated by 3.8 kb, are contained in two adjacent polyketide synthase genes encoded by the L518 operon (discussed in more detail below). The overall identity of repeats *pksA* and *pksC*, including the 120-bp spacer, is 95%. The polypeptide encoded by the repeat contains an acyltransferase consensus sequence, VVGHSMGE-SAAAVVAGAL, near its center. The repeats in *pksD*, *pksE*, and *pksX* share 68%, 66%, and 55% identity to the *pksA* DNA sequence.

Cosmid B2126 contains a duplicated 1.5-kb segment that encodes an amino acid transport gene similar to *aroP* (Tables 1 and 2). The two identical copies of this segment are arranged in tandem with a 122-bp spacer. The perfect nature of this repeat indicates an evolutionarily recent duplication or gene conversion.

Split and Fragmented Genes

At least three genes in *M. leprae* are likely to encode proteins that undergo autocatalytic splicing reactions to remove an intein (protein intron) from a nascent precursor molecule. The corresponding genes are believed to have been ''invaded'' by a DNA sequence, coding for a homing endonuclease, that is inserted in-frame in a protein-coding gene. These are *gyrA* (Fsihi et al. 1996), *xheA*, and *recA* (the sequence of *M. leprae* cosmid B2235 contained a *recA* gene and *recA*-associated ORF). There is a relationship between the *M. leprae* and *M. tuberculosis recA* genes (Davis et al. 1994). The sequences of the intein and the insertion points are different in the two organisms. In contrast, the *recA* exteins are 92% identical. Such divergence among inteins is common even among inteins targeting the same gene (Pietrokovski 1996). Features shared by the inteins include two homing double-stranded DNA (dsDNA) endonuclease motifs (LAGLIDGDG also found in introns and in HO endonuclease) separated by 80–121 amino acids plus protein-splicing catalytic sites at the intein amino terminus (Cys) and carboxyl terminus (His–Asn). The *M. leprae recA* intein has a match to intron-encoded DNA–endonucleases/RNA–maturases (e.g., P03873 | Cybm_Yeast, *P* = 4.5e-05 overall, *P* = 0.36 for the segment below), which is not detected in other intein sequences.

```
Tram_Yeast:   SMSYLIFYNLIKPYLIPQMMYKLPNT
                    |  || ||: | | |||
RecA_Mycle:   TAATAKFQSLIAPYVAPSMEYKLLPQ
                    : ::::||: |||:|||
Cybm_Yeast:   KESMPILTKIVSPYIIPSMKYKLGNY
```

The intein found in an 870-codon ORF, *xheA* in cosmid B1496, shows significant similarity to the protein-splicing and homing endonuclease domains of the vent polymerase intein and yeast HO proteins, respectively (Fig. 2). The part of the gene flanking this potential intein (codons 1–202, 581–870) corresponding to the N and C exteins, is homologous to ORFs from three major kingdoms—eukaryotes (*Antithamnion, Plasmodium*), archaebacteria (*Methanobacterium*), and eubacteria (*Synechocystis* and *Shigella*). The significance of the homing endonuclease motifs may be to target conserved gene sequences as hypothesized for thymidylate synthase (Sherman et al. 1995). Of the reported 20 gene families targeted by inteins, all 17 with homologs of known function are involved in metabolism of phosphorylated compounds.

About 3.5% of possible coding regions in the 1.5 Mbp described here appeared to contain multiple (three or more) frameshifts and/or in-frame termination codons relative to strongly similar, known genes. Reinspection of the raw data in these regions (from data on both strands) did not support the multiple changes that would be required to generate functional coding sequences. Of the total of 39 such regions, an average of 9 and as many as 21 changes per gene would be required. Highly fragmented gene sequences such as these were assumed to represent nonfunctional pseudogenes and were therefore not annotated as putative coding sequences. One possible explanation for their abundance is that strains of *M. leprae*, being slow-growing, obligate intracellular pathogens, have accumulated mutations in certain genes that are not essential for their survival in, or for transmission between, humans. It is even possible that there is a selective advantage associated with the loss of certain functions. No homologs of genes considered essential for all organisms (Mushegian and Koonin 1996) were found to be disrupted.

An alternative source of fragmented genes might be gene duplication and subsequent inactivation of one copy, possibly by repeat induced point mutagenesis (Ozer et al. 1993; Singer et al. 1995). However, in no case can a normal copy of a scrambled gene be found elsewhere in the genomic sequence (which now covers about two-thirds of the genome). Other possible explanations for highly fragmented genes should also be considered. Among these are mutations occurring during bacterial strain isolation and recombinant cloning. Biological processes have been described that can counteract insertions or frameshifts at the DNA, RNA, or protein levels at rates compatible with selective advantage for retaining such genomic regions. Such processes include cryptic genes (Hall and Sharp 1992; Hall and Xu 1992), which can easily switch via one or two mutations to a state expressing enzymatically active products at a high level, RNA splicing and editing (Bechhofer et al. 1994; Belfort et al. 1995), ribosomal reprogramming (Gesteland et al. 1992), and protein splicing (Davis et al. 1994; Perler et al. 1994; Belfort et al. 1995).

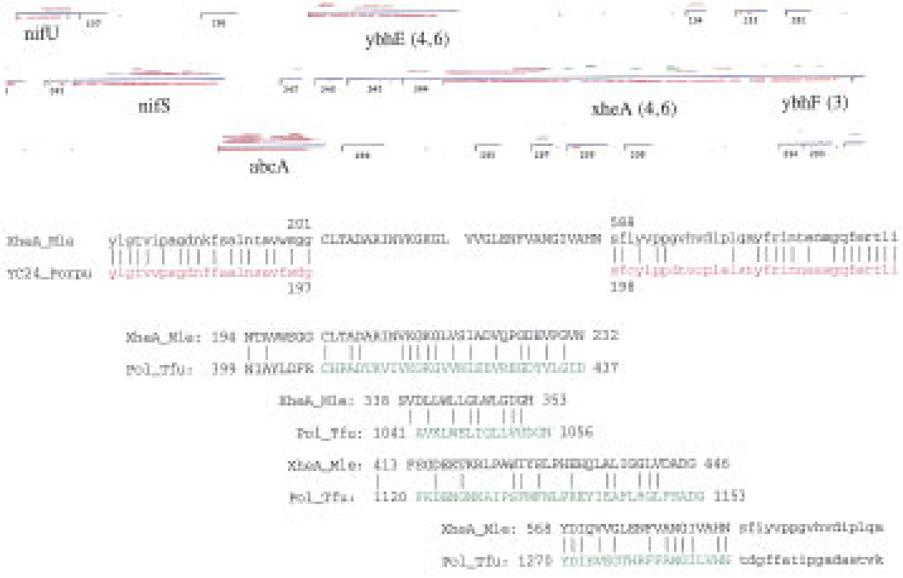

Figure 2 Analysis of the *xheA* intein. The tightly coupled operon shown is right-to-left 5′ to 3′: *ybhF, xheA, ybhE, abcA, nifS, nifU* shown in GenomeBrowser format. ORFs longer than 50 codons (blue horizontal lines) have stop codons indicated by short vertical black lines. Magenta horizontal lines above each ORF indicate matches to the NR database (Altschul et al. 1990) with significant BLASTP scores ($P < 0.001$), where the vertical displacement indicates the percent amino acid identity for that sequence segment. The red lines below the ORFs indicate quality of dicodon usage. Frame number, accession numbers, and gene names based on sequence similarity are in the text below the red lines. The *xheA* gene is located in *M. leprae* cosmid B1496 from nucleotide position 2020 to 9152. The amino- and carboxy-terminal regions have strong matches with eukaryotic, prokaryotic, and archaebacterial URFs (unknown function reading frames), including sp | P51240 | YC24_PORPU, and gi | 1742763 (*E. coli*) at 30%–42% identity ($P <$ 1E-22) as does the central intein region (where intein BLASTP segments are in green to contrast with the normal magenta). The paralogous (intragenomic *M. leprae*) *xheA–ybhF* (gi | 466874) duplication is 24% identity, $P =$ 2E-15. The numerals in parentheses represent the ORF numbers for a related cyanobacterial gene cluster (D64004). The sequence alignments (*below*) indicate the shift in amino acid identity pattern and the conserved motifs at the intein boundaries and internally.

Figure 3 illustrates an extreme case of gene fragmenting, where high amino acid sequence conservation within short blocks is seen. The sequence is derived from cosmid B2235 (4387–5673) and is homologous to *ythY*, an *M. tuberculosis* gene described in SWISSPROT and EMBL databases as encoding a putative thymidylate synthase. This assignment is probably inaccurate, as there is no significant similarity with the large thymidylate synthase (TYSY) family and there is no published evidence supporting it (the *M. leprae thyA* gene is on cosmid B1554). It is interesting to note that a *ythY* homolog on *M. tuberculosis* cosmid Y154 (Smith et al. 1996), which contains several genes in common with *M. leprae* cosmid B2235, is also fragmented. Alignment of the *M. leprae* and *M. tuberculosis ythY*-coding sequences revealed that the nucleotide spacing was identical at the position of each frameshift in the *M. tuberculosis* sequence. This suggests that loss of function preceded the divergence of these *M. leprae* and *M. tuberculosis* orthologs. However, this situation does not hold for all fragmented genes. For example, the *pyc* gene of *M. tuberculosis* is intact, whereas the *M. leprae pyc* homolog has 21 frameshifts.

Polyketide Synthase Operons

A large number of putative operons were identified in the sequences reported here based on functional relationships, collinearity, and possible translational coupling. A consistent feature of such putative operons is translational coupling between adjacent genes. A particularly long example, the polyketide operon in cosmid L518, contains at least 10 genes spanning 30 kb, most of which appear to be translationally coupled (the first gene begins at the end of the cosmid, so there may be additional genes at the 5′ end of the operon). Five genes from this operon contain a possible start codon overlapping the stop codon of the previous gene but shifted back by 1 nucleotide. In one gene the putative start is shifted back from the previous stop by 11 nucleotides, and in three others the start is shifted forward by 3, 12, and 30 bases.

The overall structure of this operon is interestingly similar to the putative mycocerosic acid synthase (*mas*) operon on cosmid B1170. The L518 operon contains six pks genes encoding large proteins (>2000 amino acids) of modular organization followed by three genes encoding components of an ABC transporter similar to the daunorubicin resistance system of *Streptomyces* (P32011) and a gene encoding a homolog of BCG (*Bacillus* Calmette-

Guerin) *mas* ORFII. The B1170 operon includes one large pks gene and genes encoding homologs of surfactin synthase (D13262) and a *Streptomyces* antibiotic transporter gene (C40046). However, there does not seem to be any translational coupling in this operon, with the downstream genes starting ~50 nucleotides after the stop codon of the previous gene. Figure 4 shows the relationship of the putative pks's from *M. leprae* to other members of this protein family. The *M. leprae* proteins contain some, or all, of the modules that are commonly found in polyketide or fatty acid synthesis that are known to effect the various functional and catalytic steps in pks genes (Fig. 4). Although the actual function of these pks genes is uncertain, it seems likely that they will be involved in the biosynthesis of cell wall components, like mycocerosic acid, as these often belong to the polyketide family.

Sequence Relationships Between *M. leprae* and *M. tuberculosis*

Approximately half of the *M. tuberculosis* genome is now available for comparison to *M. leprae* (2 Mbp of unique sequence comprised of 19 cosmids from our group and 49 from the Sanger Centre) (Barrell et al. 1996; Smith et al. 1996). Regions of similarity at the DNA level can be readily detected with an average identity of ~78%, and extending over a total of 411,800 nucleotides. These matches occur in short blocks of ~1400 nucleotides, on average, which extend over larger genomic regions (10 kb for a given pair of cosmids, on average).

The results of DNA-based cross-genome comparisons between two selected *M. leprae* cosmids and the available *M. tuberculosis* cosmids (as of October 1996) are shown in Figure 5. In the example of *M. leprae* cosmid L471 (Fig. 5A) and *M. tuberculosis* cosmids MTCY130 and MTCY373, there is a high degree of collinearity between the sequences over a 23.3-kb region. The two *M. tuberculosis* cosmids map directly adjacent to one another. A ribosomal operon has been mapped to the region containing MTCY130 (Philipp et al. 1996) but was not annotated on the sequence. The sequence beyond the *argS* gene on L471 (~7 kb) does not appear to contain any genes and is not conserved in any sequenced *M. tuberculosis* cosmids. In the second example (Fig. 5B) with *M. leprae* cosmid B32, large blocks of matching sequences occur on two *M. tuberculosis* cosmids, MTCY427 and MTCY338, which are ~650 kb apart on the genome (some of

```
                  ****  *** * ******  * ****   * *** * *          **
TYSY_MYCTU     6  TKDVNIVFRLASLPMGSEAMALLRLP^^VLPAVQIAGRIVGQGHRYHQLG
Translate         TKDVPFVFRCALLPMGSETLAFLRLP^^MLSAVQLASR.CWAWRSNDRLG
B2235       4387  aaggctgtctgttcagagacgtccccTGactggccgac ttgtctagccg
                  caatctttggcttctggactcttgtc  ttcctatcgg ggcggcaagtg
                  catcccccgtaggggccataatggac  gaggggaccg tgagaaccgtg

                   *   *** *    * **   ***        ***  *  *  **  * *
TYSY_MYCTU    54  ARPAQCAAHFGRPARADGFCGVRP^^++++++RQWWTGGGRWSPCWTDAI
Translate         SRATKCAADFS#STRTDGLISVRP^^------RQWSIGAARS#PCLIDII
B2235       4530  tcgaatgggtattaaaggcaagccCACCACGcacttagggattcttagaa
                  cgccagccatggccgcagttgtgc        gagctgccgcacgttatt
                  gatcattattcataacccccctgg        gggatcgtagaacgcttt

                             *   **     * ***   * * * * * *  * *
TYSY_MYCTU    96  STEAIPVQ.IIWGTKDV^+++LPVRHA...HMALPPSGLAIGDFRGLGTF
Translate         #PRPSQCR#LSWSAHDV^---VPIRHAWMTHSAMPSSLLKI..FEGASHC
B2235       4664  tcactctattatagcggTATGgcaccgtaacagacatccaa  tgggact
                  gcgccaggatgggcaat    tctgacgtcagctcgcttat  tagcgag
                  acgacacaaatgtattg    aactctggaccggttcaggc  taatctc

                  *    *             *****  *****     *    *** ** *
TYSY_MYCTU   138  PVSRRPCALHRHRRTLHGHTEPAEYDQAALRC.ASPGWRRTTVTGSADTR
Translate         PFNDNPRHFTDVVQRFLDATEPAE#NQAALRELLRTGGGKRTVTSSAKTH
B2235       4800  ctagaccctagggcctcggagcggtacggccgtccagggacagaatgaac
                  ctaaacgatcattagttaccaccaaaacctgattgcgggagctcgccaca
                  gtcccgctcccttgccccccgtcagtgctgtagcccccgacgccgcacc

                  ***** * *  ***   *    *          ***     ******* *
TYSY_MYCTU   187  VAVLNAIGSNERSATLITTGSVRALPQVVQSGHGIE^+++SVLRVFTDSD
Translate         VAVLNTIASDKRSAA#ATQPHARRAGP#SRA.VGIE^---TVLRVFTDAD
B2235       4950  gggcaaagtgacaggtgacccgccggctacg ggagCATTagtcgtaggg
                  tcttactccaaggccgccacacggcgcaggc tgta    cttgttcaca
                  tggaccatccacctcaacggctttaagatac gtca    gaggtcctat

                   *  *** *     *   **  **  * ** ********   *** *  *
TYSY_MYCTU   233  GNFGNPLGVINASKVEHRDRQQLA^^QSGYSETIFVDLPSPGSTTAHATI
Translate         VNSRNPLDVVDVRQVQTGDREHLA^^QLGYNETIFVDLPTADSTTVHTSI
B2235       5089  gataaccgggggacgcaggcgccgCCctgtagaatggccaggtaagcata
                  tacgactattatgatacgagaatc  atgaaactttatcccaccctacct
                  atcgtactatctgaaacgtgacga  agcttagatccgccccgtcgcgcc

                  ** *   *    *      *  **  * ****** ** * **
TYSY_MYCTU   281  HT^RTEIPFAGHP..TVGASWWLRERGTPINTLQVPAGIVQVS^++++++
Translate         HT^RNRTPVGQSPKQSERRSC#LRANGMPINTLQIPASIAQVK^------
B2235       5235  caGcacacggctcactgcctttccgagacaaaccacgaagcgaCGAAAGT
                  ac gagcctgaccaacaggcgatgcagtctactatccgtcata
                  ct gcatatcgacggggtgatggtcctgacctggcgtccggga

                               * * ** * **** **  ***  *  * ********
TYSY_MYCTU   321  +++++++++HGDLTAISARSEWAPEFAIHDLDSLDALAAADPADFPDDIA
Translate         ---------PSHRTRIRARAE#APEFTIHNFDSLGPLGSANPADFPDDIT
B2235       5369  ACCGGTGCAcaccaaaagcggtgcgtaacatgtcgccgtgacggtcggaa
                           cgagcgtgcgcagccatctaatactgctgccaccatcaatc
                           acctcgcgatgaatcaccccccgctctcccccccttcccta

                  **** * ***      ******** *** * * **  *  * *
TYSY_MYCTU   362  HYLWTWTDRSA^+++SLRARMFAANLGVTEDEATGAAAIRITDYLSRDLT
Translate         HYLWAWGDRSS^---#LRARMFAAHLGVPENEKTGSVAMQIIDSTTSATI
B2235       5501  ctctgtggctaCCGAtccgcatggccggcgagaagtggacaagtaatgaa
                  aatgcggagcg    atgcgttccatgtcaaaacgctctattacccccct
                  ccggcgctgac    ggtgagtcccacagacaggcgggggtccgtcacgc

                    ***** *** *    ****    *** *
TYSY_MYCTU   408  ..ITQGKGSLIHT^+++WSPEGWVRVAGRVV
Translate         #LITQGKRSLITT^---WSPED#AQVAGQVI
B2235       5643  tcaacgaatcaaaACCTtacggtgcgggcga
                  gttcagagcttcc    ggcaagcatcgatt
                  aaccgcgatgcac    gtcgcatgaacaca
```

Figure 3 (*See facing page for legend.*)

the coding sequences in the B32 *ftsY* region appear to be truncated, or frameshifted, relative to those on MTCY338). In this example, the region encoding the *hspD* gene on B32, which would be expected to occur on MTCY338, occurs instead on *M. tuberculosis* cosmid MTCY339 (which is located adjacent to, but not overlapping, MTCY427). Thus, there appears to have been a significant amount of gene shuffling between these two closely related species.

Another example of apparent gene shuffling between mycobacterial species involves the *mas* genes and associated ORFs of *M. leprae* and the close *M. tuberculosis* relative, *Mycobacterium bovis* BCG. In BCG these genes are in the order *orfII, orfI, mas,* and *orfIII,* with no more than 400 bp separating adjacent genes. In *M. leprae,* the apparent homologs are spread out over three regions. An *M. leprae mas* homolog with 58% identity to the BCG gene is located in cosmid B1170. A gene homologous (59% identity) to BCG *orfIII* (Q02278 | YMA2_MYCBO) occurs ~7.5 kb away as the fourth gene in a putative operon that is transcribed from the opposite strand as *mas.* A gene homologous to BCG *orfII* (Q02279), which shows 81% identity over 349 codons, occurs in cosmid L518 as the terminal gene in a 30-kb large pks operon. The closest homolog to *mas* in this putative pks operon is 8.5 kb away. Although it is not certain that these *mas*-related genes of *M. leprae* are orthologous to the BCG genes, it is quite clear that they are all members of a multigene (pks) family that may have arisen through gene duplication events.

At the protein level there are many strong similarities between *M. leprae* and *M. tuberculosis* gene products. We performed a cross comparison generating Smith–Waterman alignments between 1157 *M. leprae* proteins (reported in this study and elsewhere) and 1564 *M. tuberculosis* protein sequences reported in public databases. A plot of the percent identity for the best alignment of each *M. leprae* protein against the *M. tuberculosis* database is shown in Figure 6 (the percent identity values from long and short alignments were normalized by multiplying by the fraction of query amino acids represented in each alignment). Approximately one-quarter of the alignments (to the left of the vertical line in Fig. 6) have normalized matches ranging from ~40% to 87% identity. Most of these are likely to represent orthologous pairs, as at least 40% of the total *M. tuberculosis* proteins were represented in the target database. Most of the remaining proteins have matches ranging from 10% to 30% identity with at least one other *M. tuberculosis* protein in the data set. Although the stronger matches in this second group may represent alignments between paralogous members of protein families, the weaker ones are likely to represent only conserved motifs.

Examples of parologous mycobacterial genes include a DnaJ homolog on cosmid B1937, which shares 40% identity with the *M. tuberculosis* DnaJ protein and 38% identity with another, previously sequenced *M. leprae* DnaJ homolog that itself is 87% identical with the *M. tuberculosis* protein. Similarly, a Chaperonin 60 homolog in the overlapping cosmids B229/B1620 is 61% identical to a previously sequenced *M. leprae* Ch60 gene and 61% identical to an *M. tuberculosis* CH60 gene, whereas the latter two are 94% identical to each other. The lack of a complete genomic data set from either organism precludes a definitive analysis of orthologs and gene families.

Relationships to Other Bacterial Genomes

The current *M. leprae* genome map and sequence were examined for collinearity of genes with *E. coli, H. influenzae, M. genitalium,* and *B. subtilis.* Although patterns possibly indicative of genome duplications conserved from *B. subtilis* to *E. coli* have been described (Kunisawa 1995), the *M. leprae* data only support limited clustering at the operon level of related functions. Such clustering may be advantageous for gene transfer or gene regulation and, hence, convergent. A similar observation of widespread scrambling but consistent clustering is seen in comparison of large operons in *S. typhimurium* and *Pseudomonas denitrificans* (Roth et al. 1993). The clustering of genes in adjacent operons may reveal selective pressures to maintain proximity, for ex-

Figure 3 An extreme example of gene mangling in *M. leprae:* A region of cosmid B2235 homologous to *M. tuberculosis* TYSY_MYCTU. (*Top* line) The asterisks indicate positions of identity between the *M. tuberculosis* TYSY amino acid sequence (TYSY_MYCTU) and the conceptual translation of bases 4387–5673 of *M. leprae* cosmid B2235 (Translate). The PairWise (Birney and Thompson 1995) nucleotide triplets are displayed under the corresponding *M. leprae* amino acids. This represents only one possible *ythY* reconstruction involving 10 frameshifts (^) and 12 stop codons (#) using the results of analysis with Detect43 and PairWise programs. Additional identities covering the VGQG and AIPVQ sequences can be obtained with different hypothetical mutations. The TBLASTN probability for all GenBank nonredundant (NR) protein sequences at 376, 511, 003 amino acid residues is 6.4 e-67.

Mle	masA	..SADERC..
Mle	pksA	A......
Mle	pksB	......RC..
Mle	pksC	..SA......
Mle	pksD	..SADERC..
Mle	pksE	..SAD..C..
Mle	pksF	..SA...C..
Mbo	mas	..SADERC..
San	ole	..SA..RCT.
Sco	act	R.SAL..CMY
Sgl	tcm	..SAL..CNO
Sro	fren	..SAL.CRMY
Sgr	gris	..SAL.CRM.
Sco	whiE	..SAL..CN.
Ppa	msas	..SA..RC..
Ser	EryA-1	ACSA..RC..
Ser	EryA-2	..SA..RC..
Ser	EryA-3	..SA...C..
Ser	EryA-4	..SADERC..
Ser	EryA-5	..SA..RC..
Ser	EryA-6	..SA..RCT.
Sav	avr-1	ACSA..R...
Sav	avr-2	..SAd.Rc..
Sav	avr-3	..sA..Rc..
Sav	avr-4	..SA..rc..
Sav	avr-5	..Sa..Rc..
Sav	avr-6	..Sad.Rc..
Sav	avr-7	..SA..RC..
Sav	avr-8	..Sa...c..
Sav	avr-9	..SAD.RCT.
Sav	avr-10	..SAD.RC..
Sav	avr-11	..SAD.RC..
Sav	avr-12	..SAD.Rc..
Rno	fas	..SA..RC..

Figure 4 Modular organization and enzymatic motifs in synthetases for fatty acids (fas), mycocerosic acid (mas), methylsalicylic acid (msas), erythromycin (eryA), actinorhodin (act), tetracenomycin (tcm), frenolicin (fren), griseusin (gris), and avermectin (avr) from the following organisms: *M. leprae* (Mle, this work), *Saccharopolyspora erythraea* (Ser), *Rattus norvegicus* (Rno), *Penicillum patulum* (Ppa), *Streptomyces antibioticus* (San), *S. avermitilis* (Sav), *S. coelicolor* (Sco), *S. glaucescens* (Sgl), and *S. griseus* (Sgr). The motifs are abbreviated as follows: (A) acyl transferase; (C) acyl carrier protein; (D) dehydratase; (E) enoyl reductase; (L) chain length factor; (M) aromatase; (N) aromatase/cyclase; (O) *O*-methyltransferase; (R) ketoreductase; (S) ketoacyl–ACP synthase; (T) thioesterase; (Y) cyclase. Lowercase letters indicate uncertainty in functional assignment (Donadio et al. 1991; Mathur and Kolattukudy 1992). The 5′ end of Mle *pksA* is missing from our current sequence. Underlines indicate all of the protein domains ending in translational terminators, except for Sav avr modules, where the sequence data are incomplete. The CLF domain appears closer to pksC ($P = 8.9e\text{-}13$) than to masA ($P = 0.0018$).

ample, for recombination, coassembly, or coregulation. A cluster of seven genes in *M. leprae* are related to carbohydrate catabolism via the glycolytic and pentose phosphate pathways. Although no two of these genes is directly adjacent to each other in *E. coli,* two of them (*tpi* and *pgk*) are in the same order and orientation in *B. subtilis*. The gene order for the two transcriptionally converging operons in *M. leprae* is 5′ (*tkt tal zwf*) 3′ and 3′ (*ppc tpi pgk gap*) 5′, and the corresponding order in *B. subtilis* is 3′ (*eno pgm tpi pgk*) 5′. At least two similar clusters, including some of these genes, exist in *E. coli;* they are 5′ (*gapB pgk fda*) 3′ and 5′ (*zwf edd eda*) 3′.

The largest *M. leprae* region without identified genes covers ~7 kb (on cosmid L471, mentioned above). Such regions are rare in bacteria, are generally <2 kb (Daniels et al. 1992; M. Raha, M. Kihara, I. Kawagishi, and R.M. Macaab, unpubl.), and are often later found to contain genes (Robison et al. 1994). This particular region does not appear to have any problems with data quality, yet it contains few ORFs over 100 codons, and none over 170 codons. The ORFs have poor *M. leprae*-specific codon usage and no significant database matches.

Genes with Similarities to Eukaryotes

A number of putative genes listed in Table 1 (including *xheA, dhaP, leuA, udp, glyS, pepN, ctrA, hup, nakA, ybqR, ycjC,* and *ybtY*) have significant homology to known and hypothetical proteins from distant species (e.g., yeast, plants, human, and other eukaryotes). These sequences contain regions that could be called ancient conserved regions (ACRs) (Green et al. 1993), although a better term might be phylogenetically diverse conserved regions. The latter term has the advantage of avoiding the implication that such regions were actually maintained over the entirety of the time separating the most distantly related species, when instead they might represent remnants of more recent acquisitions by horizontal transfer. This possibility holds even in cases where the exact level of ancient in ACRs is specified (Doolittle et al. 1996).

METHODOLOGY

Genome Sequencing

Cosmids were selected from a mapped, overlapping set comprising four large segments of the *M. leprae* genome (Eiglmeier et al. 1993). The cosmid clones that were sequenced in this study were derived from three of these segments (contigs), and included cosmids depicted in Figure 1. High purity DNA preparations were made from each clone [using anion exchange columns with a procedure recommended by the manufacturer (Qiagen, Inc.) followed by equilibrium CsCl density gradient centrifugation]. The DNA was sheared by sonication or nebulization to an average size of 1–2 kb and fractionated on an agarose gel. A 1.5-kb average-size fraction was treated with T4 DNA polymerase, capped with *Bst*XI link-

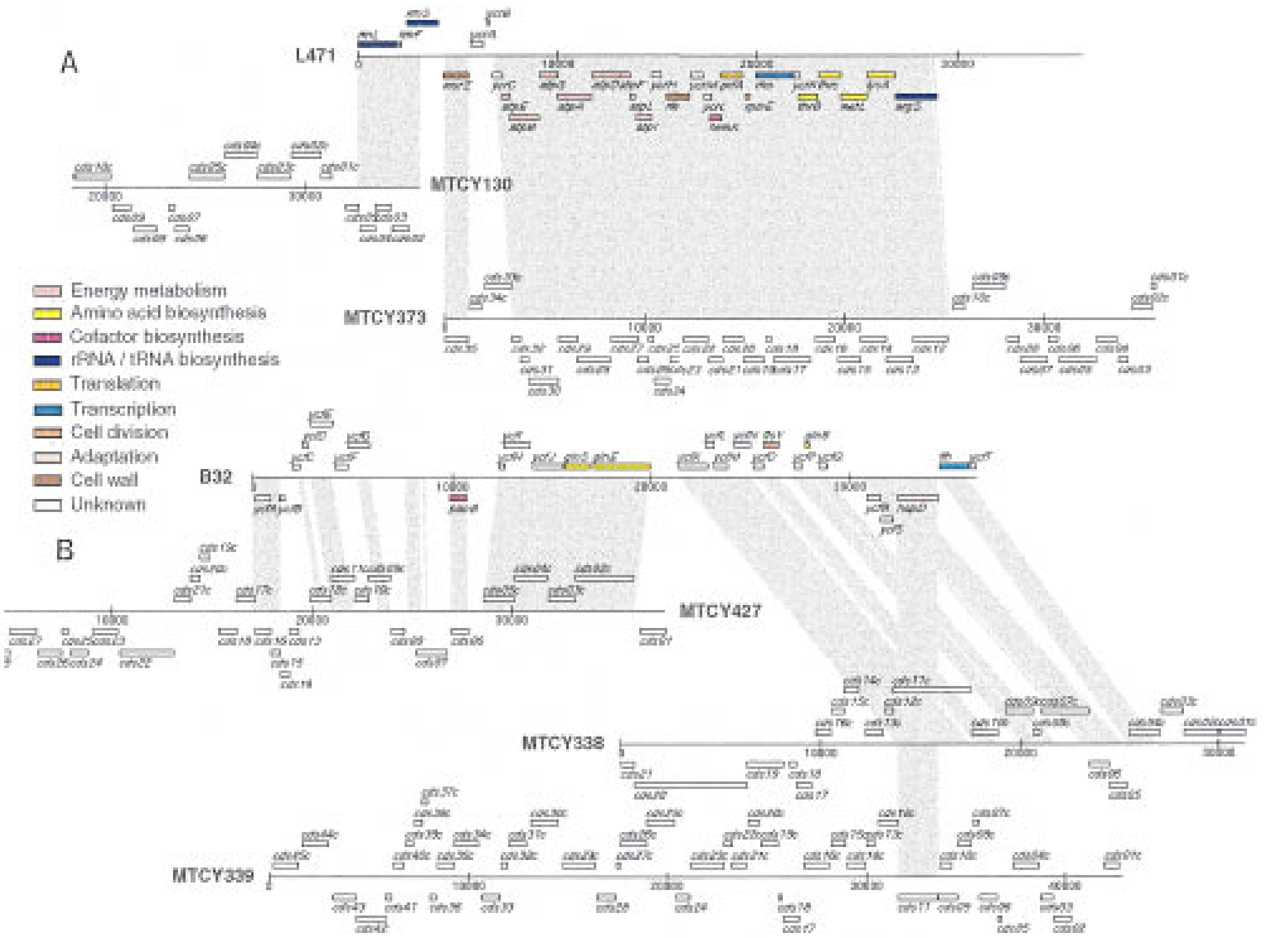

Figure 5 DNA-level matches between two *M. leprae* cosmids from this study and *M. tuberculosis* cosmids (Sanger Center, GenBank). The alignments show two particular examples from an exhaustive comparison of the set of *M. leprae* cosmids reported here against all available *M. tuberculosis* cosmids (see text). The shading indicates regions of significant similarity between each pair of cosmids. The alignments for cosmid L471 spanned a total of 26,302 nucleotides with 75%–94% identity; those for cosmid B32 spanned 24,009 nucleotides with 71%–85% identity. The *M. leprae* ORFs are color coded according to function as indicated. The sequences were compared and aligned using Cross_match, an implementation of the Smith–Waterman algorithm developed by P. Green (University of Washington, Seattle). Alignments with >60% identity were sorted using Matchtable (P. Richterich, unpubl.) and examined using a Web browser. A table summarizing the positions of aligned segments between each pair of cosmids was assembled; it was read by a Perl-tk script, Cosmid_map (R. Gibson, unpubl.) in conjunction with two other tables similar to Table 1 (but sorted by cosmid) summarizing the positions of coding frames and functional information (if available) for the *M. leprae* and *M. tuberculosis* cosmids.

ers (5′-TCTAGACCACCTGC and 5′-GTGGTCTAGA in 100- to 1000-fold molar excess), gel purified, and ligated to one of a set of 20 uniquely tagged *Bst*XI-cut plex vectors (Church and Kieffer-Higgins 1988) (M. Rubenfield, P. Rice, and D. Smith, unpubl.) to construct a series of shotgun subclone libraries. Each pool of 20 clones was picked using a 100 µl glass capillary attached to a light vacuum source to touch sequentially 20 colonies from different libraries. The capillary was then placed into a flask with growth medium to rinse out the cells. DNA was purified from a sufficient number of clones (Roach 1995) to obtain 5- to 10-fold sequence redundancy with 250- to 350-base average read lengths (typically 12 sets of 96 pools per cosmid).

DNA samples were chemically sequenced, separated on polyacrylamide gels, and transferred onto nylon membranes by electroblotting (Church and Kieffer-Higgins 1988) or by direct transfer electrophoresis from 40-cm gels (Richterich and Church 1993). In some cases, cycle sequencing reactions using Sequitherm polymerase were used. The DNA was covalently bound to the membranes by exposure to ultraviolet light and hybridized with labeled oligonucleotides complementary to tag sequences on the plex vectors (Church and Kieffer-Higgins 1988). The membranes were washed to remove nonspecifically bound probe and exposed to X-ray film to visualize individual sequence ladders. After autoradiography, the hybridized probe was removed by incubation at

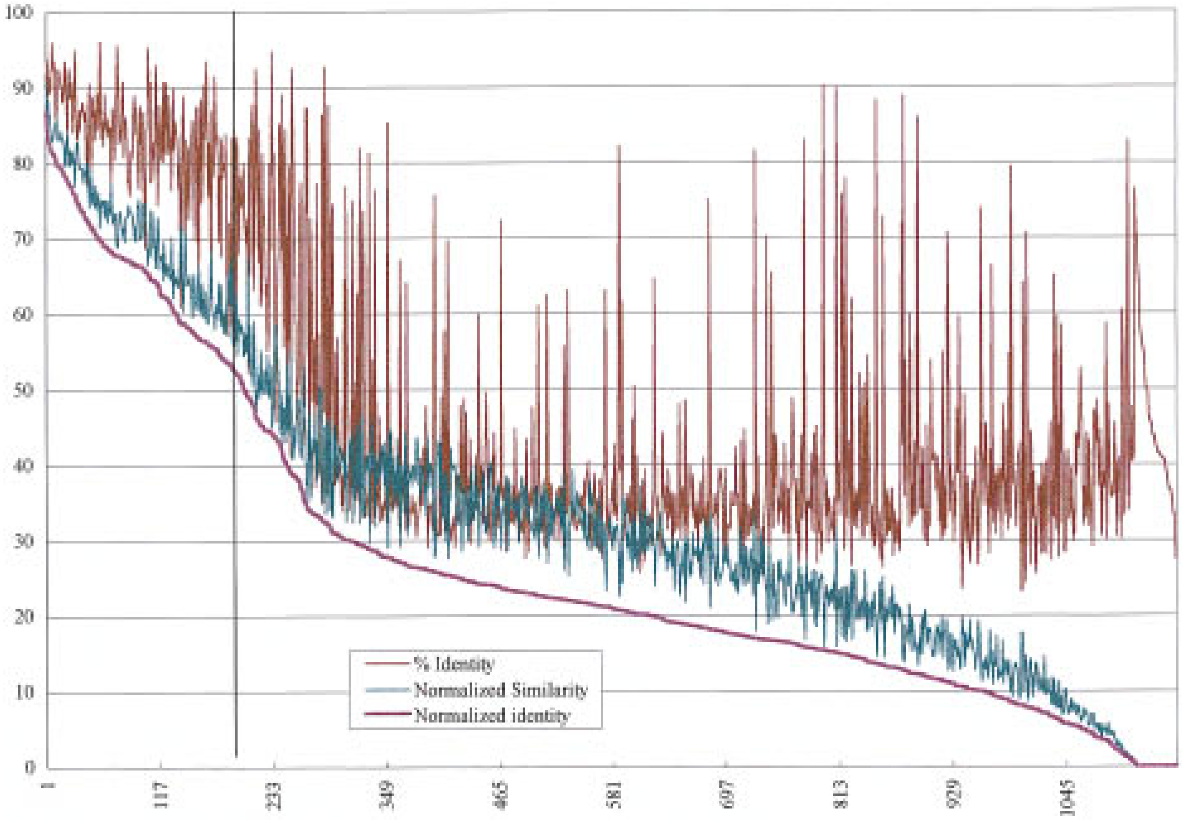

Figure 6 Summary of alignments from similarity searches between 1157 *M. leprae* proteins (including all of the gene products from this study) and 1564 *M. tuberculosis* proteins from GenPept. Each of the *M. leprae* proteins was searched against the set of *M. tuberculosis* proteins using an implementation of the Smith–Waterman algorithm with default parameters on a Biocellerator (Compugen) in conjunction with the GCG Wisconsin Package. The Normalized Similarity and % Identity values were obtained from the best alignment for each *M. leprae* protein by multiplying by the fraction of query amino acids represented in each alignment (no. of query residues in alignment/total query length). This was done to provide a better indication of the overall similarity of each *M. leprae* protein to the best *M. tuberculosis* homolog. The resulting values were termed Normalized Identity and Normalized Similarity. The pairs were sorted according to the Normalized Identity values in descending order, and the normalized values were plotted together with the raw percent identity values (for comparison) on a graph.

65°C, and the hybridization cycle was repeated with another plex oligonucleotide until the membrane had been probed 25–41 times (depending on the number of templates present). Thus, each gel produced a large number of films, each containing new sequencing information. Whenever a new blot was processed, it was initially probed for an internal standard sequence added to each of the pools.

Digital images of the films were generated using a laser-scanning densitometer (Molecular Dynamics, Sunnyvale, CA). The digitized images were processed on computer workstations (VaxStation 4000's) using the program REPLICA (Church et al. 1994). Image processing included lane straightening, contrast adjustment to smooth out intensity differences, and resolution enhancement by iterative gaussian deconvolution. The sequences were then automatically picked in REPLICA and displayed for interactive proofreading before being stored in a project database (each cosmid was saved in a separate project directory). The proofreading was accomplished by a quick visual scan of the film image followed by mouse clicks on the bands of the displayed image to modify the base calls. For most sequences derived by chemical sequencing, the error rate of the REPLICA base calling software was 2%–5%; a smaller percentage of samples had higher error rates, particularly near the end of the sequence read. Each sequence automatically receives a number corresponding to the blot number (microtiter plate and probe information) and lane set number (corresponding to microtiter plate columns). This number serves as a permanent identifier of the sequence so it is always possible to identify the origin of any particular sequence without recourse to a specialized database.

The sequences were assembled using the programs GTAC and FALCON (Church et al. 1994; Gryan 1995). These programs have proven to be fast and reliable for cosmid sequences. The assembled contigs are displayed using a modified version of GelAssemble, developed by the Genetics Computer Group (GCG) (Devereux et al. 1984) and modified by G. Church and P. Richterich to interact with REPLICA. This provides an integrated editor that allows multiple sequence gel images to be instantaneously called up from the REPLICA database and displayed to allow rapid scanning of contigs and proofreading of gel traces where discrepancies occur between different sequence reads in the assembly. Any ambiguous regions or regions with low coverage that required more coverage were resequenced by primer-directed cycled sequencing using commercially available kits and cosmid or multiplex pool templates. Each assembly was analyzed for regions with only single-strand coverage using the program SECSO (A. Graf, unpubl.).

Some of the cosmids—L518, B2126, L247, and B1170—contained large repeats that required additional analysis. In these cases, positional information associated with sequences derived from the two ends of each plasmid insert (which should have opposite orientation with respect to each other and should be separated by the average insert size of 1.5 kb) was used to remove misassembled sequences and align contigs in the proper order. This was done using the program CHECKMATES (R. Lundstrom and C. Tulig, unpubl.) which provides information on the location, spacing, and orientation of sequence pairs (mates) that do not fall within the normal range. (Comparison of cosmid restriction digests with two enzymes against predicted fragment sizes from the assembled sequence was used to verify correct assembly.) All contigs were analyzed using GenomeBrowser, and the output was examined to identify tRNAs, repetitive elements, and potential coding regions.

ACKNOWLEDGMENTS

This work was funded by National Institutes of Health grants R01HG00520 and P01HG01106 to D.R.S. and J.M. and Department of Energy grant DE-FG02-87ER60565 to G.M.C.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked ‘‘advertisement’’ in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Altschul, S.F., W. Gish, W. Miller, E.W. Myers, and D.J. Lipman. 1990. Basic local alignment search tool. *J. Mol. Biol.* 215: 403–410.

Barrell, B.G., M.A. Rajandream, and S.V. Walsh. 1996. Mycobacterium tuberculosis sequencing project. http://www.sanger.ac.uk/pathogens/.

Bechhofer, D.H., K.K. Hue, and D.A. Shub. 1994. An intron in the thymidylate synthase gene of Bacillus bacteriophage beta 22: Evidence for independent evolution of a gene, its group I intron, and the intron open reading frame. *Proc. Natl. Acad. Sci.* 91: 11669–11673.

Belfort, M., M.E. Reaban, T. Coetzee, and J.Z. Dalgaard. 1995. Prokaryotic introns and inteins: A panopoly of form and function. *J. Bacteriol.* 117: 3897–3903.

Bergh, S. and S.T. Cole. 1994. MycDB: An integrated mycobacterial database. *Mol. Microbiol.* 12: 517–534. Release 4-22 (1996): http://www.biochem.kth.se/MycDB.html.

Birney, E. and J. Thompson. 1995. PairWise. http://www.ocms.ox.ac.uk/~birney/wise/topwise.html.

Bult, C.J., O. White, G.J. Olsen, L. Zhou, R.D. Fleischmann, G.G. Sutton, J.A. Blake, L.M. Fitzgerald, R.A. Clayton, J.D. Gocayne, A.R. Kerlavage, B.A. Dougherty, J.-F. Tomb, M.D. Adams, C.I. Reich, R. Overbeek, E.F. Kirkness, K.G. Weinstock, J.M. Merrick, A. Glodek, J.L. Scott, N.S.M. Geoghagen, J.F. Weidman, J.L. Fuhrmann, D. Nguyen, T.R. Utterback, J.M. Kelley, J.D. Peterson, P.W. Sadow, M.C. Hanna, M.D. Cotton, K.M. Roberts, M.A. Hurst, B.P. Kaine, M. Borodovsky, H.-P. Klenk, C.M. Fraser, H.O. Smith, C.R. Woese, and J.C. Venter. 1996. Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii. Science* 273: 1058–1073.

Cawthon, R.M., R. Weiss, G.F. Xu, D. Viskochil, M. Culver, J. Stevens, M. Robertson, D. Dunn, R. Gesteland, and P. O’Connell. 1990. A major segment of the neurofibromatosis type 1 gene: cDNA sequence, genomic structure, and point mutations. *Cell* 62: 193–201.

Cherry, J.L., H. Young, L.J. Di Sera, F.M. Ferguson, A.W. Kimball, D.M. Dunn, R.F. Gesteland, and R.B. Weiss. 1994. Enzyme-linked fluorescent detection for automated multiplex DNA sequencing. *Genomics* 20: 68–74.

Church, G.M., G. Gryan, N. Lakey, S. Kieffer-Higgins, L. Mintz, M. Temple, M. Rubenfield, L. Jaehn, H. Ghazizadeh, K. Robison, and P. Richterich. 1994. Automated multiplex sequencing. In *Automated DNA sequencing and analysis techniques* (ed. M. Adams, C. Fields, and J.C. Venter), pp. 11–16. Academic Press, San Diego, CA.

Church, G.M. and S. Kieffer-Higgins. 1988. Multiplex DNA sequencing. *Science* 240: 185–188.

Daniels, D.L., G. Plunkett, V. Burland, and F.R. Blattner. 1992. Analysis of the *Escherichia coli* genome: DNA sequence of the region from 84.5 to 86.5 minutes. *Science* 257: 771–778.

Davis, E.O., H.S. Thangaraj, P.C. Brooks, and M.J. Colston. 1994. Evidence of selection for protein introns in the recAs of pathogenic mycobacteria. *EMBO J.* 13: 699–703.

Devereux, J., P. Haeberli, and O. Smithies. 1984. A comprehensive set of sequence analysis programs for the VAX. *Nucleic Acids Res.* 12: 387–395.

Donadio, S., M.J. Staver, J.B. AcAlpine, S.J. Swanson, and L. Katz. 1991. Modular organization of genes required for complex polyketide biosynthesis. *Science* 252: 675–679.

Doolittle, R., D. Feng, S. Tsang, G. Cho, and E. Little. 1996. Determining divergence times of the major kingdoms of living organisms with a protein clock. *Science* 271: 470–476.

Dujon, B. 1996. The yeast genome project: What did we learn? *Trends Genet.* 12: 263–270.

Durbin, R. and J. Thierry-Mieg. 1991-1995. A *C. elegans* database. Documentation, code and data available from RFP servers at lirmm.lirmm.fr, cele.mrc-lmb.cam.ac.uk and ncbi.nlm.nih.gov. Also http://probe.nalusda.gov: 8000/acedocs/index.html.

Eiglmeier, K., N. Honoré, S.A. Woods, B. Caudron, and S.T. Cole. 1993. Use of an ordered cosmid library to deduce the genomic organization of *Mycobacterium leprae. Mol. Microbiol.* 7: 197–206.

Fichant, G.A. and C. Burks. 1991. Identifying potential tRNA genes in genomic DNA sequences. *J. Mol. Biol.* 220: 659–671.

Fleischmann, R.D., M.D. Adams, O. White, R.A. Clayton, E.F. Kirkness, A.R. Kerlavage, C.J. Bult, J.-F. Tomb, B.A. Dougherty, J.M. Merrick, K. McKenney, G. Sutton, W. FitzHugh, C. Fields, J.D. Gocayne, J. Scott, R. Shirley, L.-I. Liu, A. Glodek, J.M. Kelley, J.F. Weidman, C.A. Phillips, T. Spriggs, E. Hedblom, M.D. Cotton, T.R. Utterback, M.C. Hanna, D.T. Nguyen, D.M. Saudek, R.C. Brandon, L.D. Fine, J.L. Fritchman, J.L. Furmann, N.S.M. Geoghagen, C.L. Gnehm, L.A. McDonald, K.V. Small, C.M. Fraser, H.O. Smith, and J.C. Venter. 1995. Whole genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* 269: 496–502.

Fraser, C.M., J.D. Gocayne, O. White, M.D. Adams, R.A. Clayton, R.D. Fleischmann, C.J. Bult, A.R. Kerlavage, G. Sutton, J.M. Kelley, J.L. Fritchman, J.F. Weidman, K.V. Small, M. Sandusky, J.L. Fuhrmann, D.T. Nguyen, T.R. Utterback, D.M. Saudek, C.A. Phillips, J.M. Merrick, J.-F. Tomb, B.A. Dougherty, K.F. Bott, P.-C. Hu, T.S. Lucier, S.N. Peterson, H.O. Smith, C.A.I. Hutchison, and J.C. Venter. 1995. The minimal gene complement of Mycoplasma genitalium. *Science* 270: 397–403.

Fsihi, H., V. Vincent, and S.T. Cole. 1996. Homing events in the gyrA gene of some mycobacteria. *Proc. Natl. Acad. Sci.* 93: 3410–3415.

Fsihi, H. and S.T. Cole. 1995. 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.* 16: 909–919.

Gaasterland, T. and C.W. Sensen. 1996. MAGPIE: Automated genome interpretation. *Trends Genet.* 12: 76–78. Multiple tools for automated genome interpretation in an integrated environment. http://www.mcs.anl.gov/home/gaasterl/magpie.html.

Gesteland, R.F., R.B. Weiss, and J.F. Atkins. 1992. Recoding: Programmed genetic decoding. *Science* 257: 1640–1641.

Green, P., D. Lipman, L. Hiller, R. Waterston, D. States, and J.-M. Claverie. 1993. Ancient conserved regions in new gene sequences and the protein databases. *Science* 259: 1711–1716.

Gryan, G. 1995. Falcon. ftp://rascal.med.harvard.edu./gryan/falcon/aaa_readme.falcon.

Hall, B.G. and P.M. Sharp. 1992. Molecular population genetics of Escherichia coli: DNA sequence diversity at the celC, crr, and gutB loci of natural isolates. *Mol. Biol. Evol.* 9: 654–656.

Hall, B.G. and L. Xu. 1992. Nucleotide sequence, function, activation, and evolution of the cryptic asc operon of *Escherichia coli* K12. *Mol. Biol. Evol.* 9: 688–706.

Himmelreich, R., H. Hilbert, H. Plagens, E. Pirkl, B. Li, and R. Herrmann. 1996. Complete sequence analysis of the genome of the bacterium *Mycoplasma pneumoniae. Nucleic Acids Res.* 24: 4420–4449.

Honoré, N., S. Bergh, S. Chanteau, F. Doucet-Populaire, K. Eiglmeier, T. Garnier, C. Georges, P. Launois, T. Limpaiboon, S. Newton, K. Nianag, P. del Portillo, G.R. Ramesh, P. Reddi, P.R. Ridel, N. Sittisombut, S. Wu-Hunter, and S.T. Cole. 1993. Nucleotide sequence of the first cosmid from the *Mycobacterium leprae* genome project: Structure and function of the Rif-Str regions. *Mol. Microbiol.* 7: 207–214.

Kaneko, T., S. Sato, H. Kotani, A. Tanaka, E. Asamizu, Y. Nakamura, N. Miyajima, M. Hirosawa, M. Sugiura, S. Sasamoto, T. Kimura, T. Hosouchi, A. Matsuno, A. Muraki, N. Nakazaki, K. Naruo, S. Okumura, S. Shimpo, C. Takeuchi, T. Wada, A. Watanabe, M. Yamada, M. Yasuda, and S. Tabata. 1996. Sequence analysis of the genome of the unicellular Cyanobacterium Synechocystis sp. strain PCC6803. II. Sequence determination of the entire genome and assignment of potential protein-coding regions. *DNA Res.* 3: 109–136.

Kunisawa, T. 1995. Identification and chromosomal distribution of DNA sequence segments conserved since divergence of Escherichia coli and Bacillus subtilis. *J. Mol. Evol.* 40: 585–593.

Liesack, W., C. Pitulle, S. Sela, and E. Stackebrandt. 1990. Nucleotide sequence of the 16S rRna from Mycobacterium leprae. *Nucleic Acids Res.* 18: 5558.

Mathur, M. and P.E. Kolattukudy. 1992. Molecular cloning and sequencing of the gene for mycocerosic acid synthase, a novel fatty acid elongating multifunctional enzyme, from *Mycobacterium tuberculosis* var. *bovis* Bacillus Calmette-Guerin. *J. Biol. Chem.* 267: 19388–19395.

Murray, P.J. and R.A. Young. 1992. Stress and immunological recognition in host-pathogen interaction. *J. Bacteriol.* 174: 4193–4196.

Mushegian, A.R. and E.V. Koonin. 1996. A minimal gene set for cellular life derived by comparison of complete bacterial genomes. *Proc. Natl. Acad. Sci.* 93: 10268–10273.

Ozer, J., R. Chalkley, and L. Sealy. 1993. Characterization of rat pseudogenes for enhancer factor I subunit A: Ripping provides clues to the evolution of the EFIA/dbpB/YB-1 multigene family. *Gene* 133: 187–195.

Perler, F.B., E.O. Davis, G.E. Dean, F.S. Gimble, W.E. Jack, N. Neff, C.J. Noren, J. Thorner, and M. Belfort. 1994. Protein splicing elements: Inteins and exteins—A definition of terms and recommended nomenclature. *Nucleic Acids Res.* 22: 1125–1127.

Philipp, W.J., S. Poulet, K. Eiglmeier, L. Pascopella, V.

Balasubramanian, B. Heym, S. Bergh, B.R. Bloom, W.R.J. Jacobs, and S.T. Cole. 1996. An integrated map of the genome of the tubercle bacillus, *Mycobacterium tuberculosis* H37Rv, and comparison with *Mycobacterium leprae. Proc. Natl. Acad. Sci.* 93: 3132–3137.

Pietrokovski, S. 1996. A new intein in cyanobacteria and its significance for the spread of inteins. *Trends Genet.* 12: 287–288.

Poulet, S. and S.T. Cole. 1995. Characterization of the highly abundant polymorphic GC-rich-repetitive sequence (PGRS) present in Mycobacterium tuberculosis. *Arch. Microbiol.* 163: 87–95.

Raha, M., M. Kihara, I. Kawagishi, and R.M. Macnab. 1993. Organization of the *Escherichia coli* and *Salmonella typhimurium* chromosomes between flagellar regions IIIa and IIIb, including a large non-coding region. *J. Gen. Microbiol.* 139: 1401–1407.

Richterich, P. and G.M. Church. 1993. DNA sequencing with direct transfer electrophoresis and non-radioactive detection. *Methods Enzymol.* 218: 187–222.

Roach, J. 1995. Random subcloning. *Genome Res.* 5: 464–473.

Robison, K. and G.M. Church. 1995. GenomeBrowser. http://www.belmont.com/gb.html.

Robison, K., W. Gilbert, and G.M. Church. 1994. Large-scale bacterial gene discovery by similarity search. *Nature Genet.* 7: 205–214.

Roth, J.R., J.G. Lawrence, M. Rubenfield, S. Kieffer-Higgins, and G.M. Church. 1993. Characterization of the cobalamin (vitamin B12) biosynthetic genes of *Salmonella typhimurium*. *J. Bacteriol.* 175: 3303–3316.

Sherman, D.R., P.J. Sabo, M.J. Hickey, T.M. Arain, G.G. Mahairas, Y. Yuan, C.E. Barry, and C.K. Stover. 1995. Disparate responses to oxidative stress in saprophytic and pathogenic mycobacteria. *Proc. Natl. Acad. Sci.* 92: 6625–6629.

Shine, J. and L. Dalgarno. 1975. Correlation between the 3′-terminal-polypyrimidine sequence of 16S RNA and translational specificity of the ribosome. *Eur. J. Biochem.* 57: 221–230.

Singer, M.J., B.A. Marcotte, and E.U. Selker. 1995. DNA methylation associated with repeat-induced point mutation in Neurospora crassa. *Mol. Cell. Biol.* 15: 5586–5597.

Smith, D.R., L. Doucette-Stamm, P.W. Rice, M. Rubenfield, P. Richterich, S. Toth, B. Seitz, C. Butler, H.-M. Lee, and J. Dubois. 1996. Microbial genome sequencing by integrated ABI and multiplex sequencing. *Microb. Comp. Genomics* 1: 200.

Woods, S.A. and S.T. Cole. 1990. A family of dispersed repeats in *Mycobacterium leprae. Mol. Microbiol.* 4: 1745–1751.

Received February 13, 1997; accepted in revised form June 10, 1997.