AMERICAN SOCIETY FOR MICROBIOLOGY | Journal of Clinical Microbiology®

BACTERIOLOGY

Development of a Web Tool for *Escherichia coli* Subtyping Based on *fimH* Alleles

Louise Roer,[a] Veronika Tchesnokova,[b] Rosa Allesøe,[c] Mariya Muradova,[b] Sujay Chattopadhyay,[b] Johanne Ahrenfeldt,[c] Martin C. F. Thomsen,[c] Ole Lund,[c] Frank Hansen,[a] Anette M. Hammerum,[a] Evgeni Sokurenko,[b] Henrik Hasman[a]

Department of Bacteria, Parasites and Fungi, Statens Serum Institut, Copenhagen, Denmark[a]; Department of Microbiology, University of Washington, Seattle, Washington, USA[b]; Department of Systems Biology, Technical University of Denmark, Kongens Lyngby, Denmark[c]

ABSTRACT The aim of this study was to construct a valid publicly available method for *in silico fimH* subtyping of *Escherichia coli* particularly suitable for differentiation of fine-resolution subgroups within clonal groups defined by standard multilocus sequence typing (MLST). FimTyper was constructed as a FASTA database containing all currently known *fimH* alleles. The software source code is publicly available at https://bitbucket.org/genomicepidemiology/fimtyper, the database is freely available at https://bitbucket.org/genomicepidemiology/fimtyper_db, and a service implementing the software is available at https://cge.cbs.dtu.dk/services/FimTyper. FimTyper was validated on three data sets: one containing Sanger sequences of *fimH* alleles of 42 *E. coli* isolates generated prior to the current study (data set 1), one containing whole-genome sequence (WGS) data of 243 third-generation-cephalosporin-resistant *E. coli* isolates (data set 2), and one containing a randomly chosen subset of 40 *E. coli* isolates from data set 2 that were subjected to conventional *fimH* subtyping (data set 3). The combination of the three data sets enabled an evaluation and comparison of FimTyper on both Sanger sequences and WGS data. FimTyper correctly predicted all 42 *fimH* subtypes from the Sanger sequences from data set 1 and successfully analyzed all 243 draft genomes from data set 2. FimTyper subtyping of the Sanger sequences and WGS data from data set 3 were in complete agreement. Additionally, *fimH* subtyping was evaluated on a phylogenetic network of 122 sequence type 131 (ST131) *E. coli* isolates. There was perfect concordance between the typology and *fimH*-based subclones within ST131, with accurate identification of the pandemic multidrug-resistant clonal subgroup ST131-*H*30. FimTyper provides a standardized tool, as a rapid alternative to conventional *fimH* subtyping, highly suitable for surveillance and outbreak detection.

KEYWORDS *fimH*, *Escherichia coli*, typing, whole-genome sequencing analysis

Received 5 May 2017 **Returned for modification** 29 May 2017 **Accepted** 30 May 2017

Accepted manuscript posted online 7 June 2017

Citation Roer L, Tchesnokova V, Allesøe R, Muradova M, Chattopadhyay S, Ahrenfeldt J, Thomsen MCF, Lund O, Hansen F, Hammerum AM, Sokurenko E, Hasman H. 2017. Development of a Web tool for *Escherichia coli* subtyping based on *fimH* alleles. J Clin Microbiol 55:2538–2543. https://doi.org/10.1128/JCM.00737-17.

Editor Daniel J. Diekema, University of Iowa College of Medicine

Address correspondence to Louise Roer, loro@ssi.dk.

L.R. and V.T. contributed equally to this article.

The *fimH* gene is part of the *fim* operon, which encodes a surface organelle named type 1 fimbriae found in most *Escherichia coli* strains (1). The FimH protein is located at the tip of the fimbrial structure and serves as a D-mannose-specific adhesin, which aids in immobilizing the bacterium on both biotic and abiotic surfaces (2, 3). Studies have shown only minor sequence variation within the *fimH* genes, which renders the *fimH* alleles feasible for use in high-resolution subtyping of multilocus sequence typing (MLST)-based *E. coli* clonal groups. The applicability of *fimH* subtyping has been shown to be particularly relevant within the highly virulent sequence type 131 (ST131) clonal group, where the resistant and multiresistant *H*30 subgroups carrying the *fimH30* allele have been identified (4, 5). As ST131 *E. coli* is the most dominant human-pathogenic clonal group being reported in relation to bloodstream infections, the need to perform *fimH* subtyping is undisputed. Traditionally, typing of *fimH* alleles has been obtained

TABLE 1 *fimH* subtype prediction by conventional typing versus FimTyper

Data set no.	No. of samples[a]			Concordance between conventional and FimTyper results (%)[b]
	Total	Positive by conventional typing	Detected by FimTyper	
1	42	42	42	100
2	243	ND	230	NA
3	40	37	37	100

[a]ND, not determined.
[b]NA, not available.

through PCR amplification of the approximately 900-bp *fimH* gene, followed by a single Sanger sequencing run and alignment of the 489-nucleotide (nt) typing region to an *fimH* allele database containing the currently known *fimH* typing variants or alleles. This typing could be performed rapidly and easily on whole-genome sequencing (WGS) data; thus a need to develop a solution to handle WGS data in relation to *fimH* typing of especially pathogenic *E. coli* has emerged. The aim of the present study was construction and validation of a Web tool which would enable the user to obtain *fimH* allelic information from either simple Sanger-generated sequences or raw as well as assembled WGS data.

(Part of this work was presented previously at the 11th International Meeting on Microbial Epidemiological Markers [IMMEM XI], Estoril, Portugal, 12 March 2016.)

RESULTS AND DISCUSSION

Construction of FimTyper. FimTyper was constructed to perform *fimH* subtyping using sequencing data originating from PCR and subsequent Sanger sequencing (assembled and saved in FASTA format), raw reads obtained directly from sequencing platforms such as Illumina, Ion Torrent, or Roche 454, or *de novo*-assembled draft (or complete) genomes. The FimTyper tool contains all currently known *fimH* alleles and is a BLAST-based publicly available Web-based service hosted by the Center for Genomic Epidemiology (CGE [https://cge.cbs.dtu.dk/services/FimTyper]). The default settings for FimTyper were set to a minimum identity (ID) of 95% and minimum length of 60% compared to the reference sequence to avoid noise from, e.g., gene fragments; however, FimTyper allows the user to specify similarity from 55% to 100% identity. The best-matching hit from the database was given as output, including the percent identity (%ID) between the hit in the genome and in the database and the length of the hit compared to the database record of the *fimH* allele. Additionally, the contig in which the hit was found, followed by the position in the contig, and the accession number of the *fimH* allele were reported. A detailed description of the output of FimTyper can be found at the CGE website.

Using FimTyper on Sanger sequences from PCR products. To evaluate the performance of the FimTyper Web tool versus conventional typing, multiple analysis strategies were employed. Initially, the tool was evaluated on preassembled pairs of Sanger sequences from a data set consisting of 42 samples (data set 1) that had already been subtyped in relation to their *fimH* alleles by conventional typing methods prior to the current study. The Sanger sequences covered 13 different variants of *fimH* subtypes (Table 1). The FimTyper identified *fimH* subtypes from all 42 assembled Sanger sequences correctly at a 100% identity match.

Thus, an excellent concordance between conventional typing and the FimTyper tool was found, suggesting an equally good performance for the FimTyper tool as for conventional typing in analyzing preassembled Sanger sequences uploaded as FASTA files.

Using FimTyper on whole-genome sequencing data. The FimTyper Web tool successfully analyzed all 243 draft genomes of third-generation-cephalosporin-resistant *E. coli* isolates (data set 2). FimTyper was able to identify an *fimH* allele in 230 of the 243 draft genome data sets. The 13 *fimH*-negative isolates were further verified as negative by BLAST against the complete *fim* operon, including part of its flanking regions (9,754

TABLE 2 Distribution of *fimH* subtypes identified among 243 draft genomes of *Escherichia coli* isolates using the FimTyper Web tool

fimH subtype	No. of isolates with the subtype
fimH30	98
fimH27	42
fimH5	18
fimH41	15
fimH54	9
fimH24	5
fimH106	4
fimH29	4
fimH2	3
fimH35	3
fimH65	3
fimH31	3
fimH38	2
fimH64	2
fimH34	2
fimH22	1
fimH517[a]	1
fimH103	1
fimH142	1
fimH63	1
fimH25	1
fimH32	1
fimH39	1
fimH58	1
fimH60	1
fimH43	1
fimH215	1
fimH445	1
fimH97	1
fimH483	1
fimH10	1
fimH15	1
fimH negative	13

[a]New *fimH* subtype identified by FimTyper.

nt in total) from *E. coli* K-12 MG1655. All 13 *fimH*-negative isolates showed BLAST hits to the upstream and downstream regions of the *fim* operon but no hits to any of the genes of the *fim* operon, including *fimH*, suggesting that these isolates were missing not only the *fimH* gene but the complete *fim* operon.

Among the 492 *fimH* alleles in the FimTyper database, 32 different alleles were found to match the sequences of the 243 draft genomes, including one new allele (Table 2). The most abundant hits were to the *fimH30* allele ($n = 98$), the *fimH27* allele ($n = 42$), the *fimH5* allele ($n = 17$), and the *fim41* allele ($n = 15$). The new allele was assigned number 517 (*fimH517*) and added to the database.

Among the 40 randomly chosen isolates from data set 2 that were additionally subjected to conventional *fimH* typing (comprising data set 3), three of the samples did not yield any PCR products; these results were in agreement with the FimTyper results on whole-genome sequence data, where *fimH*-negative results were predicted for the same three samples. For the remaining 37 samples, the conventional typing using DNA alignment and the FimTyper predictions using both assembled Sanger sequences and whole-genome sequence data were in 100% agreement.

MLST versus *fimH* subtype. Subtyping of *fimH* is especially relevant for the major *E. coli* clonal group ST131. Therefore, the 122 *E. coli* isolates from data set 2 previously predicted by Roer et al. (6) to belong to ST131 by the Achtman MLST scheme (7) were further analyzed in relation to their *fimH* subtype. All 122 ST131 *E. coli* isolates harbored an *fimH* allele, with *fimH30* being the most frequent ($n = 95$, 78%) and representing the pandemic multidrug-resistant clonal group ST131-*H*30, followed by *fimH27* ($n = 14$, 11%), *fimH41* ($n = 11$, 9%), *fimH22* ($n = 1$, <1%), and *fimH35* ($n = 1$, <1%). In a study by Johnson et al. (8), the same five *fimH* alleles were among the seven *fimH* subtypes

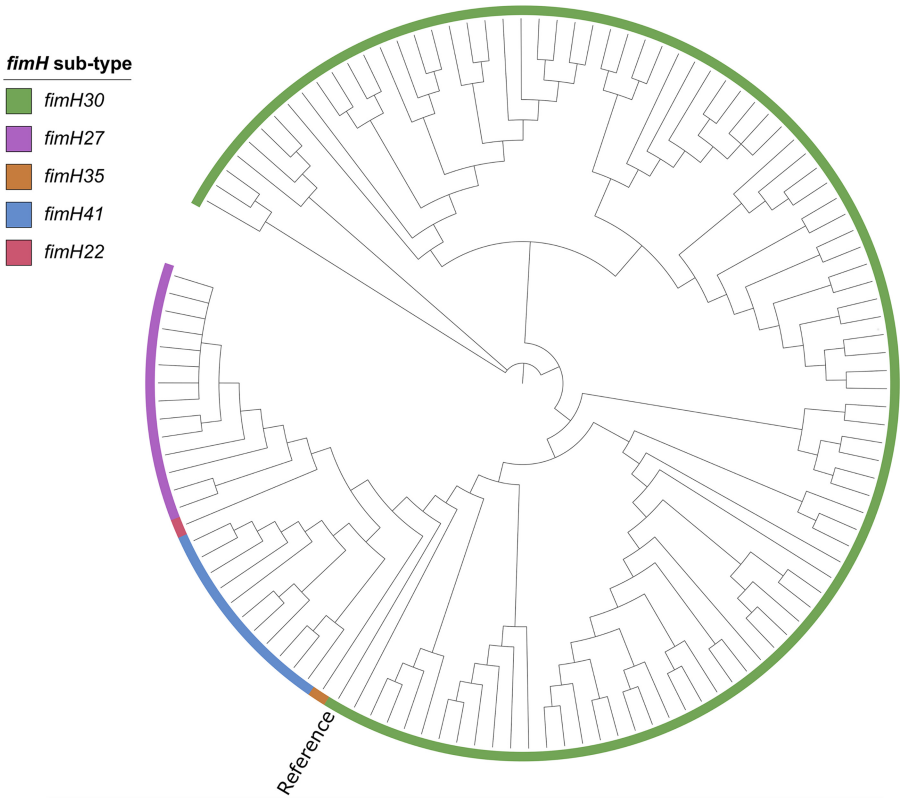

FIG 1 SNP-based phylogeny of the 122 ST131 *Escherichia coli* isolates. Phylogenetic reconstruction of the 122 ST131 *E. coli* isolates was performed with *E. coli* JJ1886 as the reference genome. The tree was constructed from 13,155 SNPs, and results are represented as a cladogram. The *fimH* subtype is marked at the branch tip for each isolate according to the color legend on the figure.

found in a collection of 352 historical and recent ST131 *E. coli* isolates, subtyped by conventional typing. Two infrequent subtypes found by Johnson et al., the *fimH15* (1/352) and *fimH94* (1/352) alleles, were not found among the 122 ST131 *E. coli* isolates tested in the current study.

In addition to the analysis above, the phylogenetic relationship between the 122 ST131 *E. coli* isolates was constructed from single nucleotide polymorphism (SNP) analysis and compared to the *fimH* subtypes, as depicted in Fig. 1. From this analysis, a clear overlap between the structure of the phylogenetic relationship and the *fimH* subtype was observed. All *fimH30* isolates clustered together in a distinct ST131-*H*30 clade, as did the *fimH41* isolates and the *fimH27* isolates. The two single isolates with *fimH35* and *fimH22* clustered in between the clades of the other *fimH* subtypes. Miyoshi-Akiyama et al. reported a similar correlation between SNP-based phylogeny and *fimH* subtype for a collection of global ST131 *E. coli* isolates (9). However, in their study, two *fimH30* isolates clustered within the distinct *fimH41* clade, whereas a clear grouping was observed in our study. This difference could be a result of mistyping of the two *fimH30* isolates or of the differences in the methods used for calling SNPs, reconstructing the phylogenetic tree, and choice of genome reference. We did not have access to either the data or the custom script for SNP concatemers used by Miyoshi-Akiyama et al.; however, to investigate the possible differences caused by the reference, a new phylogenetic reconstruction was created with *E. coli* SE15 used by Miyoshi-Akiyama et al. (GenBank accession no. NC_013654.1) as a reference (data not shown). The reference was subtyped as an *fimH41* isolate and clustered together with all our *fimH41* isolates in the phylogenetic reconstruction. The overall topology of the tree once again clustered according to their *fimH* subtypes, eliminating the choice of reference as a parameter for differences between the two studies.

However, both studies illustrate the high diversity within the ST131 clonal clade and underline the benefit of, including *fimH* analysis as a fast tool to subtype beyond the level of MLST.

In the present study, a Web tool to identify *fimH* alleles from either simple Sanger-generated sequences or raw or assembled WGS data from *E. coli* genomes has been developed, thus enabling researchers and primary investigators to rapidly detect the *fimH* allele in their data sets. The software source code for the tool is publicly available at https://bitbucket.org/genomicepidemiology/fimtyper, and the database hosted by the Center for Genomic Epidemiology (CGE) is freely available at https://bitbucket.org/genomicepidemiology/fimtyper_db. A publicly available Web service implementing the software can be found at https://cge.cbs.dtu.dk/services/FimTyper.

MATERIALS AND METHODS

Development of a Web tool for *fimH* subtyping. An *fimH* allele database was created to contain all previously identified *fimH* allele variants ($n = 492$) collected at the State University of New York and used for conventional typing. The database was constructed as a single FASTA file and implemented in a BLAST-based PERL script, originally developed by Zankari et al. for *in silico* detection of acquired resistance genes (10). The default setting for minimum percent identity and minimum length of a hit to be reported by BLAST was chosen as 95% and 60%, respectively, to reduce false-positive hits caused by reporting of small fragments unrelated to the *fimH* gene. Perfect identity hits (%ID of 100) reports the corresponding *fimH* allele whereas nonperfect hits (100 > %ID >95) report an "unknown or presumptive new variant," and the user is encouraged to contact the curator of FimTyper for updating the database with this new variant.

The new stand-alone Web tool, called FimTyper, has been made publicly available as a component of the CGE Web tools (http://cge.cbs.dtu.dk/services/).

Data sets for validation. To validate the FimTyper Web tool, two different data sets and a subset of one of these were used, covering a total of 32 *fimH* subtypes. Data set 1 was comprised of paired Sanger sequences of 42 *E. coli* isolates, where the *fimH* allele variants had previously been determined by the conventional typing method. The data set covered 13 different *fimH* subtypes. Data set 2 comprised draft genomes obtained from whole-genome sequencing using a 250-bp paired-end Illumina data set of 243 third-generation-cephalosporin-resistant *E. coli* isolates originating from blood infections and submitted to Statens Serum Institut in 2014 as part of the surveillance of third-generation-cephalosporin-resistant *E. coli* (6). These 243 *E. coli* isolates covered 49 different STs, of which 122 isolates belonged to ST131 (11). Data set 3 was comprised of a randomly chosen subset from data set 2 of 40 *E. coli* isolates belonging to 28 different STs and covering 29 different *fimH* subtypes and an *fimH*-negative fraction.

Conventional *fimH* subtyping. The 40 *E. coli* isolates of data set 3 were subjected to conventional *fimH* subtyping, performed as previously described (12). Briefly, *fimH* PCR amplification was conducted using a Qiagen Multiplex PCR kit (Qiagen, Aarhus, Denmark) with the following two *fimH* primers: fimH-F, CACTCAGGGAACCATTCAGGCA (binds 50 to 72 nucleotides upstream of the *fimH* start), and fimH-R, CTTATTGATAAACAAAAGTCAC (spans the last 21 nucleotides of *fimH*). The thermocycler program for the PCRs consisted of 1 cycle of 94°C for 5 min for heat activation, followed by 30 cycles of 94°C for 30 s (denaturation), 57°C for 90 s (annealing), and 72°C for 60 s (extension), with 1 cycle of 72°C for 60 s as a final extension. The resulting PCR products were applied on an Illustra ExoProStar 1-Step kit (GE Healthcare), a BigDye Terminator, version 3.1, cycle sequencing kit (Thermo Fisher), and a BigDye XTerminator purification kit (Thermo Fisher), followed by sequencing with an Applied Biosystems 3130 XL genetic analyzer (Thermo Fisher). Contigs were assembled based on the paired chromatograms using CLC Genomics Workbench, version 9.5.1 (Qiagen).

Validation of FimTyper. Individual FASTA assemblies of the paired Sanger sequences of data set 1 from the 42 isolates that had been subjected to conventionally *fimH* subtyping prior to the current study were analyzed with the newly developed FimTyper Web tool presented in this study (https://cge.cbs.dtu.dk/services/FimTyper-1.0/). The output results were compared with the results previously obtained by conventional typing using manual alignment analysis toward the *fimH* database.

Draft genome sequences of the 243 third-generation-cephalosporin-resistant *E. coli* isolates from data set 2 were analyzed directly using the FimTyper Web tool. In situations where FimTyper did not report BLAST hits with an identity of >95%, the draft genome sequences were additionally analyzed by BLAST against the complete *fim* operon of *E. coli* K-12 MG1655 (GenBank accession no. U00096, bases 4540457 to 4550210) including flanking regions with 500 bp upstream of *fimB* and 500 bp downstream of *fimH*, to confirm the absence of one of more *fim*-related genes. Finally, as the *fimH* subtypes of data set 2 had not been examined previously by conventional *fimH* subtyping, 40 randomly chosen isolates (data set 3) were subjected to conventional typing with PCR and Sanger sequencing and analyzed manually by multiple-alignment analysis with the known *fimH* sequences. The results were evaluated and compared to the results from FimTyper on Sanger sequences and whole-genome sequence data.

Clonal variation within ST131 analyzed by SNP analysis. SNP variants were called using NASP, version 1.0 (http://biorxiv.org/content/early/2016/01/25/037267), by aligning whole-genome sequence data from the 122 ST131 *E. coli* isolates against the chromosome of JJ1886 (GenBank accession no. NC_022648.1) using the Burrows-Wheeler aligner (BWA) after removal of duplicated regions in the reference using NUCmer. Variants were identified using the GATK Unified Genotyper, and SNPs that did

not pass a minimum coverage of 10 or SNPs that were not present in a minimum of 90% of the base calls were excluded. Phylogenetic analyses of the identified SNPs were performed by maximum-likelihood approximation with the generalized time-reversible model in FastTree, version 2.1.5 (13).

ACKNOWLEDGMENTS

We thank Karin Sixhøj Pedersen for excellent laboratory assistance.

This work was partly supported by the Danish Ministry of Health and Prevention as part of the Integrated Surveillance of ESBL/AmpC-producing *E. coli* and Carbapenemase-Producing Bacteria and partly by the Center for Genomic Epidemiology (www.genomicepidemiology.org).

We have no conflicts of interest to declare.

REFERENCES

1. Klemm P, Christiansen G. 1987. Three *fim* genes required for the regulation of length and mediation of adhesion of *Escherichia coli* type 1 fimbriae. Mol Gen Genet 208:439–445. https://doi.org/10.1007/BF00328136.
2. Bhomkar P, Materi W, Semenchenko V, Wishart DS. 2010. Transcriptional response of *E. coli* upon FimH-mediated fimbrial adhesion. Gene Regul Syst Bio 4:1–17. https://doi.org/10.4137/GRSB.S4525.
3. Cookson AL, Cooley WA, Woodward MJ. 2002. The role of type 1 and curli fimbriae of Shiga toxin-producing *Escherichia coli* in adherence to abiotic surfaces. Int J Med Microbiol 292:195–205. https://doi.org/10.1078/1438-4221-00203.
4. O'Hara JA, Hu F, Ahn C, Nelson J, Rivera JI, Pasculle AW, Doi Y. 2014. Molecular epidemiology of KPC-producing *Escherichia coli*: occurrence of ST131-*fimH30* subclone harboring pKpQIL-like IncFIIk plasmid. Antimicrob Agents Chemother 58:4234–4237. https://doi.org/10.1128/AAC.02182-13.
5. Rogers BA, Ingram PR, Runnegar N, Pitman MC, Freeman JT, Athan E, Havers S, Sidjabat HE, Gunning E, De Almeida M, Styles K, Paterson DL, ASID CRN. 2015. Sequence type 131 *fimH30* and *fimH41* subclones amongst *Escherichia coli* isolates in Australia and New Zealand. Int J Antimicrob Agents 45:351–358. https://doi.org/10.1016/j.ijantimicag.2014.11.015.
6. Roer L, Hansen F, Thomsen MCF, Knudsen JD, Hansen DS, Wang M, Samulionienė J, Justesen US, Røder BL, Schumacher H, Østergaard C, Andersen LP, Dzajic E, Søndergaard TS, Stegger M, Hammerum AM, Hasman H. 22 March 2017. WGS-based surveillance of third-generation cephalosporin-resistant *Escherichia coli* from bloodstream infections in Denmark. J Antimicrob Chemother https://doi.org/10.1093/jac/dkx092.
7. Wirth T, Falush D, Lan R, Colles F, Mensa P, Wieler LH, Karch H, Reeves PR, Maiden MCJ, Ochman H, Achtman M. 2006. Sex and virulence in *Escherichia coli*: an evolutionary perspective. Mol Microbiol 60:1136–1151. https://doi.org/10.1111/j.1365-2958.2006.05172.x.
8. Johnson JR, Tchesnokova V, Johnston B, Clabots C, Roberts PL, Billig M, Riddell K, Rogers P, Qin X, Butler-Wu S, Price LB, Aziz M, Nicolas-Chanoine M-H, Debroy C, Robicsek A, Hansen G, Urban C, Platell J, Trott DJ, Zhanel G, Weissman SJ, Cookson BT, Fang FC, Limaye AP, Scholes D, Chattopadhyay S, Hooper DC, Sokurenko EV. 2013. Abrupt emergence of a single dominant multidrug-resistant strain of *Escherichia coli*. J Infect Dis 207:919–928. https://doi.org/10.1093/infdis/jis933.
9. Miyoshi-Akiyama T, Sherchan JB, Doi Y, Nagamatsu M, Sherchand JB, Tandukar S, Ohmagari N, Kirikae T, Ohara H, Hayakawa K. 2016. Comparative genome analysis of extended-spectrum-β-lactamase-producing *Escherichia coli* sequence type 131 strains from Nepal and Japan. mSphere 1:e00289-16. https://doi.org/10.1128/mSphere.00289-16.
10. Zankari E, Hasman H, Cosentino S, Vestergaard M, Rasmussen S, Lund O, Aarestrup FM, Larsen MV. 2012. Identification of acquired antimicrobial resistance genes. J Antimicrob Chemother 67:2640–2644. https://doi.org/10.1093/jac/dks261.
11. Bager F, Birk T, Høg BB, Jensen LB, Jensen AN, Knegt L de, Korsgaard H, Dalby T, Hammerum AM, Hoffmann S, Kuhn KG, Larsen AR, Laursen M, Nielsen EM, Olsen SS, Petersen A, Sönksen UW. 2014. DANMAP 2014—use of antimicrobial agents and occurrence of antimicrobial resistance in bacteria from food animals, food and humans in Denmark. Danish Integrated Antimicrobial Resistance Monitoring and Research Programme, Copenhagen, Denmark. http://www.danmap.org/~/media/projekt%20sites/danmap/danmap%20reports/danmap%202014/danmap_2014.ashx.
12. Weissman SJ, Johnson JR, Tchesnokova V, Billig M, Dykhuizen D, Riddell K, Rogers P, Qin X, Butler-Wu S, Cookson BT, Fang FC, Scholes D, Chattopadhyay S, Sokurenko E. 2012. High-resolution two-locus clonal typing of extraintestinal pathogenic *Escherichia coli*. Appl Environ Microbiol 78:1353–1360. https://doi.org/10.1128/AEM.06663-11.
13. Price MN, Dehal PS, Arkin AP. 2010. FastTree 2: approximately maximum-likelihood trees for large alignments. PLoS One 5:e9490. https://doi.org/10.1371/journal.pone.0009490.