# Complete Genome Sequence of *Paenibacillus polymyxa* DSM 365, a Soil Bacterium of Agricultural and Industrial Importance

Santosh Kumar,[a] Victor C. Ujor[a]

[a]Metabolic Engineering and Fermentation Science Group, Department of Food Science, University of Wisconsin-Madison, Madison, Wisconsin, USA

**ABSTRACT** We report the complete genome sequence of *Paenibacillus polymyxa* DSM 365. The genome consists of a 5,788,318-bp chromosome, with a GC content of 45.48%. Annotation of the genome revealed a total of 5,246 genes (average length, 943 bp). Gene function analysis indicated the ability to fix nitrogen ($N_2$) and to produce value-added chemicals.

*P*aenibacillus polymyxa DSM 365 is a Gram-positive plant growth-promoting rhizobacterium (1) with capabilities for $N_2$ fixation and production of antimicrobials and commercially relevant chemicals (1–7). *P. polymyxa* DSM 365 was procured from the German Collection of Microorganisms and Cell Cultures GmbH (Leibniz Institute DSMZ GmbH). To isolate DNA, cultures were grown overnight in tryptic soy broth at 30 °C and 200 rpm. Genomic DNA was extracted using the Wizard high-molecular-weight (HMW) DNA extraction kit (Promega, Madison, WI, USA). Library preparation and sequencing were conducted by Novogene Inc. (Sacramento, CA) using the Illumina NovaSeq 6000 platform. To prepare the library for sequencing, genomic DNA was randomly sheared into short fragments. The fragments were end repaired, adenine tailed, and ligated with Illumina adapters. The quantified libraries (350-bp size) were pooled and sequenced to produce 6 Mb of paired-end 150-bp reads (1,800 Mb of raw data). In order to ensure accuracy and reliability, the reads were filtered using readfq software (v.10) (8) with default parameters to screen out low-quality data. The resulting 5,286,666 reads were assembled using SOAPdenovo (v.2.04) (9, 10), SPAdes (v.3.10.0) (11), and ABySS (v.1.3.7) (12) assembly software with default settings. Before assembly, the genome size was estimated by k-mer analysis (9). Assembly results from the three software tools were integrated with the Contig Integrator for Sequence Assembly (CISA) database (13). GapCloser (v.1.12) (14) was used to fill the gaps in the preliminary assembly. Fragments of less than 500 bp were filtered out, and the final result was counted for gene prediction. The assembly data revealed a total of 5,788,318 bp ($N_{50}$, 357,841 bp) in 47 scaffolds, with a GC content of 45.48% and an average read coverage of 291×.

GeneMarkS (v.4.10) (15) was used to identify coding genes, and noncoding RNAs were scanned using tRNAscan-SE, RNAmmer, and BLAST with the Rfam database (16–18). Interspersed repeats were predicted using RepeatMasker (v.4.0.9) (19), tandem repeats were predicted using Tandem Repeats Finder (v.4.09) (20), and clustered regularly interspaced short palindromic repeat (CRISPR) sequences were predicted using CRISPRFinder (v.2.0.3) (21).

The whole-genome sequence of *P. polymyxa* DSM 365 was submitted to the National Center for Biotechnology Information (NCBI) database using the Prokaryotic Genome Annotation Pipeline (PGAP) (v.6.0) (22). Homology-based gene prediction detected a total of 5,246 genes (85.43% of the total genome), with 4,966 protein coding sequences (CDSs), 156 RNA genes (tRNA, 104 genes; 5S rRNA, 13 genes; 16S rRNA, 18 genes; 23S rRNA, 17 genes; 4 ncRNA genes), and 104 pseudogenes. All of the protein sequences were aligned to the genome sequences using BLAST, and then GeneWise (23) was used to predict gene structure-based reliable alignments (E value of $<1e^{-5}$). Coding genes were

predicted by Augustus (v.2.7) (24) with homologous evidence. Several genes encoding enzymes involved in carbohydrate metabolism (e.g., rhamnogalacturonan lyase, cellulase, and cellobiohydrolase), nitrogen fixation (*nif* operon), sporulation, acetoin utilization, biosynthesis of siderophores, polyketides, exopolysaccharides, and butanediol were detected.

**Data availability.** The annotated genome sequence of *P. polymyxa* DSM 365 has been deposited in GenBank under the BioProject accession number PRJNA809744, the BioSample accession number SAMN26200526, and the Sequence Read Archive (SRA) accession number SRR18173204. The whole-genome shotgun project has been deposited in DDBJ/ENA/GenBank under the accession number JAKVDC010000000.

## REFERENCES

1. Timmusk S, Grantcharova N, Wagner EG. 2005. *Paenibacillus polymyxa* invades plant roots and forms biofilms. Appl Environ Microbiol 71:7292–7300. https://doi.org/10.1128/AEM.71.11.7292-7300.2005.

2. Weselowski B, Nathoo N, Eastman AW, MacDonald J, Yuan Z-C. 2016. Isolation, identification and characterization of *Paenibacillus polymyxa* CR1 with potentials for biopesticide, biofertilization, biomass degradation and biofuel production. BMC Microbiol 16:244. https://doi.org/10.1186/s12866-016-0860-y.

3. Xie J, Shi H, Du Z, Wang T, Liu X, Chen S. 2016. Comparative genomic and functional analysis reveal conservation of plant growth promoting traits in *Paenibacillus polymyxa* and its closely related species. Sci Rep 6:21329. https://doi.org/10.1038/srep21329.

4. Lal S, Tabacchioni S. 2009. Ecology and biotechnological potential of *Paenibacillus polymyxa*: a minireview. Indian J Microbiol 49:2–10. https://doi.org/10.1007/s12088-009-0008-y.

5. Okonkwo CC, Ujor VC, Mishra PK, Ezeji TC. 2017. Process development for enhanced 2,3-butanediol production by *Paenibacillus polymyxa* DSM 365. Fermentation 3:18. https://doi.org/10.3390/fermentation3020018.

6. Okonkwo CC, Ujor V, Ezeji TC. 2017. Investigation of relationship between 2,3-butanediol toxicity and production during growth of *Paenibacillus polymyxa*. N Biotechnol 34:23–31. https://doi.org/10.1016/j.nbt.2016.10.006.

7. Nelson DM, Glawe AJ, Labeda DP, Cann IK, Mackie RI. 2009. *Paenibacillus tundrae* sp. nov. and *Paenibacillus xylanexedens* sp. nov., psychrotolerant, xylan-degrading bacteria from Alaskan tundra. Int J Syst Evol Microbiol 59:1708–1714. https://doi.org/10.1099/ijs.0.004572-0.

8. Chen S, Zhou Y, Chen Y, Gu J. 2018. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics 34:i884–i890. https://doi.org/10.1093/bioinformatics/bty560.

9. Li R, Zhu H, Ruan J, Qian W, Fang X, Shi Z, Li Y, Li S, Shan G, Kristiansen K, Li S, Yang H, Wang J, Wang J. 2010. De novo assembly of human genomes with massively parallel short read sequencing. Genome Res 20:265–272. https://doi.org/10.1101/gr.097261.109.

10. Li R, Li Y, Kristiansen K, Wang J. 2008. SOAP: short oligonucleotide alignment program. Bioinformatics 24:713–714. https://doi.org/10.1093/bioinformatics/btn025.

11. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol 19:455–477. https://doi.org/10.1089/cmb.2012.0021.

12. Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJM, Birol I. 2009. ABySS: a parallel assembler for short read sequence data. Genome Res 19:1117–1123. https://doi.org/10.1101/gr.089532.108.

13. Lin SH, Liao YC. 2013. CISA: contig integrator for sequence assembly of bacterial genomes. PLoS One 8:e60843. https://doi.org/10.1371/journal.pone.0060843.

14. Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, He G, Chen Y, Pan Q, Liu Y, Tang J, Wu G, Zhang H, Shi Y, Liu Y, Yu C, Wang B, Lu Y, Han C, Cheung DW, Yiu S-M, Peng S, Xiaoqian Z, Liu G, Liao X, Li Y, Yang H, Wang J, Lam T-W, Wang J. 2012. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. Gigascience 1:18. https://doi.org/10.1186/2047-217X-1-18.

15. Besemer J, Lomsadze A, Borodovsky M. 2001. GeneMarkS: a self-training method for prediction of gene starts in microbial genomes: implications for finding sequence motifs in regulatory regions. Nucleic Acids Res 29:2607–2618. https://doi.org/10.1093/nar/29.12.2607.

16. 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. https://doi.org/10.1093/nar/25.5.955.

17. 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. https://doi.org/10.1093/nar/gkm160.

18. Gardner PP, Daub J, Tate JG, Nawrocki EP, Kolbe DL, Lindgreen S, Wilkinson AC, Finn RD, Griffiths-Jones S, Eddy SR, Bateman A. 2009. Rfam: updates to the RNA families database. Nucleic Acids Res 37:D136–D140. https://doi.org/10.1093/nar/gkn766.

19. Saha S, Bridges S, Magbanua ZV, Peterson DG. 2008. Empirical comparison of ab initio repeat finding programs. Nucleic Acids Res 36:2284–2294. https://doi.org/10.1093/nar/gkn064.

20. Benson G. 1999. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res 27:573–580. https://doi.org/10.1093/nar/27.2.573.

21. 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. https://doi.org/10.1093/nar/gkm360.

22. Tatusova T, DiCuccio M, Badretdin A, Chetvernin V, Nawrocki EP, Zaslavsky L, Lomsadze A, Pruitt KD, Borodovsky M, Ostell J. 2016. NCBI Prokaryotic Genome Annotation Pipeline. Nucleic Acids Res 44:6614–6624. https://doi.org/10.1093/nar/gkw569.

23. Clamp M, Durbin R, Birney E. 2004. GeneWise and GenomeWise. Genome Res 14:988–995. https://doi.org/10.1101/gr.1865504.

24. Stanke M, Diekhans M, Baertsch R, Haussler D. 2008. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. Bioinformatics 24:637–644.