JOURNAL OF BACTERIOLOGY, Oct. 2011, p. 5591–5592 Vol. 193, No. 19
0021-9193/11/$12.00 doi:10.1128/JB.05452-11

Complete Genome Sequences of *Mycobacterium tuberculosis* Strains CCDC5079 and CCDC5080, Which Belong to the Beijing Family

Yuanyuan Zhang,[1]# Chen Chen,[1,2]#* Jie Liu[1] Haijun Deng,[2] Aizhen Pan,[1] Lishui Zhang,[3] Xiuqin Zhao,[1] Mingxiang Huang,[3] Bing Lu,[1] Haiyan Dong,[1] Pengcheng Du,[1] Weijun Chen,[2] and KangLin Wan[1]*

National Institute for Communicable Disease Control and Prevention, Chinese Center for Disease Control and Prevention/State Key Laboratory for Infectious Disease Prevention and Control, Beijing 102206, China[1]*; Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 101300, China*[2]*; and Fuzhou Pulmonary Hospital (Clinical Teaching Hospital of Fujian Medical University), Fuzhou, Fujian Province 350008, China*[3]

Received 5 June 2011/Accepted 18 July 2011

***Mycobacterium tuberculosis* is one of most prevalent pathogens in the world. Drug-resistant strains of this pathogen caused by the excessive use of antibiotics have long posed serious threats to public health worldwide. A broader picture of drug resistance mechanisms at the genomic level can be obtained only with large-scale comparative genomic methodology. Two closely related Beijing family isolates, one resistant to four first-line drugs (CCDC5180) and one sensitive to them (CCDC5079), were completely sequenced. These sequences will serve as valuable references for further drug resistance site identification studies and could be of great importance for developing drugs targeting these sites.**

Mycobacterium tuberculosis is one of the best-studied pathogens, because of its prevalence and virulence (2, 6). In recent years, due to excessive antibiotic use, multidrug-resistant tuberculosis has become a serious public health threat in many countries and a major obstacle to disease control. Identifying the differences between drug-resistant and -sensitive strains will be of great help in drug resistance site identification. Toward this end, two closely related strains, CCDC5079 and CCDC5180, were sequenced using a whole-genome shotgun strategy.

CCDC5079 and CCDC5180 were collected from patients with secondary pulmonary tuberculosis at the same hospital in Fujian Province, China, in 2004. Spoligotyping results showed that they both belong to the Beijing family. However, they showed completely opposite phenotypes in drug resistance tests: CCDC5079 was sensitive to all four first-line drugs, while CCDC5180 was resistant to all four. We constructed two DNA libraries, of 1.5-kb and 3-kb fragments, for classical Sanger genome sequencing with automated capillary DNA sequencers (ABI3730 or MegaBACE1000) and constructed 200-bp DNA fragments for Solexa analysis. We obtained ~10-fold coverage with Sanger reads. The remaining gaps were filled by ~50-fold Solexa reads and PCR-amplified genomic fragments. Solexa data were first assembled with Velvet (7) and then combined with Sanger data using Phred-Phrap-Consed (3–5). Complete genome sequences were annotated by our automatic genome analysis pipeline, including gene prediction by Glimmer (1), gene annotation by BLAST against different databases, such as the GenBank nt and nr databases and Swissprot, and gene function prediction with COG and InterproScan. Gene pathways were annotated with KEGG.

The CCDC5079 genome is a circular DNA of 4,398,812 bp, which is 21,165 bp shorter than that of the virulent H37Rv strain. The most unusual feature of this genome is the presence of high-GC repetitive DNA, which might explain the difficulty in sequencing. The genome size of CCDC5180, the drug-resistant isolate, is 4,405,981 bp. Global alignment showed that the two strains are highly conserved. Genomic annotation revealed three kinds of virulence factors in the two genomes. One kind is involved in metabolism and *in vivo* growth, such as *pabB*, *ilvD*, *phoH2*, etc. This type of gene plays critical roles in lipid metabolism and signal pathways. The second includes cell envelope- and cell wall-associated genes (including *dim* and *pks*), which help the pathogen escape or invade host macrophages. The last corresponds to the PE/PPE family, which shows extensive differences between attenuated H37Ra and virulent H37Rv. Interestingly, we found that the genes responsible for first-line drug resistance (*katG*, *inhA*, *ahpC*, *kasA*, and *ndh* for isoniazid (INH) resistance, *rpoB* for rifapentine resistance, *embB* for ethambutol (EMB) resistance, and *rpsL* and *rrs* for streptomycin (SM) resistance) exist in both genomes. In all, 949 single-nucleotide polymorphisms are observed in the coding regions. Among them, 30.3% are nonsense and 69.7% change the amino acid sequences, which makes them good candidates for identifying drug resistance sites. Our study provides a foundation for drug resistance site identification.

Nucleotide sequence accession numbers. *M. tuberculosis* CCDC5079 and CCDC5180 were deposited in the GenBank database under accession numbers CP001641 and CP001642.

We thank the Beijing Institutes of Genomics for complete sequencing of two *M. tuberculosis* strains and making the results available to the public.

This work was supported by the project "Transmission Mode of

* Corresponding author. Mailing address: National Institute for Communicable Disease Control and Prevention, Chinese Center for Disease Control and Prevention/State Key Laboratory for Infectious Disease Prevention and Control, Beijing 102206, China. Phone for C. Chen: 8610-58900725. Fax: 8610-58900700. E-mail: chenchen@icdc.cn. Phone for K. Wan: 8610-58900776. Fax: 8610-58900779. E-mail: wankanglin@icdc.cn.

# These authors contributed equally to this work.

Tuberculosis" of the National Key Programme of Mega Infectious Diseases (2008ZX100/03-010) and "Study on the genetic polymorphism and strain level identification of *Mycobacterium tuberculosis* in China" of National Natural Science Funding of China (31771853). The authors have declared that no competing interests exist.

KangLin Wan, Chen Chen, and Weijun Chen conceived and designed the experiments. Yuanyuan Zhang, Jie Liu, Lishui Zhang, Xiuqin Zhao, Aizhen Pan, Mingxiang Huang, Bing Lu, and Haiyan Dong performed the experiments. Yuanyuan Zhang, Chen Chen, and Haijun Deng analyzed the data. Pengcheng Du contributed reagents, materials, and analysis tools. Yuanyuan Zhang, KangLin Wan, and Chen Chen wrote the paper.

REFERENCES

1. **Delcher, A. L., D. Harmon, S. Kasif, O. White, and S. L. Salzberg.** 1999. Improved microbial gene identification with GLIMMER. Nucleic Acids Res. **27:**4636–4641.
2. **Dye, C., S. Scheele, P. Dolin, V. Pathania, and M. C. Raviglione.** 1999. Consensus statement. Global burden of tuberculosis: estimated incidence, prevalence, and mortality by country. WHO Global Surveillance and Monitoring Project. JAMA **282:**677–686.
3. **Ewing, B., P. Green.** 1998. Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res. **8:**186–194.
4. **Ewing, B., L. Hillier, M. C. Wendl, and P. Green.** 1998. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res. **8:**175–185.
5. **Gordon, D., C. Abajian, and P. Green.** 1998. Consed: a graphical tool for sequence finishing. Genome Res. **8:**195–202.
6. **Sandgren, A., et al.** 2009. Tuberculosis drug resistance mutation database. PLoS Med. **6:**e2.
7. **Zerbino, D. R., and E. Birney.** 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. **18:**821–829.