Updated reference genome sequence and annotation of *Mycobacterium bovis* AF2122/97

Kerri M. Malone[a], Damien Farrell[a], Tod P. Stuber[b], Olga T. Schubert[c*], Reudi Aebersold[c], Suelee Robbe-Austerman[b] and Stephen V. Gordon[a,e,f,g#]

[a]School of Veterinary Medicine, University College Dublin, Ireland;

[b]Diagnostic Bacteriology Laboratory, National Veterinary Services Laboratories, Ames, Iowa, USA;

[c]Department of Biology, Institute of Molecular Systems Biology, ETH Zurich, CH-8093, Switzerland;

[e]School of Medicine, University College Dublin, Ireland;

[f]School of Biomolecular and Biomedical Science, University College Dublin, Ireland;

[g]UCD Conway Institute of Biomolecular and Biomedical Research, University College Dublin, Ireland

*Present address: Department of Human Genetics, University of California, Los Angeles, USA

Running title

Mycobacterium bovis AF2122/97: an update

[#]Address correspondence to Stephen V. Gordon, stephen.gordon@ucd.ie

Abstract

We report an update to the reference genome of the bovine tuberculosis bacillus *Mycobacterium bovis* AF2122/97 generated using an integrative multi-'omics approach. Updates include 42 new CDS, 14 modified annotations, 26 SNP corrections, and disclosure that the RD900 locus, previously described as absent from the genome, is in fact present.

Text

The *Mycobacterium tuberculosis* complex (MTBC) is a collection of genetically related mycobacterial species that cause tuberculosis (TB) in human and animal hosts. *Mycobacterium bovis,* the causative agent of bovine tuberculosis (bTB), is the most widely studied animal-adapted MTBC member; bTB exacts a tremendous global economic toll through productivity losses and disease control costs, while zoonotic transmission of *M. bovis* infection is a threat to human health (Abernethy *et al.*, 2013; De Garine-Wichatitsky *et al.*, 2013; Kazoora *et al.*, 2016; Khattak *et al.*, 2016; Malama *et al.*, 2014; Muller *et al.*, 2013).

M. bovis AF2122/97 was the first bovine MTBC strain to be fully sequenced and provided a reference genome (Garnier *et al.*, 2003). Initial comparisons of the *M. bovis* AF2122/97 genome with that of the human-adapted *M. tuberculosis* H37Rv reference genome revealed high nucleotide sequence identity (> 99%), no unique genes *per se* in *M. bovis* AF2122/97 and a number of genomic deletions that led to a reduced genome size (Garnier et al., 2003). *M. bovis* AF2122/97 continues to serve as an MTBC reference genome despite last being updated in 2003; by comparison, the genome annotation of the reference *M. tuberculosis* H37Rv strain is currently on

release 27 (Lew *et al.*, 2013). An updated reference *M. bovis* genome will provide an essential resource for the TB research community and as a basis for comparative studies into animal- and human-adapted MTBC members.

To update the *M. bovis* AF2122/97 genome and annotation, a low-passage stock taken from the original *M. bovis* AF2122/97 seed stock was re-sequenced and re-annotated using a combination of DNA-, RNA-sequencing and proteomics data. All nucleic acid and protein samples were derived from exponentially grown cultures.

Short read DNA sequencing libraries were prepared using the Nextera XT DNA Library Preparation Kit (Illumina®) and sequenced on the MiSeq® system (Illumina®), generating 250bp paired-end reads that were trimmed using Sickle (Q >30), with 60X reference coverage (Joshi NA, 2011). For PacBio RS II sequencing, enzymatically extracted DNA was prepared using large insert library (6kb-8kb) size selection (van Soolingen *et al.*, 1991). Two SMRT cells were used for an output of 542,585,804 bases, a mean read length of 8,141, and 86X reference coverage. DNA sequencing datasets were analysed using a combination of *de novo* assembly (short reads, SOAPdenovo (Xie *et al.*, 2014); long reads, Canu (Koren S, 2016)) and nucleotide variant identification methods (short reads, Stampy, SAMtools and VCFtools (Li, 2011; Li *et al.*, 2009; Lunter and Goodson, 2011); long reads, Pilon (Walker *et al.*, 2014); MUMmer (Kurtz *et al.*, 2004)). This allowed for the update of the genome nucleotide sequence and the identification of genomic regions that were misassembled, or missed entirely, in the original sequencing project. Re-annotation of the *M. bovis* AF2122/97 genome was achieved by automatic annotation transfer from *M. tuberculosis* H37Rv (Version 27) (Otto *et al.*, 2011) and a

proteogenomic analysis using both *M. bovis* AF2122/97 shotgun MS/MS, SWATH MS datasets and *M. tuberculosis* H37Rv SWATH MS datasets (Schubert *et al.*, 2013).

Overall, 26 single nucleotide polymorphisms were identified. Strikingly, the large sequence polymorphism RD900, originally described as deleted from *M. bovis* 2122/97 (Bentley *et al.*, 2012), was found to be present; recombination between repeat structures flanking the RD900 locus in clones used for the original shotgun sequencing genome project may have led to loss of RD900. Furthermore, 42 novel coding sequences were identified while 14 existing annotations were modified.

Nucleotide sequence accession number(s): This Whole Genome Shotgun project had been deposited in DDBj/ENA/Genbank under the accession no. LT708304. SWATH MS data can be found on PeptideAtlas (http://www.peptideatlas.org) under identifier: PASS00932.

Acknowledgements:

The work was supported by funding from the Department for Agriculture, Food and the Marine (MycobactDiagnosis 11/RD/EMIDA/1), Science Foundation Ireland (08/IN.1/B2038), SystemsX.ch (through the project TbX) and a research grant from Institut Mérieux. We would like to thank Dr. Gerard Cagney (UCD School of Biomolecular and Biomedical Science) and Dr. Kieran Wynne (UCD Conway Proteomics Core) for shotgun mass spectrometric analysis of *M. bovis* samples.

References:

ABERNETHY, D. A., UPTON, P., HIGGINS, I. M., MCGRATH, G., GOODCHILD, A. V., ROLFE, S. J., BROUGHAN, J. M., DOWNS, S. H., CLIFTON-HADLEY, R., MENZIES, F. D., DE LA RUA-DOMENECH, R., BLISSITT, M. J., DUIGNAN, A. & MORE, S. J. 2013. Bovine tuberculosis trends in the UK and the Republic of Ireland, 1995-2010. *Vet Rec,* 172**,** 312.

BENTLEY, S. D., COMAS, I., BRYANT, J. M., WALKER, D., SMITH, N. H., HARRIS, S. R., THURSTON, S., GAGNEUX, S., WOOD, J., ANTONIO, M., QUAIL, M. A., GEHRE, F., ADEGBOLA, 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.

DE GARINE-WICHATITSKY, M., CARON, A., KOCK, R., TSCHOPP, R., MUNYEME, M., HOFMEYR, M. & MICHEL, A. 2013. A review of bovine tuberculosis at the wildlife-livestock-human interface in sub-Saharan Africa. *Epidemiol Infect,* 141**,** 1342-56.

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,* 100**,** 7877-82.

JOSHI NA, F. J. 2011. *Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files* [Online]. Available: https://github.com/najoshi/sickle [Accessed].

KAZOORA, H. B., MAJALIJA, S., KIWANUKA, N. & KANEENE, J. B. 2016. Knowledge, Attitudes and Practices Regarding Risk to Human Infection due to Mycobacterium bovis among Cattle Farming Communities in Western Uganda. *Zoonoses Public Health*.

KHATTAK, I., MUSHTAQ, M. H., AHMAD, M. U., KHAN, M. S. & HAIDER, J. 2016. Zoonotic tuberculosis in occupationally exposed groups in Pakistan. *Occup Med (Lond)*.

KOREN S, W. B., BERLIN K, MILLER JR, PHILLIPPY AM 2016. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *bioRxiv*.

KURTZ, S., PHILLIPPY, A., DELCHER, A. L., SMOOT, M., SHUMWAY, M., ANTONESCU, C. & SALZBERG, S. L. 2004. Versatile and open software for comparing large genomes. *Genome Biology,* 5.

LEW, J. M., MAO, C., SHUKLA, M., WARREN, A., WILL, R., KUZNETSOV, D., XENARIOS, I., ROBERTSON, B. D., GORDON, S. V., SCHNAPPINGER, D., COLE, S. T. & SOBRAL, B. 2013. Database resources for the tuberculosis community. *Tuberculosis (Edinb),* 93**,** 12-7.

LI, H. 2011. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics,* 27**,** 2987-93.

LI, H., HANDSAKER, B., WYSOKER, A., FENNELL, T., RUAN, J., HOMER, N., MARTH, G., ABECASIS, G., DURBIN, R. & GENOME PROJECT DATA PROCESSING, S. 2009.

The Sequence Alignment/Map format and SAMtools. *Bioinformatics,* 25**,** 2078-9.

LUNTER, G. & GOODSON, M. 2011. Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Res,* 21**,** 936-9.

MALAMA, S., MUMA, J., MUNYEME, M., MBULO, G., MUWONGE, A., SHAMPUTA, I. C., DJONNE, B., GODFROID, J. & JOHANSEN, T. B. 2014. Isolation and molecular characterization of Mycobacterium tuberculosis from humans and cattle in Namwala District, Zambia. *Ecohealth,* 11**,** 564-70.

MULLER, B., DURR, S., ALONSO, S., HATTENDORF, J., LAISSE, C. J., PARSONS, S. D., VAN HELDEN, P. D. & ZINSSTAG, J. 2013. Zoonotic Mycobacterium bovis-induced tuberculosis in humans. *Emerg Infect Dis,* 19**,** 899-908.

OTTO, T. D., DILLON, G. P., DEGRAVE, W. S. & BERRIMAN, M. 2011. RATT: Rapid Annotation Transfer Tool. *Nucleic Acids Res,* 39**,** e57.

SCHUBERT, O. T., MOURITSEN, J., LUDWIG, C., ROST, H. L., ROSENBERGER, G., ARTHUR, P. K., CLAASSEN, M., CAMPBELL, D. S., SUN, Z., FARRAH, T., GENGENBACHER, M., MAIOLICA, A., KAUFMANN, S. H., MORITZ, R. L. & AEBERSOLD, R. 2013. The Mtb proteome library: a resource of assays to quantify the complete proteome of Mycobacterium tuberculosis. *Cell Host Microbe,* 13**,** 602-12.

VAN SOOLINGEN, D., HERMANS, P. W., DE HAAS, P. E., SOLL, D. R. & VAN EMBDEN, J. D. 1991. Occurrence and stability of insertion sequences in Mycobacterium tuberculosis complex strains: evaluation of an insertion sequence-dependent DNA polymorphism as a tool in the epidemiology of tuberculosis. *J Clin Microbiol,* 29**,** 2578-86.

WALKER, B. J., ABEEL, T., SHEA, T., PRIEST, M., ABOUELLIEL, A., SAKTHIKUMAR, S., CUOMO, C. A., ZENG, Q. D., WORTMAN, J., YOUNG, S. K. & EARL, A. M. 2014. Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. *Plos One,* 9.

XIE, Y., WU, G., TANG, J., LUO, R., PATTERSON, J., LIU, S., HUANG, W., HE, G., GU, S., LI, S., ZHOU, X., LAM, T. W., LI, Y., XU, X., WONG, G. K. & WANG, J. 2014. SOAPdenovo-Trans: de novo transcriptome assembly with short RNA-Seq reads. *Bioinformatics,* 30**,** 1660-6.