Complete Genome Sequence of *Spiroplasma turonicum* Tab4cT, a Bacterium Isolated from Horse Flies (*Haematopota* sp.)

Wen-Sui Lo (羅文穗),[a,b,c] Gail E. Gasparich,[d,e] Chih-Horng Kuo (郭志鴻)[a,b,f]

Institute of Plant and Microbial Biology, Academia Sinica, Taipei, Taiwan[a]; Molecular and Biological Agricultural Sciences Program, Taiwan International Graduate Program, National Chung Hsing University and Academia Sinica, Taipei, Taiwan[b]; Graduate Institute of Biotechnology, National Chung Hsing University, Taichung, Taiwan[c]; Department of Biological Sciences, Towson University, Towson, Maryland, USA[d]; College of Arts and Sciences, Salem State University, Salem, Massachusetts, USA[e]; Biotechnology Center, National Chung Hsing University, Taichung, Taiwan[f]

Spiroplasma turonicum Tab4cT was isolated from a horse fly (*Haematopota* sp.; probably *Haematopota pluvialis*) collected at Champchevrier, Indre-et-Loire, Touraine, France, in 1991. Here, we report the complete genome sequence of this bacterium to facilitate the investigation of its biology and the comparative genomics among *Spiroplasma* spp.

Address correspondence to Chih-Horng Kuo, chk@gate.sinica.edu.tw.

Spiroplasma turonicum is a bacterium associated with *Haematopota* sp. horse flies. The type strain Tab4cT was isolated from a single fly, probably *Haematopota pluvialis*, collected at Champchevrier (Indre-et-Loire, Touraine, France) in 1991 and was assigned to group XVII within the genus (1). As part of our ongoing effort to investigate *Spiroplasma* genome evolution (2), we determined the complete genome sequence of *S. turonicum* Tab4cT.

The DNA sample was prepared from the strain maintained in Gail Gasparich's laboratory at Towson University, which was acquired from the USDA/ARS *Spiroplasma* Culture Collection of Robert Whitcomb in 1996. This strain had been lyophilized after 17 passes from the original isolation. Prior to our completion of this project, a complete genome sequence of this bacterium was published based on another subculture of the same strain using the Pacific Biosciences platform (3). Because of this, we utilized this published sequence (GenBank accession no. CP012328.1) as the reference for a resequencing analysis. We chose the Illumina MiSeq platform to generate 301-bp reads from one paired-end library (~510-bp insert, 1,206,242 reads, ~288-fold coverage). The raw reads were mapped to the reference genome using BWA version 0.7.12 (4), programmatically checked using SAMTOOLS version 1.2 (5), and visually inspected using IGV version 2.3.57 (6).

The procedures for genome annotation were based on those described in our previous studies on *Spiroplasma* genomes (7–15). The programs RNAmmer (16), tRNAscan-SE (17), and Prodigal (18) were used for gene prediction. The gene names and product descriptions were first annotated based on the homologous genes in other *Spiroplasma* genomes, as identified by OrthoMCL (19). Subsequent manual curation was based on BLASTp (20) searches against the NCBI nonredundant database (21) and the KEGG database (22, 23). Putative clustered regularly interspaced short palindromic repeats (CRISPRs) were identified using CRISPRFinder (24).

Our resequencing analysis identified 13 polymorphic sites, including 12 single-nucleotide polymorphisms and one 1-bp indel in a homopolymeric region. It is unclear if these polymorphisms reflect true genetic variations or are artifacts of the sequencing technologies used. After correcting for these polymorphisms, the *S. turonicum* Tab4cT chromosome described in this work is 1,261,375 bp in size and has a G+C content of 24.2%. The two *S. turonicum* genomes both have one set of 16S-23S-5S rRNA genes, 29 tRNA genes (covering all 20 amino acids), and one 2,940-bp CRISPR locus (containing 44 spacers). However, the annotation of protein-coding genes differs between the two genomes. In CP012328.1, the annotation includes 1,085 protein-coding genes and no pseudogenes. Several of these predicted protein-coding genes appeared to be fragments of disrupted open reading frames and were merged into pseudogenes in our annotation. In the first version of our annotation, the *S. turonicum* Tab4cT genome contains 1,066 protein-coding genes and eight pseudogenes. Finally, the annotation of gene name and product description in this newly reported *S. turonicum* genome is more consistent with the majority of published *Spiroplasma* genomes (7–15).

Accession number(s). The complete genome sequence of *S. turonicum* Tab4cT has been deposited at DDBJ/EMBL/GenBank under the accession number CP013860.

REFERENCES

1. **Hélias C, Vazeille-Falcoz M, Le Goff F, Abalain-Colloc M-L, Rodhain F, Carle P, Whitcomb RF, Williamson DL, Tully JG, Bové JM, Chastel C.** 1998. *Spiroplasma turonicum* sp. nov. from *Haematopota* horse flies (Diptera: Tabanidae) in France. Int J Syst Bacteriol **48:**457–461. http://dx.doi.org/10.1099/00207713-48-2-457.

2. **Lo W-S, Huang Y-Y, Kuo C-H.** Winding paths to simplicity: genome evolution in facultative insect symbionts. FEMS Microbiol Rev, in press. http://dx.doi.org/10.1093/femsre/fuw028.

3. **Davis RE, Shao J, Zhao Y, Gasparich GE, Gaynor BJ, Donofrio N.** 2015. Complete genome sequence of *Spiroplasma turonicum* strain Tab4c^T, a parasite of a horse fly, *Haematopota* sp. (Diptera: Tabanidae). Genome Announc **3**(6)**:**e01367-15. http://dx.doi.org/10.1128/genomeA.01367-15.

4. **Li H, Durbin R.** 2009. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics **25:**1754–1760. http://dx.doi.org/10.1093/bioinformatics/btp324.

5. **Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R; 1000 Genome Project Data Processing Subgroup.** 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics **25:**2078–2079. http://dx.doi.org/10.1093/bioinformatics/btp352.

6. **Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP.** 2011. Integrative genomics viewer. Nat Biotechnol **29:**24–26. http://dx.doi.org/10.1038/nbt.1754.

7. **Lo W-S, Chen L-L, Chung W-C, Gasparich GE, Kuo C-H.** 2013. Comparative genome analysis of *Spiroplasma melliferum* IPMB4A, a honeybee-associated bacterium. BMC Genomics **14:**22. http://dx.doi.org/10.1186/1471-2164-14-22.

8. **Ku C, Lo W-S, Chen L-L, Kuo C-H.** 2013. Complete genomes of two dipteran-associated spiroplasmas provided insights into the origin, dynamics, and impacts of viral invasion in *Spiroplasma*. Genome Biol Evol **5:**1151–1164. http://dx.doi.org/10.1093/gbe/evt084.

9. **Lo W-S, Ku C, Chen L-L, Chang T-H, Kuo C-H.** 2013. Comparison of metabolic capacities and inference of gene content evolution in mosquito-associated *Spiroplasma diminutum* and *S. Taiwanense*. Genome Biol Evol **5:**1512–1523. http://dx.doi.org/10.1093/gbe/evt108.

10. **Ku C, Lo W-S, Chen L-L, Kuo C-H.** 2014. Complete genome sequence of *Spiroplasma apis* B31^T (ATCC 33834), a bacterium associated with May disease of honeybees (*Apis mellifera*). Genome Announc **2**(1)**:**e01151-13. http://dx.doi.org/10.1128/genomeA.01151-13.

11. **Chang T-H, Lo W-S, Ku C, Chen L-L, Kuo C-H.** 2014. Molecular evolution of the substrate utilization strategies and putative virulence factors in mosquito-associated *Spiroplasma* species. Genome Biol Evol **6:**500–509. http://dx.doi.org/10.1093/gbe/evu033.

12. **Paredes JC, Herren JK, Schüpfer F, Marin R, Claverol S, Kuo C-H, Lemaitre B, Béven L.** 2015. Genome sequence of the *Drosophila melanogaster* male-killing *Spiroplasma* strain MSRO endosymbiont. mBio **6:**e02437-14. http://dx.doi.org/10.1128/mBio.02437-14.

13. **Lo W-S, Gasparich GE, Kuo C-H.** 2015. Found and lost: the fates of horizontally acquired genes in arthropod-symbiotic *Spiroplasma*. Genome Biol Evol **7:**2458–2472. http://dx.doi.org/10.1093/gbe/evv160.

14. **Lo W-S, Lai Y-C, Lien Y-W, Wang T-H, Kuo C-H.** 2015. Complete genome sequence of *Spiroplasma litorale* TN-1^T (DSM 21781), a bacterium isolated from a green-eyed horsefly (*Tabanus nigrovittatus*). Genome Announc **3**(5)**:**e01116-15. http://dx.doi.org/10.1128/genomeA.01116-15.

15. **Lo W-S, Liu P-Y, Kuo C-H.** 2015. Complete genome sequence of *Spiroplasma cantharicola* CC-1^T (DSM 21588), a bacterium isolated from soldier beetle (*Cantharis carolinus*). Genome Announc **3**(5)**:**e01253-15. http://dx.doi.org/10.1128/genomeA.01253-15.

16. **Lagesen K, Hallin P, Rødland EA, Staerfeldt H-H, Rognes T, Ussery DW.** 2007. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Res **35:**3100–3108. http://dx.doi.org/10.1093/nar/gkm160.

17. **Lowe TM, Eddy SR.** 1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res **25:**955–964. http://dx.doi.org/10.1093/nar/25.5.0955.

18. **Hyatt D, Chen G-L, LoCascio PF, Land ML, Larimer FW, Hauser LJ.** 2010. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics **11:**119. http://dx.doi.org/10.1186/1471-2105-11-119.

19. **Li L, Stoeckert CJ, Roos DS.** 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res **13:**2178–2189. http://dx.doi.org/10.1101/gr.1224503.

20. **Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL.** 2009. BLAST+: architecture and applications. BMC Bioinformatics **10:**421. http://dx.doi.org/10.1186/1471-2105-10-421.

21. **Benson DA, Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW.** 2015. GenBank. Nucleic Acids Res **43:**D30–D35. http://dx.doi.org/10.1093/nar/gku1216.

22. **Kanehisa M, Goto S.** 2000. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res **28:**27–30. http://dx.doi.org/10.1093/nar/28.1.27.

23. **Kanehisa M, Goto S, Furumichi M, Tanabe M, Hirakawa M.** 2010. KEGG for representation and analysis of molecular networks involving diseases and drugs. Nucleic Acids Res **38:**D355–D360. http://dx.doi.org/10.1093/nar/gkp896.

24. **Grissa I, Vergnaud G, Pourcel C.** 2007. CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. Nucleic Acids Res **35:**W52–W57. http://dx.doi.org/10.1093/nar/gkm360.