Research Journal of Microbiology

ISSN 1816-4935

Academic Journals Inc.

www.academicjournals.com

Research Journal of Microbiology 4 (3): 112-121, 2009
ISSN 1816-4935

Genome Wide Single Nucleotide Polymorphism Analysis of *Mycobacterium* Species and Subspecies

S.K. Srivastava, M. Agrawal and M. Grover
Amity Institute of Biotechnology, Amity University, Sec-125, Expressway,
Noida 210304, Uttar Pradesh, India

Abstract: In this study we report the reannotation of the genome of seven *Mycobacterial* species and subspecies. We have used bioinformatics tools for annotation and reevaluated each of the Protein-Coding Sequences (CDS) previously annotated and presented the combined results of recent database searches. We have also used comparative genomics tools to focus on comparative analysis as an effective strategy. Pair wise comparison between the various *Mycobacterium* strains was performed so as to predict the relationships between them. Among the wide variety of mycobacterium strains present, we selected seven and showed how their genome is interrelated by studying synteny with the genomes of various strains studied. The genome wide SNP analysis in the seven genomes of *Mycobacterium* sp. was also done in this study and the base by base changes in the genome of these seven subspecies were identified. The gene based SNPs were further classified into the marker SNPs (SNPs which are unique amongst all the seven studied species). Out of a total of 2073 SNPs, 966 were identified as marker SNPS. This study may be used for further analysis of host pathogen interactions at the pathway and product level. The present investigation will also be useful for study of evolutionary relationship.

Key words: Gene annotation, *Mycobacterium tuberculosis*, SNPs, synteny

INTRODUCTION

One third of the world's population-nearly two billion people are infected with the bacterium that causes tuberculosis. Two million people die from it each year. Effective drugs to treat and cure the disease have been available for more than 50 years, yet every 15 sec, someone in the world dies from TB. In addition to this, a person is reported to be newly infected with TB every second of every day. Left untreated, a person with active TB will infect an average of 10 to 15 other people every year. (http://www.worldlungfoundation.org/tuberculosis.php). The etiologic agent of tuberculosis in humans is bacterium *Mycobacterium tuberculosis*.

The sequenced genomes of pathogenic bacteria provide useful information for understanding genomic traits. The complete genome sequence of *M. tuberculosis* strain H37Rv was obtained by Cole *et al*. (1998). The genome comprises 4,411,529 base pairs, containing around 4,000 genes. The sequence of the genome and its comparison to sequences of other microorganisms reported in several databases has allowed the assignation of precise functions to 40% of the predicted proteins and the identification of 44% of orthologues (genes with very similar functions in a different species), leaving 16% as unique unknown proteins.

Some unique characteristics of the biology of the tubercle bacillus, such as its characteristic slow growth, the nature of its complex cell wall, certain genes related to its virulence and persistence and the apparent stability of its genome have been elucidated with the help of genome sequence of

Corresponding Author: Subodh K. Srivastava, Amity Institute of Biotecnology, Amity University, Sec-125, Expressway, Nodia 210304, Uttar Pradesh, India

M. tuberculosis H37Rv. The availability of the complete genome sequences has also led to the development of downstream sciences that take advantage of this sequence information. It is critical, therefore, that genome annotations are frequently updated if the information they contain is to remain accurate, relevant and useful. Several other *Mycobacterial* genomes have thus been reannotated recently (Dandekar *et al*., 2000; Serres *et al*., 2001; Gaasterland and Opera, 2001; Bocs *et al*., 2002).

In this study we report the reannotation of seven *Mycobacterial* species and subspecies genomes. We have used bioinformatic tools for annotation and reevaluated each of the Protein-Coding Sequences (CDS) previously annotated and presented the combined results of recent database searches.

In the current investigation we have also used the bioinformatic tools for comparative genomics. Pair wise comparison between the various mycobacterium strains was performed so as to predict the relationships between them. Among the wide variety of *Mycobacterium* strains present, we chose seven and showed how their genome is interrelated by studying synteny with the genomes of various strains taken. The genome wide SNP analysis in the seven genomes of *Mycobacterium* sp. was also done in this study and the base by base changes in the genome of these seven subspecies were identified. The gene based SNPs were further classified into the marker SNPs (SNPs which are unique amongst all the seven studied species). Out of a total of 2073 SNPs, 966 were identified as marker SNPS. This study may be used for further analysis of host pathogen interactions at the pathway and product level. The present investigation will also be useful for study of evolutionary relationships.

MATERIALS AND METHODS

The present investigation was carried out in June 2008.

Sequences

The following sets of *Mycobacterium* subspecies were used in this study:

- *Mycobacterium tuberculosis* H37Rv, Accession No. AL123456 (Cole *et al*., 1998, 2001)
- *Mycobacterium leprae* TN, Accession No. AL450380 (Geluk *et al*., 2005)
- *Mycobacterium bovis* BCG Pasteur, Accession No. AM408590 (Garnier *et al*., 2003)
- *Mycobacterium bovis* subsp. *bovis* AF2122/97, Accession No. BX248333 (Garnier *et al*., 2003)
- *Mycobacterium* CDC 1551, Accession No. AE000516 (Fleischmann *et al*., 2002)
- *Mycobacterium avium* subsp. *Paratuberculosis*, Accession No. AE016958 (Li *et al*., 2005)
- *Mycobacterium smegmatis str.* MC2 155, Accession No. CP000480 (TIGR)

The whole genome sequence of each of the above sub-species of *Mycobacterium* made available by National Center for Biotechnology Information (NCBI) was obtained from http://www.ncbi.nlm.nih.gov/ in a FASTA format, a text-based format for representing either nucleic acid sequences or peptide sequences, in which base pairs or amino acids were represented using single-letter codes. The format also allowed for sequence names and comments to precede the sequences.

Sequence Analysis and Re-Annotation

Each of the downloaded sequence was split up into parts by using JR Split File (http://www.spadixbd.com/freetools/jsplit.htm). The File Split function allowed us to split a large file into smaller files and create a standard bat file that could be used to reconstruct a copy of the original file. Then each of the split sequence was re-annotated. Re-annotation was supported by FGENESB software (http://www.softberry.com/berry.phtml ?topic=fgenesb& group=programs&subgroup=gfindb), a tool for Bacterial Operon and Gene Prediction. Each of the *mycobacterium* sub-species coding genes were predicted and annotated. Further analysis of putative

proteins was performed, based on the results of similarity searches using BLASTP with the default parameters and Fasta sequence comparisons using non-redundant (nr) data from the GenBank database. Once we determined the total number of hits to nr, the best top BLAST hit for each gene was obtained.

Identification of SNPs

Computational strategy was used to identify SNPs present in seven *mycobacterium* genomes. The sequence data sets for each sub-species were aligned using SNPs Finder software (http://snpsfinder.lanl.gov/) to detect polymorphic sites (taking *Mycobacterium* tuberculosis H37Rv as a reference sequence) which were differentiated as SNPs and cSNPs. SNPs were differentiated from cSNPs by marking the coding and non coding sequence in the reference sequence (*Mycobacterium tuberculosis* 37 Rv).

RESULTS AND DISCUSSION

The comparative genomic studies focuses primarily on biotechnological areas and potential biomedical benefits include insights into the specialized, shared systems used by disease-causing organisms (pathogens) to disable or destroy human cells. Comparing these genomic data with those of other microbes may help understand a diverse range of pathogens that have remarkably similar methods for infiltrating organisms with protein-coding genes capable of sneaking past human defense systems. Hence, comparative genomics and related technologies are helping to unravel the molecular basis of the pathogenesis, host range, evolution and phenotypic differences of the slow-growing mycobacteria.

Selection of Species

In this post-genomic era, an understanding of strain-to-strain variation is expected to fast-forward the rational design and development of effective diagnostic reagents, vaccines and drugs in the case of the difficult-to-crack TB pathogen. Therefore, not surprisingly the present era of *Mycobacterium* genomics is rapidly evolving into a period of comparative genomics.

Several genomes including two strains of *M. tuberculosis* (i.e., H37Rv and CDC1551), *M. bovis*, *M. leprae*, *M. smegmatis*, *M. ulcerans*, *Mycobacterium* sp. MCS, *Mycobacterium* sp. KMS, *M. vanbaalenii* PYR-1, *M. avium* and *M. avium* subsp. *paratuberculosis* have been sequenced and annotated completely (www.ncbi.nlm.nih.gov/genomes/lproks.cgi). Fifteen additional *Mycobacterial* genome sequencing projects have either been completed and awaiting annotation or are underway.

On the basis of this, we chose seven sub-species of *Mycobacterium* for our whole genome comparative study. A combined approach of genomics and bio-informatics was adopted for this study, which was undertaken with the widely used reference strain *M. tuberculosis* H37Rv. This comparative genome analysis revealed the similarities and differences across various genomes.

***In silico* Gene Prediction and Re-Annotation**

The initial genome annotation in *M. tuberculosis* H37Rv strain soon became out of date. For this reason, a re-annotation of that genome sequence was published in 2002. This re-annotation incorporated eighty-two new Protein-Coding Sequences (CDS) and 22 of these have a predicted function. According to the above study, it is possible to assign a function to 2058 proteins (52% of the 3995 proteins predicted) and only 376 putative proteins share no homology with known proteins and thus could be unique to *M. tuberculosis*.

Here, in this paper complete re-annotation of the genome sequence of *Mycobacterium tuberculosis* sub-species is presented. The tool (FGENESB) was used on seven different organisms' viz., *Mycobacterium tuberculosis* H37Rv, *Mycobacterium leprae* TN, *Mycobacterium bovis* BCG Pasteur, *Mycobacterium bovis* subsp. *bovis* AF2122/97, *Mycobacterium* CDC 1551, *Mycobacterium avium* subsp. *paratuberculosis* and *Mycobacterium smegmatis* MC2 155.

The input submitted to the tool was a set of protein sequences in FASTA format, which were made available from NCBI. The tool was run at default parameters and against the closest organism *Mycobacterium tuberculosis* H37Rv, taken as a reference. The tool identifies genes based on similarities with the genes present in the database. The results obtained deciphered the number of genes predicted (CDS), along with their starting and ending positions on the genome, position on the chain and the sequence of each gene annotated in each of the species taken. The scores were also obtained due to matching of predicted genes with the database.

Revising the Number of Genes in the Genome

When the results were compared with the first annotation of the genome, it was found that there is a rise in the number of annotated putative genes in all the organisms taken (Table 1). All of the new genes are believed to encode polypeptides. The numbering of the new CDS has not interfered with the labeling of the existing genes.

The gene sequences obtained after gene annotation were re-analysed based on the results of similarity searches using BLASTP with the default parameters. The tool identifies essential genes based on similarities with the genes present in the database. The tool matchesall the genes in the input data set with the database using BLAST and displayed hits at the specified E-value and a bit score, using top hit.

Genome Wide SNP Analysis

The genome wide SNP analysis in the seven genomes of *Mycobacterium* sp. was done in this study and the base by base changes in the genome of these seven subspecies were identified which is shown in Table 2. The gene based SNPs were further classified into the marker SNPs (SNPs which are unique amongst all the seven studied species). The genome wide SNP analysis was extended to find out the variations at the gene level. The annotated genes of *Mycobacterium tuberculosis* Rv were taken as a reference sequence and the variations falling in the genes of Rv strain were coded as gene SNPs (Table 2).

On analysis of SNP discovery, we observed that a total of 2073 SNPs amongst seven *mycobacterium* subspecies are spread throughout the genome out of which 966 represent the unique/ marker SNPs, which means that marker SNPs (46.6% of the total SNPs) have a greater importance in the genomic studies.

The new genomic annotation of *M. tuberculosis* H37Rv as a reference sequence and other species are used in this study to know the various changes at the nucleotide level. The above study was used to know the base by base changes in the genome of seven different species in their genic and intergenic region. In a similar study a combination of previously and newly identified SNPs based on whole-genome comparisons of *M. tuberculosis* strains H37Rv, CDC1551 and 210 and *M. bovis* AF2122/97

Table 1: Total numbers of genes found in each of the *Mycobacterium* species and subspecies and probable SNP markers in numbers with reference to *Mycobacterium tuberculosis* H37Rv, control sequence

Species	Accession No.	Size of genome (Mbp)	Total No. of genes	Total SNPs	Marker SNPs
Mycobacterium tuberculosis H37Rv (Ref. Sequence)	AL123456	4.40000	4356	0	0
Mycobacterium leprae TN	AL450380	3.26820	5253	160	108
Mycobacterium bovis BCG Pasteur	AM408590	4.40000	4293	1238	202
Mycobacterium bovis subsp. *bovis* AF2122/97	BX248333	4.34549	4280	1235	237
Mycobacterium CDC 1551	AE000516	4.40384	4352	616	302
Mycobacterium avium subsp. *Paratuberculosis*	AE016958	4.80000	1578	160	117
Mycobacterium smegmatis str. MC2 155	CP000480	7.00000	7005		

Table 2: The annotated genes and SNPs in *Mycobacterium* species and sub species

GENE Rv	Genome_Coordinate		Anchor	Mtb_adam	Mtb_bovis	Mtb_bovisAF2122	Mtb_CDC1551	Mtb_kpnw	Mtb_smeg	Mtb_H37Rv
1	467	A		G	A					A
1	1057	G		A	G					G
2	2347	A		G	A					A
2	2532	T		C	T					T
	3751	T		G	T					T
	4480	C		T	C					C
5	5752	G		A	G					G
5	6406	C		T	C					C
5	6446	G		T	G					G
6	8285	C		T	C					C
6	8741	C		T	C					C
6	9143	T		C	T					T
6	9217	A		C	A					A
9	10727	A		G	A					A
	11370	C		C	T					C
17	13197	G		C	G					G
17	13460	A		G	A					A
19	14401	A		G	A					A
20	15117	C		G	C					C
22	17857	T	T	C	T					T
27	24679	A	A	G	A					A
27	25306	C	C	G	C					C
31	27714	C	T	C	C					C
34	30519	C	C	C	T					C
34	30688	T	T	T	G					T
34	30943	C	C	C	T					C
	31081	C	C	C	T					C
	33804	G	G	A	G					G
41	35342	C	T	C	C					C
45	39093	G	A	G	G					G
45	39273	G	A	G	G					G
46	40478	C	C	C	A					C
	41244	C	C	A	C					C
47	41453	C	C	T	C					C
	50005	G	G	A	G					G
57	50766	A	G	A	A					A
58	51649	G	G	G	C					G
59	51954	T	T	T	A					T
61	53785	C	C	C	G					C
61	55552	G	T	G	G					G
61	55557	C	C	C	T					C
63	57393	A	A	A	T					A
	59871	G	G	A	G					G
69	62657	G	G	G	A					G
	65786	A	C	A	A					A
74	65867	G	G	A	G					G
74	66285	C	C	C	G					C
77	69072	C	T	C	C					C
77	69871	C	C	C	T					C
77	69923	C	C	T	C					C
77	70092	A	A	G	A					A
77	71396	C	C	C	G					C
	75264	C	C	C	T					C
88	83982	A	A	A	G					A
89	84972	A	G	A	A					A
92	87079	T	C	T	T					T
96	89124	C	T	C	C					C
99	90929	C	C	T	C					C
103	95591	G	G	G	T					G
104	97388	G	G	G	T					G
119	114093	C	A	C	C					C
127	126042	G	C	G	G					G
127	126043	G	C	G	G					G
127	127410	C	C	G	C					C
127	128264	A	G	A	A					A
127	130200	C	C	T	C					C
131	135548	G	G	G	C					G
	141137	T	T	T	C					T
140	144390	G	G	G	A					G
141	145966	G	G	A	G					G
141	147455	C	C	A	C					C
142	147919	A	A	A	G					A
142	147985	C	C	C	T					C
145	150327	A	A	A	C					A
145	150334	C	C	C	G					C
145	150338	G	G	G	C					G
148	155293	G	G	G	A					G
153	159179	G	G	A	G					G
154	160335	T	T	T	C					T
154	160344	C	C	T	C					C
155	161351	C	A	C	C					C
156	162309	G	G	A	G					G
158	163573	A	A	A	C					A
160	165609	A	A	G	A					A
161	166506	G	G	A	G					G
	166852	C	C	C	T					C
169	172901	C	C	A	C					C
170	174684	C	C	C	A					C
174	178946	C	C	C	T					C
184	187738	G	G	G	A					G
187	191987	G	G	A	G					G
187	191988	G	G	A	G					G
187	193014	T	C	T	T					T
188	193381	C	C	C	G					C
196	202369	G	G	G	A					G
201	208305	A	G	A	A					A
211	217779	T	G	T	T					T
212	219144	C	T	C	C					C
214	220050	C	C	C	T					C
217	224441	T	C	T	T					T
220	227022	A	A	G	A					A
223	231994	C	C	G	C					C
224	232574	G	G	G	T					G
224	233979	C	T	C	C					C
230	241776	A	C	A	A					A
237	251468	G	G	G	T					G
238	253032	C	T	C	C					C
239	254012	G	G	G	A					G
242	258128	T	G	T	T					T
245	261235	G	G	G	A					G
247	263195	C	C	C	T					C
247	264021	T	C	T	T					T
249	265654	A			C					A
249	265968	C			G					C
250	266859	G	G	A	G					G
250	266860	G	G	A	G					G
254	271665	C	C	T	C					C
	277885	C	C	C	G					C
264	284623	G	G	G	A					G
267	288952	C	C	T	C					C
270	290454	C	C	T	C					C
274	294127	C	T	C	C					C
274	294128	C	T	C	C					C
277	297194	C	C	T	C					C
283	304679	G	G	G	T					G
287	308167	C	C	C	A					C
287	308337	G	G	G	A					G
291	313086	C	T	C	C					C
293	315502	C	C	C	A					C
295	316566	T	C	T	T					T
296	318068	C	T	C	C					C
296	320038	G	G	G	A					G
296	320502	C	C	C	T					C
297	322684	C	C	T	C					C
	323335	C	C	C	T					C
301	324902	C	G	C	C					C
304	328641	G	G	G	T					G
304	329218	G	G	A	G					G
313	338020	A		C						A
313	338100	T		C						T
319	346034	A	G	A	A					A
328	357278	C	T	C	C					C
331	361069	G	G	G	C					G
332	362667	C	C	C	T					C
334	363513	G	G	A	G					G
334	363757	C	C	T	C					C
337	364975	C	C	T	C					C
337	364976	C	C	T	C					C
349	382984	G	G	G	A					G
365	394808	C	C	C	T					C
367	397275	G	G	G	C					G
371	399660	T	G	T	T					T
374	405400	A	A	A	G					A
374	405750	T	T	T	G					T
376	407507	T	C	T	T					T
377	408515	G	A	G	G					G
379	409965	C	C	C	G					C
390	423289	A	G	A	A					A
393	424424	C	C	C	A					C
393	424425	C	C	C	A					C
394	426909	A	C	C						
394	428221	C	C	C	G					C
394	428228	G	G	G	T					G
394	428231	A	A	A	G					A
394	429060	T	T	C	T					T
398	437465	G	G	C	G					G
402	440745	T	T	C	T					T
408	445415	C	C	T	C					C
408	446406	C	C	C	G					C
412	448905	C	C	T	C					C
416	453394	T	T	T	G					T
416	453992	C	C	C	T					C
416	453993	C	C	C	T					C
421	456949	A	A	A	G					A
426	466326	T	T	C	T					T
426	466385	A	A	C	A					A
427	466923	C	T	C	C					C
445	483292	A	A	A	T					A
447	484874	G	T	G	G					G
448	487878	T	T	G	T					T
448	489073	C	C	C	G					C
448	489859	A	A	G	A					A
452	494270	C	T	C	C					C
458	501615	C	C	C	G					C
458	501620	C	C	T	C					C
463	505302	G	A	G	G					G
	506910	C	C	T	C					C
469	514245	C	C	C	T					C
473	518393	C	C	T	C					C
477	519806	G	G	A	G					G
480	524327	C	C	C	G					C
487	531550	C	G	C	C					C
490	534394	C	T	C	C					C
495	540552	C	C	T	C					C
498	543161	T	T	C	T					T
499	544461	C	C	C	T					C
	546914	G	G	G	A					G
504	549144	T	T	T	C					T
512	555330	C	C	T	C					C
514	558720	C	T	C	C					C
515	559153	C	T	C	C					C
521	562943	C	T	C	C					C
523	565155	C	C	A	C					C
540	579674	T	C	T	T					T
541	580860	G	A	G	G					G
548	587976	G	T	G	G					G
554	592059	C	A	C	C					C
	595321	T	T	T	G					T
560	597553	G	G	C	G					G
560	599165	A	A	A	C					A
567	606216	C	G	C	C					C
570	609926	G	G	A	G					G
585	621236	G	G	C	G					G
588	623931	G	G	G	T					G
593	628895	G	G	G	A					G
	631689	A	A	A	G					A
608	643082	G	G	C	G					G
608	643319	T	C	T	T					T
611	646331	G	G	G	A					G
613	648667	C	C	T	C					C
621	657269	A	A	A	G					A
625	659808	G	A	G	G					G
629	663451	G	G	G	C					G
629	664045	G	T	G	G					G
639	672637	C	C	T	C					C
639	674741	A	G	A	A					A
639	674744	C	G	C	C					C
639	674830	C	C	C	G					C
646	679173	T	T	C	T					T
647	680171	G	G	G	C					G
648	682667	C	C	C	T					C
648	682712	G	T	G	G					G
648	682964	T	C	T	T					T
648	683826	C	C	T	C					C
651	686655	C	C	C	T					C
657	692588	C	C	T	C					C
	697855	T	C	T	T					T
668	700084	C	C	G	C					C
668	700141	G	G	A	G					G
669	700943	C	T	C	C					C
684	711074	C	C	T	C					C
685	712082	C	C	G	C					C
697	721924	A	A	G	A					A
697	723473	C	T	C	C					C
698	725734	G	G	A	G					G
698	726816	G	G	G	C					G
699	728729	C	T	C	C					C
713	737857	C	T	C	C					C
718	746123	G	G	G	A					G
738	763575	G	C	G	G					G
738	765150	G	G	G	A					G
754	781568	C	T	C	C					C
755	782246	G	G	G	A					G
760	788259	G	G	G	T					G
	800219	T	T	T	C					T
778	803175	C	C	T	C					C
785	806630	T	C	T	T					T
785	806779	G	G	T	G					G
	809915	A	A	C	A					A
787	810887	A	A	G	A					A
805	822822	T	C	T	T					T
809	826181	G	G	C	G					G
	835012	A	A	T	A					A
822	835375	C	T	C	C					C
823	836487	C	C	C	A					C
823	837033	A			G					A
823	837434	G	G	G	A					G
823	837580	G	G	A	G					G
824	838990	C			G					C
824	839012	G			A					G
824	839123	A			G					A
824	839129	C			G					C
824	839194	A			G					A
824	839263	C			A					C
824	839334	A			G					A
824	839348	A			G					A

Table 2: Continued

GENE Rv	Genome Coordinate		Author	Mtb sukam	Mtb Jeck	Mtb JeckAF2122	Mtb CDC1551	Mtb RV
824	839381	C				T		C
824	839463	C				T		C
824	839579	G				C		G
829	842450	A		C	A	A		A
831	844629	A		A	C	A		A
831	845950	T		T	T	C		T
832	847648	C		C	C	T		C
833	849633	G		A	G	G		G
	850691	C		C	C	T		C
835	851133	T		T	T	G		T
840	855557	C		T	C	C		C
849	863632	T		T	G	T		T
850	864517	C		T	C	C		C
853	867927	T		T	G	T		T
854	868488	G		G	G	A		G
855	869440	C		C	C	T		C
856	872108	C		C	C	G		C
857	872431	A		G	A	A		A
864	877224	C		C	C	A		C
885	896356	C		C	C	T		C
885	897209	G		G	G	C		G
890	902413	C		C	C	T		C
892	905082	G		G	C	G		G
	916429	C		A	C	C		C
908	918316	T		T	T	C		T
	919574	T		T	T	C		T
910	919831	C		C	T	C		C
910	919923	A		A	G	A		A
915	923065	T		T	T	A		T
917	926801	G				C		G
917	927110	A				G		A
917	927385	A				G		A
918	927921	G		G	A			G
918	928220	C		C	T	C		C
926	935452	G		C	G	G		G
931	940313	C		C	A	C		C
932	940890	C		C	G	C		C
934	942964	A		A	G	A		A
936	945347	T		T	G	T		T
937	946958	C		T	C	C		C
937	947267	C		C	T	C		C
939	948219	C		C	A	C		C
940	949221	C		C	C	T		C
941	950230	C		C	C	T		C
947	955220	G		G	G	A		G
948	957533	A		A	A	C		A
950	961009	C		C	C	G		C
950	961941	C		G	C	C		C
955	964831	G		T		G		G
955	965069	G		G	A	G		G
958	967529	C		C	T	C		C
963	974138	A		A	G	A		A
	975861	T		T	C	T		T
	978373	T		T	T	C		T
969	979859	T		T	T	C		T
973	983462	C		T	C	C		C
973	983485	T		T	T	G		T
	997718	G		G	C	G		G
1000	1013936	G		G	G	C		G
	1014607	C		C	C	T		C
1008	1018028	G		G	A	G		G
1018	1028944	C		C	C	T		C
1020	1031219	C		G	C	C		C
	1039927	T		C	T	T		T
1034	1047827	C		C	T	C		C
1047	1064815	T		C	T	T		T
1054	1071673	C		C	C	T		C
1059	1075326	C		T	C	C		C
1072	1090293	C		C	C	G		C

GENE Rv	Genome Coordinate		Author	Mtb sukam	Mtb Jeck	Mtb JeckAF2122	Mtb CDC1551	Mtb RV
1081	1101047	C		A	C	C		C
1083	1102788	G		G	G	T		G
1083	1103540	C		C	T	C		C
1083	1103812	G		G	A	G		G
1084	1105607	G		A	G	G		G
1085	1106877	T		T	T	C		T
1088	1108769	C		T	C	C		C
	1112259	C		T	C	C		C
1092	1112658	A		A	G	A		A
1099	1118006	G		G	G	C		G
1101	1120396	G		G	G	C		G
1102	1121982	C		C	C	A		C
1103	1123319	G		G	G	C		G
1104	1123787	A		A	G	A		A
	1123788	A		A	G	A		A
1105	1125845	G		A	G	G		G
1117	1137325	C		C	G	C		C
1117	1137582	T		T	C	T		T
1119	1139014	C		C	T	C		C
1127	1148255	G		G	G	A		G
1129	1149361	T		C	T	T		T
1129	1151440	G		A	G	G		G
	1151706	C		C	T	C		C
1142	1161948	G		T	G	G		G
1146	1164361	G		G	G	A		G
1150	1166020	C		C	A	C		C
1152	1167982	T		T	G	T		T
1155	1171637	C		T	C	C		C
1155	1172044	T		G	T	T		T
1179	1192832	G		A	G	G		G
1183	1196508	C		T	C	C		C
1192	1205508	G		G	A	G		G
1202	1215076	C		A	C	C		C
1206	1218136	G		G	G	T		G
1206	1218141	G		G	G	T		G
1206	1218145	G		G	G	C		G
1206	1218147	A		A	A	T		A
1215	1228743	A		C	A	A		A
1216	1229475	T		T	G	T		T
1219	1231588	C		C	C	T		C
1224	1235657	C		T	C	C		C
1226	1237562	G		G	G	C		G
1238	1243347	C		C	G	C		C
1240	1244655	G		G	G	T		G
1244	1248619	C		C	C	G		C
1247	1251711	C		C	C	T		C
1251	1256806	C		C	C	T		C
1252	1257851	C		C	T	C		C
1256	1263869	G		G	G	A		G
	1275608	G		T	G	G		G
1280	1281650	C		T	C	C		C
1284	1283465	T		G	T	T		T
1284	1283552	G			G	C		G
1291	1292464	C		C	C	T		C
1292	1293114	C		T	C	C		C
1294	1294398	C		C	T	C		C
	1305531	C		C	G	C		C
	1305593	C		G		C		C
1308	1309527	G		G	G	A		G
1311	1317991	G		C	G	G		G
1320	1333814	C			T	C		C
1325	1338831	A		A	C	A		A
1332	1344741	A		A	G	A		A
1333	1346226	G		A	G	G		G
1344	1354643	A		C	A	A		A
1344	1354644	A		C	A	A		A
1349	1361375	C		C	T	C		C
1350	1362093	A		A	A	C		A
1351	1363023	C		G	C	C		C

GENE Rv	Genome Coordinate		Author	Mtb sukam	Mtb Jeck	Mtb JeckAF2122	Mtb CDC1551	Mtb RV
1353	1364961	A		C	A	A		A
1354	1365410	G		G	G	A		G
1361	1371606	A		A	C	A		A
1366	1375131	C		C	C	G		C
1373	1382459	T		T	C	T		T
1376	1384693	T		T	T	C		T
	1393994	C		C	C	T		C
1401	1414560	C		C	T	C		C
1401	1414564	A		A	T	A		A
1401	1414629	T		A	T	T		T
1405	1419349	C		C	T	C		C
1408	1423932	G		G	G	T		G
1413	1426079	C		C	C	T		C
1423	1439639	G		G	T	G		G
1423	1439938	C		C	T	C		C
1427	1444202	C		C	C	G		C
1439	1455207	C		G	C	C		C
1441	1458926	C		G	C	C		C
1448	1461915	C		C	C	G		C
1448	1462220	G		G	C	G		G
1457	1471923	T	C	T	T	T		T
1457	1471934	A	G	A	A	A	A	A
1457	1471969	T	C	T	T	T	T	T
	1471997	C	C	C	C	C	T	C
	1472030	A	T	A	A	A	A	A
	1472031	A	T	A	A	A	A	A
	1472046	G	C	G	G	G	G	G
	1472047	G	C	G	G	G	G	G
	1472059	C	C	C	C	C	T	C
	1472060	C	C	C	C	C	T	C
1458	1472070	T	T	T	T	T	C	T
1458	1472071	T	T	T	T	T	C	T
	1472153	C	C	C	C	C	A	C
	1472154	C	C	C	C	C	A	C
	1472265	C	C	C	C	C	T	C
	1472266	C	C	C	C	C	T	C
	1472303	C	C	C	C	C	T	C
	1472304	C	C	C	C	C	T	C
	1472489	A	A	A	A	A	G	A
	1472491	A	A	A	A	A	G	A
	1472557	G	G	G	G	G	A	G
1459	1472559	G	G	G	G	G	A	G
1459	1472616	G	G	G	G	G	A	G
1459	1472618	G	G	G	G	G	A	G
	1472844	C	C	C	C	C	T	C
	1472861	G	G	G	G	G	A	G
	1472969	A	G	A	A	A	A	A
	1472977	G	C	G	G	G	G	G
	1473001	G	G	G	G	G	C	G
1461	1473104	C	T	C	C	C	C	C
1461	1473105	G		G	G	G	A	G
1461	1473146	C	T	C	C	C	C	C
1461	1473167	G	A	G	G	G	G	G
	1473306	G		G	G	G	C	G
	1473668	C	T	C	C	C		C
	1473756	G	A	G	G	G		G
	1473758	G	T	G	G	G		G
	1473806	C	T	C	C	C		C
	1473813	C	T	C	C	C		C
	1473839	G	T	G	G	G		G
	1473877	G	A	G	G	G		G
	1473938	C	T	C	C	C		C
1463	1473945	G	A	G	G	G		G
1463	1473982	G	T	G	G	G		G
1463	1473997	G	T	G	G	G		G
1463	1474003	C	T	C	C	C		C
1463	1474011	A	G	A	A	A		A
1463	1474014	C	T	C	C	C		C
1463	1474053	T	C	T	T	T		T

GENE Rv	Genome Coordinate		Author	Mtb sukam	Mtb Jeck	Mtb JeckAF2122	Mtb CDC1551	Mtb RV
1463	1474056	T	C	T	T	T		T
1463	1474058	T	C	T	T	T		T
	1474114	T	C	T	T	T		T
1464	1474143	C	T	C	C	C		C
1464	1474154	C	T	C	C	C		C
1464	1474184	C	T	C	C	C		C
	1474252	G	T	G	G	G		G
	1474530	G	G	G	G	G	A	G
	1474539	C	C	C	C	C	T	C
	1474552	C	C	C	C	C	T	C
	1474714	G	G	G	G	G	T	G
1466	1474823	C	T	C	C	C	C	C
1466	1474824	C	T	C	C	C	C	C
1466	1474831	A	G	A	A	A	A	A
1466	1474832	A	G	A	A	A	A	A
	1474904	G	A	G	G	G	G	G
	1474905	G	A	G	G	G	G	G
	1474960	C	T	C	C	C	C	C
	1474961	C	T	C	C	C	C	C
	1474964	G	A	G	G	G	C	G
	1474991	C	C	C	C	C	T	C
	1474992	C	C	C	C	C	T	C
	1474994	T	C	T	T	T	T	T
	1474995	T	C	T	T	T	T	T
	1475016	G	G	G	G	G	A	G
	1475017	G	G	G	G	G	A	G
	1475018	C	T	C	C	C	C	C
	1475019	C	T	C	C	C	C	C
	1475037	G	G	G	G	G	A	G
	1475038	G	G	G	G	G	A	G
1467	1475202	G	A	G	G	G	G	G
1467	1475205	G	G	G	G	G	T	G
1467	1475229	A	A	A	A	A	C	A
1467	1475232	A	A	A	A	A	C	A
1467	1475269	T	C	T	T	T	T	T
1467	1475272	T	C	T	T	T	T	T
1467	1475274	A	C	A	A	A	A	A
1467	1475277	A	C	A	A	A	A	A
1467	1475296	A	C	A	A	A	A	A
1467	1475299	A	C	A	A	A	A	A
1467	1475302	C	C	C	C	C	T	C
	1475305	C	C	C	C	C	T	C
1468	1475320	A	A	A	A	A	T	A
1468	1475323	A	A	A	A	A	T	A
1468	1475339	T	C	T	T	T	T	T
1468	1475342	T	C	T	T	T	T	T
1468	1475367	G	G	G	G	G	A	G
1468	1475370	G	G	G	G	G	A	G
1468	1475404	C	C	C	C	C	T	C
1468	1475407	A	A	A	A	A	G	A
1468	1475410	A	A	A	A	A	G	A
1468	1475436	G	G	G	G	G	A	G
1468	1475439	G	G	G	G	G	A	G
1468	1475453	G	G	G	G	G	A	G
1468	1475455	T	T	T	T	T	G	T
1468	1475456	G	G	G	G	G	A	G
1468	1475458	T	T	T	T	T	G	T
1468	1475488	C	C	C	C	C	T	C
1468	1475491	C	C	C	C	C	T	C
1468	1475492	C	C	C	C	C	T	C
1468	1475493	C	T	C	C	C	C	C
1468	1475495	C	C	C	C	C	T	C
1468	1475512	G	A	G	G	G	G	G
1468	1475515	G	A	G	G	G	G	G
	1475533	C	A	C	C	C	C	C
	1475536	C	A	C	C	C	C	C
	1475583	C	C	C	C	C	T	C
	1475586	C	C	C	C	C	T	C
	1475608	T	C	T	T	T	T	T

Table 2: Continued

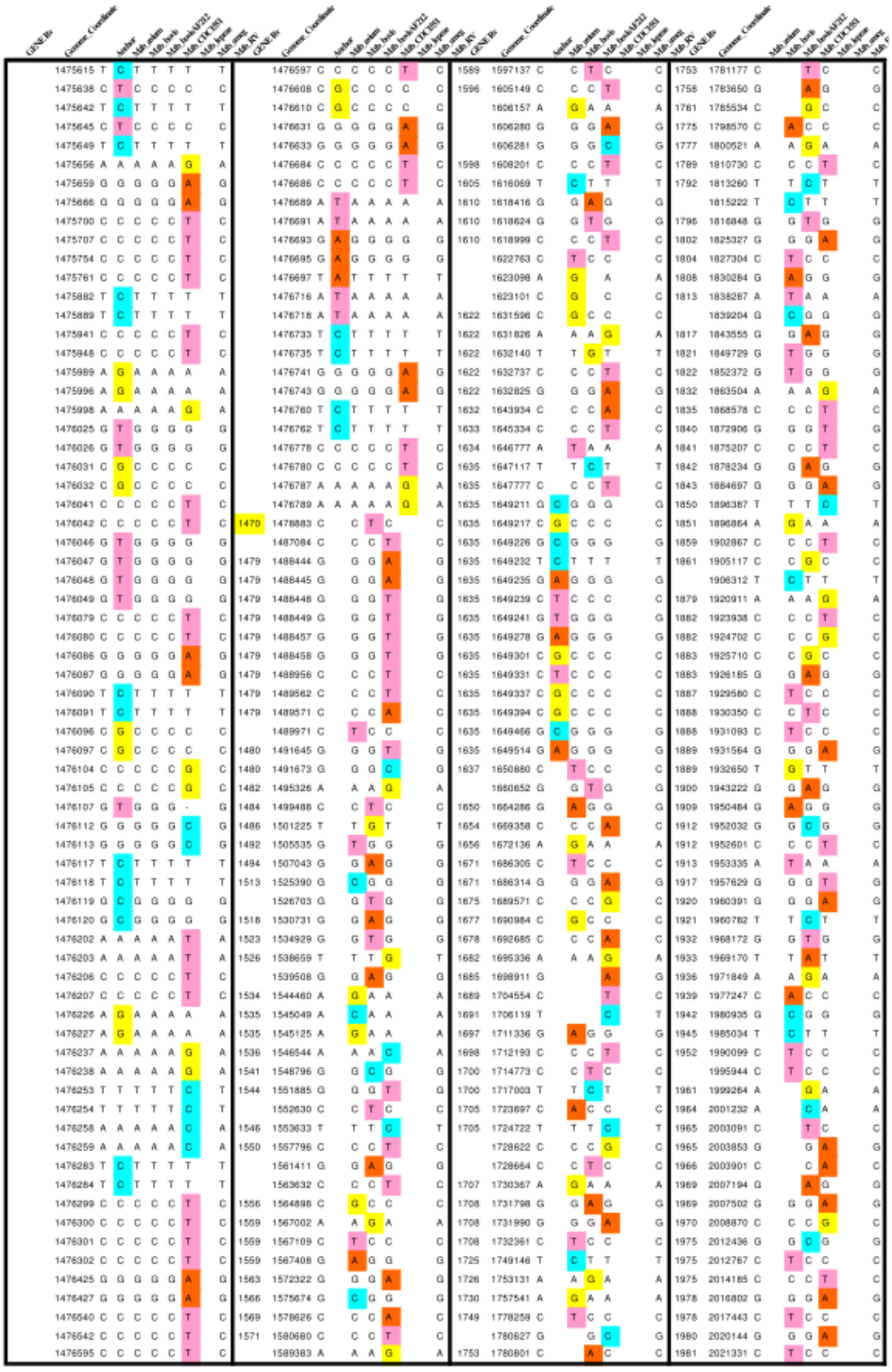
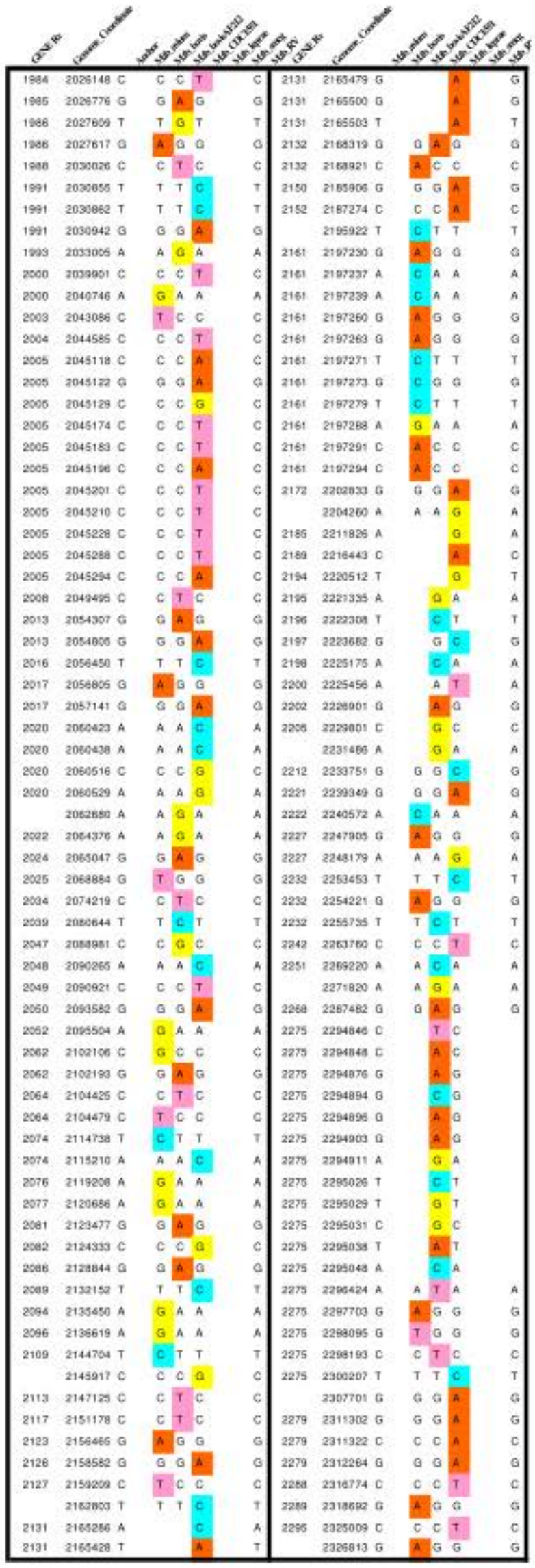


Table 2: Continued



were used to investigate *M. tuberculosis* evolution and phylogeny. These authors examined 212 SNPs (159 synonymous SNPs [sSNPs], 35 nonsynonymous SNPs [nsSNPs] and 18 intergenic SNPs [igSNPs]) discovered through pairwise comparisons of the *M. tuberculosis* H37Rv, CDC1551 and strain 210 and *M. bovis* AF2122/97 genomes (Filliol *et al*., 2006). In another study, a polymerase chain reaction and sequencing strategy was used to identify single nucleotide polymorphisms (SNPs) in 25 genes in the sheep (S) and cattle (C) strains of *Mycobacterium avium* subsp. *paratuberculosis* (*M. a. paratuberculosis*) and between *M. a. paratuberculosis* and *M. a. avium*. From 12,117 bp of sequence representing 26 loci across 25 genes, 11 SNPs were identified between the S and C strains in eight genes: *hsp65*, *sodA*, *dnaA*, *dnaN*, *recF*, *gyrB*, *inhA* and *pks8*. An *in silico* comparison of these *M. a. paratuberculosis* sequences and the *M. a. avium* 104 genome revealed 86 SNPs. (Marsh and Whittington 2007). Hershberg *et al*. (2008) discovered a total of 488 SNPs in 108 strains of *Mycobacterium tuberculosis* complex. The M. canettii strain differed from each of the other *Mycobacterium tuberculosis* complex strains at 129–145 sites (0.2% of the examined sites),while the maximum number of SNPs between any two other *Mycobacterium tuberculosis* complex strains was 46 (0.07% of the examined sites).

In this study we have generated useful information, relevant to analysis of the variations of protein formation at genic level, which directly affect the pathogenicity of organisms to the host. This study may be used for further analysis of host pathogen interactions at the pathway and product level. The whole genome SNP information can also be used for analysis of evolutionary relationships.

REFERENCES

Bocs, S., A. Danchin and C. Medigue, 2002. Re-annotation of genome microbial coding sequences: Finding new genes and inaccurately annotated genes. BMC. Bionform., 3: 1-5.

Cole, S.T., 1998. Comparative mycobacterial genomics. Curr. Opin. Microbiol., 1: 567-571.

Cole, S.T., R. Brosch, J. Parkhill, T. Garnier and C. Churcher *et al*., 1998. Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. Nature, 393: 537-544.

Cole, S.T., K. Eiglmeier, J. Parkhill, K.D. James and N.R. Thomson *et al*., 2001. Massive gene decay in the leprosy bacillus. Nature, 409: 1007-1011.

Dandekar, T., M. Huynen, J.T. Regula, B. Ueberle and C.U. Zimmermann *et al*., 2000. Re-annotating the *Mycoplasma pneumoniae* genome sequence: Adding value, function and reading frames. Nucl. Acids Res., 28: 3278-3288.

Filliol, I., A.S. Motiwala, M. Cavatore, W. Qi and M.H. Hazbón *et al*., 2006. Global phylogeny of *Mycobacterium tuberculosis* based on single nucleotide polymorphism (SNP) analysis: Insights into tuberculosis evolution, phylogenetic accuracy of other DNA fingerprinting systems and recommendations for a minimal standard SNP set. J. Bacteriol., 188: 759-772.

Fleischmann, R.D., D. Alland, J.A. Eisen, L. Carpenter and O. White *et al*., 2002. Whole-genome comparison of *Mycobacterium tuberculosis* clinical and laboratory strains. J. Bacteriol., 184: 5479-5490.

Gaasterland, T. and M. Opera, 2001. Whole genome analysis: Annotations and updates. Curr. Opin. Struct. Biol., 11: 377-381.

Garnier, T., K. Eiglmeier, J.C. Camus, M. Medina and H. Mansoor *et al*., 2003. The complete genome sequence of *Mycobacterium bovis*. Proc. Nat. Acad. Sci. USA., 100: 7877-7882.

Geluk, A., M.R. Klein, K.L. Franken, K.E. van Meijgaarden and B. Wieles *et al*., 2005. Postgenomic approach to identify novel *Mycobacterium leprae* antigens with potential to improve immunodiagnosis of infection. Infect. Immunol., 73: 5636-5644.

Hershberg, R., M. Lipatov, P.M. Small, H. Sheffer and S. Niemann, 2008. High functional diversity in *Mycobacterium tuberculosis* driven by genetic drift and human demography. PLoS Biol., 6: e311-e311.

Li, L., J.P. Bannantine, Q. Zhang, A. Amonsin and B.J. May *et al*., 2005. The complete genome sequence of *Mycobacterium avium* subspecies *paratuberculosis*. Proc. Nat. Acad. Sci. USA., 102: 12344-12349.

Marsh, I.B. and R.J. Whittington, 2007. Genomic diversity in *Mycobacterium avium*: Single nucleotide polymorphisms between the S and C strains of *M. avium* subsp. *paratuberculosis* and with *M. a. avium*. Mol. Cell. Probes, 21: 66-75.

Serres, M.H., S. Gopal, L.A. Nahum, P. Liang, T. Gaasterland and M. Riley, 2001. A functional update of the *Escherichia coli* K-12 genome. Genome Biol., 2: 0035.1-0035.7.