genomeA
Journals.ASM.org

Genome Sequence of *Proteus mirabilis* Clinical Isolate C05028

Xiaolu Shi,[a,b] Yuanfang Zhu,[c] Yinghui Li,[b] Min Jiang,[b] Yiman Lin,[b] Yaqun Qiu,[b] Qiongcheng Chen,[b] Yanting Yuan,[c] Peixiang Ni,[c] Qinghua Hu,[b] Shenghe Huang[a]

Southern Medical University, Guangzhou, China[a]; Shenzhen Center for Disease Control and Prevention, Shenzhen, China[b]; BGI-Shenzhen, Shenzhen, People's Republic of China[c]

Genomic DNA of *Proteus mirabilis* C05028 was sequenced by an Illumina HiSeq platform and was assembled to 39 scaffolds with a total length of 3.8 Mb. Next, open reading frames (ORFs) were identified and were annotated by the KEGG, COG, and NR databases. Finally, we found special virulence factors only existing in *P. mirabilis* C05028.

Received 12 February 2014 **Accepted** 10 March 2014 **Published** 27 March 2014

Citation Shi X, Zhu Y, Li Y, Jiang M, Lin Y, Qiu Y, Chen Q, Yuan Y, Ni P, Hu Q, Huang S. 2014. Genome sequence of *Proteus mirabilis* clinical isolate C05028. Genome Announc. 2(2):e00167-14. doi:10.1128/genomeA.00167-14.

Address correspondence to Qinghua Hu, huqinghua03@163.com, or Shenghe Huang, shhuang18@yahoo.com.

Proteus mirabilis is found in soil, water, and the human intestinal tract and is characterized by its swarming motility, ability to ferment maltose, and inability to ferment lactose (1). The most common infection occurs when *P. mirabilis* moves to the urethra and bladder. However, we isolated 27 *P. mirabilis* strains from stool samples of patients infected in the food-borne disease outbreak in 2005 in Shenzhen, Guangdong, China. Here, we report the whole-genome sequence of a new *P. mirabilis* strain (C05028) that was isolated from patients suffering from diarrhea in this outbreak.

The genomic DNA of *P. mirabilis* C05028 was sequenced by the Illumina HiSeq 2000 and was used to construct two libraries of 500 bp and 2 kb. A total of 133 million bp reads were generated, reaching a depth of ~350-fold genome coverage, and were assembled into 39 scaffolds (≥200 bp in size), with a total length of 3,817,619 bp and containing 24,529-bp gap regions.

Open reading frames (ORFs) were identified with Glimmer version 3.0 (2), and 3,475 protein-coding sequences (CDSs) were predicted, with an average gene length of 928 bp. Repeat regions were predicted, including transposon sequences and tandem repeat sequences using RepeatMasker (3), RepeatProteinMasker, and the TRF software. Finally, we found 5.6-kb different transposable element (TE)-related sequences, consisting of 0.15% of the genome. The gene functions were annotated into the KEGG (4), COG (5), Swiss-Prot, TrEMBL (6), and NR databases using BLASTp. Homologous proteins were identified by BLASTp, with the criteria of an *E* value cutoff of 1e - 5 and a minimum aligned sequence length coverage of 50% of a query sequence. Using the above criteria yielded 2,595 protein families, with 2,582 single-copy protein families.

Our ultimate goal was to find special virulence factors of *P. mirabilis* C05028 by comparison with other nonpathogenic bacteria. We identified about 32,000 single-nucleotide polymorphisms (SNPs) using the MUMmer tool (7) and found that some SNPs were located in predicted genes of *P. mirabilis* C05028 related to virulence factors. Meanwhile, we found several virulence factors existing only in strain *P. mirabilis* C05028.

Nucleotide sequence accession number. This whole genome of *P. mirabilis* C05028 has been deposited at DDBJ/EMBL/GenBank under the accession no. ANBT00000000. The version described in this paper is the first version.

ACKNOWLEDGMENTS

This work was supported by grants from the National Natural Science Foundation of China (grants 81071433, 81302434, 81370740), the Medical Scientific Research Foundation of Guangdong Province, China (grant B2012323), and Science and Technology Planning Project of Shenzhen, Guangdong, China (grants 201302239 and JCYJ20130329103949651).

We declare no conflicts of interest.

REFERENCES

1. **Armbruster CE, Mobley HL.** 2012. Merging mythology and morphology: the multifaceted lifestyle of *Proteus mirabilis*. Nat. Rev. Microbiol. **10:** 743–754. http://dx.doi.org/10.1038/nrmicro2890.
2. **Delcher AL, Bratke KA, Powers EC, Salzberg SL.** 2007. Identifying bacterial genes and endosymbiont DNA with Glimmer. Bioinformatics **23:** 673–679. http://dx.doi.org/10.1093/bioinformatics/btm009.
3. **Tempel S.** 2012. Using and understanding RepeatMasker. Methods Mol. Biol. **859:**29–51. http://dx.doi.org/10.1007/978-1-61779-603-6_2.
4. **Kanehisa M.** 2008. The KEGG database, p 91–101, discussion 101–103, 119–128, 244–152. *In* Bock G, Goode JA (ed), '*In silico*' simulation of biological processes: Novartis Foundation Symposium 247. John Wiley & Sons, Ltd., Chichester, United Kingdom. http://dx.doi.org/10.1002/047085789 7.ch8.
5. **Tatusov RL, Galperin MY, Natale DA, Koonin EV.** 2000. The COG database: a tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Res. **28:**33–36. http://dx.doi.org/10.1093/nar/28.1.33.
6. **Lang F.** 1997. Swiss-Prot + Trembl. Trends Genet. **13:**417. http://dx.doi.org/10.1016/S0168-9525(97)01258-4.
7. **Delcher AL, Salzberg SL, Phillippy AM.** 2003. Using MUMmer to identify similar regions in large sequence sets. Curr. Protoc. Bioinformatics **10:**10.3. http://dx.doi.org/10.1002/0471250953.bi1003s00.