

BIROn - Birkbeck Institutional Research Online

Phelan, J. and Maitra, A. and McNerney, R. and Nair, M. and Gupta, Antima and Coll, F. and Pain, A. and Bhakta, Sanjib and Clark, T.G. (2015) The draft genome of *Mycobacterium aurum*, a potential model organism for investigating drugs against *Mycobacterium tuberculosis* and *Mycobacterium leprae*. *International Journal of Mycobacteriology* 4 (3), pp. 207-216. ISSN 2212-5531.

Downloaded from: <https://eprints.bbk.ac.uk/id/eprint/12316/>

Usage Guidelines:

Please refer to usage guidelines at <https://eprints.bbk.ac.uk/policies.html>
contact lib-eprints@bbk.ac.uk.

or alternatively



Available at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/IJMYCO



The draft genome of *Mycobacterium aurum*, a potential model organism for investigating drugs against *Mycobacterium tuberculosis* and *Mycobacterium leprae*

Jody Phelan ^{a,*}, Arundhati Maitra ^{b,1}, Ruth McNerney ^{a,1}, Mridul Nair ^c, Antima Gupta ^b,
Francesc Coll ^a, Arnab Pain ^{c,2}, Sanjib Bhakta ^{b,2}, Taane G. Clark ^{a,d,2}

^a Faculty of Infectious and Tropical Diseases, London School of Hygiene & Tropical Medicine, Keppel Street, London WC1E 7HT, United Kingdom

^b Mycobacteria Research Laboratory, Institute of Structural and Molecular Biology, Department of Biological Sciences, Birkbeck College, University of London, Malet Street, London WC1E 7HX, United Kingdom

^c Biological and Environmental Sciences and Engineering Division, King Abdullah University of Science and Technology, Thuwal 23955-6900, Saudi Arabia

^d Faculty of Epidemiology and Population Health, London School of Hygiene & Tropical Medicine, Keppel Street, London WC1E 7HT, United Kingdom

ARTICLE INFO

Article history:
Received 27 April 2015
Accepted 3 May 2015
Available online xxxx

Keywords:
Mycobacteria
M. aurum
M. tuberculosis
M. leprae
Drug screening
Genome

ABSTRACT

Mycobacterium aurum (*M. aurum*) is an environmental mycobacteria that has previously been used in studies of anti-mycobacterial drugs due to its fast growth rate and low pathogenicity. The *M. aurum* genome has been sequenced and assembled into 46 contigs, with a total length of 6.02 Mb containing 5684 annotated protein-coding genes. A phylogenetic analysis using whole genome alignments positioned *M. aurum* close to *Mycobacterium vaccae* and *Mycobacterium vanbaalenii*, within a clade related to fast-growing mycobacteria. Large-scale genomic rearrangements were identified by comparing the *M. aurum* genome to those of *Mycobacterium tuberculosis* and *Mycobacterium leprae*. *M. aurum* orthologous genes implicated in resistance to anti-tuberculosis drugs in *M. tuberculosis* were observed. The sequence identity at the DNA level varied from 68.6% for *pncA* (pyrazinamide drug-related) to 96.2% for *rrs* (streptomycin, capreomycin). We observed two homologous genes encoding the catalase-peroxidase enzyme (*katG*) that is associated with resistance to isoniazid. Similarly, two *embB* homologues were identified in the *M. aurum* genome. In addition to describing for the first time the genome of *M. aurum*, this work provides a resource to aid the use of *M. aurum* in studies to develop improved drugs for the pathogenic mycobacteria *M. tuberculosis* and *M. leprae*.

© 2015 Production and hosting by Elsevier Ltd. on behalf of Asian African Society for Mycobacteriology.

* Corresponding author at: Pathogen Molecular Biology Department, London School of Hygiene & Tropical Medicine, Keppel Street, London WC1E 7HT, United Kingdom. Tel.: +44 (0) 20 7636 8636.

E-mail address: jody.phelan@lshtm.ac.uk (J. Phelan).

¹ Joint authors.

² Joint authors.

Peer review under responsibility of Asian African Society for Mycobacteriology.

<http://dx.doi.org/10.1016/j.ijmyco.2015.05.001>

2212-5531/© 2015 Production and hosting by Elsevier Ltd. on behalf of Asian African Society for Mycobacteriology.

Introduction

Mycobacterium aurum (*M. aurum*) is an acid-fast, gram-positive environmental bacteria typically found in damp conditions [1,2]. It is a fast-growing mycobacterium with an *in vitro* doubling time of 2–3 h that rarely causes infections in humans [2–6]. The *M. aurum* cell wall contains mycolic acids which are analogous to those found in *Mycobacterium tuberculosis* [7], and there are similarities between the antibiotic susceptibility profiles of the two organisms [8,9]. The fast growth rate and low pathogenicity of *M. aurum* have encouraged its use as a surrogate for the highly pathogenic *M. tuberculosis* in studies of anti-microbial activity of anti-tubercular drugs [6,10,11]. Unlike other fast-growing mycobacteria, such as *Mycobacterium smegmatis*, *M. aurum* has the ability to survive within macrophages [12,13] and has been used for high throughput intracellular drug screening, allowing assessment of the ability of compounds to permeate the cell membrane and their stability within the cell [14,15]. The emergence of strains of *M. tuberculosis* resistant to multiple first- and second-line drugs threatens efforts to control tuberculosis (TB) and has renewed interest in the search for new anti-tubercular agents [16]. Rapid-growing models for screening putative anti-tubercular compounds are needed to accelerate drug discovery studies. Similarly, surrogate bacteria are needed to enable studies on drugs that may improve treatment for infection with non-culturable *Mycobacterium leprae*. Knowledge of the bacterial genome could enhance understanding of the molecular basis for drug resistance, and to this end, the genome of *M. aurum* has been sequenced and annotated. The genome was placed in a mycobacterium phylogeny, and comparisons with *M. tuberculosis*, *M. leprae* and *M. smegmatis* genomes were made in relation to susceptibility towards anti-tubercular drugs.

Materials and methods

M. aurum sample and DNA extraction

The *M. aurum* (NCTC 10437) was grown in 7H9 Middlebrook broth (Becton Dickinson, USA) supplemented with 10% albumin–dextrose–catalase (ADC) at 35 °C. DNA was extracted using the Biltoven RFLP protocol [17]. In brief, log phase growth bacteria were treated with lysozyme, sodium dodecyl sulphate, proteinase K, N-cetyl-N,N,N-trimethyl ammonium bromide (CTAB) and chloroform-isoamyl alcohol prior to precipitation with isopropanol. Minimum inhibitory concentration (MIC) values for ethambutol, isoniazid, pyrazinamide and rifampicin drugs for the same *M. aurum* strain are available [18]. Duplications in *M. aurum* of *embB* and *katG* loci were confirmed by Sanger sequencing. For details of primers used, see Supplementary Table 1.

DNA sequencing and genome assembly

The *M. aurum* genomic DNA was sequenced using a 101 bp paired-end library on the Illumina HiSeq2000 platform. The

raw sequence data (size 0.55 Gb, ~5.5 million paired reads, available from ENA ERP009288, minimum base call accuracy greater than 99%) underwent *de novo* assembly using SPAdes software [19]. The SSPACE software [20] was applied to scaffold the assembly, and a combination of IMAGE [21] and GapFiller [22] routines were used to further close or reduce the length of remaining gaps. An alternative approach using Velvet assembly software [23] led to a near identical assembly. Genomic annotation was transferred to the draft genome using the Prokka pipeline [24]. The pipeline searches for genes present in contigs and compares them with protein and DNA databases to annotate them. The *cd-hit* software [25,26] was used to integrate the annotation from 8 mycobacterial species to create a non-redundant blast “primary” database used by the Prokka pipeline. To validate the draft assembly and annotation pipeline, the transferred annotation was compared against the *kas* operon sequence (GenBank: DQ268649.2). All 5 genes from the GenBank entry (*fabD*, *acpM*, *kasA*, *kasB*, *accD6*) were annotated in the correct order and orientation in the assembly.

Comparative genomics

Genomes from 27 species used in whole genome comparisons were downloaded from *ensembl* (bacteria.ensembl.org), and the Uniprot taxon identification numbers are listed in Table 1. Gene multiple alignments were constructed using *clustalw2* [27] for 16S rRNA and MACSE software [28] for *rpoB* sequences. *Raxml* software [29] was used to construct the best scoring maximum likelihood tree, which was rooted using the *Corynebacterium glutamicum* (strain: ATCC 13032) reference sequence, an organism closely related to the mycobacterium genus [30]. Pairwise gene alignments were constructed using MACSE software, which uses the translated amino acid sequence and accounts for frame shifts and premature stop codons. Sequence identities were calculated using the SIAS webserver. Gaps were not used in the calculation of the percent identity. Whole genome alignments were constructed using *mercator* and *mauid* programs [31], and the resulting homology map was inspected and drawn using CIRCOS [32]. Orthologue clusters were created using *OrthoMCL* [33]. To identify any protein coding genes under selective pressure across *M. aurum*, *M. tuberculosis*, *Mycobacterium bovis* – BCG, *M. smegmatis*, and *M. leprae*, the *Ka/Ks* ratio was calculated, where *Ka* is the number of non-synonymous substitutions per non-synonymous site, and *Ks* is the number of synonymous substitutions per synonymous site. Ratio values less than one imply stabilizing or purifying selection, whilst values greater than one imply positive selection. To measure the degree of polymorphism across the genes, the nucleotide diversity (π) was also calculated using the same mycobacterial sample alignments. The *Ka/Ks* and π metrics were calculated using *variscan* (<http://www.ub.edu/softevol/variscan>) and PAML (<http://abacus.gene.ucl.ac.uk/software/paml.html>) software, respectively.

Table 1 – Genomic characteristics of *M. aurum* in the context of related species.

Organism	Chromosome accession number	Uniprot Strain taxon id	Assembled genome (bp)	G + C content	No. genes	Relative in vitro growth rate	ACDP risk class ^a
<i>M. leprae</i>	AL450380.1	272631	3268203	57.80	1605	Unculturable	3
<i>C. glutamicum</i>	HE802067.1	1204414	3309401	53.81	3099	Rapid	1
<i>M. bovis</i>	BX248333	233413	4345492	65.63	3952	Slow	3
<i>M. tuberculosis</i>	AL123456.3	83332	4411532	65.61	4047	Slow	3
<i>M. xenopi</i>	AJFI01000000	1150591	4434836	66.11	4281	Slow	2
<i>M. canetti</i>	HE572590.1	1048245	4482059	65.62	3981	Slow	3
<i>M. thermoresistibile</i>	AGVE01000000	1078020	4870742	69.02	4614	Rapid	1
<i>M. hassiacum</i>	AMRA01000000	1122247	5000164	69.46	4959	Rapid	1
<i>M. abscessus</i>	CU458896.1	36809	5067172	64.15	4942	Rapid	1
<i>M. intracellulare</i>	CP003322.1	487521	5402402	68.10	5144	Slow	2
<i>M. neoaurum</i>	CP006936.1	700508	5438192	66.88	4217	Rapid	1
<i>M. avium</i>	CP000479.1	243243	5475491	68.99	5120	Slow	2
<i>M. gilvum</i>	CP002385.1	278137	5547747	67.86	5349	Rapid	1
<i>M. colombiense</i>	AFVW02000000	1041522	5579559	68.09	5197	Slow	1
<i>M. indicus pranii</i>	CP002275.1	1232724	5589007	68.03	5254	Rapid	1
<i>M. ulcerans</i>	CP000325.1	362242	5631606	65.47	4160	Slow	3
<i>M. yongonense</i>	CP003347.1	1138871	5662088	67.90	5390	Slow	1
<i>M. phlei</i>	AJFJ01000000	1150599	5681954	69.21	5435	Rapid	1
<i>M. aurum</i>	TBA^b	TBA	6019822	67.52	5684	Rapid	1
<i>M. vaccae</i>	ALQA01000000	1194972	6245372	68.60	5949	Rapid	1
<i>M. chubuense</i>	CP003053.1	710421	6342624	68.29	5843	Rapid	1
<i>M. fortuitum</i>	ALQB01000000	1214102	6349738	66.21	6241	Rapid	2
<i>M. vanbaalenii</i>	CP000511.1	350058	6491865	67.79	5979	Rapid	1
<i>M. parascrofulaceum</i>	ADNV01000000	525368	6564171	68.45	6456	Slow	1
<i>M. kansasii</i>	CP006835.1	557599	6577228	66.23	5866	Slow	2
<i>M. marinum</i>	CP000854.1	216594	6636827	65.73	5452	Slow	2
<i>M. smegmatis</i>	CP000480.1	246196	6988209	67.40	6938	Rapid	1
<i>M. rhodesiae</i>	AGIQ01000000	931627	7281599	66.07	7024	Rapid	1

a UK Advisory Committee on Dangerous Pathogens (ACDP) <http://www.hse.gov.uk/pubns/misc208.pdf>.

b ENA number ERP009288.

Results

The *M. aurum* genome

A total of ~5.5 million high quality paired end (101 bp) reads were used to assemble the *M. aurum* genome. The final *M. aurum* assembly consisted of 46 contigs, 43 of which were over 500 bp in length. The median contig length (N50) was 265 Kb (minimum 315 bp, maximum 742,983 bp). The total genome length (~6.02 Mb, G + C content 67.52%) is longer than *M. tuberculosis* (4.4 Mbp) and *Mycobacterium canettii* (4.5 Mbp), but shorter than *Mycobacterium marinum* (6.6 Mbp) and *M. smegmatis* (7.0 Mbp) (Table 1). A total of 5684 coding sequences, 1 tmRNA, 4 rRNA and 51 tRNA features were annotated, and of these 4306 (75%) were assigned a function (Fig. 1). The final contigs and annotation are available for download (pathogenseq.lshhtm.ac.uk/m_aurum).

***M. aurum* and the mycobacteria phylogeny**

A phylogenetic analysis using 27 mycobacterial whole genome sequences revealed that *M. aurum* clustered with *Mycobacterium vaccae* and *Mycobacterium vanbaalenii* within a clade related to fast-growing mycobacteria (Fig. 2). Slow-growing bacteria, including *M. tuberculosis*, clustered within

a distinct clade. However, *Mycobacterium indicus pranii*, a fast-growing mycobacterium and immunotherapy and vaccine candidate for leprosy and tuberculosis [34], clustered within the slow-growing clade. The very high bootstrap support values for the phylogenetic tree (median 100%, range 77–100%) indicates the high precision afforded when using whole genome data. Previously, *hsp65*, *sodA*, *recA*, *rpoB* and 16S rRNA gene sequence data were used to barcode bacteria, with the latter approach being adopted widely [35]. The assembled 16S rRNA sequence for *M. aurum* had the highest identity with *M. vanbaalenii* (99%), *Mycobacterium rhodesiae* (99%), and *Mycobacterium austroafricanum* (99%), in concordance with previous reports [36,37]. The phylogenetic tree constructed using 16S rRNA sequences was broadly similar to that from whole genome data (Supplementary Fig. 1). However, *M. aurum* and *M. vanbaalenii* clustered closer to *Mycobacterium abscessus* rather than *Mycobacterium gilvum*, and the topology was less robust with lower bootstrap support values.

Comparison to the *M. tuberculosis* and *M. leprae* genomes

The *M. aurum* assembled contigs were ordered according to the *M. tuberculosis* H37Rv reference genome (AL123456.3), leading to 10 gapped scaffolds. Most of the *M. tuberculosis*

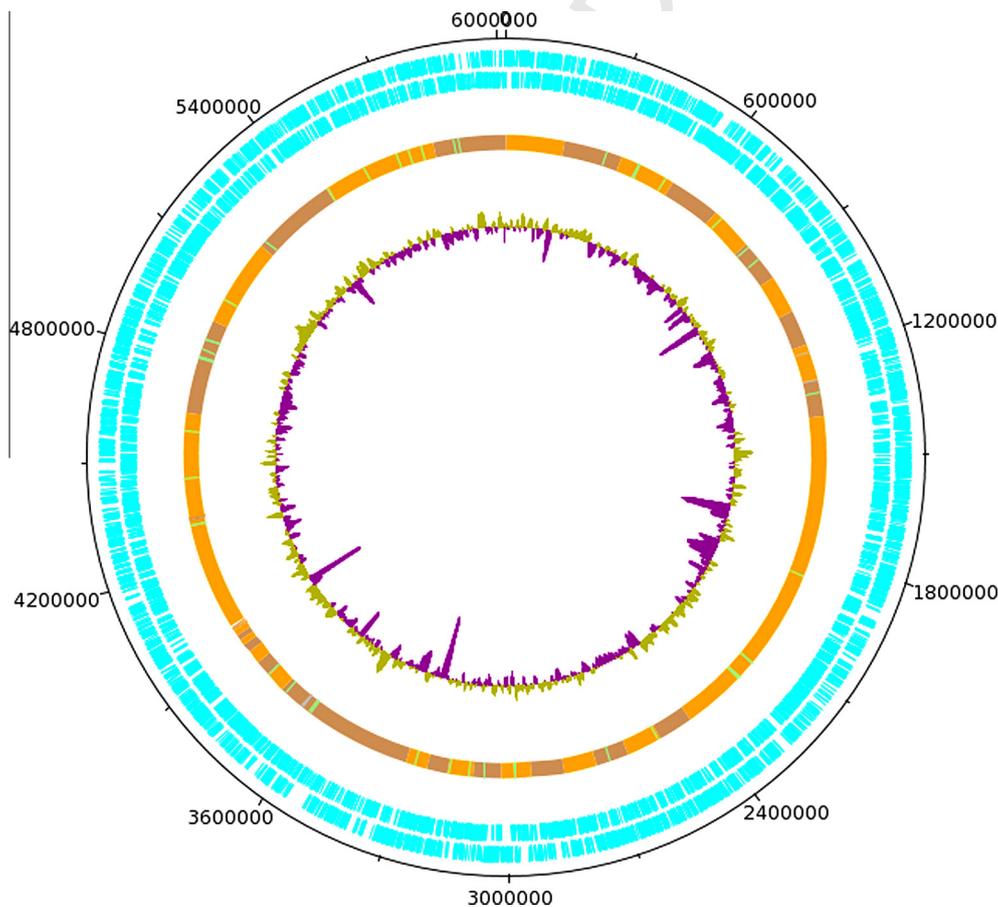


Fig. 1 – An annotated circular view of the *M. aurum* genome (length ~6.02 Mb). Innermost track: G + C% content; middle track: the 46 contigs, alternating between brown and orange with green and grey lines representing tRNA and rRNA, respectively; outer track: the 5684 forward and reverse genes.

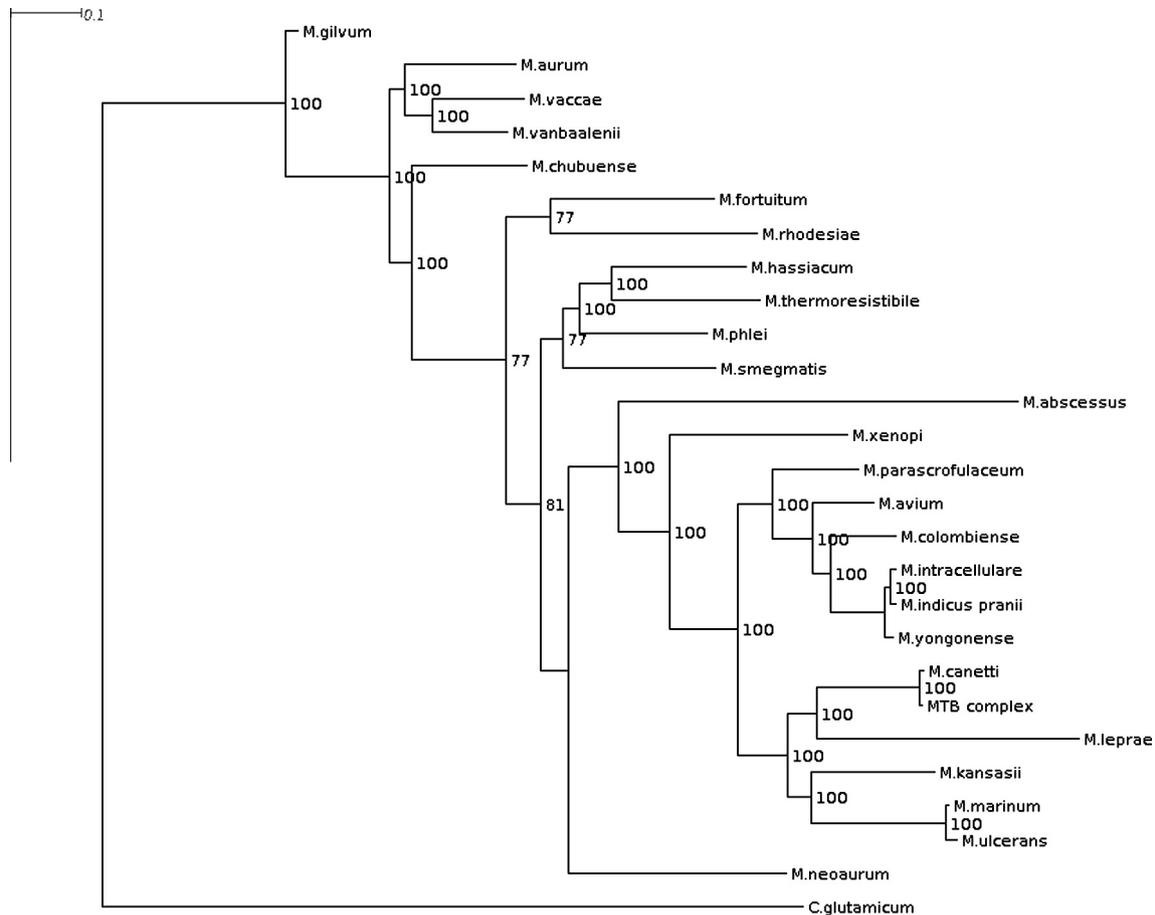


Fig. 2 – *M. aurum* and the mycobacterium phylogeny’ constructed using 27 whole genome reference sequences. ‘Constructed using RaXML and statistic support for lineages was based on 100 bootstrap samples. 27 reference sequences used are described in Table 1.

205 genome (86%) had regions with synteny in *M. aurum*. The map
206 of homology between the 10 *M. aurum* scaffolds and the *M.*
207 *tuberculosis* genome consisted of 67 regions of synteny
208 (Supplementary Table 2 and Fig. 3a). Although there was high
209 similarity between *M. aurum* and *M. tuberculosis*, there was
210 evidence for large-scale rearrangements (Fig. 3a). Twenty-
211 eight genes required for survival within macrophages were
212 observed, but a further two (*lpqY* and *eccA1*) could not be
213 found [38] (Supplementary Table 3). The putative proteome
214 for *M. aurum* suggests it lacks 1002 proteins present in *M.*
215 *tuberculosis*, but has an additional 2090 proteins not seen in
216 *M. tuberculosis* (see Table 2).

217 The map of homology between the 10 *M. aurum* scaffolds
218 and the *M. leprae* genome consisted of 73 segments of synteny
219 (Supplementary Table 2 and Fig. 3b). For *M. aurum* and *M.*
220 *leprae* there were 2047 and 222 unique proteins, respectively,
221 which had no orthologue in the other mycobacteria (see
222 Supplementary Table 2). *M. smegmatis* is often used as a
223 fast-growing model of *M. tuberculosis*. A similar analysis car-
224 ried out between *M. tuberculosis* and *M. smegmatis* revealed
225 979 and 2314 unique proteins for each, respectively, which
226 had no orthologue in the other mycobacteria. When com-
227 pared with the *M. aurum*-*M. tuberculosis* analysis, the number
228 of apparently unique proteins in *M. smegmatis* was higher by
229 224 proteins.

Drug resistance candidate genes

230

231 Pairwise alignments were constructed for the known drug tar-
232 get genes to establish the degree of homology between *M.*
233 *aurum* and *M. tuberculosis* (Table 2). The sequence identity at
234 the DNA level varied from 68.6% for *pncA* (pyrazinamide
235 drug-related) to 96.2% for *rrs* (streptomycin, capreomycin).
236 The percentage of amino acid identity was higher than the
237 sequence identity, being high among all drug resistance candi-
238 date genes analysed (range 90.6–99.2%). Interestingly, two
239 genes at different locations were annotated as *katG* in the
240 *M. aurum* genome, and denoted as *katG1* and *katG2*. The per-
241 cent identity between the two genes and their *M. tuberculosis*
242 homologue at the DNA level are 73.6% and 68.8%
243 (Supplementary Fig. 2) The putative *M. aurum katG1* found
244 in contig 20 (*aurum03417*) demonstrated the highest homol-
245 ogy to the *M. tuberculosis katG* gene (Rv1908c) and *M. smegma-*
246 *tis MSMEG_6384*. The second *M. aurum katG2* (*aurum 02416*)
247 located in contig 2 (*katG2*) was most homologous with *M.*
248 *smegmatis MSMEG_3461*. A third *M. smegmatis* gene,
249 *MSMEG_3729*, showed weak homology to each of the *katG*
250 genes in *M. aurum* and *M. tuberculosis*. Two copies of *embB*, a
251 gene associated with ethambutol in *M. tuberculosis*, were also
252 found in different locations in *M. aurum* (72.3% and 47.7%
253 identity). The semi-identical duplications for each of *katG*

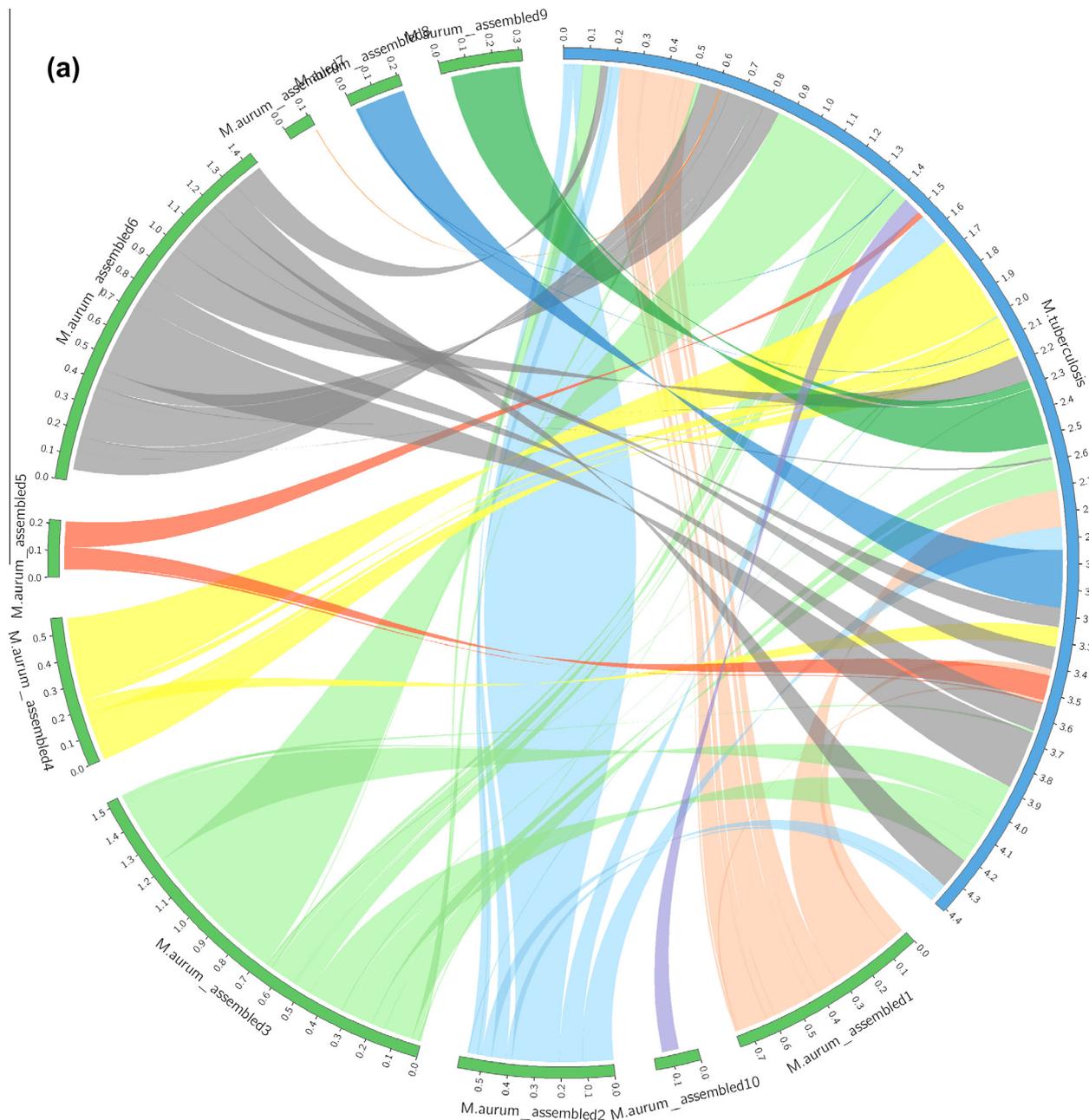


Fig. 3 – Homology between *M. aurum* and *M. tuberculosis* and *M. leprae*. (a) *M. aurum* (green) and *M. tuberculosis* H37Rv (blue). The ten contigs provide 67 segments of synteny with *M. tuberculosis* H37Rv. The segments range from 2,266 bp to 391,674 bp in length. (b) *M. aurum* (green) and *M. leprae* (blue). The ten contigs provide 73 segments of synteny with *M. leprae*. The segments range from 2,495 bp to 193,922 bp in length.

254 and *embB* were confirmed by PCR and Sanger sequencing
255 (Supplementary Table 4).

256 Across a range of therapeutic agents, potential differences
257 in minimum inhibitory concentration (MIC) levels between *M.*
258 *tuberculosis* (H37Rv) and *M. aurum* for isoniazid, ethambutol
259 and ofloxacin (Table 2) are available [8,18], with the biggest
260 difference for isoniazid. The MIC values for isoniazid were
261 greatest in *M. smegmatis* (2 mg/L), followed by *M. aurum* (0.4)
262 and *M. tuberculosis* (0.02–0.2). No known *M. tuberculosis* muta-
263 tions were identified in the *katG*, *inhA* (isoniazid), *ethA*, *ethR*
264 (ethambutol), and *gyrA/B* (ofloxacin) orthologues in *M. aurum*.

265 Homologues of *ahpC* and *embR* genes, associated with isoni-
266 azid and ethambutol drug resistance respectively, were not
267 observed in the *M. aurum* genome.

268 The alignments were compared across *M. aurum*, *M. tuber-*
269 *tuberculosis*, *M. bovis* – BCG, *M. smegmatis*, and *M. leprae* at the loci
270 considered drug targets or those loci considered to have
271 important functional roles (Table 3). All loci had a high per-
272 centage (~90%) of their nucleotides analyzable across the
273 mycobacteria, except *fas* and *gyrA* where there were large
274 insertions in *M. aurum* and *M. leprae*, respectively. Only three
275 loci did not have alignment gaps: *inhA* (isoniazid drug-

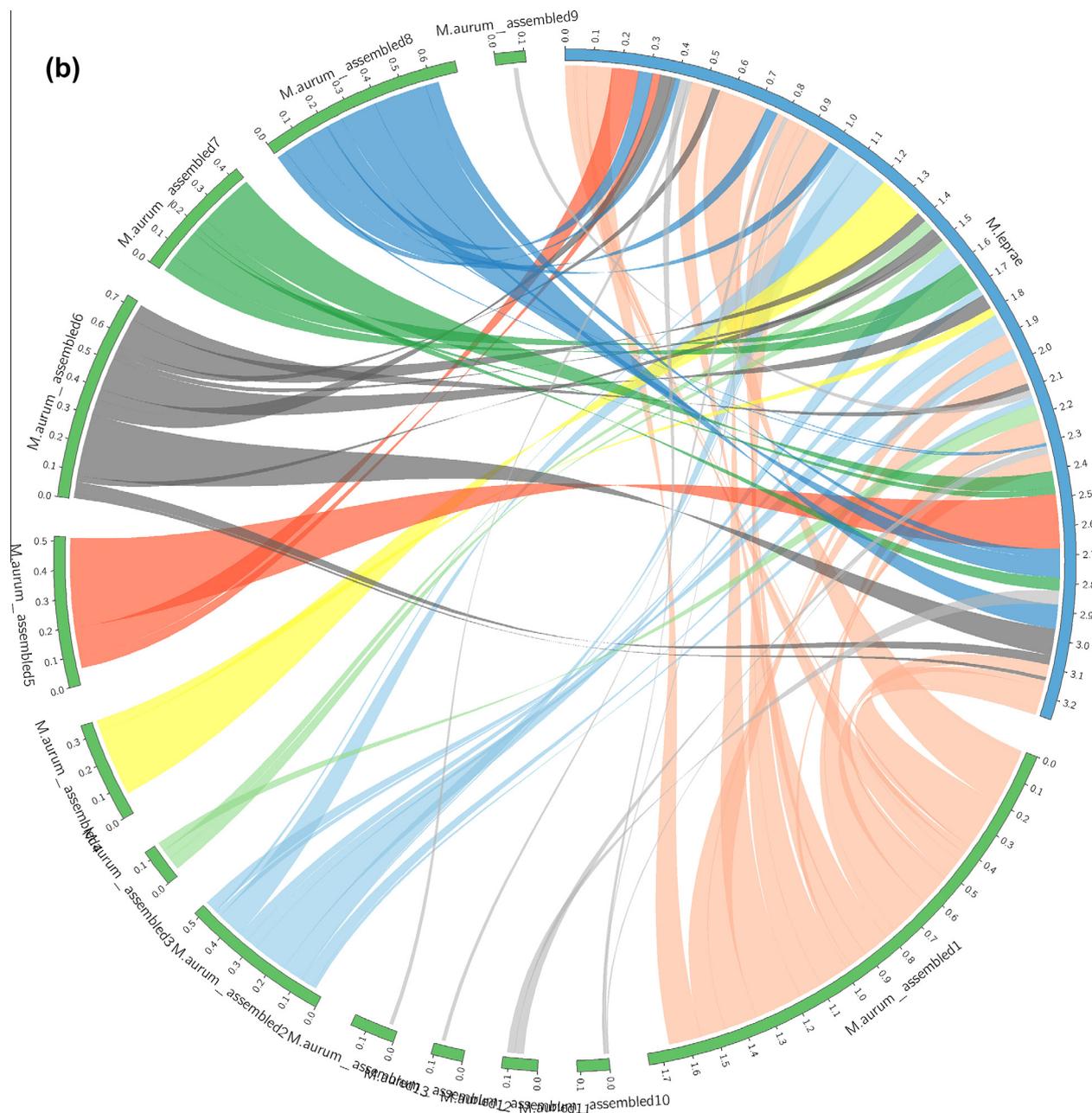


Fig 3. (continued)

276 related); *rpsL* (streptomycin); and *kasA* (thiolactomycin). The
 277 *ddn* (delamanid), *fpo1* (para-aminosalicylic acid), *murC/D/E/F*
 278 family (isoquinolines), and *nat* (cholesterol metabolism) loci
 279 were the most polymorphic (>40% sites segregating, nucleo-
 280 tide diversity $\pi > 0.2$). In contrast, the *rrs* gene associated with
 281 streptomycin drug resistance was the most conserved (2.9%
 282 segregating sites, pairwise diversity $\pi = 0.029$). In general,
 283 there was a modest degree of conservation in most genes
 284 (all with >50% of sequence conserved), which would be
 285 expected given the known synergistic drug effects across
 286 mycobacteria. All candidate genes reported *Ka/Ks* values
 287 much lower than 1, consistent with the selective removal of
 288 alleles that are deleterious (purifying selection). The highest
 289 *Ka/Ks* value was observed for *nat* (Rv3566c), a gene encoding

arylamine acetylase that is associated with resistance to iso-
 niazid [39].

Discussion

The draft genome sequence of *M. aurum* (length ~6.02 Mb,
 G + C content 67.52%) has been assembled. The genome
 assembly consists of 46 contigs and provides the first insight
 into the genetic code of *M. aurum*. Lack of alternative
 sequence data for this bacterium, particularly from technolo-
 gies with longer reads, prevents closure of the gaps at this
 time. Using whole genome alignments, the placement of *M. aurum*
 within the mycobacterial phylogeny, close to *M. vaccae*
 and *M. vanbaalenii*, was confirmed. The analysis of loci

Table 2 – Drug minimum inhibitory concentrations (MICs) and candidate resistance gene identity between *M. aurum* and *M. tuberculosis* at drug resistance loci.

Drug	MIC ^b <i>M. aurum</i> mg/L (μM)	MIC ^b H37Rv mg/L (μM)	<i>M. tb</i> loci	Gene homology with <i>M. aurum</i> (%)	Protein similarity score (%)	<i>M. aurum</i> feature
Isoniazid	0.40 (3.65)	0.02–0.2 (0.15–1.46)	<i>katG</i> ^a	72.06	93.23	2 loci
			<i>inhA</i> ^a	87.40	98.14	Absent
			<i>ahpC</i>	–	–	
Rifampicin	0.10 (0.12)	0.10 (0.12)	<i>kasA</i>	86.33	99.04	2 loci
			<i>rpoB</i> ^a	90.74	97.62	
Ethambutol	0.5 (2.45)	0.47 (2.30)	<i>rpoC</i>	90.20	98.10	Absent
			<i>embB</i>	69.87	91.46	
			<i>embA</i>	68.87	94.28	
			<i>embC</i>	73.83	93.78	
Streptomycin, aminoglycosides, capreomycin	0.2 (0.34) – –	0.1–0.5 (0.17–0.86) – –	<i>embR</i>	–	–	Absent
			<i>rrs</i>	–	–	
			<i>rpsL</i>	96.00	99.20	
Pyrazinamide	>100, (812.26)	>100 (812.26)	<i>tlyA</i>	73.58	92.98	2 loci
			<i>pncA</i>	64.17	90.62	
Ethionamide	5 (30.08)	0.6–2.5 (3.6–15.04)	<i>rpsA</i>	93.56	98.55	Absent
			<i>ethA</i>	65.71	94.30	
Ofloxacin	0.2 (0.55)	1–2 (2.77–5.53)	<i>ethR</i>	69.15	93.11	2 loci
			<i>gyrA</i>	90.34	97.75	
			<i>gyrB</i>	86.39	92.30	

Homology as calculated using protein alignment. Protein similarity is quite high for most proteins analysed.

a Selected alignments can be found in http://pathogenseq.lshtm.ac.uk/m_aurum/.

b MIC value Ref. [18].

Table 3 – A comparison across *M. aurum*, *M. tuberculosis*, *M. bovis* – BCG, *M. smegmatis*, and *M. leprae* alignments at drug targets or other important loci.

Drug resistance or function	Gene name	Alignment length ^a	% Sites analysed ^b	Gaps	% Segregating sites	% Conserved sites	π^c	Ka/Ks ^d
Bedaquiline (TMC207)	<i>atpE</i>	261	94.25	15	28.0	72.0	0.150	0.089
BTZ043, DNB1, VI-9376, 377790, TCA1	<i>dprE1</i>	1410	98.09	27	34.6	65.4	0.210	0.128
Cholesterol metabolism	<i>hsaA</i>	1191	99.50	6	27.6	72.4	0.166	0.150
	<i>hsaB</i>	570	98.95	6	26.2	73.8	0.157	0.182
	<i>hsaC</i>	903	99.67	3	30.1	69.9	0.182	0.104
	<i>hsaD</i>	921	93.81	57	30.9	69.1	0.189	0.129
	<i>nat</i>	861	95.82	36	43.8	56.2	0.268	0.287
Fluoro-quinolones	<i>gyrA</i>	3807	65.33	1320	32.2	67.8	0.171	0.070
	<i>gyrB</i>	2157	93.88	132	36.2	63.8	0.193	0.077
Isoniazid/pyridomycin	<i>inhA</i>	810	100	0	28.1	71.9	0.152	0.101
Isoquinolines	<i>murC</i>	1512	94.64	81	43.1	56.9	0.235	0.170
	<i>murD</i>	1509	96.02	60	46.1	53.9	0.254	0.240
	<i>murE</i>	1653	91.83	135	48.1	51.9	0.267	0.228
	<i>murF</i>	1617	91.65	135	42.7	57.3	0.230	0.167
Isoxyl (thiocarlide)	<i>fas</i>	10701	85.79	1521	37.0	63.0	0.195	0.163
PA-824, delamanid (OPC67683)	<i>ddn</i>	513	88.30	60	44.6	55.4	0.280	0.331
para-aminosalicylic acid	<i>folP1</i>	882	92.18	69	42.9	57.1	0.241	0.144
	<i>folP2</i>	957	88.71	108	35.8	64.2	0.198	0.129
Q203, IP3	<i>qcrB</i>	1695	96.46	60	31.5	68.5	0.174	0.126
Rifampicin	<i>rpoB</i>	3537	98.47	54	23.7	76.3	0.128	0.084
Streptomycin	<i>rpsL</i>	375	100	0	23.2	76.8	0.122	0.028
	<i>rrs</i>	1563	95.84	65	2.9	97.1	0.029	0.029
Thiolactomycin	<i>kasA</i>	1251	100	0	29.1	70.9	0.159	0.100
	<i>kasB</i>	1326	91.63	111	36.4	63.6	0.195	0.117

Selected alignments can be found at pathogenseq.lshtm.ac.uk/m_aurum/.

a The total number of columns in the alignment including gaps.

b A function of the number of sites used in determining the number of segregating and conserved sites.

c π nucleotide diversity.

d The Ka/Ks is the ratio of the number of non-synonymous substitutions per non-synonymous site (Ka) to the number of synonymous substitutions per synonymous site (Ks).

involved in drug resistance demonstrated homology with *M. tuberculosis* and *M. leprae*. This insight corroborates earlier investigations of *inhA* gene mutants of *M. aurum* that showed similarity in drug resistance mechanisms against isoniazid and ethionamide between *M. aurum* and *M. tuberculosis* [6,40]. The draft *M. aurum* genome is larger than that of *M. tuberculosis* with an additional 2090 genes not observed in *M. tuberculosis*; it is also lacking 1002 of the genes found in *M. tuberculosis*. Multiple copies of some homologous genes were observed. Of particular interest are two putative copies of *embB*, a gene involved in the biosynthesis of the mycobacterial cell wall component arabinan and that is associated with resistance to ethambutol in *M. tuberculosis*. Similarly, two annotated catalase-peroxidase (*katG*) genes that may be involved in the activation of the anti-tuberculosis pro-drug isoniazid were identified and confirmed. Multiple *katG* genes have been reported in other mycobacteria, for example in *Mycobacterium fortuitum* [41]. It could be hypothesized that the duplications of *katG* in *M. aurum* and *M. smegmatis* could have an effect on the MIC values. Further laboratory work is underway to elucidate the endogenous function of the observed duplications.

In summary, these genomic analyses support the use of *M. aurum* as a potential model organism for providing insights into *M. tuberculosis* biology, particularly for new drug development, with the possibility of leading to new control measures for tuberculosis disease. Further insight may be gained from the genome sequence of additional strains and related mycobacteria.

Conflicts of interest

The authors declare no conflict of interests.

Acknowledgements

T.G.C. is funded by the Medical Research Council UK (Grant No. MR/K000551/1). J.P. is supported by a BBSRC UK Ph.D. studentship. The project was supported by the KAUST faculty baseline research fund to A.P.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at <http://dx.doi.org/10.1016/j.ijmyco.2015.05.001>.

REFERENCES

- [1] S. Hartmans, J.A.M. de Bont, E. Stackebrandt, The genus mycobacterium – non-medical, in: M. Dworkin, S. Falkow (Eds.), *The Prokaryotes*: Vol. 3: Archaea. Bacteria: Firmicutes, Actinomycetes, Springer Science & Business Media, New York, 2006, pp. 889–918.
- [2] B. Honarvar, H. Movahedan, M. Mahmoodi, F.M. Sheikholeslami, P. Farnia, *Mycobacterium aurum* keratitis: an unusual etiology of a sight-threatening infection, *Braz. J. Infect. Dis.* 16 (2012) 204–208.
- [3] J. Esteban, R. Fernandez-Roblas, A. Roman, A. Molleja, M.S. Jimenez, F. Soriano, Catheter-related bacteremia due to *Mycobacterium aurum* in an immunocompromised host, *Clin. Infect. Dis.* 26 (1998) 496–497.
- [4] K.I. Koranyi, M.A. Ranalli, *Mycobacterium aurum* bacteremia in an immunocompromised child, *Pediatr. Infect. Dis. J.* 22 (2003) 1108–1109.
- [5] A. Martin-Aspas, F. Guerrero-Sanchez, P. Garcia-Martos, E. Gonzalez-Moya, F. Medina-Varo, J.A. Giron Gonzalez, Bilateral pneumonia by *Mycobacterium aurum* in a patient receiving infliximab therapy, *J. Infect.* 57 (2008) 167–169.
- [6] A. Gupta, S. Bhakta, S. Kundu, M. Gupta, B.S. Srivastava, R. Srivastava, Fast-growing, non-infectious and intracellularly surviving drug-resistant *Mycobacterium aurum*: a model for high-throughput antituberculosis drug screening, *J. Antimicrob. Chemother.* 64 (2009) 774–781.
- [7] J.T. Belisle, V.D. Vissa, T. Sievert, K. Takayama, P.J. Brennan, G.S. Besra, Role of the major antigen of *Mycobacterium tuberculosis* in cell wall biogenesis, *Science* 276 (1997) 1420–1422.
- [8] B. Phetsuksiri, A.R. Baulard, A.M. Cooper, D.E. Minnikin, J.D. Douglas, G.S. Besra, et al, Antimycobacterial activities of isoxyl and new derivatives through the inhibition of mycolic acid synthesis, *Antimicrob. Agents Chemother.* 43 (1999) 1042–1051.
- [9] F.G. Winder, Mode of action of the antimycobacterial agents and associated aspects of the molecular biology of mycobacteria, in: C. Ratledge, J. Standford (Eds.), *The Biology of Mycobacteria*, Academic Press Inc., New York, 1982, pp. 353–438.
- [10] R. Srivastava, D. Kumar, B.S. Srivastava, Recombinant *Mycobacterium aurum* expressing *Escherichia coli* beta-galactosidase in high throughput screening of antituberculosis drugs, *Biochem. Biophys. Res. Commun.* 240 (1997) 536–539.
- [11] G.A. Chung, Z. Aktar, S. Jackson, K. Duncan, High-throughput screen for detecting antimycobacterial agents, *Antimicrob. Agents Chemother.* 39 (1995) 2235–2238.
- [12] A. Gupta, A. Kaul, A.G. Tsolaki, U. Kishore, S. Bhakta, *Mycobacterium tuberculosis*: immune evasion, latency and reactivation, *Immunobiology* 217 (2012) 363–374.
- [13] R. Srivastava, D.K. Deb, K.K. Srivastava, C. Lochter, B.S. Srivastava, Green fluorescent protein as a reporter in rapid screening of antituberculosis compounds in vitro and in macrophages, *Biochem. Biophys. Res. Commun.* 253 (1998) 431–436.
- [14] A. Gupta, S. Bhakta, An integrated surrogate model for screening of drugs against *Mycobacterium tuberculosis*, *J. Antimicrob. Chemother.* 67 (2012) 1380–1391.
- [15] D.K. Deb, K.K. Srivastava, R. Srivastava, B.S. Srivastava, Bioluminescent *Mycobacterium aurum* expressing firefly luciferase for rapid and high throughput screening of antimycobacterial drugs in vitro and in infected macrophages, *Biochem. Biophys. Res. Commun.* 279 (2000) 457–461.
- [16] A. Zumla, I. Abubakar, M. Raviglione, M. Hoelscher, L. Ditiu, T.D. McHugh, et al, Drug-resistant tuberculosis-current dilemmas, unanswered questions, challenges, and priority needs, *J. Infect. Dis.* 205 (Suppl. 2) (2012) S228–240.
- [17] J.D. van Embden, M.D. Cave, J.T. Crawford, J.W. Dale, K.D. Eisenach, B. Gicquel, et al, Strain identification of *Mycobacterium tuberculosis* by DNA fingerprinting: recommendations for a standardized methodology, *J. Clin. Microbiol.* 31 (1993) 406–409.
- [18] J.D. Guzman, D. Evangelopoulos, A. Gupta, K. Birchall, S. Mwaigwisa, B. Saxty, et al, Antitubercular specific activity of ibuprofen and the other 2-arylpropanoic acids using the HT-SPOTi whole-cell phenotypic assay, *BMJ Open* 3 (2013).

- 420 [19] A. Bankevich, S. Nurk, D. Antipov, A.A. Gurevich, M. Dvorkin, 462
 421 A.S. Kulikov, et al, SPAdes: a new genome assembly 463
 422 algorithm and its applications to single-cell sequencing, J. 464
 423 Comput. Biol. 19 (2012) 455–477. 465
- 424 [20] M. Boetzer, C.V. Henkel, H.J. Jansen, D. Butler, W. Pirovano, 466
 425 Scaffolding pre-assembled contigs using SSPACE, 467
 426 Bioinformatics 27 (2011) 578–579. 468
- 427 [21] I.J. Tsai, T.D. Otto, M. Berriman, Improving draft assemblies 469
 428 by iterative mapping and assembly of short reads to 470
 429 eliminate gaps, Genome Biol. 11 (2010) R41. 471
- 430 [22] M. Boetzer, W. Pirovano, Toward almost closed genomes with 472
 431 GapFiller, Genome Biol. 13 (2012) R56. 473
- 432 [23] D.R. Zerbino, E. Birney, Velvet: algorithms for de novo short 474
 433 read assembly using de Bruijn graphs, Genome Res. 18 (2008) 475
 434 821–829. 476
- 435 [24] T. Seemann, Prokka: rapid prokaryotic genome annotation, 477
 436 Bioinformatics 30 (2014) 2068–2069. 478
- 437 [25] L. Fu, B. Niu, Z. Zhu, S. Wu, W. Li, CD-HIT: accelerated for 479
 438 clustering the next-generation sequencing data, 480
 439 Bioinformatics 28 (2012) 3150–3152. 481
- 440 [26] W. Li, A. Godzik, Cd-hit: a fast program for clustering and 482
 441 comparing large sets of protein or nucleotide sequences, 483
 442 Bioinformatics 22 (2006) 1658–1659. 484
- 443 [27] M.A. Larkin, G. Blackshields, N.P. Brown, R. Chenna, P.A. 485
 444 McGettigan, H. McWilliam, et al, Clustal W and Clustal X 486
 445 version 2.0, Bioinformatics 23 (2007) 2947–2948. 487
- 446 [28] V. Ranwez, S. Harispe, F. Delsuc, E.J. Douzery, MACSE: 488
 447 Multiple Alignment of Coding Sequences accounting for 489
 448 frameshifts and stop codons, PLoS One 6 (2011) e22594. 490
- 449 [29] A. Stamatakis, RAxML version 8: a tool for phylogenetic 491
 450 analysis and post-analysis of large phylogenies, 492
 451 Bioinformatics 30 (2014) 1312–1313. 493
- 452 [30] M.V. Omelchenko, Y.I. Wolf, E.K. Gaidamakova, V.Y. 494
 453 Matrosova, A. Vasilenko, M. Zhai, et al, Comparative 495
 454 genomics of *Thermus thermophilus* and *Deinococcus* 496
 455 *radiodurans*: divergent routes of adaptation to thermophily 497
 456 and radiation resistance, BMC Evol. Biol. 5 (2005) 57. 498
- 457 [31] C.N. Dewey, Aligning multiple whole genomes with Mercator 499
 458 and MAVID, Methods Mol. Biol. 395 (2007) 221–236. 500
- 459 [32] S.J. McKay, I.A. Vergara, J.E. Stajich, Using the Generic 500
 460 Synteny Browser (GBrowse_syn), Curr. Protoc. Bioinformatics 500
 461 (2010) (Chapter 9: Unit 9 12).
- [33] S. Fischer, B.P. Brunk, F. Chen, X. Gao, O.S. Harb, J.B. Iodice, 462
 et al, Using OrthoMCL to assign proteins to OrthoMCL-DB 463
 groups or to cluster proteomes into new ortholog groups, 464
 Curr. Protoc. Bioinformatics (2011) (Chapter 6: Unit 6 12 11– 465
 19). 466
- [34] S.A. Zaheer, R. Mukherjee, B. Ramkumar, R.S. Misra, A.K. 467
 Sharma, H.K. Kar, et al, Combined multidrug and 468
Mycobacterium w vaccine therapy in patients with 469
 multibacillary leprosy, J. Infect. Dis. 167 (1993) 401–410. 470
- [35] T. Adekambi, M. Drancourt, Dissection of phylogenetic 471
 relationships among 19 rapidly growing *Mycobacterium* 472
 species by 16S rRNA, hsp65, sodA, recA and rpoB gene 473
 sequencing, Int. J. Syst. Evol. Microbiol. 54 (2004) 2095–2105. 474
- [36] N.C. Gey van Pittius, S.L. Sampson, H. Lee, Y. Kim, P.D. van 475
 Helden, R.M. Warren, Evolution and expansion of the 476
Mycobacterium tuberculosis PE and PPE multigene families and 477
 their association with the duplication of the ESAT-6 (esx) 478
 gene cluster regions, BMC Evol. Biol. 6 (2006) 95. 479
- [37] M. Goodfellow, J.G. Magee, Taxonomy of mycobacteria, in: 480
 P.R.J. Gangadharam, P.A. Jenkins (Eds.), *Mycobacteria: Basic* 481
Aspects, Chapman and Hall, New York, 1998, pp. 1–53. 482
- [38] J. Rengarajan, B.R. Bloom, E.J. Rubin, Genome-wide 483
 requirements for *Mycobacterium tuberculosis* adaptation and 484
 survival in macrophages, Proc. Natl. Acad. Sci. USA 102 (2005) 485
 8327–8332. 486
- [39] C. Vilcheze, W.R. Jacobs Jr., Resistance to isoniazid and 487
 ethionamide in mycobacterium tuberculosis: genes, 488
 mutations and causalities, in: G.F. Hatfull, W.R. Jacobs Jr. 489
 (Eds.), *Molecular Genetics of Mycobacteria*, 2nd ed., ASM 490
 Press, USA, 2014. 491
- [40] F. Bardou, A. Quemard, M.A. Dupont, C. Horn, G. Marchal, M. 492
 Daffe, Effects of isoniazid on ultrastructure of *Mycobacterium* 493
aurum and *Mycobacterium tuberculosis* and on production of 494
 secreted proteins, Antimicrob. Agents Chemother. 40 (1996) 495
 2459–2467. 496
- [41] M.C. Menendez, J.A. Ainsa, C. Martin, M.J. Garcia, KatGI and 497
 katGII encode two different catalases-peroxidases in 498
Mycobacterium fortuitum, J. Bacteriol. 179 (1997) 6880–6886. 499
 500