Development of a web tool for Escherichia coli subtyping based on fimH alleles

Running title: Development of E. coli fimH sub-typing web-tool

Roer, Louise; Tchesnokova, Veronika; Allesøe, Rosa Lundbye; Muradova, Mariya; Chattopadhyay, Sujay; Ahrenfeldt, Johanne; Thomsen, Martin Christen Frølund; Lund, Ole; Hansen, Frank; Hammerum, Anette Marie

Total number of authors:
12

Published in:
Journal of Clinical Microbiology

Link to article, DOI:
10.1128/JCM.00737-17

Publication date:
2017

Document Version
Peer reviewed version

JCM Accepted Manuscript Posted Online 7 June 2017
J. Clin. Microbiol. doi:10.1128/JCM.00737-17

Development of a web tool for *Escherichia coli* sub-typing based on *fimH* alleles

Running title: Development of *E. coli fimH* sub-typing web-tool

Louise ROER[1#], Veronika TCHESNOKOVA[2], Rosa ALLESØE[3], Mariya MURADOVA[2], Sujay CHATTOPADHYAY[2], Johanne AHRENFELDT[3], Martin C. F. THOMSEN[3], Ole LUND[3], Frank HANSEN[1], Anette M. HAMMERUM[1], Evgeni SOKURENKO[2] and Henrik HASMAN[1]

[1]Department of Bacteria, Parasites and Fungi, Statens Serum Institut, Copenhagen, Denmark

[2]Department of Microbiology, University of Washington, Seattle, USA

[3]Department of Systems Biology, Technical University of Denmark, Kongens Lyngby, Denmark

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

LR and VT contributed equally to this work.

Keywords: *fimH*, *E. coli*, typing, whole genome sequencing analysis

Abstract

The aim of this study was to construct a valid publicly available method for *in silico fimH* sub-typing of *Escherichia coli* particularly suitable for differentiation of fine-resolution subgroups within clonal groups defined by standard multi-locus sequence typing (MLST). FimTyper was constructed as a FASTA database containing all currently known *fimH* alleles. The software source code is publicly available on https://bitbucket.org/genomicepidemiology/fimtyper, the database freely available at https://bitbucket.org/genomicepidemiology/fimtyper_db, and a service implementing the software available at https://cge.cbs.dtu.dk/services/FimTyper.

FimTyper was validated on three datasets; (i) containing Sanger sequences of *fimH* alleles of 42 *E. coli* isolates generated prior to the current study, (ii) whole-genome sequence data of 243 third-generation cephalosporins-resistant *E. coli* isolates, and (iii) a randomly chosen subset of 40 *E. coli* isolates from dataset (ii), which were subjected to conventional *fimH* sub-typing. The combination of the three datasets enabled an evaluation and comparison of FimTyper on both Sanger sequences and WGS data. FimTyper correctly predicted all 40 *fimH* sub-types from the Sanger sequences from dataset (i), and successfully analyzed all 243 drafted genomes from dataset (ii). FimTyper sub-typing of the Sanger sequences and WGS data from dataset (iii) were in complete agreement. Additionally, *fimH* sub-typing was evaluated on a phylogenetic network of 122 ST131 *E. coli* isolates. There were perfect concordance between the typology and *fimH*–based sub-clones 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* sub-typing, highly suitable for surveillance and outbreak detection.

Journal of Clinical Microbiology

Introduction

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 to immobilize 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 to be used for high-resolution sub-typing of MLST-based *E. coli* clonal group. The applicability of *fimH* sub-typing has shown to been particularly relevant within the highly virulent ST131 clonal group, where resistant and multi-resistant *H*30 sub-group 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* sub-typing is undisputed. Traditionally, typing of *fimH* alleles have been obtained through PCR amplification of the approximately 900-bp *fimH* gene followed by a single Sanger sequencing run and alignment of the 489-nucleotide typing region to a *fimH* allele database containing the currently known *fimH* typing variants or alleles. This typing could be performed rapidly and easily on 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 enables the user to obtain *fimH* allelic information from either simple Sanger generated sequences or from raw as well as assembled WGS data.

Materials and Methods

Development of a web tools for *fimH* sub-typing

A *fimH* allele database was created to contain all previously identified *fimH* allele variants (n=492) collected at State University of New York and used for conventional typing. The database was

constructed as a single FASTA-file, and implemented into a BLAST-based PERL script, originally developed by Zankari *et al*. for *in silico* detection of acquired resistance genes (6). The default setting for minimum %ID 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 = 100) reports the corresponding *fimH* allele where as non-perfect hits (100 > %ID > 95) are reported as "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 datasets and a subset hereof were used, covering a total of 32 *fimH* subtypes: (i) Paired Sanger sequences of 42 *E. coli* isolates, where the *fimH* allele variants had previously been determined by the conventional typing method. The dataset covered 13 different *fimH* subtypes. (ii) a dataset of draft genomes obtained from whole genome sequencing using 250 bp paired-end Illumina data of 243 third-generation cephalosporins-resistant *E. coli* isolates originating from blood infections and submitted to Statens Serum Institut in 2014 as part of Surveillance of third-generation cephalosporins-resistant *E. coli* (7). These 243 *E. coli* isolates covered 49 different STs, of which 122 isolates belonged to ST131 (8). (iii) A randomly chosen subset from dataset (ii) of 40 *E. coli* isolates belonging to 28 different STs, and covering 29 different *fimH* subtypes and a *fimH* negative fraction.

Conventional *fimH* sub-typing

Journal of Clinical Microbiology

The 40 *E. coli* isolates from subset (iii), was subjected to conventional *fimH* sub-typing, performed as previously described (9). Briefly, *fimH* PCR amplification was conducted using the Qiagen Multiplex PCR kit (Qiagen, Aarhus, Denmark) with the following two *fimH* primers: *fimH*-F, CACTCAGGGAACCATTCAGGCA (binds 50 to 72 nucleotides up-stream of *fimH* start), and *fimH*-R, CTTATTGATAAACAAAAGTCAC (spans the last 21 nucleotides of *fimH*). The thermocycler program for the PCR reactions consisted of 1 cycle of 94^{o}C for 5 minutes, for heat activation, followed by 30 cycles of 94^{o}C for 30 s (denaturation), 57^{o}C for 90 s (annealing), 72^{o}C for 60 s (extension), and finally 1 cycle of 72^{o}C for 60 s as final extension. The resulting PCR products were applied to Illustra ExoProStar 1-step kit (GE Healthcare), BigDye® Terminator v3.1 Cycle Sequencing Kit (Thermo Fischer) and BigDye XTerminator® Purification Kit (Thermo Fischer), with following sequencing at the Applied Biosystems 3130 XL Genetic Analyzer (Thermo Fisher). Contigs were assembled based on the paired chromatograms using CLC Genomics Workbench 9.5.1 (Qiagen).

Validation of FimTyper

Individual FASTA assemblies of the paired Sanger sequences of dataset (i) from the 42 isolates, which had been subjected to conventionally *fimH* sub-typing 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 towards the *fimH* database.

Draft genome sequences of the 243 third-generation cephalosporins-resistant *E. coli* isolates from the second dataset (ii) were analyzed directly using the FimTyper web-tool. In situations, where FimTyper did not report BLAST hits with an %ID > 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, region 4540457-4550210) 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* sub-types of dataset (ii) had not been examined previously by conventional *fimH* sub-typing, 40 randomly chosen isolates (iii) were subjected to conventional typing with PCR, 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 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 GATK Unified Genotyper, and SNPs that did not pass a minimum coverage of 10 or SNPs that were not present in minimum 90% of the base calls were excluded. Phylogenetic analyses of the identified SNPs was performed by maximum-likelihood approximation with the generalized time-reversible model in FastTree 2.1.5 (10).

Results and Discussion

Construction of FimTyper

FimTyper was constructed to perform *fimH* sub-typing on sequencing data originating from PCR and subsequent Sanger sequencing (assembled and saved in FASTA format), as well as raw reads directly from sequencing platforms such as Illumina, Ion Torrent or Roche 454, or as *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 CGE. The default settings for FimTyper were set to a minimum ID of 95% and minimum length of 60% compared to the reference, to avoid noise from e.g. gene fragments, however FimTyper allows the user to specify similarity from 55%-100% identity. The best matching hit from the database was given as output, including the percent identity 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 web service.

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 pre-assembled pairs of Sanger sequences from a dataset (i) consisting of 42 samples, which had already been sub-typed in relation to their *fimH* allele by conventional typing methods prior to the current study. The Sanger sequences covered 13 different variants of *fimH* sub-types (Table 1). The FimTyper identified *fimH* sub-types from all 42 assembled Sanger sequences correctly at a 100% identity match.

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

Using FimTyper on Whole-Genome Sequencing Data

The FimTyper web-tool successfully analyzed all 243 drafted genomes from dataset (ii). FimTyper was able to identify a *fimH* allele in 230 of the 243 draft genome datasets. The 13 *fimH* negative

152 isolates were further verified as negative by BLAST against the complete *fim*-operon including part
153 of its flanking regions (9.754 nt in total) from *E. coli* K-12 MG1655. All 13 *fimH* negative isolates
154 showed BLAST hits to the upstream and downstream regions of the *fim* operon, but no hits to any
155 of the genes of the *fim*-operon including *fimH* suggesting that these isolates were missing not only
156 the *fimH*, but the complete *fim* operon.

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

161 Among the 40 randomly chosen isolates from dataset (iii), which were additionally subjected to
162 conventional typing, three of the samples did not yield any PCR products, which were in agreement
163 with the FimTyper results on whole-genome sequence data, where *fimH* negative results were
164 predicted for the same three samples. For the remaining 37 samples, the conventional typing using
165 DNA alignment and the FimTyper predictions obtained using both assembled Sanger sequences and
166 whole-genome sequence data were in 100% agreement.

167 **MLST vs *fimH* sub-type**

168 Sub-typing of *fimH* is especially relevant for the major *E. coli* clonal group ST131. Therefore, the
169 122 *E. coli* isolates from dataset (ii) previously predicted by Roer *et al.* (7) to belong to ST131 by
170 the Achtman MLST scheme (11) were further analyzed in relation to their *fimH* sub-type. All 122
171 ST131 *E. coli* isolates harbored a *fimH* allele, with *fimH30* being the most frequent (n = 95, 78%)
172 and representing the pandemic multi-drug resistant clonal group ST131-*H*30, followed by *fimH27*
173 (n = 14, 11%), *fimH41* (n = 11, 9%), *fimH22* (n = 1, < 1%) and *fimH35* (n = 1, < 1%). In a study by
174 Johnson *et al*. (12), the same 5 *fimH* alleles were all among the seven *fimH* sub-types found in a

collection of 352 historical and resent ST131 *E. coli* isolates, sub-typed by conventional typing. Two infrequent sub-types found by Johnson *et al.*, *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 SNP analysis and compared to the *fimH* sub-types as depicted in Figure 1. From this analysis, a clear overlap between the structure of the phylogenetic relationship and the *fimH* sub-type 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*, respectively, clustered in-between the clades of the other *fimH* sub-types. Miyoshi-Akiyama *et al*. reported a similar correlation between SNP based phylogeny and *fimH* sub-type for a collection of global ST131 *E. coli* isolates (13). However, in their study, two *fimH30* isolates clustered within the *fimH41* distinct clade, whereas a clear grouping was observed in our study. This difference could be a result of mistyping of the two *fimH30* isolates, or caused by 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 data or the custom script for SNP concatemers used by Miyoshi-Akiyama *et al*., however, for investigating the possible differences caused by the reference, a new phylogenetic reconstruction was created with *E. coli* SE15 used by Miyoshi-Akiyama *et al*., (accession no. NC_013654.1) as reference (data not shown). The reference was sub-typed as a *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* sub-types, eliminating the choice of reference as a parameter for differences between the two studies.

However, both studies illustrates the high diversity within the ST131 clonal clade and underlines the benefit to include *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, as well as 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 datasets. The software source code for the tool is publicly available on 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 on https://cge.cbs.dtu.dk/services/FimTyper.

Acknowledgment

The authors wish to 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). Part of this work was presented at the IMMEM XI conference in Estoril, Portugal, 2016.

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 208:439–45.

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.

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.

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 58:4234–7.

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 45:351–8.

6. 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 67:2640–4.

7. Roer L, Hansen F, Thomsen MCF, Knudsen JD, Hansen DS, Wang M, Samulioniené J, Justesen US, Røder BL, Schumacher H, Østergaard C, Andersen LP, Dzajic E, Søndergaard TS, Stegger M, Hammerum AM, Hasman H. 2017. WGS-based surveillance of third-generation cephalosporin-resistant *Escherichia coli* from bloodstream infections in Denmark.

J Antimicrob Chemother.

8. 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.

9. 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 78:1353–60.

10. Price MN, Dehal PS, Arkin AP. 2010. FastTree 2--approximately maximum-likelihood trees for large alignments. PLoS One 5:e9490.

11. 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.

12. 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 E V. 2013. Abrupt emergence of a single dominant multidrug-resistant strain of *Escherichia coli*. J Infect Dis 207:919–28.

13. Miyoshi-Akiyama T, Sherchan JB, Doi Y, Nagamatsu M, Sherchand JB, Tandukar S,

257 Ohmagari N, Kirikae T, Ohara H, Hayakawa K. 2016. Comparative Genome Analysis of
258 Extended-Spectrum-β-Lactamase-Producing *Escherichia coli* Sequence Type 131 Strains
259 from Nepal and Japan. mSphere 1:e00289-16.

260

261

Journal of Clinical Microbiology

262 **Table 1. *fimH* Sub-type Prediction by Conventional Typing versus FimTyper**

Dataset	Numbers of samples	Conventional Positive	Detected by FimTyper	Concordance between conventional and FimTyper
(i)	42	42	42	100%
(ii)	243	ND	230	NA
(iii)	40	37	37	100%

263 ND; Not Determined, NA; Not Available.

264 **Table 2. Distribution of *fimH* sub-types identified among the 243 whole-genome sequenced**

265 ***Escherichia coli* isolates by using the FimTyper web-tool**

fimH sub-type	No. of isolates with *fimH* sub-type
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

266 [a] New *fimH* sub-type identified by FimTyper.

267

Journal of Clinical Microbiology

Figure 1. SNP based phylogeny of the 122 ST131 *Escherichia coli* isolates. Phylogenetic reconstruction of the 122 ST131 *E. coli* isolates, with *E. coli* JJ1886 as reference genome. The tree was constructed from 13,155 SNPs, and represented as a cladogram. The *fimH* sub-type is marked at the branch tip for each isolate: Green; *fimH30*, Purple; *fimH27*, Orange; *fimH35*, Blue; *fimH41*, Pink; *fimH22*.