Journal of Microbiological Methods 190 (2021) 106329

Contents lists available at ScienceDirect

Journal of Microbiological Methods

journal homepage: www.elsevier.com/locate/jmicmeth

Note

Comparison of four enzymatic library preparation kits for sequencing Shiga toxin-producing *Escherichia coli* for surveillance and outbreak detection

Jenny Truong, Angela Poates, Yoo Jin Joung [1], Ashley Sabol, Taylor Griswold, Amanda J. Williams-Newkirk, Rebecca Lindsey, Eija Trees [*,2]

Centers for Disease Control and Prevention, 1600 Clifton Road, Mail stop H23-7, Atlanta, GA 30329, USA

ARTICLE INFO

Keywords:
Escherichia
Library preparation
WGS
GC bias
Allele detection

ABSTRACT

Four enzymatic DNA library preparation kits were compared for sequencing Shiga toxin-producing *E. coli*. All kits produced high quality sequence data which performed equally well in the downstream analyses for surveillance and outbreak detection. Important differences were noted in the workflow user-friendliness and per sample cost.

Whole genome sequencing (WGS) library preparation based on enzymatic fragmentation has increased in popularity, particularly in smaller clinical and public health laboratories. Unlike DNA library preparation based on mechanical fragmentation, the enzymatic procedure requires a minimal amount of input DNA and has no need for dedicated equipment for DNA fragmentation (Deurenberg et al., 2017). Several recent reports have indicated that one of the earliest enzymatic library preparation kits, the Illumina Nextera XT, has a substantial GC bias resulting in uneven sequence coverage in extreme AT and GC rich regions which in turn negatively affects the downstream bioinformatic analyses (Sato et al., 2019; Seth-Smith et al., 2019; Li et al., 2020; Uelze et al., 2020; Haendiges et al., 2021). In this study, we compared the performance of four more recent enzymatic library preparation kits by sequencing a diverse set of Shiga toxin-producing *E. coli* (STEC) strains. STEC have an average GC content of 51%, but several genomic regions containing critical genes for strain characterization such as the *rfb* gene cluster encoding the O serogroup and the pathogenicity island containing the type III secretion system have a GC content as low as 30% (Wang and Reeves, 2000; Brown and Finlay, 2011).

Twenty-five STEC strains representing nine different serotypes from past outbreaks and sporadic cases were included in the study (Suppl. Table S1 Tab 1). DNA was extracted using the DNeasy Blood and Tissue kit (Qiagen, Germantown, MD) following the PulseNet standard operating procedure (SOP) PNL33 (https://www.cdc.gov/pulsenet/pathogens/wgs.html). The following four library preparation kits were used: Illumina DNA Prep (formerly Nextera DNA Flex, Illumina, Inc., San Diego, CA), Qiagen QIAseq FX (Qiagen), KAPA HyperPlus (Roche, Indianapolis, IN) and NEBNext Ultra II FS (New England Biolabs, Ipswich, MA). Manufacturers' instructions were followed to obtain the desired fragment size of 500–1000 bp except for modifications outlined in Supplemental Table S2. All four DNA libraries for 25 strains each were sequenced to the minimum average coverage of 40× on the Illumina MiSeq using both 300 and 500 cycle chemistries. All 200 sequences were analyzed using BioNumerics 7.6.3 (Applied Maths, Sint-Martens-Latem, Belgium) following the PulseNet workflow for core genome multi-locus sequence typing (cgMLST) and genotyping as described by Tolar et al. (2019) and Ribot et al. (2019). Sequence reads for one strain, 2015EL-1671a (original coverage 54-93× depending on the library prep/sequencing chemistry combination), were down-sampled to 50×, 40×, 30×, 20× and 10× coverage levels using the Computational Genomics Pipeline (CG-Pipeline; https://github.com/lskatz/CG-Pipeline; (Katz et al., 2011)) in order to determine the minimum coverage level for accurate detection of key genes and relatedness. Sequence coverage was visualized for the *wzx* gene (part of the *rfb* gene cluster and a common target for O antigen determination) of two strains by generating read mappings against reference assemblies (Spades 3.7.1 draft assemblies exported from BioNumerics) for each isolate using bowtie2, followed by coverage determination with Samtools mpileup (www.htslib.org) and

* Corresponding author at: Association of Public Health Laboratories, 8515 Georgia Avenue, Suite 700, Silver Springs, MD 20910, USA.
E-mail address: EHyytia-Trees@cdc.gov (E. Trees).
[1] Present address: Morehouse School of Medicine, 720 Westview Drive SW, Atlanta, GA 30310, USA.
[2] Present address: Association of Public Health Laboratories, 8515 Georgia Avenue, Suite 700, Silver Springs, MD 20910, USA.

https://doi.org/10.1016/j.mimet.2021.106329
Received 26 July 2021; Received in revised form 13 September 2021; Accepted 14 September 2021
Available online 22 September 2021

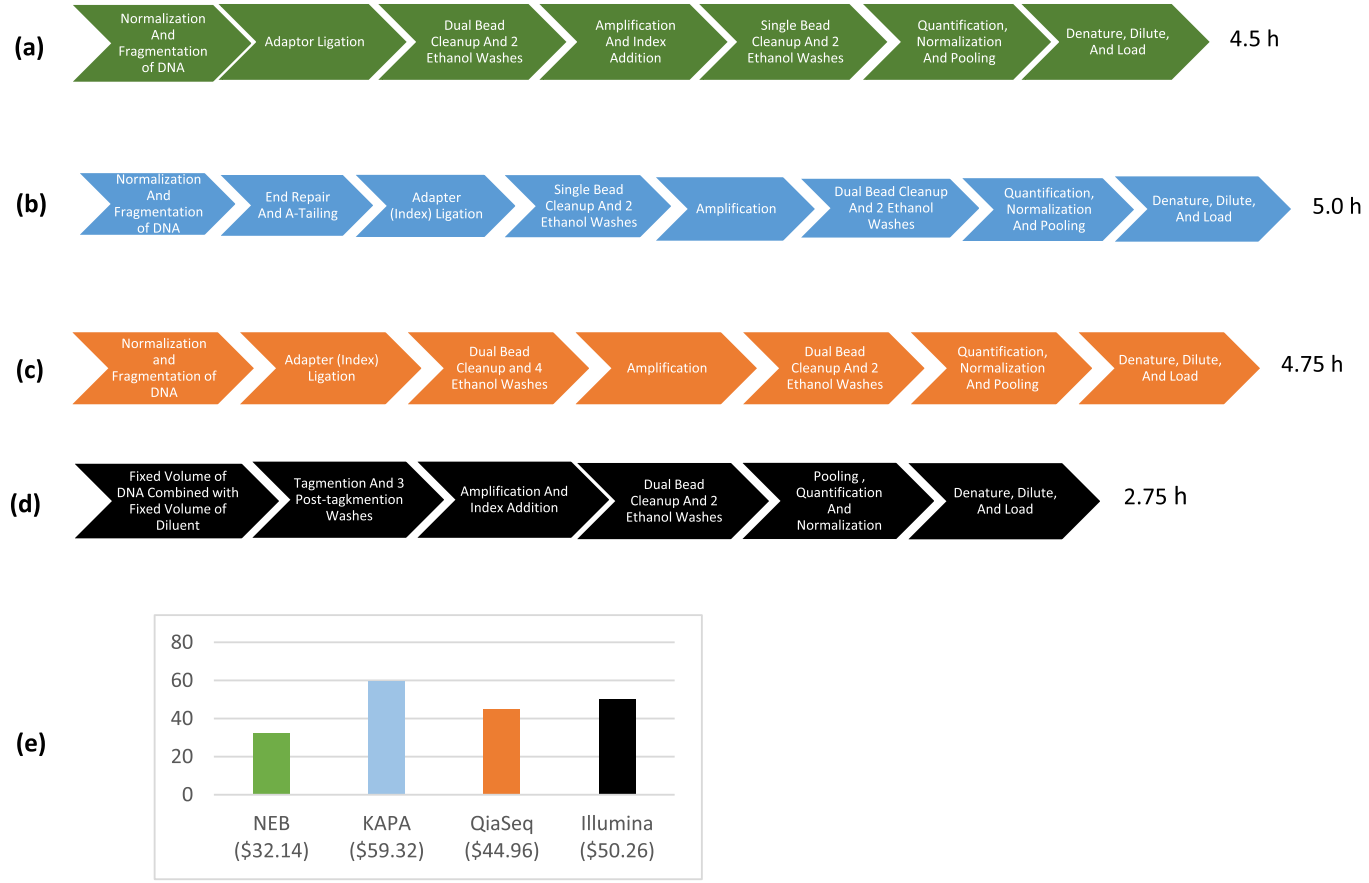

Fig. 1. Comparison of laboratory workflows and per sample cost for the four library preparation kits evaluated. (a) NEB Ultra II FS Workflow: 1 quantification step, 4 thermal cycler (TC) steps, 3 bead steps. Approximately 4.5 h [Hands-on: 3 h 10 min, TC: 80 min]. (b) KAPA HyperPlus Workflow: 1 quantification step, 4 thermal cycler steps, 3 bead steps. Approximately 5 h [Hands-on: 3 h 45 min, TC: 80 min]. (c) Qiagen QIAseq FX Workflow: 1 quantification step, 3 thermal cycler steps, 4 bead steps. Approximately 4.75 h [Hands-on: 3 h 30 min, TC: 80 min]. (d) Illumina DNA Prep workflow: 1 quantification step, 3 thermal cycler steps, 2 bead steps. Approximately 2.75 h [Hands-on: 1 h 45 min, TC: 60 min]. (e) Per sample cost includes the library preparation kit, indexes and beads. Quoted prices are based on the company list prices in US dollars in June 2021 and do not include ancillary reagents, supplies or labor.

plotting of coverage data in R v.4.0.5 (www.r-project.org/ject.org) using ggplot2 v.3.3.3 and egg v.0.4.5. The accession numbers for the raw sequences deposited in the National Center for Biotechnology Information are listed in Supplemental Table S1 Tab 1.

The workflows and the pricing of the four library preparation kits are compared in Fig. 1a–e. The Illumina DNA Prep provided the most streamlined and fastest workflow with fewest processing steps, no normalization of the input DNA concentration and no quantification of the individual libraries resulting in a total processing time (includes hands-on time and thermal cycler steps) of 2.75 h (Fig. 1d). The workflow of the KAPA HyperPlus kit, on the other hand, was most complex and slowest of the four (Fig. 1b). The per sample cost without ancillary supplies and labor (Fig. 1e) was lowest for the NEBNext Ultra II kit at $32 and highest for the Kapa HyperPlus kit at $59. The quality of the raw reads and assemblies is summarized in Table 1. All kits produced raw reads with high quality scores. The average read lengths were close to the maximum for all datasets except for the Illumina DNA Prep 500 cycle sequences where 11 out of 25 sequences had an average read length < 210 bp (Suppl. Table S1 Tab 3). The KAPA HyperPlus and QIAseq FX library preps produced the highest quality assemblies in terms of number of contiqs and N50 while the assembly lengths were comparable across all four kits (Table 1). While all sequences passed the minimum threshold of 95% for percentage of core genome alleles detected, ten out of 25 Illumina DNA Prep 300 cycle sequences had 1–2 percentage points lower values compared to the other kits (Suppl. Table S1 Tab 3). The seven-gene multi-locus sequence type (MLST) was assigned for all sequences except one NEBNext Ultra II 500 cycle sequence which had different alleles detected from assembly-free and assembly-based allele calling for the *icd* gene. The difference between the alleles was a single C to A mutation (Suppl. Table S1 Tab 2). There were no discrepancies detected in the serotype (Suppl. Table S1 Tab 1) or antimicrobial resistance profile (Suppl. Table S1 Tab 5) prediction among the four library preps. The main STEC virulence genes (*stx* subtype, *eaeA, ehxA*) were also correctly detected from all sequences using in-silico PCR (Suppl. Table S1 Tab 1). The blast-based virulence profile prediction missed the *toxB* gene from one KAPA HyperPlus 300 cycle sequence and the *tccP* gene from one NEBNext Ultra II 300 cycle sequence (Suppl. Table S1 Tab 6). Plasmid profile prediction was concordant from all sequences with the exception of one small col-like plasmid missing from one Illumina DNA Prep 300 and 500 sequence each (Suppl. Table S1 Tab 7). Clustering by cgMLST showed no allele differences between the four library preps for twenty strains while four strains exhibited up to one allele difference between the library preps and for one strain the sequences from the different library preps differed by three alleles (Suppl. Table S1 Tab 4). The coverage down-sampling did not reveal significant differences between the kits with a minimum of 30× coverage required for reliable downstream analysis (Suppl. Table S3 Tabs 1–7). The read distribution graphs for the *wzx* gene (Fig. 2a–d) revealed 10× or higher coverage for over 90% of the gene length with the highest overall coverage provided by the NEBNext kit and lowest by the KAPA HyperPlus kit.

The Illumina DNA Prep kit provided the fastest and most user-

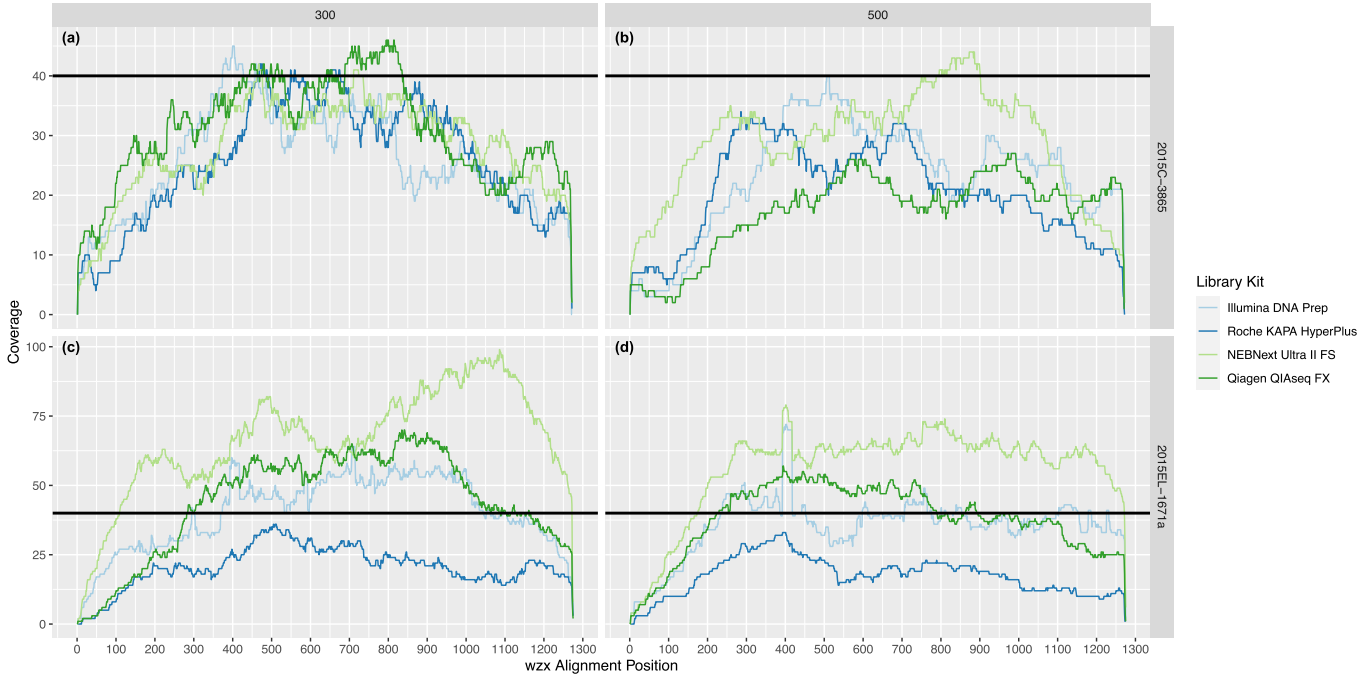

Fig. 2. Depth of coverage graphs for the *wzx* gene for two isolates sequenced using the four library preparation kits evaluated and both 300 and 500 cycle MiSeq chemistries. (a) Isolate 2015C-3865 (O181:H49) sequenced using 300 cycle chemistry (b) Isolate 2015C-3865 sequenced using 500 cycle chemistry (c) Isolate 2015EL-1671a (O130:H11) sequenced using 300 cycle chemistry (d) Isolate 2015EL-1671a sequenced using 500 cycle chemistry. X-axis displays the *wzx* gene position in base pairs and Y-axis the coverage in each nucleotide position. The black horizontal line denotes the 40× minimum average coverage used by PulseNet for *Escherichia.*

Table 1
Average values for key quality control (QC) metrics by library preparation method and MiSeq sequencing chemistry for 25 sequenced samples per library prep method.

Library prep & seq. chemistry	Read length[a]	Read quality[b]	No. of contigs	N50	Assembly length	No. of alleles	% core present	No. of bases N[c]	De novo coverage
FLEX_300	146	36	229	121,486	5,239,838	5776	98.7	605	58
KAPA_300	149	36	193	123,899	5,246,121	5794	99.1	639	61
NEB_300	149	37	218	120,801	5,246,111	5803	99.0	480	61
QIA_300	150	36	200	123,137	5,245,933	5803	99.2	426	62
FLEX_500	221	35	228	138,305	5,302,858	5811	99.4	374	91
KAPA_500	245	35	192	143,117	5,306,647	5795	99.0	877	64
NEB_500	243	35	217	137,128	5,305,692	5817	99.3	265	138
QIA_500	247	35	196	143,502	5,307,218	5812	99.3	377	71

The QC metrics were generated in the BioNumerics 7.6.3 software and averages calculated in Excel.
[a] Read lengths averaged for reads 1 and 2 in base pairs.
[b] Quality scores averaged for reads 1 and 2.
[c] Ambiguous bases in the de novo assembly.

friendly workflow. The only drawback was having to handle and pipet beads throughout the entire procedure which may initially make the kit more challenging for novice users resulting in suboptimal read lengths as was seen particularly with the 500 cycle dataset. The high per sample cost for the Illumina DNA Prep may be offset by less hands-on technician time needed compared to other kits. All four kits generated high quality data that performed equally well in the downstream analyses with a comparable number of alleles and percentage of core genome detected despite the slight differences in the quality of assemblies. The excellent performance can be attributed to the even coverage all four kits generated for genomic regions that are challenging to sequence due to high or low GC content, a finding that was also recently reported by Sato et al. (2019). In fact, only 10× coverage was needed in the coverage downsampling experiment to accurately predict the serotype, which is considerably less than the 40× minimum coverage requirement originally set by PulseNet for STEC sequenced using the Nextera XT kit (Lindsey et al., 2016). The small differences noted in the overall detection of core genome alleles, virulence profile genes and small plasmids do not impact surveillance and outbreak detection since in the PulseNet surveillance a cluster is coded if three or more clinical isolates are within ten alleles and at least two of them are within five alleles (Tolar et al., 2019). We hypothesize that the missed 7-gene MLST type for one NEBNext Ultra II sequence may have been caused by a sequencing error amplified by the unusually high coverage (273×) for that particular sequence as C substitutions are a known error profile for the MiSeq (Schirmer et al., 2015).

In conclusion, Illumina DNA Prep, KAPA HyperPlus, NEBNext Ultra II FS and QIAseq FX all generate high quality sequence data for STEC surveillance with negligible differences in downstream analysis. The choice of the kit comes down to availability, local pricing and individual laboratory workflow considerations such as automation. The findings of this study can also be extrapolated to other surveillance methods, such as high-quality single nucleotide polymorphism (hqSNP) and whole genome MLST (wgMLST). The quality of the data produced by all four kits exceeds the minimum quality requirements set forth by the quality assurance program of the PulseNet and GenomeTrakr networks for

hqSNP analysis (Timme et al., 2020). Additionally, based on the total number of genes detected (5500-6000) per strain, the accessory genome is adequately captured to perform wgMLST (Rasko et al., 2008).

The findings and conclusions in this presentation are those of the authors and do not necessarily represent the views of the Centers for Disease Control and Prevention. Use of trade names is for identification only and does not imply endorsement by the Centers for Disease Control and Prevention or by the U.S. Department of Health and Human Services.

The authors wish to thank Dr. Heather Carleton, Dr. Patricia Fields and Dr. Efrain Ribot for critically reviewing the manuscript.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.mimet.2021.106329.

Funding statement

This work was made possible through support from the Advanced Molecular Detection (AMD) Initiative grand number AMD-21 at the Centers for Disease Control and Prevention. Aside the funding, the Office of AMD did not have any further involvement in the study. This project was supported in part by an appointment to the Research Participation Program at the Centers for Disease Control and Prevention administered by the Oak Ridge Institute for Science and Education through an inter-agency agreement between the U.S. Department of Energy and the Centers for Disease Control and Prevention.

Declaration of interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Brown, N.F., Finlay, B.B., 2011. Potential origins and horizontal transfer of type III secretion systems and effectors. Mob. Genet. Elem. 1, 118–121. https://doi.org/10.4161/mge.1.2.16733.

Deurenberg, R.H., Bathoorn, E., Chlebowicz, M.A., Couto, N., Ferdous, M., García-Cobos, S., Kooistra-Smid, A.M., Raangs, E.C., Rosema, S., Veloo, A.C., Zhou, K., Friedrich, A.W., Rossen, J.W., 2017. Application of next generation sequencing in clinical microbiology and infection prevention. J. Biotechnol. 243, 16–24. https://doi.org/10.1016/j.jbiotec.2016.12.022.

Haendiges, J., Jinneman, K., Gonzalez-Escalona, N., 2021. Choice of library preparation affects sequence quality, genome assembly, and precise in silico prediction of virulence genes in Shiga toxin-producing *Escherichia coli*. PLoS One 16, e0242294, https://doi.org/10.1371/journal.pone.0242294.

Katz, L.S., Humphrey, J.C., Conley, A.B., Nelakuditi, V., Kislyuk, A.O., Agrawal, S., Jayaraman, P., Harcourt, B.H., Olsen-Rasmussen, M.A., Frace, M., Sharma, N.V., Mayer, L.W., Jordan, I.K., 2011. *Neisseria* Base: a comparative genomics database for *Neisseria meningitidis*. Database (Oxford) 2011. https://doi.org/10.1093/database/bar035 bar035.

Li, S., Zhang, S., Deng, X., 2020. GC content-associated sequencing bias caused by library preparation method may infrequently affect *Salmonella* serotype prediction using SeqSero2. Appl. Environ. Microbiol. 86, e00614–e00620. https://doi.org/10.1128/AEM.00614-20.

Lindsey, R.L., Pouseele, H., Chen, J.C., Strockbine, N.A., Carleton, H.A., 2016. Implementation of whole genome sequencing (WGS) for identification and characterization of Shiga toxin-producing *Escherichia coli* (STEC) in the United States. Front. Microbiol. 7, 766. https://doi.org/10.3389/fmicb.2016.00766.

Rasko, D.A., Rosovitz, M.J., Myers, G.S., Mongodin, E.F., Fricke, W.F., Gajer, P., Crabtree, J., Sebaihia, M., Thomson, N.R., Chaudhuri, R., Henderson, I.R., Sperandio, V., Ravel, J., 2008. The pangenome structure of *Escherichia coli:* comparative genomic analysis of *E. coli* commensal and pathogenic isolates. J. Bacteriol. 190, 6881–6893. https://doi.org/10.1128/JB.00619-08.

Ribot, E.M., Freeman, M., Hise, K.B., Gerner-Smidt, P., 2019. PulseNet: entering the age of next-generation sequencing. Foodborne Pathog. Dis. 16, 451–456. https://doi.org/10.1089/fpd.2019.2634.

Sato, M.P., Ogura, Y., Nakamura, K., Nishida, R., Gotoh, Y., Hayashi, M., Hisatsune, J., Sugai, M., Takehiko, I., Hayashi, T., 2019. Comparison of the sequencing bias of currently available library preparation kits for Illumina sequencing of bacterial genomes and metagenomes. DNA Res. 26, 391–398. https://doi.org/10.1093/dnares/dsz017.

Schirmer, M., Ijaz, U.Z., D'Amore, R., Hall, N., Sloan, W.T., Quince, C., 2015. Insight into biases and sequencing errors for amplicon sequencing with the Illumina MiSeq platform. Nucleic Acids Res. 43, e37, https://doi.org/10.1093/nar/gku1341.

Seth-Smith, H.M.B., Bonfiglio, F., Cuenod, A., Reist, J., Egli, A., Wuthrich, D., 2019. Evaluation of rapid library preparation protocols for whole genome sequencing based outbreak investigation. Front. Public Health 7, 241. https://doi.org/10.3389/fpubh.2019.00241.

Timme, R.E., Lafon, P.C., Balkey, M., Adams, J.K., Wagner, D., Carleton, H., Strain, E., Hoffmann, M., Sabol, A., Rand, H., Lindsey, R., Sheehan, D., Baugher, J.D., Trees, E., 2020. Gen-FS coordinated proficiency test data for genomic foodborne pathogen surveillance, 2017 and 2018 exercises. Sci. Data 7, 402. https://doi.org/10.1038/s41597-020-00740-7.

Tolar, B., Joseph, L.A., Schroeder, M.N., Stroika, S., Ribot, E.M., Hise, K.B., Gerner-Smidt, P., 2019. An overview of PulseNet USA databases. Foodborne Pathog. Dis. 16, 457–462. https://doi.org/10.1089/fpd.2019.2637.

Uelze, L., Borowiak, M., Deneke, C., Szabó, I., Fischer, J., Tausch, S.H., Malorny, B., 2020. Performance and accuracy of four open-source tools for in silico serotyping of *Salmonella* spp. based on whole-genome short-read sequencing data. Appl. Environ. Microbiol. 86 https://doi.org/10.1128/AEM.02265-19 e02265–19.

Wang, L., Reeves, P.R., 2000. The *Escherichia coli* O111 and *Salmonella enterica* O35 gene clusters: gene clusters encoding the same colitose-containing O antigen are highly conserved. J. Bacteriol. 182, 5256–5261. https://doi.org/10.1128/JB.182.18.5256-5261.2000.