ACCESS MICROBIOLOGY

SHORT COMMUNICATION
Farrell *et al.*, *Access Microbiology* 2020;2
DOI 10.1099/acmi.0.000129

MICROBIOLOGY SOCIETY

OPEN MICROBIOLOGY

Updated functional annotation of the *Mycobacterium bovis* AF2122/97 reference genome

Damien Farrell[1,*], Joseph Crispell[1] and Stephen V. Gordon[1,2,3,4]

Abstract

Mycobacterium bovis AF2122/97 is the reference strain for the bovine tuberculosis bacillus. Here we report an update to the *M. bovis* AF2122/97 genome annotation to reflect 616 new protein identifications that replace many of the old hypothetical coding sequences and proteins of unknown function in the genome. These changes integrate information from functional assignments of orthologous coding sequences in the *Mycobacterium tuberculosis* H37Rv genome. We have also added 69 additional new gene names.

DATA SUMMARY

The sequence and annotation data referred to in this article are available under the DDBJ/ENA/GenBank accession no. LT708304 and can be viewed at this address: https://www.ebi.ac.uk/ena/browser/view/LT708304. All data sources, output files and the Jupyter notebook used to produce this analysis are archived at Zenodo with the DOI 10.5281/zenodo.3741935.

INTRODUCTION

Mycobacterium bovis is a causative agent of bovine tuberculosis (bTB) and the most widely studied animal-adapted member of the *Mycobacterium tuberculosis* complex (MTBC). The genome of the *M. bovis* AF2122/97 strain was first sequenced in 2003 [1] and is considered to be the reference sequence for this species. *Mycobacterium bovis* has considerable importance as the basis for comparative studies into animal- and human-adapted species of the MTBC. This genome was revised in 2017 [2] with an updated sequence that included the previously missing RD900 region and added 42 new coding sequences. These revisions brought the annotation in line with updates to the *M. tuberculosis* H37Rv genome, to which *M. bovis* shares high identity.

Hypothetical and unknown proteins

Proteins encoded by genes with no clear functional activity have traditionally been annotated with designations such as ‘unknown protein’, ‘hypothetical protein’ or ‘conserved hypothetical’. These labels are usually placed in the /product field of the annotation file (in this context we refer to the GenBank file format). This /product qualifier is not automatically updated when new protein functions are discovered, and hence genome annotation files become out of date over time if not regularly updated. Cross-references to databases with functional information are automatically added in a /db_xref qualifier during updates on the DDBJ/EMBL/GenBank system; however, these are not human readable. Therefore, it is a valuable and necessary exercise to update the protein product information directly in genome annotations of reference species.

New sources of annotation

Since the original annotation of *M. bovis* AF2122/97, the function of many hypothetical and newly identified proteins has been recognized. These data were collated by Doerks *et al.* in 2012 [3], who added approximately 620 new functional assignments to the remaining unknowns in the *M. tuberculosis* H37Rv reference genome. This was done using orthology and genomic context evidence. If a hypothetical protein was a member of a known orthologous

Received 19 February 2020; Accepted 16 March 2020; Published 30 April 2020
Author affiliations: [1]UCD School of Veterinary Medicine, University College Dublin, Dublin, Ireland; [2]UCD Conway Institute of Biomolecular and Biomedical Research, University College Dublin, Dublin, Ireland; [3]UCD School of Medicine, University College Dublin, Dublin 4, Ireland; [4]UCD School of Biomolecular and Biomedical Science, University College Dublin, Dublin 4, Ireland.
*Correspondence: Damien Farrell, damien.farrell@ucd.ie
Keywords: *Mycobacterium bovis*; annotation; genome; TB.
Abbreviations: DDBJ, DNA Data Bank of Japan; ENA, European Nucleotide Archive; PDB, Protein Data Bank; STRING, *Search Tool for the Retrieval of Interacting Genes/Proteins*.

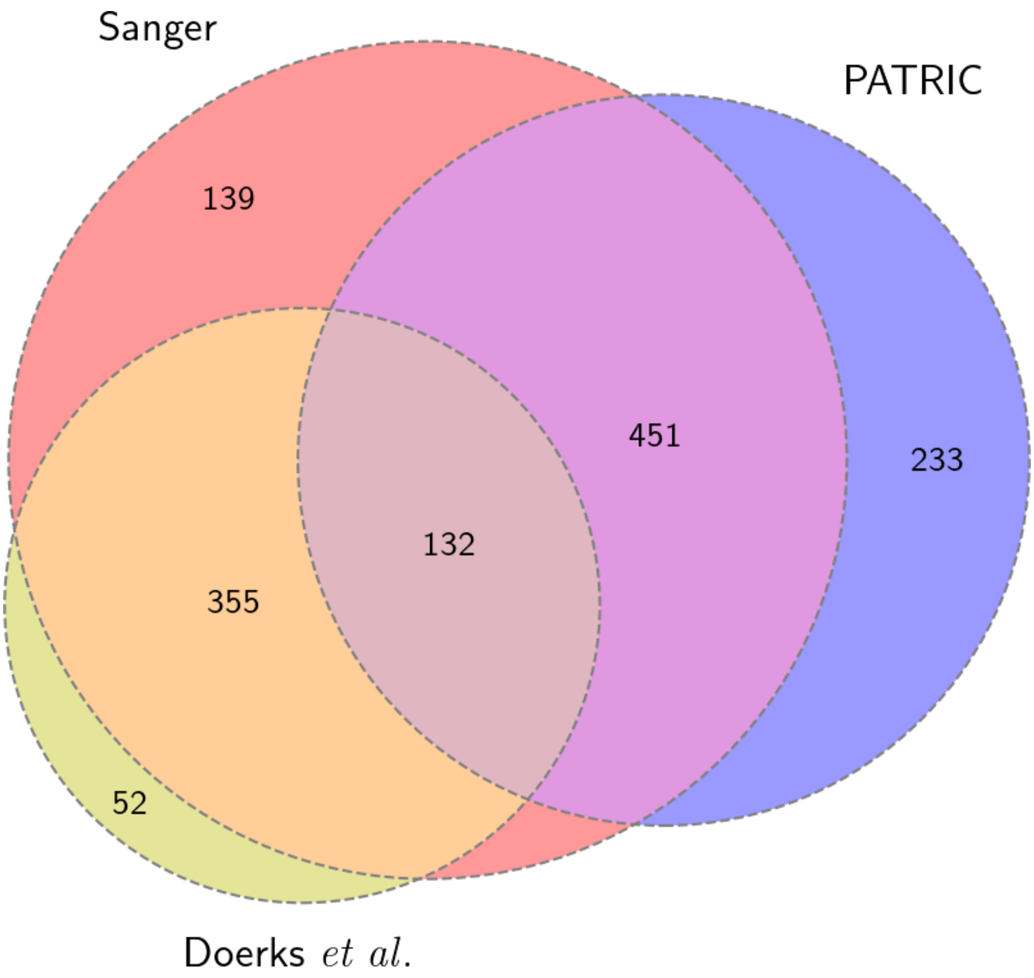

Fig. 1. Overlap between the hypothetical/unknown proteins in the reference (Sanger) and PATRIC H37Rv annotations and those found by Doerks *et al.*

group in the eggNOG database, this annotation was transferred to the corresponding *M. tuberculosis* protein. The STRING tool was also used, combining gene fusion events and significant co-occurrence to predict links with known proteins. The annotations vary in specificity; those predicted through association are more 'functional hints' as to the nature of the protein function rather than ascribing a specific function.

The PATRIC database [4] used their own pipeline based on RASTtk to reannotate H37Rv. The PATRIC annotation has some additional functional assignments and small genes that were not in the *M. tuberculosis* H37Rv reference. Some of the genes found by Doerks *et al.* are also assigned in PATRIC and a few have since been assigned to the reference. Fig. 1 shows the overlap in these two sets with the existing unknown proteins in the H37Rv reference.

UniProt [5] stores up-to-date protein annotations for the *M. tuberculosis* H37Rv strain, several of which are recent and were not present in either the PATRIC or the Doerks *et al.* data. These three sources were used to make an integrated table that was used to update the protein product field of the *M. bovis* AF2122/97 genome annotation.

METHODS

Data from three sources were used: (1) Doerks *et al.* (2012) [3]; (2) the PATRIC H37Rv annotation (https://www.patricbrc.org/view/Genome/83332.12); and (3) the UniProt *M. tuberculosis* H37Rv proteome (https://www.uniprot.org/proteomes/UP000001584) were downloaded as csv files and combined together by matching corresponding entries on the H37Rv locus tags (/locus_tag field). For all the unknown proteins in the H37Rv genome we selected updated annotations from these combined sources: if an annotation was present in the Doerk *et al.* dataset this was used preferentially, then PATRIC and finally UniProt. The order of preference was not significant. The majority of the data were taken from the Doerks *et al.* dataset, although 62 proteins from this dataset were excluded due to lack of a specific functional description.

Table 1. Current annotation statistics for the *M. bovis* and H37Rv genomes. The revised numbers for the updated annotations for AF2122/97 are shown in parentheses

	Coding sequences	Coding sequences with gene names	Hypothetical	Pseudogenes
Mbovis AF2122/97	3989	1964 (2026)	1097 (488)	11
MTB-H37Rv (reference)	4018	1953	1097	27
MTB-H37Rv (broad)	4143	16	843	92

The resulting table with new protein products was then matched to the *M. bovis* orthologous genes using a mapping between *M. tuberculosis* H37Rv and *M. bovis* locus tags. In a final step we performed a BLAST [6] of the remaining unknowns to the Protein Data Bank and found five additional proteins that have structures and function ascribed, and these were also added. Analysis was performed in Python, utilizing the Biopython [7] and pandas [8] libraries.

RESULTS

The current *M. bovis* genome annotation contains 3989 protein coding genes [2]. Of these, 1097 were marked as hypothetical, conserved or unknown proteins. These data are summarized in Table 1, with two *M. tuberculosis* H37Rv annotations for comparison. We have now added a total of 616 new protein product annotations to the genome, including five products from a PDB search. Sixty-nine new gene names were added from the UniProt data. There are now 488 hypothetical/unknowns remaining in the updated *M. bovis* AF2122/97 annotation. This revised annotation has been submitted to DDBJ/ENA/GenBank and is available under the accession no. LT708304.

CONCLUSION

Reference genomes stored on the International Nucleotide Sequence Database Collaboration (INSDC) provide the primary sources of annotation for virtually all bacterial species. Although there are now multiple alternative information sources for bacterial genomes, most are specialist databases and not universally known. The proliferation of data sources risks fragmentation of genome annotation and linked functional information; it is vitally important that reference sequence annotation in GenBank and Ensembl, the first port of call for the majority of researchers, are as up to date as possible. The *M. bovis* AF2122/97 strain is well established as a reference for *M. bovis* and the MTBC, and will remain a research cornerstone for the foreseeable future; maintaining an updated genome annotation for the research community drove our current work. We note that the latest reference annotation of *M. tuberculosis* H37Rv also lacks many of the updates we have added here to hypothetical proteins, underlining the need for constant curation of reference sequence annotation.

Funding information

This work was supported by the Irish Department of Agriculture Food and the Marine grant 15/S/651 (NEXUSMAP) and Science Foundation Ireland (SFI) grant 16/BBSRC/3390, part of the SFI–Biotechnology and Biological Sciences Research Council joint funding partnership.

Conflicts of interest

The authors declare that there are no conflicts of interest.

References

1. Garnier T, Eiglmeier K, Camus J-C, Medina N, Mansoor H *et al*. The complete genome sequence of *Mycobacterium bovis*. *Proc Natl Acad Sci U S A* 2003;100:7877–7882.
2. Malone KM, Farrell D, Stuber TP, Schubert OT, Aebersold R *et al*. Updated reference genome sequence and annotation of *Mycobacterium bovis* AF2122/97. *Genome Announc* 2017;5:17–18.
3. Doerks T, van Noort V, Minguez P, Bork P. Annotation of the *M. tuberculosis* hypothetical ORFeome: adding functional information to more than half of the uncharacterized proteins. *PLoS One* 2012;7:e34302.
4. Wattam AR, Abraham D, Dalay O, Disz TL, Driscoll T *et al*. PATRIC, the bacterial bioinformatics database and analysis resource. *Nucleic Acids Res* 2014;42:D581–.
5. Bateman A, The UniProt Consortium. UniProt: the universal protein knowledgebase. *Nucleic Acids Res* 2017;45:D158-D169.
6. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol* 1990;215:403–410.
7. Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ *et al*. Biopython: freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics* 2009;25:1422–1423.
8. Mckinney W. Pandas, python data analysis library. [Online]. Available: http://pandas.pydata.org/ 2015.