JOURNAL OF BACTERIOLOGY, Oct. 2011, p. 5866–5867 Vol. 193, No. 20
0021-9193/11/$12.00 doi:10.1128/JB.05928-11

Genome Sequence of the *Mycobacterium colombiense* Type Strain, CECT 3035

Mónica González-Pérez,[1,2,3] Martha I. Murcia,[1,2] David Landsman,[3] I. King Jordan,[2,4] and Leonardo Mariño-Ramírez[1,2,3]*

Departamento de Microbiología, Facultad de Medicina, Universidad Nacional de Colombia, Bogotá, Colombia[1]; PanAmerican Bioinformatics Institute, Santa Marta, Magdalena, Colombia[2]; Computational Biology Branch, National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, 8600 Rockville Pike, Bethesda, Maryland 20894[3]; and School of Biology, Georgia Institute of Technology, Atlanta, Georgia 30332[4]

Received 2 August 2011/Accepted 4 August 2011

We report the first whole-genome sequence of the *Mycobacterium colombiense* type strain, CECT 3035, which was initially isolated from Colombian HIV-positive patients and causes respiratory and disseminated infections. Preliminary comparative analyses indicate that the *M. colombiense* lineage has experienced a substantial genome expansion, possibly contributing to its distinct pathogenic capacity.

The genus *Mycobacterium* comprises nearly 150 species (2, 3), including a number of human pathogens that pose major challenges to public health. *Mycobacterium colombiense* is a slow-growing, urease-positive, nontuberculous mycobacterium (NTM) that belongs to the *Mycobacterium avium* complex (MAC). *M. colombiense* was originally isolated from HIV-positive individuals in Bogotá, Colombia, and the patient isolates were determined to represent a distinct species by virtue of sequence comparisons with closely related *Mycobacterium* species (7). Since the discovery of this new species in 2006, *M. colombiense* has been confirmed to cause respiratory disease and disseminated infection in immunocompromised HIV patients, as well as lymphadenopathy in immunocompetent children (1, 8). Nevertheless, very little is currently known about the molecular mechanisms that underlie *M. colombiense* infection and pathogenesis. We have characterized the complete genome sequence of *M. colombiense* in an effort to better understand its virulence mechanisms.

The *M. colombiense* genome was sequenced by a whole-genome shotgun strategy using Roche 454 GS-FLX titanium pyrosequencing technology. A total of 720,174 sequence reads were generated, with an average read length of 375 bp, yielding more than 270 Mb of total sequence. This represents 45× coverage for the estimated 5.6-Mb genome size. A *de novo* assembly of the 454 single-end data was created using the Newbler assembler (Roche), version 2.6, resulting in 27 large contigs with an N_{50} of 436 kb. Genome annotation was performed using the NCBI Prokaryotic Genomes Automatic Annotation Pipeline (PGAAP), which produces functional annotation using the NCBI nonredundant protein and protein cluster databases with functional domain assignments for each protein by RPS-BLAST (5) against the NCBI Conserved Domain Database (6). The *M. colombiense* genome was predicted to encode 5,230 coding sequences (CDS).

M. colombiense was previously shown to be most closely related to *M. avium*, based on 16S rRNA sequence analysis along with DNA-DNA hybridization experiments (7). Here, we show that *M. colombiense* is most closely related to *M. avium* subsp. *paratuberculosis* (4) and confirm these results via sequence comparisons of *M. colombiense* contigs against the NCBI microbial sequence database. Despite the close relationship between these two species, reference-based assembly of the *M. colombiense* genome using *M. avium* subsp. *paratuberculosis* produced a highly fragmented assembly, with markedly lower quality than seen for the *de novo* assembly (1,914 large contigs with an N_{50} of 1,253), indicating that numerous genome rearrangements have occurred since the two species diverged. Furthermore, our characterization of the *M. colombiense* genome shows it to be substantially larger (5.6 Mb) than the genome of *M. avium* (4.8 Mb) and to encode many more genes (5,230 versus 4,400). Sequence alignments between the two species revealed that these differences could be attributed to large genomic insertions specific to the *M. colombiense* lineage. We hypothesize that a genome expansion may have allowed for the elaboration of novel pathways that contribute to the virulence of this emerging opportunistic pathogen. Additional genomic and functional analyses are needed to interrogate this hypothesis.

Nucleotide sequence accession number. The *M. colombiense* Whole Genome Shotgun project has been deposited at DDBJ/EMBL/GenBank under the accession number AFVW00000000.

This work was supported by an Alfred P. Sloan Research Fellowship in Computational and Evolutionary Molecular Biology (BR-4839 to I.K.J.) and the NCBI Scientific Visitors Program (ORISE to M.G.-P.). The research was supported in part by the Intramural Research Program of the NIH, NLM, NCBI.

We thank the Spanish Type Culture Collection (CECT) for providing strains.

* Corresponding author. Mailing address: Computational Biology Branch, Building 38A, Room 6S614M, 8600 Rockville Pike, MSC 6075, Bethesda, MD 20894-6075. Phone: (301) 402-3708. Fax: (301) 480-2288. E-mail: marino@ncbi.nlm.nih.gov.

REFERENCES

1. **Esparcia, O., F. Navarro, M. Quer, and P. Coll.** 2008. Lymphadenopathy caused by Mycobacterium colombiense. J. Clin. Microbiol. **46:**1885–1887.
2. **Euzéby, J. P.** 2011, posting date. List of Prokaryotic names with Standing in Nomenclature, genus Mycobacterium. J. P. Euzéby, Société de Bactériologie Systématique et Vétérinaire (SBSV), France. http://www.bacterio.cict.fr/m/mycobacterium.html.
3. **Euzéby, J. P.** 1997. List of Bacterial Names with Standing in Nomenclature: a folder available on the Internet. Int. J. Syst. Bacteriol. **47:**590–592.
4. **Li, L., et al.** 2005. The complete genome sequence of Mycobacterium avium subspecies paratuberculosis. Proc. Natl. Acad. Sci. U. S. A. **102:**12344–12349.
5. **Marchler-Bauer, A., and S. H. Bryant.** 2004. CD-Search: protein domain annotations on the fly. Nucleic Acids Res. **32:**W327–W331.
6. **Marchler-Bauer, A., et al.** 2011. CDD: a Conserved Domain Database for the functional annotation of proteins. Nucleic Acids Res. **39:**D225–D229.
7. **Murcia, M. I., E. Tortoli, M. C. Menendez, E. Palenque, and M. J. Garcia.** 2006. Mycobacterium colombiense sp. nov., a novel member of the Mycobacterium avium complex and description of MAC-X as a new ITS genetic variant. Int. J. Syst. Evol. Microbiol. **56:**2049–2054.
8. **Vuorenmaa, K., I. Ben Salah, V. Barlogis, H. Chambost, and M. Drancourt.** 2009. Mycobacterium colombiense and pseudotuberculous lymphadenopathy. Emerg. Infect. Dis. **15:**619–620.