Resequencing the *Mycobacterium avium* subsp. *paratuberculosis* K10 Genome: Improved Annotation and Revised Genome Sequence[▽]

James W. Wynne,[1]* Torsten Seemann,[2] Dieter M. Bulach,[1] Scott A. Coutts,[3]
Adel M. Talaat,[4] and Wojtek P. Michalski[1]

*CSIRO Livestock Industries, Australian Animal Health Laboratory, East Geelong, Victoria 3219, Australia[1];
Victorian Bioinformatics Consortium, Monash University, Clayton, Victoria 3800, Australia[2];
Micromon, Monash University, Clayton, Victoria 3800, Australia[3]; and The Laboratory of
Bacterial Genomics, Department of Pathobiological Sciences, University of
Wisconsin, Madison, Wisconsin[4]*

We report the resequencing and revised annotation of the *Mycobacterium avium* subsp. *paratuberculosis* K10 genome. A total of 90 single-nucleotide errors and a 51-bp indel in the original K10 genome were corrected, and the whole genome annotation was revised. Correction of these sequencing errors resulted in 28 frameshift alterations. The amended genome sequence is accessible via the supplemental section of study SRR060191 in the NCBI Sequence Read Archive and will serve as a valuable reference genome for future studies.

The American bovine isolate K10 remains the only *Mycobacterium avium* subsp. *paratuberculosis* genome to be fully sequenced and published to date (1). Although this 4.8-Mbp genome likely contains some assembly errors (3), it has provided, and will continue to provide, an invaluable resource for *Mycobacterium* research. The assembly errors were identified through optical mapping of related *M. avium* subsp. *paratuberculosis* strain ATCC 19698, which revealed a 648-kb inversion around the origin of replication and two additional copies of the insertion sequences IS*1311* and IS_*MAP03* (3). These findings were subsequently validated via PCR, Southern blotting, and (Sanger) sequence analysis in ATCC 19698 and were also confirmed to be present in K10 (3). We designate this interim corrected genome *M. avium* subsp. *paratuberculosis* K10′. To further improve this resource, we undertook a resequencing project of the original *M. avium* subsp. *paratuberculosis* K10 genome.

Whole-genome sequencing was performed on the Illumina GAIIx platform using one flow cell lane with 36-cycle paired-end chemistry. Reads were variably trimmed at the 3′ end based on the Illumina Read Segment Quality Indicator (Illumina manual), and read pairs containing ambiguous bases were removed. Read mapping onto the K10′ genome sequence was performed using SHRiMP (ver. 1.3.2) (2), and single-nucleotide polymorphisms and indels (deletion and insertion polymorphisms [DIPs]) were called using Nesoni (ver. 0.29; Monash University Victorian Bioinformatics Consortium) with default parameters. Read mapping determined that the data set comprised an average sequence coverage of 72.6 across the K10′ genome. This high sequence coverage allowed differences between K10\K10′ and the resequenced version of the genome, designated K10″, to be identified with high confidence. Ninety single-nucleotide differences and one 51-bp indel

were identified in the K10″ genome. As confirmation that these differences are likely to represent errors in the original genome sequence, we have also detected these polymorphisms in two additional bovine *M. avium* subsp. *paratuberculosis* genomes recently sequenced and assembled within our laboratory (data not shown). Seven of the 90 differences and the 51-bp indel were subjected to PCR and Sanger sequencing for verification. All of the polymorphisms were confirmed to be present in K10″ compared to the original genome sequence.

Thirty-six single-nucleotide deletions and four nucleotide insertions were identified in K10″ compared to the reference. These DIPs resulted in 27 frameshift mutations of protein coding loci. As a consequence of these frameshifts, one complete coding sequence (CDS) feature was removed (MAPK_3751), one novel CDS was created (MAPK_2081b), and one pseudogene was repaired (MAPK_4158-4159). In almost all of the other cases, the frameshifts resulted in proteins which more closely resembled their orthologs in *M. avium* subsp. *hominissuis* and *M. intracellulare*. Other frameshifts of biological interest include the truncation of a PPE family protein (MAPK_1173) and the extension of an MCE (mammalian cell entry) family protein (MAPK_4086). Compared to the reference, K10″ also had a 51-bp indel within a possible MCE family protein (MAPK_1575). This indel consisted of an 11-bp deletion (bases 2436510 to 2436520 in the original K10 sequence) and an insertion of 51 bp. The resulting protein sequence now more closely resembles orthologs of the MCE family in other *Mycobacterium* spp. In conclusion, the fact that so many of the amended bases have resulted in revised coding regions indicates the underlying importance of this exercise.

Nucleotide sequence accession number. The revised K10″ genome sequence and updated annotation have been deposited in the NCBI Sequence Read Archive as study SRR060191.

---

* Corresponding author. Mailing address: CSIRO Livestock Industries, Australian Animal Health Laboratory, 5 Portarlington Rd., East Geelong, Victoria 3219, Australia. Phone: 61352275602. Fax: 61352275000. E-mail: james.wynne@csiro.au.

REFERENCES

1. **Li, L. L., J. P. Bannantine, Q. Zhang, A. Amonsin, B. J. May, D. Alt, N. Banerji, S. Kanjilal, and V. Kapur.** 2005. The complete genome sequence of *Mycobacterium avium* subspecies *paratuberculosis*. Proc. Natl. Acad. Sci. U. S. A. **102:**12344–12349.
2. **Rumble, S. M., P. Lacroute, A. V. Dalca, M. Fiume, A. Sidow, and M. Brudno.** 2009. SHRiMP: accurate mapping of short color-space reads. PLoS Comput. Biol. **5:**e1000386.
3. **Wu, C. W., T. M. Schramm, S. G. Zhou, D. C. Schwartz, and A. M. Talaat.** 2009. Optical mapping of the *Mycobacterium avium* subspecies *paratuberculosis* genome. BMC Genomics **10:**25.