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; Sokurenko, Evgeni; Hasman, Henrik

Published in:
Journal of Clinical Microbiology

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

Publication date:
2017

Document Version
Peer reviewed version

Link back to DTU Orbit

Citation (APA):
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

1 Louise ROER1#, Veronika TCHESNOKOVA2, Rosa ALLESØE3, Mariya MURADOVA2, Sujay CHATTOPADHYAY2, Johanne AHRENFELDT3, Martin C. F. THOMSEN3, Ole LUND3, Frank HANSEN1, Anette M. HAMMERUM1, Evgeni SOKURENKO2 and Henrik HASMAN1

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](https://bitbucket.org/genomie...), the database freely available at [https://bitbucket.org/genomie...](https://bitbucket.org/...), and a service implementing the software available at [https://cge.cbs.dtu.dk/services/FimTyper](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-H30. FimTyper provides a standardized tool, as a rapid alternative to conventional fimH sub-typing, highly suitable for surveillance and outbreak detection.
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 be particularly relevant within the highly virulent ST131 clonal group, where resistant and multi-resistant H30 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/](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**
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°C for 5 minutes, for heat activation, followed by 30 cycles of 94°C for 30 s (denaturation), 57°C for 90 s (annealing), 72°C for 60 s (extension), and finally 1 cycle of 72°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 Fisher), 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 \textit{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 \textit{fimH} allele. Additionally, the contig in which the hit was found, followed by the position in the contig, and the accession number of the \textit{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 \textit{fimH} allele by conventional typing methods prior to the current study. The Sanger sequences covered 13 different variants of \textit{fimH} sub-types (Table 1). The FimTyper identified \textit{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 \textit{fimH} allele in 230 of the 243 draft genome datasets. The 13 \textit{fimH} negative
isolates were further verified as negative by BLAST against the complete \textit{fim}\text{-operon} including part of its flanking regions (9,754 nt in total) from \textit{E. coli} K-12 MG1655. All 13 \textit{fimH} negative isolates showed BLAST hits to the upstream and downstream regions of the \textit{fim} operon, but no hits to any of the genes of the \textit{fim}\text{-operon} including \textit{fimH} suggesting that these isolates were missing not only the \textit{fimH}, but the complete \textit{fim} operon.

Among the 492 \textit{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 \textit{fimH30} allele (n=98), the \textit{fimH27} allele (n=42), the \textit{fimH5} allele (n=17), and the \textit{fim41} allele (n=15). The new allele was assigned number 517 (\textit{fimH517}) and added to the database.

Among the 40 randomly chosen isolates from dataset (iii), which were additionally subjected to conventional typing, three of the samples did not yield any PCR products, which were in agreement with the FimTyper results on whole-genome sequence data, where \textit{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 obtained using both assembled Sanger sequences and whole-genome sequence data were in 100\% agreement.

**MLST vs \textit{fimH} sub-type**

Sub-typing of \textit{fimH} is especially relevant for the major \textit{E. coli} clonal group ST131. Therefore, the 122 \textit{E. coli} isolates from dataset (ii) previously predicted by Roer \textit{et al.} (7) to belong to ST131 by the Achtman MLST scheme (11) were further analyzed in relation to their \textit{fimH} sub-type. All 122 ST131 \textit{E. coli} isolates harbored a \textit{fimH} allele, with \textit{fimH30} being the most frequent (n = 95, 78\%) and representing the pandemic multi-drug resistant clonal group ST131-\textit{H30}, followed by \textit{fimH27} (n = 14, 11\%), \textit{fimH41} (n = 11, 9\%), \textit{fimH22} (n = 1, < 1\%) and \textit{fimH35} (n = 1, < 1\%). In a study by Johnson \textit{et al.} (12), the same 5 \textit{fimH} alleles were all among the seven \textit{fimH} sub-types found in a
A collection of 352 historical and recent ST131 *E. coli* isolates, sub-typed by conventional typing.

Two infrequent sub-types found by Johnson et al., *fimH*15 (1/352) and *fimH*94 (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 *fimH*30 isolates clustered together in a distinct ST131-H30 clade, as did the *fimH*41 isolates and the *fimH*27 isolates. The two single isolates with *fimH*35 and *fimH*22, 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 *fimH*30 isolates clustered within the *fimH*41 distinct clade, whereas a clear grouping was observed in our study. This difference could be a result of mistyping of the two *fimH*30 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 *fimH*41 isolate and clustered together with all our *fimH*41 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 \textit{fimH} alleles from either simple Sanger generated sequences, as well as raw or assembled WGS data from \textit{E. coli} genomes has been developed, thus enabling researchers and primary investigators to rapidly detect the \textit{fimH} allele in their datasets. The software source code for the tool is publicly available on \url{https://bitbucket.org/genomicepidemiology/fimtyper} and the database hosted by the Center for Genomic Epidemiology (CGE), is freely available at \url{https://bitbucket.org/genomicepidemiology/fimtyper_db}. A publicly available web service implementing the software can be found on \url{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 \textit{E. coli} and Carbapenemase Producing Bacteria and partly by the Center for Genomic Epidemiology (\url{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


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

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Numbers of samples</th>
<th>Conventional Positive</th>
<th>Detected by FimTyper</th>
<th>Concordance between conventional and FimTyper</th>
</tr>
</thead>
<tbody>
<tr>
<td>(i)</td>
<td>42</td>
<td>42</td>
<td>42</td>
<td>100%</td>
</tr>
<tr>
<td>(ii)</td>
<td>243</td>
<td>ND</td>
<td>230</td>
<td>NA</td>
</tr>
<tr>
<td>(iii)</td>
<td>40</td>
<td>37</td>
<td>37</td>
<td>100%</td>
</tr>
</tbody>
</table>

ND; Not Determined, NA; Not Available.
Table 2. Distribution of fimH sub-types identified among the 243 whole-genome sequenced Escherichia coli isolates by using the FimTyper web-tool

<table>
<thead>
<tr>
<th>fimH sub-type</th>
<th>No. of isolates with fimH sub-type</th>
</tr>
</thead>
<tbody>
<tr>
<td>fimH30</td>
<td>98</td>
</tr>
<tr>
<td>fimH27</td>
<td>42</td>
</tr>
<tr>
<td>fimH5</td>
<td>18</td>
</tr>
<tr>
<td>fimH41</td>
<td>15</td>
</tr>
<tr>
<td>fimH54</td>
<td>9</td>
</tr>
<tr>
<td>fimH24</td>
<td>5</td>
</tr>
<tr>
<td>fimH106</td>
<td>4</td>
</tr>
<tr>
<td>fimH29</td>
<td>4</td>
</tr>
<tr>
<td>fimH2</td>
<td>3</td>
</tr>
<tr>
<td>fimH35</td>
<td>3</td>
</tr>
<tr>
<td>fimH65</td>
<td>3</td>
</tr>
<tr>
<td>fimH31</td>
<td>3</td>
</tr>
<tr>
<td>fimH38</td>
<td>2</td>
</tr>
<tr>
<td>fimH64</td>
<td>2</td>
</tr>
<tr>
<td>fimH34</td>
<td>2</td>
</tr>
<tr>
<td>fimH22</td>
<td>1</td>
</tr>
<tr>
<td>fimH517 (a)</td>
<td>1</td>
</tr>
<tr>
<td>fimH103</td>
<td>1</td>
</tr>
<tr>
<td>fimH142</td>
<td>1</td>
</tr>
<tr>
<td>fimH63</td>
<td>1</td>
</tr>
<tr>
<td>fimH25</td>
<td>1</td>
</tr>
<tr>
<td>fimH32</td>
<td>1</td>
</tr>
<tr>
<td>fimH39</td>
<td>1</td>
</tr>
<tr>
<td>fimH58</td>
<td>1</td>
</tr>
<tr>
<td>fimH60</td>
<td>1</td>
</tr>
<tr>
<td>fimH43</td>
<td>1</td>
</tr>
<tr>
<td>fimH215</td>
<td>1</td>
</tr>
<tr>
<td>fimH445</td>
<td>1</td>
</tr>
<tr>
<td>fimH97</td>
<td>1</td>
</tr>
<tr>
<td>fimH483</td>
<td>1</td>
</tr>
<tr>
<td>fimH10</td>
<td>1</td>
</tr>
<tr>
<td>fimH15</td>
<td>1</td>
</tr>
<tr>
<td>fimH-negative</td>
<td>13</td>
</tr>
</tbody>
</table>

\(a\) New fimH sub-type identified by FimTyper.
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; *fimH*30, Purple; *fimH*27, Orange; *fimH*35, Blue; *fimH*41, Pink; *fimH*22.