Microb. Resour. Syst. **32**(1): 25-37, 2016

Core housekeeping proteins useful for identification and classification of mycobacteria

Takuya Mizuno[1, 2)]*, Tatsuya Natori[1)], Izumi Kanazawa[1)], Ibrahim Eldesouky[3)], Hajime Fukunaga[1)] and Takayuki Ezaki[1)]

[1)]Department of Microbiology, Regeneration and Advanced Medical Science, Gifu University Graduate School of Medicine, 1-1 Yanagido, Gifu, 501-1194, Japan
[2)]Gifu Prefectural Institute of Health and Environmental Sciences
1-1 Naka-fudogaoka, Kakamigahara, 504-0838, Japan
[3)]Department of Bacteriology, Mycology and Immunology, Faculty of Veterinary Medicine, Kafrelsheikh University, 33516, Egypt

We generated and analyzed draft genomes for 42 *Mycobacterium* strains representing 30 different species to select core housekeeping proteins (HKPs) that would be useful for differentiating among closely related *Mycobacterium* species.

HKPs of *Mycobacterium tuberculosis* H37Rv were selected as reference proteins, and values (designated diversity values) representing the amino acid differences between these H37Rv HKPs and those of other individual *Mycobacterium* species were calculated for each HKP. From seven NCBI protein categories, we analyzed the 107 proteins commonly found in all 30 *Mycobacterium* species. We then selected the 12 most variable HKPs to construct a concatenated protein sequence (designated C12HKP). The average C12HKP diversity value for these 30 *Mycobacterium* species was 22.50%. Phylogenetic trees constructed with either C12HKP or C50RP (50 concatenated sequences from 50S and 30S ribosomal proteins) had reliable bootstrap values that were higher than those of a 16S rRNA gene tree. Of the three entities (C12HKP, 16S rRNA gene, and C50RP), C12HKP exhibited the greatest diversity. To differentiate among closely related species within the genus *Mycobacterium*, the C12HKP entity provided the most powerful and discriminating dataset.

Key words: housekeeping proteins, mycobacterium, classification, identification, MLSA

INTRODUCTION

To date, 160 species have been classified into the genus *Mycobacterium* (http://www.bacterio.net/mycobacterium.html); however, the interspecies variation of the 16S rRNA gene sequence among these 160 species is less than 6%. In this genus, 16S rRNA gene sequence identities between several closely related species are higher than 99%. Many authors have used housekeeping genes (HKGs) such as *dnaJ1* (Yamada-Noda *et al.*, 2007), *gyrB* (Kasai *et al.*, 2000; Niemann *et al.*, 2000), *rpoB* (Kim *et al.*, 1999), *hsp65* (Brunello *et al.*, 2001; McNabb *et al.*, 2004; Ringuet *et al.*, 1999), *recA* (Blackwood *et al.*, 2000), or *sodA* (Adékambi & Drancourt, 2004), or some combination thereof, to discriminate between closely related species. The 16S rRNA gene sequence identity between *Mycobacterium avium* and *M. intracellulare* is 100%, whereas that of *rpoB* is 95.8%. Therefore, *rpoB* seems to be more powerful than the 16S rRNA gene for discriminating among *Mycobacterium* species, but the data obtained by using this single HKG are not variable enough to fully evaluate phylogenetic relationships in this genus (Kim *et al.*, 1999).

More recently, multilocus sequence analysis (MLSA) has been used as a powerful alternative for identifying and classifying bacteria (Gevers *et al.*, 2005; Schleifer, 2009). This multigenic approach fulfills the recommendation of the *ad hoc* committee for the re-evaluation of the definition of bacterial species (Stackebrandt *et al.*, 2002). However, researchers have used many different sets of genes for MLSA (Adékambi & Drancourt, 2004; Dai *et al.*, 2011; Devulder *et al.*, 2005). As a result, the proposed phylogenetic relationships determined among closely related species by using different MLSA datasets differ substantially.

*Corresponding author
E-mail: mizuno-takuya@pref.gifu.lg.jp
Accepted: January 29, 2016

Draft genome sequencing has become a relatively easy and inexpensive procedure, and complete protein sequences are available from these draft sequences. In 2004, Cole *et al*. (1998) published a complete genome of *M. tuberculosis* H37Rv; this was the first such report for the genus *Mycobacterium*. Recently, genome analysis of many mycobacterial strains has been completed, but most of the sequenced strains listed on the National Center for Biotechnology Information (NCBI) home page are *M. tuberculosis* strains.

Determination of average nucleotide identity (ANI) with whole genome information is a powerful method for analyzing the relationships among strains within a species. ANI values of >95% correspond to the traditional >70% DNA–DNA reassociation standard that has been used to define a species (Konstantinidis & Tiedje, 2005). ANI data have been used for species identification (Cho *et al.*, 2013) and can distinguish species more clearly than can MLSA data (Adékambi & Drancourt, 2004; Dai *et al.*, 2011; Devulder *et al.*, 2005). However, ANI values do not offer enough information to determine interspecies relationships.

Here, we generated draft genome sequences for 19 *Mycobacterium* species and then used these genome data and publicly available data on other *Mycobacterium* species to analyze 107 housekeeping proteins (HKPs) that are commonly found in, and used to analyze, *Mycobacterium* species. We then generated datasets based on two different concatenated protein sequences. One concatenated sequence was constructed by using 50 proteins from 50S or 30S ribosomal proteins and designated C50RP. The other was constructed from the 12 most variable HKPs and designated C12HKP; these 12 HKPs were selected from among 107 proteins belonging to seven protein categories of NCBI classification, namely ribosomal proteins (small and large subunit), molecular chaperones, DNA/RNA replication or modification enzymes, tRNA-synthesis proteins, and cell division-associated proteins.

Datasets generated by using either of the concatenated protein sequences or the 16S rRNA gene were compared with regard to their usefulness in identifying and differentiating species in the genus *Mycobacterium*.

MATERIALS AND METHODS

Bacterial strains and DNA preparation

Table 1 lists the 27 strains (9 of which were type strains) that represented 19 non-tuberculosis mycobacteria and were used to generate the draft genome sequence. Strains were provided by Gifu University School of Medicine as the Gifu Type Culture Collection (GTC), these strains were deposited in the Japan National Collection of Bacterial Pathogen (JNBP) Data Base. Each strain was cultured on Middlebrook 7H11 Agar supplemented with OADC (BD Co., Franklin Lakes, NJ, USA) and incubated at 30℃. EZ beads (AMR Co., Ltd., Gifu, Japan) were used according to the manufacturer's instructions to disrupt bacterial cells, and DNA purification was performed as described previously (Boom *et al.*, 1990).

Draft genome sequence

All genome sequences were determined by using the Illumina HiSeq system (Illumina, Inc., San Diego, CA). Each genome sequence was automatically annotated with the Microbial Genome Annotation Pipeline (MiGAP) ver. 2.18 (http:/www.migap.org/). Start codon positions and any additional genes missing from the MiGAP annotation were manually inspected and corrected.

Sequence diversity among *Mycobacterium* species was calculated for 30 species and 42 strains (Table 1). The DNA sequence of the 16S rRNA gene and the amino acid sequences of 107 HKPs were used to calculated percent diversities. *Mycobacterium tuberculosis* H37Rv was used as a reference strain to generate multiple sequence alignments and calculate percent diversities for each of the other strains; these multiple sequence alignments were generated with Clustal W (Thompson *et al.*, 2002), implemented in DNASIS Pro version 3.0 software (Hitachi Software Engineering Co., Ltd., Tokyo, Japan). Likewise, intraspecies diversities were calculated for *M. abscessus* and *M. tuberculosis*. Strains for these sequences were obtained from the Patric database (http://www.patricbrc.org).

The set of 107 HKP proteins (Fig. 1) comprised 50S and 30S ribosome proteins (n=50), molecular chaperones (n=7), proteins associated with replication and modification (n=21), tRNA-synthesis proteins (n=24), proteins associated with cell division (n=4), and RecA.

Table 1 List of *Mycobacterium* species analyzed

Strain name	JNBP No.[a]	History	Genome size (Mbp)	G+C mol%	tRNA	CDS[b]	Reference
M. abscessus	JNBP_03165	<GTC 15113<Sputum, Osaka, Japan 1990	5.1	64.1	47	5019	This study
M. abscessus	JNBP_03167	<GTC 15115<Sputum, Osaka, Japan 1990	5.1	64.1	47	5032	This study
M. avium	JNBP_03535	<GTC 15594<Sputum, Kyoto, Japan 1991	5.2	69.1	47	4979	This study
M. branderi	JNBP_03209^T	<GTC 00811< ATCC 51789	5.9	66.5	46	5800	This study
M. chelonae	JNBP_03213	<GTC 12637<Sputum, Mie, Japan 1990	4.9	64.2	46	4746	This study
M. chelonae	JNBP_03238	<GTC 15348<Sputum, Osaka, Japan 1988	5	63.9	47	4960	This study
M. flavescens	JNBP_03250	<GTC 15352<Sputum, Osaka, Japan 1988	5.4	66.8	46	5154	This study
M. fortuitum	JNBP_03252	<GTC 12780<Sputum, Japan 1992	6.6	66.1	54	6545	This study
M. fortuitum	JNBP_03300^T	<GTC 15396<KPM4015	6.3	66.2	54	6137	This study
M. gordonae	JNBP_03328	<GTC 15401<Sputum, Kyoto, Japan 1990	6.6	66.8	49	6348	This study
M. kansasii	JNBP_03473	<GTC 15535<Sputum, Kyoto, Japan 1990	6.5	66	43	6031	This study
M. kansasii	JNBP_03464	<GTC 03844<Sputum, Tokyo, Japan 2007	6.4	66.1	46	5861	This study
M. lentiflavum	JNBP_03532	<GTC 03852<Sputum, Tokyo, Japan 1990	6.5	65.8	48	6117	This study
M. malmoense	JNBP_03536^T	<GTC 15595<ATCC 29571	5.3	67	45	5001	This study
M. mucogenicum	JNBP_03551^T	<GTC 03155<CCUG 47451	6.2	67.2	57	6052	This study
M. sphagni	JNBP_03654^T	<GTC 01619<ATCC 33027	5.5	66.7	46	5241	This study
M. peregrinum	JNBP_07211^T	<GTC 01725<ATCC 14467	7	66.3	48	6879	This study
M. peregrinum	JNBP_03578	<GTC 15642<Sputum, Kyoto, Japan 1990	7.4	66.1	72	7299	This study
M. peregrinum	JNBP_03577	<GTC 15641<Sputum, Kyoto, Japan 1991	7	66.3	48	6902	This study
M. peregrinum	JNBP_03881	<GTC16404<Abscess left leg, Japan 2010	6.5	66.3	53	6340	This study
M. interjectum	JNBP_03801	<GTC 12928<IWGMT strain, USA 1992	7	67.9	47	6846	This study
M. scrofulaceum	JNBP_03599	<GTC 12927<IWGMT90529, USA 1992	5.4	67.6	48	5149	This study
M. senegalense	JNBP_03637^T	<GTC 01622<ATCC 35796	6.1	66.6	47	5832	This study
M. szulgai	JNBP_03662^T	<GTC 15712<JCM6383	6.8	65.6	49	6248	This study
M. triplex	JNBP_03854	<GTC 12949<IWGMT90553, USA 1992	5.5	68.4	48	5284	This study
M. vaccae	JNBP_03729	<GTC 15849<clinical, Osaka, Japan 1988	6.2	68.5	49	6007	This study
M. vaccae	JNBP_03730^T	<GTC15850<KPM4714	6.3	68.5	49	6040	This study
NCBI[c] Database							
M. africanum		GM041182	4.4	65.6	45	4038	Bentley, et al., 2012
M. avium		strain 104	5.5	69	46	5280	Horan, et al., 2006
M. avium subsp. *avium*	JNBP_07180^T	ATCC 25291^T	4.9	69.2	47	4792	Unpublished
M. avium subsp. *paratuberculosis*		strain K-10	4.8	69.3	46	4624	Li, et al., 2005
M. bovis subsp. *bovis*		AF2122/97	4.3	65.6	45	4004	Garnier, et al., 2003

Table 1 Continued

Strain name	JNBP No.[a]	History	Genome size (Mbp)	G+C mol%	tRNA	CDS[b]	Reference
M. bovis BCG		strain Pasteur1173P2	4.4	65.6	47	4023	Brosch, et al., 2007
M. canettii		CIPT140010059	4.5	65.6	43	4049	Supply, et al., 2013
M. gilvum		PYR-GCK	5.6	67.9	46	5429	Badejo, et al., 2013
M. intracellulare	JNBP_03815^T	ATCC 13950^T	5.4	68.1	46	5104	Kim, et al., 2012
M. leprae		Br4923	3.3	57.8	45	3050	Monot, et al., 2009
M. marinum		strain M	6.6	65.7	46	5600	Stinear, et al., 2008
M. smegmatis		strain MC2 155	7	67.4	46	6769	Gallien, et al., 2009
M. tuberculosis	JNBP_03875^T	H37Rv^T	4.4	65.6	45	4066	Cole, et al., 1998
M. ulcerans		Agy99	5.6	65.4	46	5513	Stinear, et al., 2007
M. vanbaalenii		PYR-1^T	6.5	67.8	49	6218	Unpublished

[a]JNBP; Japan National Collection of Bacterial Pathogen, National Bioresource Project (http://www.nbrp.jp/)
[b]CDS; coding sequence
[c]NCBI; National Center for Biotechnology Information

Phylogenetic analysis

The amino acid sequences were edited with DNASIS Pro ver. 3.0 Software. The phylogenetic tree based on the amino acid sequence alignment had higher bootstrap values than the values based on the nucleotide sequence alignment. Therefore, we used amino acid sequence alignment. Phylogenetic analysis of the 16S rRNA gene, C50RP, and C12HKP datasets were aligned with Clustal W and the percent diversities calculated. Neighbor-joining was performed with MEGA (Molecular Evolutionary Genetics Analysis) 6 software (Tamura *et al.*, 2013). Two different models were used for neighbor-joining analysis, namely the Kimura 2-parameters model for nucleotide substitution and the Poisson model for amino acid substitution. The bootstrap values of the phylogenetic trees were calculated on the basis of 1000 replicates.

RESULTS

The DDBJ accession numbers for the gene sequences we determined are LC077734–LC077793, LC079090–LC080649, and LC082309–LC82335.

Characteristics of the draft genomes are summarized in Table 1. The *Mycobacterium* genomes ranged in size from 4.9 to 7.4 Mbp; the % G+C content ranged from 63.9% to 69.1%, and the number of predicted coding sequences ranged from 4746 to 7299.

Protein diversities of *Mycobacterium* species

Mycobacterium tuberculosis H37Rv NC_000962.3 data were used as reference sequences to calculate protein sequence diversity values for each other *Mycobacterium* species. The percentage amino acid difference for each protein from each species was calculated relative to the *M. tuberculosis* H37Rv NC_000962.3 dataset, and each such value was designated the "diversity value" for the respective HKP in the respective species. The diversity values for each protein are shown in Fig. 1. We assessed the range of diversity values for seven groups of proteins. Those from: (i) 50S ribosome proteins (large subunit) ranged from 2.23% to 19.02%; (ii) 30S ribosome proteins (small subunit) ranged from 1.94% to 12.71%; (iii) molecular chaperones ranged from 1.86% to 23.53%, (iv) DNA/RNA modification proteins ranged from 0.12% to 16.36%, (v) DNA/RNA replication proteins ranged from 6.83% to 24.12%, (vi) tRNA-synthesis proteins ranged from 6.96% to 19.58%, and (vii) cell division-associated proteins ranged from 14.53% to 22.41%.

The protein with the largest average interspecies diversity value (24.12%) was DnaQ. The 12 most variable HKPs (with regard to comparisons among species within the genus *Mycobacterium*) were the ribosomal proteins RplY (19.02%) and RplI (18.27%); the modification enzymes GrpE (23.53%), DnaQ (24.12%), DnaN (17.24%), HolA (17.05%), PheT (19.58%), MiaA (18.35%), and MetG (17.89%); and the FTS proteins FtsQ (22.41%), FtsY (21.43%), and FtsX (18.21%). Notably, among the 107 HKPs, intraspecies protein variation was small.

Analysis of a concatenated sequence comprising 50 ribosome proteins (C50RP)

The C50RP sequence of *M. tuberculosis* H37Rv (the type strain) comprised 6802 residues. The

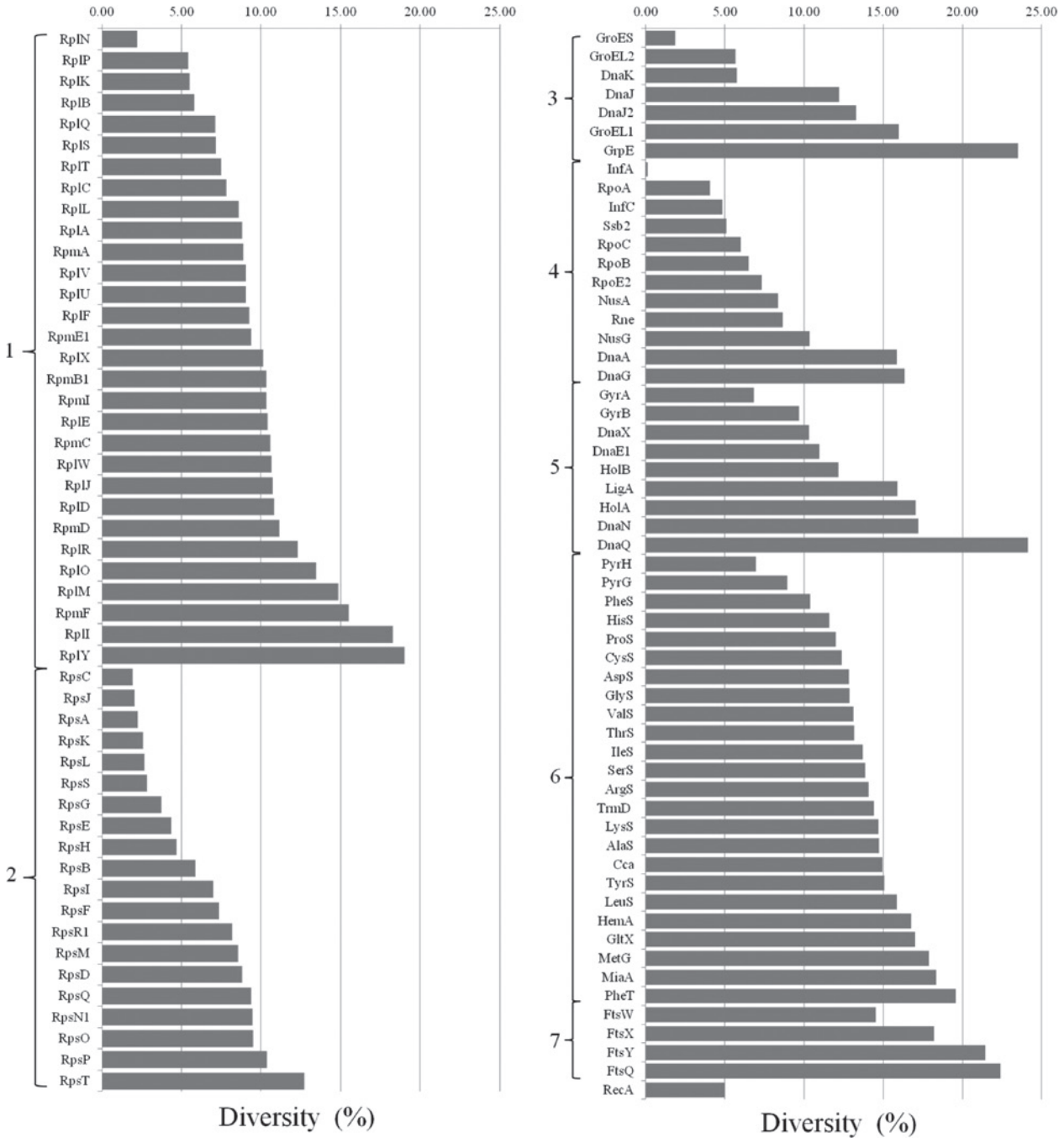


Fig. 1 Graphical depiction of mean diversity value for each of 107 *Mycobacterium* housekeeping proteins (HKPs), with *M. tuberculosis* used as the reference species. HKPs are listed along the Y-axis. The X-axis shows the average % amino acid differences between the *M. tuberculosis* H37Rv reference sequence and other *Mycobacterium* sequences. The HKPs are classified into seven groups: Group 1, 50S ribosome proteins; Group 2, 30S ribosome proteins; Group 3, molecular chaperones; Group 4, DNA/RNA modification proteins; Group 5, DNA/RNA replication proteins; Group 6, tRNA-synthesis proteins; and Group 7, cell division proteins.

C50RP diversity values among the 30 species ranged from 0.06% to 14.65% (mean diversity 9.08%) (Table 2).

Intraspecies diversity of the C50RP values ranged from 0% to 1.26% (data not shown); these values were about double those obtained with the 16S rRNA gene data (Fig. 2a and Table 2). On the basis of the C50RP sequence analysis, phylogenetic trees were constructed by using a neighbor-joining algorithm (Fig. 3b). The bootstrap values at each node in the C50RP tree were higher than those for any node in the 16S rRNA gene tree (Fig. 3a).

The phylogenetic relationships determined by using the C50RP data corresponded well with those determined by using the 16S rRNA gene tree. The correlation coefficient between C50RP and 16S rRNA gene sequence diversity was 0.75 (Fig. 4a).

Phylogenic analysis based on a concatenated protein sequence (C12HKP) generated by using the 12 most variable HKPs

The C12HKP sequence obtained by using the *M. tuberculosis* H37Rv data comprised 3928 residues. With the C12HKP data from 30 species, the diversity values ranged from 0.08% to 37.23% (mean diversity 22.50%) (Table 2). The degree of diversity with the C12HKP data was larger than that with the 16S rRNA gene, C50RP, DnaJ1, RpoB, RecA, or GyrB data (Fig. 2a and 2b and Table 2).

On the basis of the C12HKP sequence analysis, phylogenetic trees were constructed by using a neighbor-joining algorithm (Fig. 3c). The bootstrap values of the C12HKP tree were similar to those of the C50RP tree, but the diversity values calculated with the C12HKP data were larger than those calculated with the C50RP data (Fig. 2a and 2b and Table 2).

The correlation coefficient between the C12HKP and C50RP sequence diversities was 0.94 (Fig. 4b); notably, this value was bigger than the correlation coefficient between the C12HKP and 16S rRNA gene sequence diversities (0.66) (Fig. 4a).

Differentiation between closely related species

Data from individual HKPs, the 16S rRNA gene, C50RP, or C12HKP were used to compare between closely related, slow-growing *Mycobacterium* species; we defined "closely related" as species that exhibited >99% 16S rRNA gene sequence identity (Table 2). Discrimination by using C12HKP was better than that with C50RP or the 16S rRNA gene.

DISCUSSION

Differentiation among closely related species by

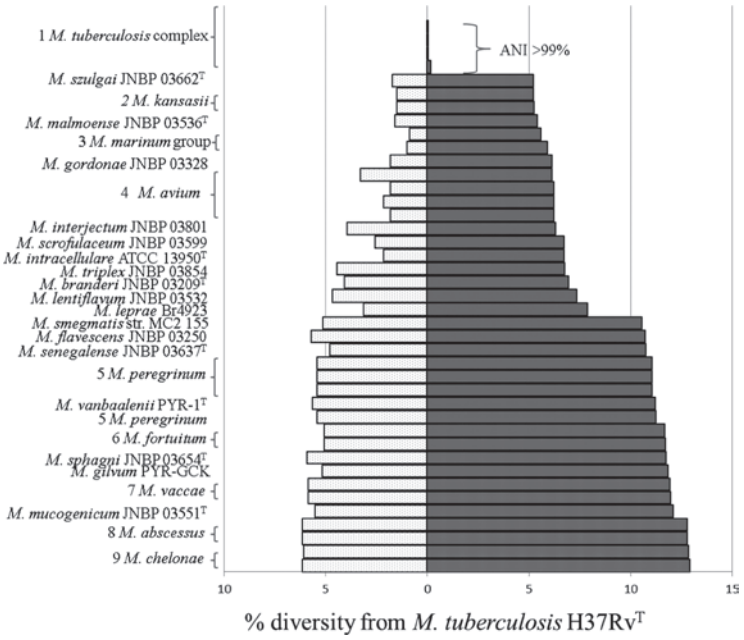
Table 2 ANI values and diversity values based on 16S rRNA gene, C50RP, C12HKP, DnaJ1, RpoB, RecA, or GyrB data

Mycobacterium species	Diversity (%)							ANI[a] value (%)
	16S rRNA gene	C50RP	C12HKP	DnaJ1	RpoB	RecA	GyrB	
Average diversities among 30 species (Range)	4.01 (0.00–7.88)	9.08 (0.06–14.65)	22.50 (0.08–37.23)	14.45 (0.00–26.84)	6.08 (0.00–10.30)	4.66 (0–10.08)	10.49 (0.00–17.53)	/
M. tuberculosis complex[b]	0.00	0.00–0.19	0.00–0.36	0.00–0.51	0.00–0.37	0.00	0.00–0.15	99.86–99.96
M. tuberculosis-M. marinum	0.79	5.48	16.84	11.03	4.29	2.82	8.86	80.1
M. tuberculosis-M. ulcerans	0.93	5.76	17.44	11.53	4.38	3.23	9.01	80.12
M. marunum-M. ulcerans	0.14	0.98	1.14	0.51	0.09	0.40	0.15	98.97
M. avium complex[c]	0.00–0.36	0.09–1.27	1.09–3.95	0.00–2.81	0.00–1.10	0.00	0.00–0.15	98.64–99.23
M. avium-M.intracellulare	0.65–1.00	1.99–3.17	6.21–8.48	0.00–2.81	0.00–1.10	0.00	1.92–2.07	86.01–88.58
M. avium-M. szulgai	1.00	5.33–5.57	15.46–17.08	9.41–10.43	2.75–2.84	2.02	6.06–6.20	80.51–80.69
M. intracellulare-M. szulgai	0.86	5.80	15.79	9.41	2.84	2.02	5.17	80.35
M. kansasii-M. malmoense	0.86	4.95–4.97	13.87	10.58	4.72	2.42	7.98	81.65–81.69
M. kansasii-M. szulgai	1.00	4.78–4.79	12.89	12.37	4.12	0.81	9.16	80.68

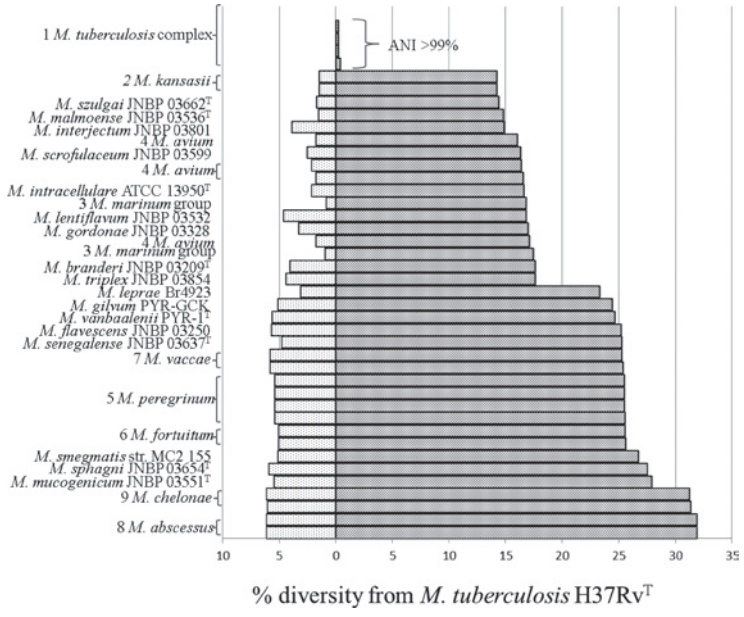
[a]ANI; Average nucleotide identity

[b]Diversity values calculated based on comparisons among species within the *M. tuberculosis* complex (*M. tuberculosis* H37Rv, *M. bovis* AF2122/97, *M. bovis* BCG str. Pasteur1173P2, *M. africanum* GM041182 and *M. canettii* CIPT1400100590).

[c]Diversity values calculated based on comparisons among *M. avium* subspecies (*M. avium* subsp. *avium* ATCC 25291, *M. avium* 104, *M. avium* JNBP_03535, and *M. avium* subsp. *paratuberculosis* K-10).



a: 16S rRNA gene/C50RP



b: 16S rRNA gene/C12HKP

Fig. 2 Comparison between 16S rRNA gene and C50RP or C12HKP sequences with regard to diversity values
X-axis indicates the percent difference (diversity) for each *Mycobacterium* sequence relative to the *M. tuberculosis* H37Rv reference sequence. Y-axis shows each individual *Mycobacterium* strain. **a**, Comparison between 16S rRNA gene and C50RP data. **b**, Comparison between 16S rRNA gene and C12HKP data. C50RP diversity was double the 16S rRNA gene diversity. C12HKP diversity was four times the 16S rRNA gene diversity. Group 1: *M. tuberculosis* complex (*M. tuberculosis* H37Rv^T, *M. africanum* GM041182, *M. bovis* subsp. *bovis* AF2122/97, *M. bovis* BCG str. Pasteur 1173, *M. canettii* CIPT140010059); Group 2: *M. kansasii* (JNBP03473, JNBP03464); Group 3: *M. marinum* group (*M. marinum* M strain *M. ulcerans* Agy99); Group 4: *M. avium* (*M. avium* 104, *M. avium* JNBP03535, *M. avium*. subsp. *paratuberculosis* K10, *M. avium* ATCC25291^T); Group 5: *M. peregrinum* (JNBP 03881, JNBP 03577, JNBP 03578, JNBP03576^T); Group 6: *M. fortuitum* (JNBP 03252, JNBP 03300^T); Group 7: *M. vaccae* (JNBP 03730^T, JNBP 03729); Group 8: *M. abscessus* (JNBP 03165, JNBP 03167); Group 9: *M. chelonae* (JNBP 03238, JNBP 03213)

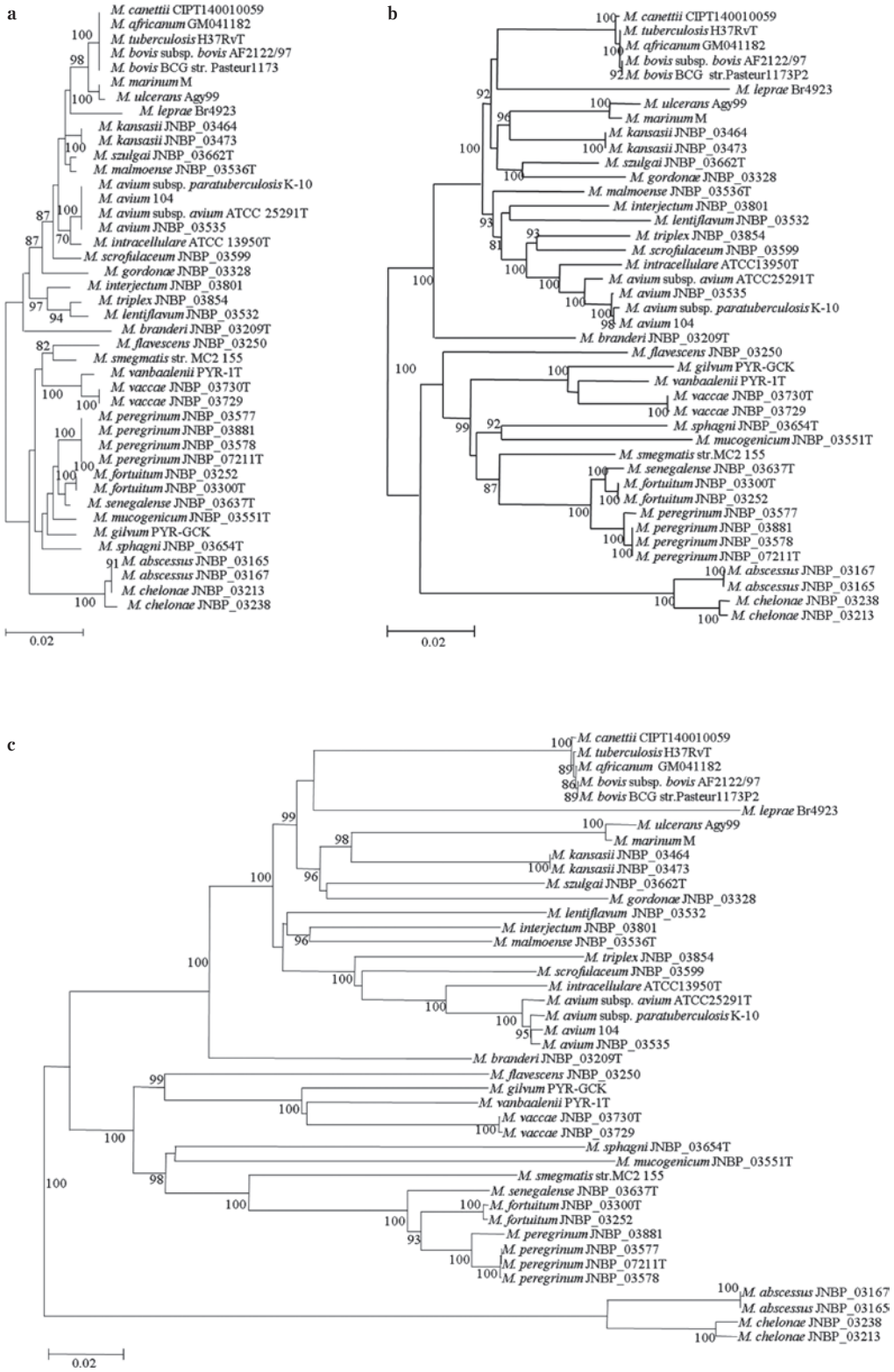


Fig. 3 Neighbor-joining phylogenetic trees based on 16S rRNA gene (a), C50RP (b), or C12HKP (c) sequences and obtained by using 42 *Mycobacterium* strains. Percentage at each node represents a bootstrap value (1000 trials). Scale bar represents 2% sequence divergence. Only bootstrap values >80% are shown.

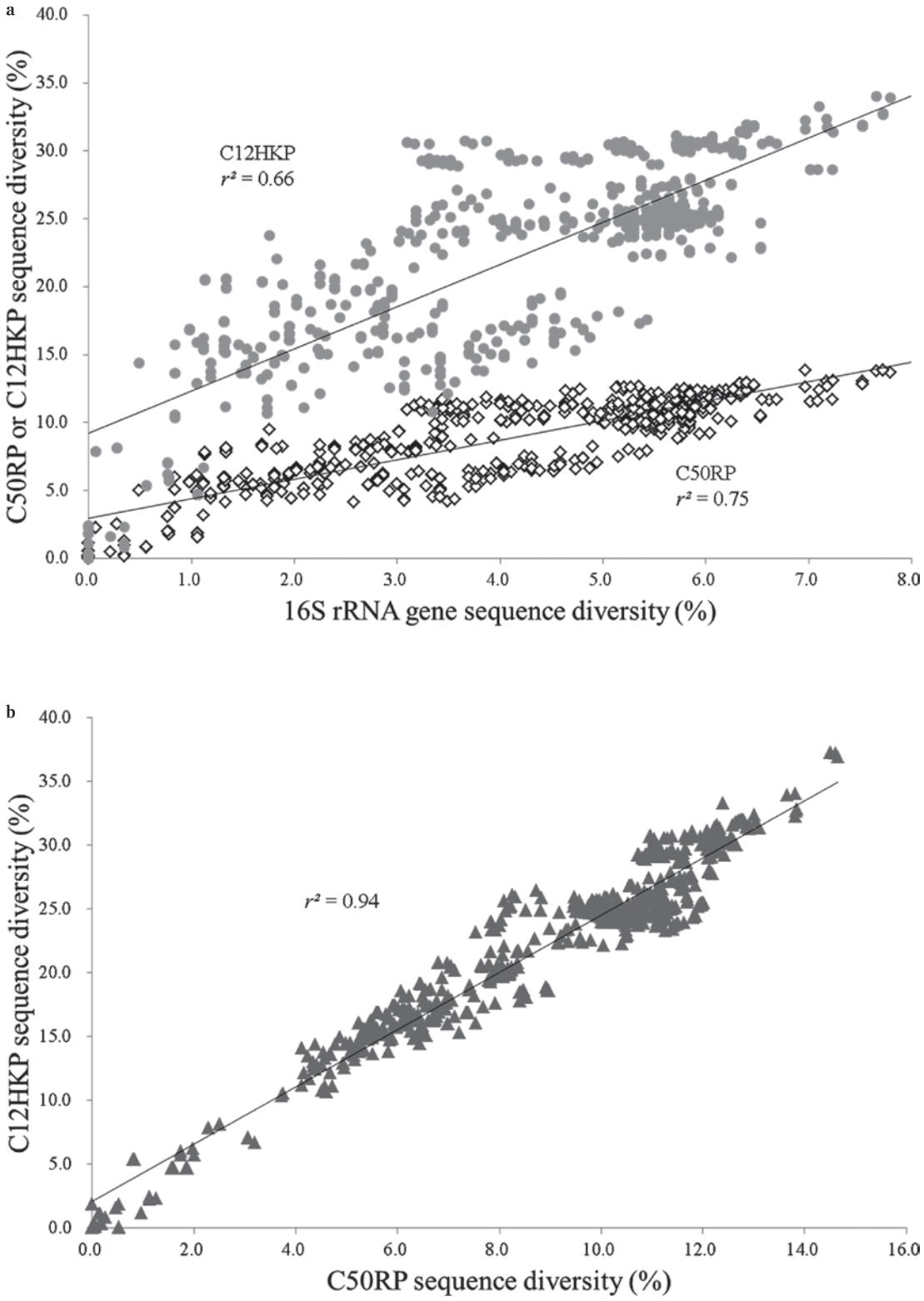


Fig. 4 Scatter plots of C50RP (◇) or C12HKP (●) diversity vs. 16S rRNA gene diversity (a) and C50RP diversity (▲) vs. C12HKP diversity (b). Each data point represents a pair of taxa and is plotted according to the pairwise sequence difference. Each solid line represents a regression line. Each r^2 value represents the correlation coefficient.

using C12HKP

The phylogenetic trees constructed with either the C12HKP data or the C50RP data had very high bootstrap values; therefore, either of these trees was presumably more reliable than the 16S rRNA gene tree. The correlation coefficient between the C12HKP and C50RP sequence diversities was 0.94. Nevertheless, the mean diversity (22.50%) and diversity range (0.08% to 37.23%) calculated with the C12HKP data were higher than the mean diversity (9.08%) and range (0.06% to 14.65%) calculated with the C50RP data. ANI can significantly distinguish closely related species, but it cannot be used to analyze phylogenetic relationships among distantly related species. C12HKP could be used for both types of analysis. Therefore, we used the C12HKP data to analyze and compare three genetically related species complexes, namely the *M. tuberculosis* complex, the *M. avium-intracellulare* complex, and the *M. marinum-M. ulcerans* group.

Within the *M. tuberculosis* complex, five established species (*M. tuberculosis*, *M. bovis*, *M. africanum*, *M. caprae*, and *M. microti*) and a recently described putative progenitor species "*M. canettii*" have almost identical 16S rRNA gene sequences. The ANI for four of the genomes analyzed (*M. tuberculosis*, *M. bovis*, *M. africanum*, and "*M. canettii*") was higher than 99% (Fig. 2a and Table 2). The C12HKP variation within the *M. tuberculosis* complex was less than 0.36% (Fig. 2b and Table 2). Therefore, these species could not be differentiated from one another, even with the C12HKP dataset. ANI values of more than 95% correspond to the traditional >70% DNA–DNA reassociation standard currently used for definition of a species (Konstantinidis & Tiedje, 2005). The close relatedness between *M. tuberculosis* complex bacteria has been established by DNA–DNA hybridization, 16S rRNA gene analysis, and housekeeping gene analysis (Baess, 1979; Feizabadi *et al.*, 1996). *Mycobacterium tuberculosis* complex may be considered a subspecies of *M. tuberculosis* (Tsukamura *et al.*, 1985; Wayne and Kubica, 1986). Our data supported the conclusion that the species in the *M. tuberculosis* complex can be classified as a single species.

The strains that constitute the *M. avium-M. intracellulare* complex have very similar 16S rRNA gene sequences. Variation in the 16S rRNA gene sequence ranged from 0.65% to 1.00% (Table 2), and it was difficult to differentiate between any two species on the basis of the 16S rRNA gene data. However, the ANI value for *M. avium* ATCC 25291 and *M. intracellulare* ATCC 13950 was 88.58%; this value indicated that these two strains were independent species; moreover, the C12HKP variation between these strains was 6.21%. Notably, it was possible to differentiate between these two species on the basis of the C12HKP-sequence comparison.

Another closely related pair of species, *M. marinum* and *M. ulcerans*, had 99.86% identity in 16S rRNA gene sequences. The ANI value for these two species was 98.97%. These data indicated that the two species were identical. Similarly, the C12HKP diversity between the two species was only 1.14%. Stinear *et al.* (2000, 2007) indicated that *M. marinum* and *M. ulcerans* should be classified as a single species. Our data, as in Table 2, supported this opinion.

Effective use of the three data sets

Polymorphism in the three data sets —16S rRNA gene, C12HKP, and C50RP— has different meanings. Current bacterial systematics is based on 16S rRNA gene sequence diversity. Because the 16S rRNA genes are shared among all prokaryotes and have commonly shared conserved sequences, it is possible to use them to compare all microorganisms.

The C12HKP data set was effective for differentiating closely related species within a genus. However, because the types of genes held in common differ at the taxonomic level of family or higher, this data set cannot be applied at these higher taxonomic levels. The diversity of the C50RP data set was mid-way between those of the 16S rRNA gene and C12HKP data sets. Ribosomal proteins in this data set are found in the higher taxa, but each gene sequence polymorphism is difficult to compare because the variations are too big and conserved sequences are not found in each ribosomal protein. Within the family Corynebacteriaceae, however, members of the genus *Mycobacterium* shared common amino acid mutations in many ribosomal proteins (data not shown). This suggests that C50RP might be helpful in re-evaluating the taxonomic position of current species, although genome-based genus criteria have not yet been fully discussed.

ACKNOWLEDGMENTS

This work was supported mainly by the Japanese National Bioresource Project and the NBRP Genome Information Upgrading Program.

REFERENCES

Adékambi, T. & Drancourt, M. 2004. Dissection of phylogenetic relationships among 19 rapidly growing *Mycobacterium* species by 16S rRNA, *hsp65*, *sodA*, *recA* and *rpoB* gene sequencing. Int. J. Syst. Evol. Microbiol. **54**: 2095–2105.

Badejo, A.C., Badejo, A.O., Shin, K.H. & Chai, Y.G. 2013. A gene expression study of the activities of aromatic ring-cleavage dioxygenases in *Mycobacterium gilvum* PYR-GCK to changes in salinity and pH during pyrene degradation. PLoS One **8**: e58066.

Baess, I. 1979. Deoxyribonucleic acid relatedness among species of slowly growing mycobacteria. Acta Pathol. Microbiol. Scand. **87**: 221–226.

Bentley, S.D., Comas, I., Bryant, J.M., Walker, D., Smith, N.H., Harris, S.R., Thurston, S., Gaqneux, S., Wood, J., Antonio, M., Quail, M.A., Gehre, F., Adeqbola, R.A., Parkhill, J. & de Jong, B.C. 2012. The genome of *Mycobacterium africanum* West African 2 reveals a lineage-specific locus and genome erosion common to the *M. tuberculosis* complex. PLoS Negl. Trop. Dis. **6**: e1552.

Blackwood, K.S., He, C., Gunton, J., Turenne, C.Y., Wolfe, J. & Kabani, A.M. 2000. Evaluation of *recA* sequences for identification of *Mycobacterium* species. J. Clin. Microbiol. **38**: 2846–2852.

Boom, R., Sol, C.J.A., Salimans, M.M.M., Jansen, C.L., Wertheim-van Dillen, P.M.E. & van der Noordaa, J. 1990. Rapid and simple method for purification of nucleic acids. J. Clin. Microbiol. **28**: 495–503.

Brosch, R., Gordon, S.V., Garnier, T., Eiglmeier, K., Frigui, W., Valenti, P., Dos Santos, S., Duthoy, S., Lacroix, C., Garcia-Pelayo, C., Inwald, J.K., Golby, P., Garcia, J.N., Hewinson, R.G., Behr, M.A., Quail, M.A., Churcher, C., Barrell, B.G., Parkhill, J. & Cole, S.T. 2007. Genome plasticity of BCG and impact on vaccine efficacy. Proc. Natl. Acad. Sci. U.S.A. **104**: 5596–5601.

Brunello, F., Ligozzi, M., Cristelli, E., Bonora, S., Tortoli, E. & Fontana, R. 2001. Identification of 54 mycobacterial species by PCR-restriction fragment length polymorphism analysis of the *hsp65* gene. J. Clin. Microbiol. **39**: 2799–2806.

Cho, Y.-J., Yi, H., Chun, J., Cho, S.-N., Daley, C.L., Koh, W.-J. & Shin, S.J. 2013. The genome sequence of '*Mycobacterium massiliense*' strain CIP 108297 suggests the independent taxonomic status of the *Mycobacterium abscessus* complex at the subspecies level. PLoS One **8**: e81560.

Cole, S.T., Brosch, R., Parkhill, J., Garnier, T., Churcher, C., Harris, D., Gordon, S.V., Eiglmeier, K., Gas, S., Barry 3rd, C.E., Tekaia, F., Badcock, K., Basham, D., Brown, D., Chillingworth, T., Connor, R., Davies, R., Devlin, K., Feltwell, T., Gentles, S., Hamlin, N., Holroyd, S., Hornsby, T., Jagels, K., Krogh, A., McLean, J., Moule, S., Murphy, L., Oliver, K., Osborne, J., Quail, M.A., Rajandream, M.-A., Rogers, J., Rutter, S., Seeger, K., Skelton, J., Squares, R., Squares, S., Sulston, J.E., Taylor, K., Whitehead, S. & Barrell, B.G. 1998. Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. Nature **393**: 537–544.

Dai, J., Chen, Y., Dean, S., Morris, J.G., Salfinger, M. & Johnson, J.A. 2011. Multiple-genome comparison reveals new loci for *Mycobacterium* species identification. J. Clin. Microbiol. **49**: 144–153.

Devulder, G., Pérouse de Montclos, M. & Flandrois, J.P. 2005. A multigene approach to phylogenetic analysis using the genus *Mycobacterium* as a model. Int. J. Syst. Evol. Microbiol. **55**: 293–302.

Feizabadi, M.M., Robertson, I.D., Cousins, D.V. & Hampson, D.J. 1996. Genomic analysis of *Mycobacterium bovis* and other members of the *Mycobacterium tuberculosis* complex by isoenzyme analysis and pulsed-field gel electrophoresis. J. Clin. Microbiol. **34**: 1136–1142.

Gallien, S., Perrodou, E., Carapito, C., Deshayes, C., Reyrat, J.-M., van Dorsselaer, A., Poch, O., Schaeffer, C. & Lecompte, O. 2009. Ortho-proteogenomics: multiple proteomes investigation through orthology and a new MS-based protocol. Genome Res. **19**: 128–135.

Garnier, T., Eiglmeier, K., Camus, J.-C., Medina, N., Mansoor, H., Pryor, M., Duthoy, S., Grondin, S., Lacroix, C., Monsempe, C., Simon, S., Harris, B., Atkin, R., Doggett, J., Mayes, R., Keating, L., Wheeler, P.R., Parkhill, J., Barrell, B.G., Cole, S.T., Gordon, S.V. & Hewinson, R.G. 2003. The complete genome sequence of *Mycobacterium bovis*. Proc. Natl. Acad. Sci. U.S.A. **103**: 7877–7882.

Gevers, D., Cohan, F.M., Lawrence, J.G., Spratt, B.G., Coenye, T., Feil, E.J., Stackebrandt, E., Van de Peer, Y., Vandamme, P., Thompson, F.L. & Swings,

J. 2005. Opinion: Re-evaluating prokaryotic species. Nat. Rev. Microbiol. **3**: 733–739.

Horan, K.L., Freeman, R., Weigel, K., Semret, M., Pfaller, S., Covert, T.C., van Soolingen, D., Leão, S.C., Behr, M.A. & Cangelosi, G.A. 2006. Isolation of the genome sequence strain *Mycobacterium avium* 104 from multiple patients over a 17-year period. J. Clin. Microbiol. **44**: 783–789.

Kasai, H., Ezaki, T. & Harayama, S. 2000. Differentiation of phylogenetically related slowly growing mycobacteria by their *gyrB* sequences. J. Clin. Microbiol. **38**: 301–308.

Kim, B.J., Lee, S.H., Lyu, M.A., Kim, S.J., Bai, G.H., Kim, S.J., Chae, G.T., Kim, E.C., Cha, C.Y. & Kook, Y.H. 1999. Identification of mycobacterial species by comparative sequence analysis of the RNA polymerase gene (*rpoB*). J. Clin. Microbiol. **37**: 1714–1720.

Kim, B.J., Choi, B.S., Lim, J.S., Choi, I.Y., Lee, J.H., Chun, J., Kook, Y.H. & Kim, B.J. 2012. Complete genome sequence of *Mycobacterium intracellulare* strain ATCC 13950^{T}. J. Bacteriol. **194**: 2750.

Konstantinidis, K.T. & Tiedje, J.M. 2005. Genomic insights that advance the species definition for prokaryotes. Proc. Natl. Acad. Sci. U.S.A. **102**: 2567–2572.

Li, L., Bannantine, J.P., Zhang, Q., Amonsin, A., May, B.J., Alt, D., Banerji, N., Kanjilal, S. & Kapur, V. 2005. The complete genome sequence of *Mycobacterium avium* subspecies *paratuberculosis*. Proc. Natl. Acad. Sci. U.S.A. **102**: 12344–12349.

McNabb, A., Eisler, D., Adie, K., Amos, M., Rodrigues, M., Stephens, G., Black, W.A. & Isaac-Renton, J. 2004. Assessment of partial sequencing of the 65-kilodalton heat shock protein gene (*hsp65*) for routine identification of *Mycobacterium* species isolated from clinical sources. J. Clin. Microbiol. **42**: 3000–3011.

Monot, M., Honoré, N., Garnier, T., Zidane, N., Sherafi, D., Paniz-Mondolfi, A., Matsuoka, M., Taylor, G.M., Donoghue, H.D., Bouwman, A., Mays, S., Watson, C., Lockwood, D., Khamesipour, A., Dowlati, Y., Jianping, S., Rea, T.H., Vera-Cabrera, L., Stefani, M.M., Banu, S., Macdonald, M., Sapkota, B.R., Spencer, J.S., Thomas, J., Harshman, K., Singh, P., Busso, P., Gattiker, A., Rougemont, J., Brennan, P.J. & Cole, S.T. 2009. Comparative genomic and phylogeographic analysis of *Mycobacterium leprae*. Nat. Genet. **41**: 1282–1289.

Niemann, S., Harmsen, D., Rüsch-Gerdes, S. & Richter, E. 2000. Differentiation of clinical *Mycrobacterium tuberculosis* complex isolates by *gyrB* DNA sequence polymorphism analysis. J. Clin. Microbiol. **38**: 3231–3234.

Ringuet, H., Akoua-Koffi, C., Honore, S., Varnerot, A., Vincent, V., Berche, P., Gaillard, J.L. & Pierre-Audigier, C. 1999. *hsp65* sequencing for identification of rapidly growing mycobacteria. J. Clin. Microbiol. **37**: 852–857.

Schleifer, K.H. 2009. Classification of Bacteria and Archaea: past, present andfuture. Syst. Appl. Microbiol. **32**: 533–542.

Stackebrandt, E., Frederiksen, W., Garrity, G.M., Grimont, P.A.D., Kämpfer, P., Maiden, M.C.J., Nesme, X., Rosselló-Mora, R., Swings, J., Trüper, H.G., Vauterin, L., Ward, A.C. & Whitman, W.B. 2002. Report of the ad hoc committee for the re-evaluation of the species definition in bacteriology. Int. J. Syst. Evol. Microbiol. **52**: 1043–1047.

Stinear, T.P., Jenkin, G.A., Johnson, P.D.R. & Davies, J.K. 2000. Comparative genetic analysis of *Mycobacterium ulcerans* and *Mycobacterium marinum* reveals evidence of recent divergence. J. Bacteriol. **182**: 6322–6330.

Stinear, T.P., Seemann, T., Pidot, S., Frigui, W., Reysset, G., Garnier, T., Meurice, G., Simon, D., Bouchier, C., Ma, L., Tichit, M., Porter, J.L., Ryan, J., Johnson, P.D.R., Davies, J.K., Jenkin, G.A., Small, P.L.C., Jones, L.M., Tekaia, F., Laval, F., Daffé, M., Parkhill, J. & Cole, S.T. 2007. Reductive evolution and niche adaptation inferred from the genome of *Mycobacterium ulcerans*, the causative agent of Buruli ulcer. Genome Res. **17**: 192–120.

Stinear, T.P., Seemann, T., Harrison, P.F., Jenkin, G.A., Davies, J.K., Johnson, P.D.R., Abdellah, Z., Arrowsmith, C., Chillingworth, T., Churcher, C., Clarke, K., Cronin, A., Davis, P., Goodhead, I., Holroyd, N., Jagels, K., Lord, A., Moule, S., Mungall, K., Norbertczak, H., Quail, M.A., Rabbinowitsch, E., Walker, D., White, B., Whitehead, S., Small, P.L.C., Brosch, R., Ramakrischnan, L., Fischbach, M.A., Parkhill, J. & Cole, S.T. 2008. Insights from the complete genome sequence of *Mycobacterium marinum* on the evolution of *Mycobacterium tuberculosis*. Genome Res. **18**: 729–741.

Supply, P., Marceau, M., Mangenot, S., Roche, D., Rouanet, C., Khanna, V., Majlessi, L., Criscuolo, A., Tap, J., Pawlik, A., Fiette, L., Orgeur, M., Fabre, M., Parmentier, C., Frigui, W., Simeone, R.,

Boritsch, E.C., Debrie, A.-S., Willery, E., Walker, D., Quail, M.A., Ma, L., Bouchier, C., Salvignol, G., Sayes, F., Cascioferro, A., Seemann, T., Barbe, V., Locht, C., Gutierrez, M.-C., Leclerc, C., Bentley, S., Stinear, T.P., Brisse, S., Médigue, C., Parkhill, L., Cruveiller, S. & Brosch, R. 2013. Genome analysis of smooth tubercle bacilli provides insights into ancestry and pathoadaptation of the etiologic agent of tuberculosis. Nat. Genet. **45**: 172–179.

Tamura, K., Stecher, G., Peterson, D., Filipski, A. & Kumar, S. 2013. MEGA6: molecular evolutionary genetics analysis version 6.0. Mol. Biol. Evol. **30**: 2725–2729.

Thompson, J.D., Gibson, T.J. & Higgins, D.G. 2002. Multiple sequence alignment using ClustalW and ClustalX. Curr. Protoc. Bioinformatics **Chapter 2**: Unit 2.3.

Tsukamura, M., Mizuno, S. & Toyama, M. 1985. Taxonomic studies on the *Mycobacterium tuberculosis* series. Microbiol. Immunol. **29**: 285–299.

Wayne, L.G. & Kubica. G.P. 1986. The mycobacteria, *In* Sneath, P.H.A. & Holt, J.G. (eds.), Bergey's Manual of Systematic Bacteriology vol. 2, p. 1435–1457, Williams and Wilkins, Baltimore.

Yamada-Noda, M., Ohkusu, K., Hata, H., Shah, M.M., Nhung, P.H., Sun, X.S., Hayashi, M. & Ezaki, T. 2007. *Mycobacterium* species identification —A new approach via *dnaJ* gene sequencing. Syst. Appl. Microbiol. **30**: 453–462.

抗酸菌の分類同定指標として有効なコアーハウスキーピング遺伝子

水野卓也[1,2]，名取達矢[1]，金澤　泉[1]，イブラヒム・エルデソウキー[3]，福永　肇[1]，江崎孝行[1]

[1] 岐阜大学大学院再生医科学病原体制御分野，[2] 岐阜県保健環境研究所，
[3] カーフレルシェイク大学獣医学部細菌菌類免疫学講座

Mycobacterium 属 30 菌種 42 株のドラフトゲノム解析を行い，分類と同定に有効な housekeeping protein (HKP) を選択し解析した．NCBI の遺伝子の各分類区分から解析菌種に共通に保有され，多型の大きい 107 タンパクを選択した．そこからさらに多型が大きく，大きな脱落と挿入がある遺伝子を除外し，菌種識別に有用な 12 個の完全長の HKP を連結し C12HKP として解析した．C12HKP は 16S rRNA 遺伝子との相関は 0.66 で，比較に使用した 50 個の ribosomal protein（C50RP）とは 0.94 と高い相関があった．C12HKP の解析菌種間での多型の大きさは 22.50%，C50RP では 9.08%，16S rRNA 遺伝子では 4.01％であり，C12HKP が最も大きな多型を示した．このことから C12HKP は類縁菌種の識別と同定に最も有用性が高いことが証明された．