

Complete Genomes of *E. coli* with Diverse K antigens

Carine Roese Mores

ETH Zurich <https://orcid.org/0000-0001-8622-2461>

Christopher M. Field

University Hospital Basel <https://orcid.org/0000-0002-6434-3745>

Hans-Joachim Ruscheweyh

ETH Zurich <https://orcid.org/0000-0001-7473-6086>

Emma Slack

ETH Zurich <https://orcid.org/0000-0002-2473-1145>

Timothy G. Keys

ETH Zurich <https://orcid.org/0000-0003-2921-0362>

Shinichi Sunagawa

ssunagawa@ethz.ch

ETH Zurich <https://orcid.org/0000-0003-3065-0314>

Data Note

Keywords:

Posted Date: March 12th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-9087730/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

Additional Declarations: The authors declare no competing interests.

Abstract

Escherichia coli is a diverse species with commensal and pathogenic lifestyles. Capsular polysaccharides (K antigens) are prominent surface structures, important virulence factors and targets for immune recognition. The complete genomes of 68 strains with serologically distinct capsule types were sequenced using a combination of PacBio/Oxford Nanopore and Illumina technologies.

ANNOUNCEMENT

Escherichia coli is a common and generally benign colonizer of the human gut. However, some strains can cause serious infections and pose a significant health risk (1, 2). The capsular polysaccharide (CPS), a major surface antigen (K antigen), stands as an important virulence factor of pathogenic strains. Despite this, our understanding of CPS epidemiology in *E. coli* infections is limited due to the labor-intensive and costly nature of *in vitro* serotyping (3). Clinical and diagnostic laboratories have increasingly adopted whole genome sequencing (WGS) to phenotype bacterial isolates *in silico*. However, the scarcity of links between CPS serotyping and genome data limits the inference of *E. coli* capsule types based on WGS. Here, we sequenced the genomes of 68 *E. coli* strains from a K-antigen reference strain collection.

Cryostocked strains were streaked on LB agar plates and incubated at 37°C for 24 h. Single colonies were transferred to LB, incubated at 37°C for 24 h and genomic DNA was extracted from cell pellets using NucleoBond AXG 20 columns with the buffer set III (Macherey-Nagel, Düren, Germany). Concentration, purity, and fragment size were measured using a Qubit fluorometer, a Nanodrop spectrophotometer and a Femto Pulse system, respectively (Thermo Fisher Scientific, USA, Agilent Technologies, USA). For Illumina sequencing, libraries were generated with the Nextera XT kit and sequenced (2 x 150 bp) on the NovaSeq 6000 system. Raw reads were quality-controlled to remove adapter and PhiX sequences, and to trim low-quality bases using bbmap v38.87 (4) with the "trimq" option of 14 and "minlength" of 45. For PacBio sequencing, DNA was sheared to an average size of 7–10 kb using a Megaruptor 3 device (Diagenode) and was not size selected. Libraries were constructed using the SMRTbell Prep Kit 3.0 and sequenced on the Sequel IIe system using the Sequencing Kit 2.0 (one SMRT Cell 8 M, HiFi mode, 15 h). QC was performed during the run using the Control SW Version 11.0.0.144466. Oxford Nanopore Technology (ONT) libraries were prepared using the rapid barcoding kit (SQK-RBK114) and sequenced using a MinION FLO-MIN114 flow cell (MinION Mk1C device, 72 h). Guppy v6.4.6 (5) was used for basecalling, demultiplexing and quality-controlling the ONT reads. Assemblies were performed with Flye v2.9 (6) with the "pacbio-hifi" and "nano-raw" options for PacBio ("Q20+") and ONT ("pass") reads, respectively. Quality-controlled Illumina reads were used to polish all assemblies using Pilon v1.23 (7). All tools were run with default parameters unless otherwise specified.

The genomes exhibited a size range from 4.5 to 5.2 Mb (Table 1) and were annotated with the NCBI PGAP (8). Visualization of assembly graphs and verification of completeness was performed using Bandage (9). Overall, the reported genomes provide unprecedented access to study the molecular basis underlying the diversity of K antigens in *E. coli*.

Declarations

ACKNOWLEDGMENTS

This work was supported by funding from the Basel Research Centre for Child Health (BRCC) to E.S. and S.S., the Swiss National Science Foundation through the project grants 40B2-0_180953 and 310030_185128 to E.S. and the NCCR Microbiomes (51NF40_180575) to E.S. and S.S, and core funding from ETH Zürich. ONT basecalling was performed on the ETH Euler cluster. We thank Gregor Rot for bioinformatics support and acknowledge the IT Service and HPC facilities of ETH Zürich.

Data availability.

The genomes and sequencing data reported in this study have been deposited at the European Nucleotide Archive under the project PRJEB83217 with accession numbers listed in Table 1. All scripts used to generate the reported assemblies are available on GitHub (<https://github.com/carinemores/ecoliCPS>).

References

1. Bonten M, Johnson JR, van den Biggelaar AHJ, Georgalis L, Geurtsen J, De Ibarra P, Gravenstein S, Verstraeten T, Hermans P, Poolman JT Epidemiology of *Escherichia coli* Bacteremia: A Systematic Literature Review <https://doi.org/10.1093/cid/ciaa210>
2. MacKinnon MC, McEwen SA, Pearl DL, Lyytikäinen O, Jacobsson G, Collignon P, Gregson DB, Valiquette L, Laupland KB (2021) Mortality in *Escherichia coli* bloodstream infections: a multinational population-based cohort study. *BMC Infect Dis* 21:1–10
3. Fratamico PM, DebRoy C, Liu Y, Needleman DS, Baranzoni GM, Feng P (2016) Advances in Molecular Serotyping and Subtyping of *Escherichia coli*. *Front Microbiol* 7:644
4. Bushnell B (2014) BBMap: A Fast, Accurate, Splice-Aware Aligner. United States. <https://www.osti.gov/servlets/purl/1241166>
5. Wick RR, Judd LM, Holt KE Performance of neural network basecalling tools for Oxford Nanopore sequencing <https://doi.org/10.1186/s13059-019-1727-y>
6. Kolmogorov M, Yuan J, Lin Y, Pevzner PA Assembly of long, error-prone reads using repeat graphs <https://doi.org/10.1038/s41587-019-0072-8>
7. Walker BJ, Abeel TD, Shea T, Priest M, Abouelliel A (2014) Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. *PLoS ONE* 9:112963

8. 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

9. Wick RR, Schultz MB, Zobel J, Holt KE (2015) Bandage: interactive visualization of de novo genome assemblies. *Bioinformatics* 31(20):3350–3352. 10.1093/bioinformatics/btv383

Table 1

Table 1. Genome assembly statistics and accession numbers

Capsular serotype	Source	No. of short reads	Long read technology	No. of long reads	Long read N50 (bp)	Genome length (bp)	GC content (%)	Genome coverage	ECEs	GenBank accession no.	Short read accession no.	
K1	SSI 85361	8'843'528	PacBio	37'200	9'485	4'961'657	50.71	61	1	GCA_977947085	ERR14050185	I
K2a	CCUG 27	11'751'454	ONT	57'349	25'061	5'125'785	50.54	49	1	GCA_977947715	ERR14050194	I
K2ab	SSI 85362	9'401'764	PacBio	38'443	9'226	4'700'954	50.53	68	1	GCA_977947245	ERR14050196	I
K3	SSI 85363	11'370'150	PacBio	34'889	9'383	5'141'176	50.41	57	2	GCA_977947575	ERR14050206	I
K4	SSI 85364	14'322'598	PacBio	36'172	9'928	4'763'144	50.71	66	1	GCA_977947105	ERR14050210	I
K5	SSI 85365	10'278'192	ONT	88'384	17'421	5'227'725	50.75	41	3	GCA_977947595	ERR14050213	I
K6	SSI 81964	8'440'706	ONT	30'422	21'162	5'167'108	50.42	31	1	GCA_977947035	ERR14050216	I
K7	SSI 81965	11'578'402	PacBio	41'401	9'538	4'736'535	50.63	76	1	GCA_977947585	ERR14050218	I
K8	SSI 85366	10'442'820	PacBio	45'287	9'376	4'750'763	50.92	81	0	GCA_977947505	ERR14050223	I
K9	SSI 85367	7'893'366	ONT	92'184	20'616	4'913'262	50.78	50	0	GCA_977946355	ERR14050225	I
K10	SSI 85368	9'330'748	PacBio	30'281	9'653	4'883'361	50.79	53	2	GCA_977947665	ERR14050229	I
K11	SSI 85369	9'423'246	ONT	88'383	19'801	4'573'209	50.74	53	4	GCA_977947875	ERR14050233	I
K12	SSI 81966	3'639'474	ONT	96'876	15'401	5'286'215	50.45	40	2	GCA_977946375	ERR14050238	I
K13	SSI 81967	11'463'720	PacBio	26'200	9'413	5'078'894	50.46	42	0	GCA_977947965	ERR14050242	I
K14	SSI 85370	9'456'546	PacBio	35'501	9'568	4'581'132	50.71	68	0	GCA_977946545	ERR14050246	I
K16	SSI 85371	8'768'754	PacBio	37'824	9'719	5'007'734	50.49	62	2	GCA_977947945	ERR14050249	I
K17	SSI 85372	8'458'344	PacBio	37'922	9'152	4'600'220	50.8	69	1	GCA_977947915	ERR14050252	I
K18a	SSI 81968	8'888'756	PacBio	35'780	9'325	5'068'856	50.71	57	2	GCA_977946865	ERR14050254	I
K18ab	SSI 85373	486'230	ONT	27'461	15'050	5'069'166	50.72	35	2	GCA_977946875	ERR14050255	I
K19	SSI 85374	8'343'316	PacBio	41'642	9'692	4'880'777	50.7	75	0	GCA_977947905	ERR14050258	I
K20	SSI 85375	10'728'454	PacBio	38'318	9'860	4'991'724	50.86	70	0	GCA_977947155	ERR14050261	I
K22	CCUG 11329	7'852'506	PacBio	38'024	8'664	5'085'183	50.72	56	2	GCA_977947955	ERR14050263	I
K23	SSI 81970	8'639'116	ONT	95'956	21'074	5'125'131	50.42	92	0	GCA_977946625	ERR14050266	I
K24	SSI 81971	11'535'816	ONT	58'864	34'261	5'105'026	50.46	59	0	GCA_977946345	ERR14050269	I
K26	SSI 81972	9'819'104	PacBio	36'939	10'507	4'555'035	50.91	73	1	GCA_977946425	ERR14050272	I
K27	SSI 81973	10'544'150	PacBio	31'751	11'071	4'646'608	50.93	68	0	GCA_977946285	ERR14050275	I
K28	SSI 81974	7'338'250	PacBio	31'365	10'747	4'654'761	51.05	65	0	GCA_977946325	ERR14050278	I

K29	SSI 81975	3'251'212	ONT	23'279	37'588	4'828'683	50.79	39	2	GCA_977946365	ERR14050279	I
K30	SSI 81976	9'417'840	PacBio	37'295	9'715	4'808'187	50.87	64	2	GCA_977946635	ERR14050282	I
K31	SSI 81977	10'874'260	PacBio	32'419	11'105	4'521'692	51.04	71	1	GCA_977947365	ERR14050286	I
K32	SSI 81978	7'818'130	PacBio	31'234	10'974	4'943'816	50.84	62	1	GCA_977946405	ERR14050288	I
K33	SSI 81979	8'403'202	PacBio	30'205	11'803	4'597'013	50.98	67	1	GCA_977946605	ERR14050291	I
K34	SSI 81980	5'957'044	PacBio	29'718	11'398	4'529'990	50.91	67	0	GCA_977946575	ERR14050294	I
K35	SSI 81981	6'432'850	PacBio	30'641	9'039	4'610'049	50.91	52	1	GCA_977946905	ERR14050296	I
K36	SSI 81982	6'876'172	PacBio	42'137	9'185	4'897'268	50.83	72	0	GCA_977946735	ERR14050324	I
K37	SSI 81983	8'387'472	PacBio	39'222	8'648	4'597'789	50.97	68	0	GCA_977946935	ERR14050363	I
K38	SSI 81984	9'360'838	PacBio	29'217	8'653	4'842'863	51.08	47	0	GCA_977946685	ERR14050365	I
K39	SSI 81985	10'294'304	PacBio	37'282	9'274	4'682'936	50.84	67	1	GCA_977946655	ERR14050368	I
K40	SSI 81986	10'043'060	PacBio	33'985	9'822	4'542'529	50.81	66	1	GCA_977946775	ERR14051037	I
K41	SSI 81987	8'383'774	PacBio	37'163	9'549	4'797'978	50.89	67	1	GCA_977946705	ERR14051039	I
K42	SSI 81988	7'951'674	PacBio	37'978	9'526	4'791'211	50.89	67	1	GCA_977946795	ERR14051042	I
K43	SSI 81989	4'254'874	PacBio	37'266	9'959	4'827'059	50.83	69	1	GCA_977946825	ERR14051044	I
K44	SSI 81990	8'643'794	PacBio	28'320	11'510	4'938'150	50.73	56	2	GCA_977947275	ERR14051047	I
K45	SSI 81991	10'978'198	PacBio	32'373	10'729	5'010'117	50.62	62	1	GCA_977947375	ERR14051121	I
K46	SSI 81992	9'280'636	PacBio	32'459	10'445	4'805'835	50.8	63	0	GCA_977946895	ERR14051122	I
K47	SSI 81993	10'409'706	PacBio	33'154	11'599	4'883'445	50.75	69	1	GCA_977947135	ERR14051123	I
K48	SSI 81994	10'165'850	PacBio	23'045	10'255	4'960'040	50.74	41	1	GCA_977946915	ERR14051160	I
K49	SSI 81995	9'169'518	PacBio	31'890	10'362	4'725'875	50.8	63	0	GCA_977946475	ERR14051161	I
K50	SSI 81996	10'168'774	PacBio	31'925	10'689	5'081'262	50.66	59	1	GCA_977947205	ERR14051162	I
K51	SSI 81997	8'834'586	PacBio	33'349	10'048	5'214'915	50.6	56	1	GCA_977947185	ERR14051202	I
K52	SSI 81998	8'566'464	ONT	17'920	22'161	5'093'197	50.81	28	1	GCA_977947195	ERR14051208	I
K53	SSI 81999	5'866'276	PacBio	35'373	9'496	5'110'598	50.55	60	0	GCA_977946495	ERR14051209	I
K54	SSI 82000	11'910'714	PacBio	34'639	9'683	4'597'013	50.89	62	3	GCA_977946555	ERR14051210	I
K55	SSI 82001	10'047'660	PacBio	37'216	9'746	4'661'405	51.02	70	1	GCA_977947805	ERR14051217	I
K83	SSI 82003	9'893'698	PacBio	32'658	9'627	4'957'187	50.78	57	0	GCA_977946385	ERR14051218	I
K84	SSI 82004	10'079'880	PacBio	35'163	10'084	4'949'990	50.74	62	3	GCA_977947815	ERR14051219	I

K87	SSI 82005	8'400'154	PacBio	39'286	10'039	4'943'518	50.79	69	4	GCA_977946395	ERR14051220	I
K92	SSI 82006	10'654'126	PacBio	33'033	10'439	5'156'189	50.58	58	2	GCA_977947695	ERR14051221	I
K93	SSI 85376	7'285'978	PacBio	25'459	9'987	4'906'586	50.55	43	2	GCA_977946505	ERR14051222	I
K94	SSI 85377	9'715'314	PacBio	35'202	10'259	4'886'020	50.71	66	0	GCA_977947735	ERR14051223	I
K95	SSI 85378	13'358'338	PacBio	33'676	11'076	5'220'776	50.3	65	0	GCA_977947145	ERR14051224	I
K96	SSI 85379	6'724'984	PacBio	28'338	11'097	5'063'914	50.63	54	1	GCA_977947645	ERR14051225	I
K97	SSI 85380	6'641'816	ONT	30'822	26'522	5'298'499	50.74	30	5	GCA_977946295	ERR14051226	I
K98	SSI 85381	13'814'536	PacBio	33'350	10'737	4'741'888	50.88	69	0	GCA_977947765	ERR14051229	I
K100	SSI 82007	10'314'626	PacBio	33'594	9'707	5'098'252	50.62	57	1	GCA_977946755	ERR14051230	I
K101	SSI 82008	10'089'768	PacBio	37'418	8'992	4'764'323	50.75	62	2	GCA_977946665	ERR14051231	I
K102	SSI 82009	10'424'902	PacBio	37'305	9'394	4'726'166	50.73	64	3	GCA_977947395	ERR14051246	I
K103	SSI 82010	10'278'752	PacBio	26'680	9'068	4'798'391	50.78	45	0	GCA_977946695	ERR14051248	I