



Genome Sequence of *Escherichia coli* Stbl4, a Versatile Genetic Tool for Heterologous Expression

Lina Assad,^a Karolis Matjošaitis,^b  Harald Gross^{a,c}

^aDepartment of Pharmaceutical Biology, Pharmaceutical Institute, University of Tübingen, Tübingen, Germany

^bThermo Fisher Scientific Baltics, Vilnius, Lithuania

^cGerman Centre for Infection Research (DZIF), Partner Site Tübingen, Tübingen, Germany

ABSTRACT *Escherichia coli* Stbl4 is widely used as a laboratory strain for heterologous expression of large gene clusters. Since no genome sequence has been publicly available, we here report the draft sequence of Stbl4, including its F-plasmid. It should serve as a useful reference for researchers working with Stbl4.

Historically, strain Stbl4 represents an advancement of Stbl2, which is derived from JM109/J5 (1). The latter strain was in turn constructed from strain DH1 (2, 3) via JM106 to JM108 (4). DH1 itself stems from the parent strain lineage strains MM294 (5), 1100 (6, 7), 1000 (7) and HfrCxW208 (8) from the original K-12 wild-type strain (8), which was isolated from the stool of a convalescent diphtheria patient in 1922 (9). The strain Stbl4 (shortcut for “stable 4”) is recommended for cloning of unstable DNA mediated by the introduction of repetitive DNA, retroviral sequences, or large inserts which otherwise would lead to undesired DNA rearrangements (10–12). For the latter reason, Stbl4 is also used as a host for the discovery of complex secondary metabolites (12).

During a heterologous expression study of a biosynthetic gene cluster (BGC) using Stbl4, we experienced unexpected cross talk between the inserted gene cluster and an existing BGC and needed detailed information about the genetic architecture of the host to investigate this phenomenon. Therefore, we describe the complete genome sequence of Stbl4.

ElectroMAX Stbl4 competent cells were purchased from Invitrogen and streaked onto Luria-Bertani (LB) agar plates containing tetracycline (12.5 μ g/ml). After 16 h of incubation at 37°C, two single colonies were picked and inoculated into liquid LB at 37°C on a rotary shaker (200 rpm) until saturation. Genomic DNA extraction and PacBio library preparation were conducted as previously described (13). The 6-kb library was sequenced on a PacBio Sequel instrument using one single-molecule real-time (SMRT) cell. An aliquot of the same DNA preparation was used to create a genomic Nextera XT paired-end library for sequencing using an Illumina NovaSeq platform. The results of both sequencing platforms enabled a *de novo* hybrid assembly. The PacBio data were processed and filtered using the SMRT Link software suite, whereby subreads shorter than 50 bp were discarded. The remaining PacBio long reads were assembled using SMRT Link v7.0.1 and HGAP4 (14, 15). The initial quality assessment of the Illumina data was based on data passing the Illumina Chastity filter. Afterwards, reads containing a PhiX control signal were removed, and reads containing adapters were clipped. Further quality assessment was based on the remaining reads using the FASTQC tool v0.11.8 (16). Subsequently, the Illumina data were aligned to the abovementioned HGAP4 assembly genome using BMap v36.77 (<https://sourceforge.net/projects/bbmap/>). Finally, assembly errors and the nucleotide disagreements between the Illumina reads and scaffold sequences were corrected using Pilon v1.21.1 (17). All software settings were kept at their defaults except for the HGAP4 genome size estimate parameter, which was set to 4.7 Mbp. Overall, the hybrid *de novo* assembly resulted in a nucleotide draft

Citation Assad L, Matjošaitis K, Gross H. 2021. Genome sequence of *Escherichia coli* Stbl4, a versatile genetic tool for heterologous expression. *Microbiol Resour Announc* 10: e00823-21. <https://doi.org/10.1128/MRA.00823-21>.

Editor David Rasko, University of Maryland School of Medicine

Copyright © 2021 Assad et al. This is an open-access article distributed under the terms of the [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/).

Address correspondence to Harald Gross, harald.gross@uni-tuebingen.de.

Received 17 August 2021

Accepted 19 September 2021

Published 7 October 2021

TABLE 1 Sequencing metrics for *E. coli* Stbl4

Statistic or characteristic	Data for <i>E. coli</i> Stbl4
PacBio sequencing	
No. of reads	540,276
Mean read length (bp)	4,673
No. of mapped reads	534,707
Avg coverage (×)	486
No. of contigs	6
N_{50} (bp)	4,404,505
No. of gaps	0
Illumina sequencing	
Read length (nt) ^a	2 × 150
No. of initial reads	14,301,826
Yield (Mbp)	2,049
No. of reads passing quality control	12,622,970 (88.3%)
Avg quality	35.68
Avg coverage (×)	431
Median insert size (bp)	354.00
<i>De novo</i> hybrid assembly	
Genome size (bp)	4,502,782
GC content (%)	50.7
No. of contigs	2
Contig 1 (chromosome) size (bp)	4,404,514
Contig 2 (F-plasmid) size (bp)	98,268
No. of genes (total)	4,353
No. of genes (coding)	4,060
No. of predicted biosynthetic gene clusters	2 (1 encoding thiopeptide, 1 encoding enterochelin)

^ant, nucleotides.

sequence consisting of two contigs, representing one chromosome and its F-plasmid (18, 19). The assembled contigs were annotated using the PGAP v5.2 pipeline (20, 21), while an automated secondary metabolism analysis was conducted employing antiSMASH v6.0.1 (22). The essential genome features are summarized in Table 1.

Data availability. The genome sequences of both contigs have been deposited in the NCBI GenBank database under accession numbers [CP076043.1](#) and [CP076044.1](#) for the chromosome and plasmid, respectively. The SRA accession numbers are [SRX11660715](#) and [SRR15358205](#) for the PacBio and Illumina reads, respectively.

REFERENCES

- Trinh T, Jessee J, Bloom FR, Hirsch V. 1994. STBL2TM: an *Escherichia coli* strain for the stable propagation of retroviral clones and direct repeat sequences. *Focus* 16:78–80.
- Hanahan D. 1983. Studies on transformation of *Escherichia coli* with plasmids. *J Mol Biol* 166:557–580. [https://doi.org/10.1016/s0022-2836\(83\)80284-8](https://doi.org/10.1016/s0022-2836(83)80284-8).
- Suzuki S, Ono N, Furusawa C, Ying B-W, Yomo T. 2011. Comparison of sequence reads obtained from three next-generation sequencing platforms. *PLoS One* 6:e19534. <https://doi.org/10.1371/journal.pone.0019534>.
- Yanisch-Perron C, Vieira J, Messing J. 1985. Improved M13 phage cloning vectors and host strains: nucleotide sequences of the M13mp18 and pUC19 vectors. *Gene* 33:103–119. [https://doi.org/10.1016/0378-1119\(85\)90120-9](https://doi.org/10.1016/0378-1119(85)90120-9).
- Meselson M, Yuan R. 1968. DNA restriction enzyme from *E. coli*. *Nature* 217: 1110–1114. <https://doi.org/10.1038/2171110a0>.
- Fox CF, Wilson G. 1968. The role of a phosphoenolpyruvate-dependent kinase system in 3-glucoside catabolism in *Escherichia coli*. *Proc Natl Acad Sci U S A* 59:988–995. <https://doi.org/10.1073/pnas.59.3.988>.
- Dürwald H, Hoffmann-Berling H. 1968. Endonuclease I-deficient and ribonuclease I-deficient *Escherichia coli* mutants. *J Mol Biol* 34:331–346. [https://doi.org/10.1016/0022-2836\(68\)90257-x](https://doi.org/10.1016/0022-2836(68)90257-x).
- Bachmann BJ. 1972. Pedigrees of some mutant strains of *Escherichia coli* K-12. *Bacteriol Rev* 36:525–557. <https://doi.org/10.1128/br.36.4.525-557.1972>.
- Lederberg J. 1951. Genetic studies with bacteria, p 263–289. *In* Dunn LC (ed), *Genetics in the 20th century*. Macmillan, New York, NY.
- Bzymek M, Lovett ST. 2001. Instability of repetitive DNA sequences: the role of replication in multiple mechanisms. *Proc Natl Acad Sci U S A* 98: 8319–8325. <https://doi.org/10.1073/pnas.111008398>.
- Mostafa A, Kanrai P, Petersen H, Ibrahim S, Rautenschlein S, Pleschka S. 2015. Efficient generation of recombinant influenza A viruses employing a new approach to overcome the genetic instability of HA segments. *PLoS One* 10:e0116917. <https://doi.org/10.1371/journal.pone.0116917>.
- Zhang JJ, Yamanaka K, Tang X, Moore BS. 2019. Direct cloning and heterologous expression of natural product biosynthetic gene clusters by transformation-associated recombination. *Methods Enzymol* 621:87–110. <https://doi.org/10.1016/bs.mie.2019.02.026>.
- Buchmann A, Gross H. 2020. Improved *de novo* draft genome sequence of the nocavionin-producing type strain *Nocardia terpenica* IFM 0706 and comparative genomics with the closely related strain *Nocardia terpenica* IFM 0406. *Microbiol Resour Announc* 9:e00689-20. <https://doi.org/10.1128/MRA.00689-20>.
- Chin C-S, Alexander DH, Marks P, Klammer AA, Drake J, Heiner C, Clum A, Copeland A, Huddleston J, Eichler EE, Turner SW, Korlach J. 2013. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat Methods* 10:563–569. <https://doi.org/10.1038/nmeth.2474>.
- Chin C-S, Peluso P, Sedlazeck FJ, Nattestad M, Concepcion GT, Clum A, Dunn C, O'Malley R, Figueroa-Balderas R, Morales-Cruz A, Cramer GR, Delledonne M, Luo C, Ecker JR, Cantu D, Rank DR, Schatz MC. 2016. Phased diploid

- genome assembly with single-molecule real-time sequencing. *Nat Methods* 13:1050–1054. <https://doi.org/10.1038/nmeth.4035>.
16. Andrews S. 2010. FastQC: a quality control tool for high throughput sequence data. <https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>.
 17. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q, Wortman J, Young SK, Earl AM. 2014. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* 9:e112963. <https://doi.org/10.1371/journal.pone.0112963>.
 18. Firth N, Ippen-Ihler K, Skurray RA. 1996. Structure and function of the F factor and mechanism of conjugation, p 2377–2401. *In* Neidhardt FC, Curtiss R, III, Ingraham JL, Lin ECC, Low KB, Magasanik B, Reznikoff WS, Riley M, Schaechter M, Umberger HE (ed), *Escherichia coli* and *Salmonella*: cellular and molecular biology, 2nd ed. ASM Press, Washington, DC.
 19. Virolle C, Goldlust K, Djermoun S, Bigot S, Lesterlin C. 2020. Plasmid transfer by conjugation in Gram-negative bacteria: from the cellular to the community level. *Genes* 11:1239. <https://doi.org/10.3390/genes11111239>.
 20. Tatusova T, DiCuccio M, Badretdin A, Chetvermin 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>.
 21. Li W, O'Neill KR, Haft DH, DiCuccio M, Chetvermin V, Badretdin A, Coulouris G, Chitsaz F, Derbyshire MK, Durkin AS, Gonzales NR, Gwadz M, Lanczycki CJ, Song JS, Thanki N, Wang J, Yamashita RA, Yang M, Zheng C, Marchler-Bauer A, Thibaud-Nissen F. 2021. RefSeq: expanding the Prokaryotic Genome Annotation Pipeline reach with protein family model curation. *Nucleic Acids Res* 49:D1020–D1028. <https://doi.org/10.1093/nar/gkaa1105>.
 22. Blin K, Shaw S, Kloosterman AM, Charlop-Powers Z, van Wezel GP, Medema MH, Weber T. 2021. antiSMASH 6.0: improving cluster detection and comparison capabilities. *Nucleic Acids Res* 49:W29–W35. <https://doi.org/10.1093/nar/gkab335>.