Rapid and Easy In Silico Serotyping of Escherichia coli Isolates by Use of Whole-Genome Sequencing Data

Joensen, Katrine Grimstrup; Tetzschner, Anna M. M.; Iguchi, Atsushi; Aarestrup, Frank Møller; Scheutz, Flemming

Published in:
Journal of Clinical Microbiology

Link to article, DOI:
10.1128/JCM.00008-15

Publication date:
2015

Document Version
Publisher's PDF, also known as Version of record

JCM
Journals.ASM.org

Rapid and Easy *In Silico* Serotyping of *Escherichia coli* Isolates by Use of Whole-Genome Sequencing Data

Katrine G. Joensen,[a,b] Anna M. M. Tetzschner,[b] Atsushi Iguchi,[c] Frank M. Aarestrup,[a] Flemming Scheutz[b]

National Food Institute, Division for Epidemiology and Microbial Genomics, Technical University of Denmark, Kgs. Lyngby, Denmark[a]; WHO Collaborating Centre for Reference and Research on Escherichia and Klebsiella, Department of Microbiology and Infection Control, Statens Serum Institut, Copenhagen, Denmark[b]; Department of Animal and Grassland Sciences, Faculty of Agriculture, University of Miyazaki, Miyazaki, Japan[c]

Accurate and rapid typing of pathogens is essential for effective surveillance and outbreak detection. Conventional serotyping of *Escherichia coli* is a delicate, laborious, time-consuming, and expensive procedure. With whole-genome sequencing (WGS) becoming cheaper, it has vast potential in routine typing and surveillance. The aim of this study was to establish a valid and publicly available tool for WGS-based *in silico* serotyping of *E. coli* applicable for routine typing and surveillance. A FASTA database of specific O-antigen processing system genes for O typing and flagellin genes for H typing was created as a component of the publicly available Web tools hosted by the Center for Genomic Epidemiology (CGE) (www.genomicepidemiology.org). All *E. coli* isolates available with WGS data and conventional serotype information were subjected to WGS-based serotyping employing this specific SerotypeFinder CGE tool. SerotypeFinder was evaluated on 682 *E. coli* genomes, 108 of which were sequenced for this study, where both the whole genome and the serotype were available. In total, 601 and 509 isolates were included for O and H typing, respectively. The O-antigen genes *wzx*, *wzy*, *wzm*, and *wzt* and the flagellin genes *fliC*, *flkA*, *fllA*, *flmA*, and *flnA* were detected in 569 and 508 genome sequences, respectively. SerotypeFinder for WGS-based O and H typing predicted 560 of 569 O types and 504 of 508 H types, consistent with conventional serotyping. In combination with other available WGS typing tools, *E. coli* serotyping can be performed solely from WGS data, providing faster and cheaper typing than current routine procedures and making WGS typing a superior alternative to conventional typing strategies.

Escherichia coli is usually a harmless commensal, but some strains have evolved the capability to cause disease in humans and/or animals by specific particular pathogenic mechanisms. In some cases, infection can be fatal (1).

Serotyping is a method for classification of *E. coli* that has existed since the 1940s and has since been developed into standardized procedures (2–4). Performance of serotyping requires a high level of expertise and access to cross-absorbed antisera. It is a time-consuming and laborious procedure.

O:K:H serotyping is based on a combination of the three immunogenic structures: the lipopolysaccharide (LPS) (O antigen), the capsular antigen (K), and the flagellar (H) antigen.

Since few laboratories are able to perform K typing, O:H serotyping has become the gold standard for characterization of pathogenic *E. coli*. O:H serotyping is crucial in the detection of outbreaks, for epidemiological surveillance, for taxonomic differentiation of *E. coli*, for detecting pathogenic serotypes within the species, and for clonal and evolutionary studies. In contrast to several more recently developed molecular typing methods, such as pulsed-field gel electrophoresis (PFGE), ribotyping and to some extent multilocus sequence typing (MLST), serotyping provides information that is directly associated with the antigenic response and as such is of importance for the ecology of the isolates. At present, this typing method cannot be replaced by any other method.

The current serotyping scheme comprises 188 O groups designated O1 to O188 (publication of O182 to O188 is pending), with the O groups O31, O47, O67, O72, O94, and O122 having been withdrawn from the scheme (3, 5). Fifty-three H antigens are included in the scheme, designated H1 to H56, with the exception of H13, H22, and H50, which have been withdrawn (5, 6).

Lipopolysaccharide (LPS) is embedded in the outer bacterial membrane and consists of three components: lipid A, a core oligosaccharide, and the O-specific polysaccharide chain, referred to as the O antigen. O antigens usually consist of 10 to 25 repeating units of oligosaccharides (the O unit) comprising 2 to 7 sugar residues from a broad range of sugars and are the most variable region of the bacterial cell. The majority of the O units are translocated across the membrane and polymerized by the O-antigen processing proteins, which are encoded by *wzx* (O-antigen flippase) and *wzy* (O-antigen polymerase). This translocation is referred to as the Wzx/Wzy-dependent pathway. Acidic capsule K antigens belonging to group 1 and 4 also employ the Wzx/Wzy-dependent pathway for translocation and are often coexpressed with one of the neutral LPS-linked polymers of O group O8, O9, O20, or O101 (7). In the absence of these neutral LPS-linked O groups, the group 1 and 4 K antigens are given an O designation; e.g., K87 is designated O32, K85 is O141, and K9 is O104 (8). An alternative ABC transporter pathway responsible for translocation

Received 8 January 2015 **Returned for modification** 12 February 2015
Accepted 2 May 2015

Accepted manuscript posted online 13 May 2015

Citation Joensen KG, Tetzschner AMM, Iguchi A, Aarestrup FM, Scheutz F. 2015. Rapid and easy *in silico* serotyping of *Escherichia coli* isolates by use of whole-genome sequencing data. J Clin Microbiol 53:2410–2426. doi:10.1128/JCM.00008-15.

Editor: K. C. Carroll

Address correspondence to Flemming Scheutz, FSC@ssi.dk.

Supplemental material for this article may be found at http://dx.doi.org/10.1128/JCM.00008-15.

doi:10.1128/JCM.00008-15

of the O antigen has been described for the O groups O8, O9, O52, and O99, where translocation is facilitated by a transporter protein encoded by *wzm* and an ATP-binding component encoded by *wzt* (9).

The H antigen is a homopolymer filament composed of repeated molecules of the protein flagellin, which facilitate bacterial motility (10, 11). The serological specificity of the H antigen is determined by the central region of the flagellin molecule. This central region may vary considerably, resulting in serologically different flagellar H antigens.

The H antigen is generally conferred by expression of the flagellin gene, *fliC*, which is the case in 44 of the 53 known H types, whereas the H antigen in the remaining 9 H types is encoded by non-*fliC* flagellin genes: H3, H35, H36, H47, and H53 are encoded by *flkA*, H44 and H55 by *fllA*, H54 by *flmA*, and H17 by *flnA* (12–16). Strains with non-*fliC*-encoded H antigens sometimes also contain a nonexpressed or cryptic *fliC* gene. This has been found for the reference strains for H3 (Bi7372-41), H47 (1755-58), H53 (E480-68), H54 (E223-69), and H55 (E2987-73), which have been shown to carry the *fliC* genes for H16, H21, H16, H21, and H38, respectively, while the reference strains for H17 (P12b) and H44 (781-55) both carry *fliC*, encoding H4 (15, 16). Furthermore, a nonreference strain, 1-391, has been described to contain a nonexpressed H2 *fliC* gene in addition to the *flkA* gene, encoding its H3 phenotype (17). In addition, the H40 reference strain E49 has been shown to contain an additional, nonexpressed H8 *fliC* gene, in addition to its H-type-determining H40 *fliC* gene (16). The coexistence of two flagellar genes as well as the presence of two sets of genes from two O-antigen pathways in the same strain poses challenges to molecular typing methods.

Previously, we presented the tool VirulenceFinder for detection of *E. coli* virulence genes from whole-genome sequencing (WGS) data and demonstrated it as a superior alternative to current routine typing of Vero cytotoxin-producing *E. coli* (VTEC) infections, offering rapid and comparable typing results (18).

To facilitate complete WGS-based typing of all *E. coli* strains, including VTEC, we have now enabled WGS-based serotyping of *E. coli*. Here, we present SerotypeFinder, a publicly available Center for Genomic Epidemiology (CGE) Web tool, constructed for serotype prediction of *E. coli* from WGS data, on the basis of the O-antigen processing genes *wzx*, *wzy*, *wzm*, and *wzt* and the flagellin genes *fliC*, *flkA*, *fllA*, *flmA*, and *flnA*. In addition, we demonstrate the comparability of the results obtained by SerotypeFinder to those obtained via conventional serotyping and thus the superiority of the WGS-based typing, which can be performed much more rapidly and cost effectively than conventional serotyping.

MATERIALS AND METHODS

Study design and isolates. The use of WGS for *in silico* serotyping of *E. coli* was evaluated by comparing the serotypes predicted on the basis of WGS data by SerotypeFinder, which was constructed as part of this study, to the O and H types obtained by conventional serotyping.

A total of 682 *E. coli* genomes with available serotype information were included in the study, most of which were of human origin and were isolated as pathogenic strains and thus clinically relevant. Some of these genome sequences, 379, were publicly availably genomes with serotype information collected from the NCBI. Of these, 249 were assembled genomes from the PulseNet STEC genome reference library, deposited by the Centers for Disease Control and Prevention (CDC), Atlanta, GA (accession no. PRJNA218110), and 130 were genome sequences included in a previous study on comparative genome analysis of *E. coli* (19, 20). Four *Shigella* genomes were available and included from this study. Additionally, 44 *E. coli* genome assemblies were included from a previous study we conducted on real-time typing and surveillance of VTEC from WGS data (18). Raw sequence data (Illumina HiSeq; two 100-bp reads) from 146 *E. coli* reference strains were obtained from CDC (accession no. PRJNA186441).

One hundred eight isolates from the strain collection at Statens Serum Institute (SSI), Copenhagen, Denmark, were included in the study for validation. These isolates were subjected to conventional serotyping and subsequent sequencing on a MiSeq sequencer (Illumina, San Diego, CA) at the WHO Collaborating Centre for Reference and Research on *Escherichia* and *Klebsiella* (WHOCC) at Statens Serum Institute. Four isolates and corresponding genome assemblies (ASTS01, ANSR01, ANWO01, and ASVR01), were kindly provided by Karen Leth Nielsen (SSI, Denmark), and one isolate with WGS assembly was provided by Yvonne Agersø (DTU Food, Denmark). These isolates were subjected to phenotypic confirmation of serotypes at SSI as described below.

The complete set of isolates used in this study is listed in Table S1 in the supplemental material.

For the validation of the SerotypeFinder database, we chose to include at least three *E. coli* isolates per H type and further tried to cover as many different O groups as possible.

Conventional serotyping. The isolates included from the strain collections at SSI were O:H serotyped at the WHOCC at Statens Serum Institute according to standard procedures as previously described (3, 5).

Whole-genome sequencing. Genomic DNA (gDNA) was purified from the isolates using the Genetra Puregene Yeast/Bact. kit (Qiagen, Netherlands) or the Wizard genomic DNA purification kit (Promega, Fitchburg, WI), and DNA concentrations were determined with the Qubit double-stranded DNA (dsDNA) BR assay kit (Invitrogen, Carlsbad, CA). DNA libraries were prepared according to the Illumina Nextera XT DNA sample preparation kit (Illumina) protocol, consisting of fragmentation by tagmentation, PCR for amplification, addition of index sequences, and bead-based normalization of library concentrations prior to pooling of libraries. Pooled libraries were sequenced on the MiSeq benchtop sequencer (Illumina) using the MiSeq Reagent kit v2 (Illumina) for 300 or 500 cycles, producing two 150- bp or two 250- bp sequence reads.

Raw data were quality trimmed using CutAdapt (21), keeping reads of at least 40 bp with a quality score of Q20. *De novo* assembly was done with the Velvet algorithm (22), with a *k*-mer length of 35 to 95, collecting contigs with a minimum length of 200 bp.

SerotypeFinder gene databases. A FASTA database was constructed for automatic, *in silico* serotyping from WGS data, as a component of the publicly available CGE Web tools (http://cge.cbs.dtu.dk/services/).

The database was based on the O-antigen processing system genes *wzx*, *wzy*, *wzm*, and *wzt* for *in silico* O typing and the flagellin genes *fliC*, *flkA*, *flmA*, *flnA*, and *fllA* for *in silico* H typing. The content for the databases was obtained by searching the NCBI nucleotide collection for all the above-mentioned genes in *E. coli* and collecting only complete genes from the entries with assigned O types or H types, in accordance with the specific gene. All unique gene variants were stored in the databases. In addition, a sequence set of O-antigen processing genes from O1 to O187 reported by Iguchi et al. (23) was also employed for database construction. Finally, the *E. coli* reference strains Su4411-41 (O14), F10018-41 (O18ab), F8198-41 (O57), 2745-53 (125ab), 2129-54 (125ac), 56-54 (128ab), 5564-64 (128ac), E68 (O141ab), and RVC2907 (O141ac) were sequenced at SSI as described above, and the gene variants were extracted and added to the database. Similarly, the H24 reference strain K72 (H25w) was sequenced, and the *fliC* variant added to the H database.

Multiple alignment of database gene variants was performed using MUSCLE (24) with default parameters, and relationships among *wzx*, *wzy*, and flagellin variants were visualized by construction of neighbor-joining trees based on the percent identity (ID) (25). Furthermore, to identify possible difficulties in the prediction by SerotypeFinder, all data-

base variants of the same gene were compared by pairwise BLAST against each other, to consider full-length hits of high similarity.

Using SerotypeFinder for analysis of sequence data. SerotypeFinder was constructed for prediction of *E. coli* serotypes from WGS data, like the existing CGE Web tools, with the purpose of being user-friendly and simple. The tool accepts sequence data in various formats, both as assembled genomes and as raw data from the different sequencing platforms. The assembly of raw data is incorporated into the tool as previously described for other CGE tools (18, 26, 27). All database genes are subjected to a BLAST search against the genome assembly. The tool outputs the best-matching genes from the BLAST analysis, as well as percent ID between the database gene and the gene detected in the query genome, the length of query and database gene, the position of the hit in the query sequence, and the accession number of the best-matching database gene. In addition, the tool outputs the predicted O and H types, based on the best-matching genes.

The user can select the threshold of percent ID between the best matching database gene and the input sequence and can also choose the minimum length of the database gene a gene hit has to cover. The default is set to 60%.

In this study, genome assemblies were submitted to SerotypeFinder using a threshold of 85% ID and a minimum length of 60%

Evaluation of the SerotypeFinder database for *in silico* serotyping. The WGS-based O typing performed by SerotypeFinder was done on the basis of a combination of the *wzx* and *wzy* gene variants or the *wzm* and *wzt* variants, whereas the H typing was based on the gene variants of *fliC*, or less commonly, a variant of another flagellin gene, *flkA*, *fllA*, *flmA*, or *flnA*. The O and H types assigned by SerotypeFinder were compared to the types determined by conventional tests.

RESULTS

A total of 682 isolates with available WGS data and conventional serotype information were obtained and used for evaluation and validation of the CGE Web tool, with the O typing and H typing being validated on 601 and 509 isolates, respectively.

SerotypeFinder O-typing database. The O-typing database covered all valid O types from O1 to O187, except O14 and O57, since neither *wzx*, *wzy*, *wzm*, nor *wzt* could be detected in the two *E. coli* reference strains, Su4411-41 (O14) and F8198-41 (O57), that were sequenced for database construction. The gene content of the O-typing database is listed in Table 1.

One hundred sixty-nine O types were represented by *wzx* and *wzy*, whereas 10 types, O9, O52, O60, O89, O92, O95, O97, O99, O101, and O162, were determined by *wzm* and *wzt*. The only O type represented by *wzx*, *wzy*, *wzm*, and *wzt* was O8. In total, the database contained 202 *wzx* variants, 209 *wzy* variants, 14 *wzm* variants, and 14 *wzt* variants. Most O groups were represented only by a single *wzx* and *wzy* variant, whereas O6, O7, O45, O55, O103, O111, O126, O127, O145, and O157 were represented by more variants of both *wzx* and *wzy*. O8 and O9 were represented by more variants of *wzm* and *wzt*. The O groups O18, O28, O112, O125, O128, and O141 were also represented by several gene variants, due to their ab/ac variations.

The relationships between the *wzx* variants and those between the *wzy* variants are illustrated in Fig. 1A and B, respectively.

Overall, the *wzx* and *wzy* variants clustered according to their O type. The O groups with ab/ac variation, O18ab/ac, O125ab/ac, O128ab/ac, and O141ab/ac, clustered together for *wzx* and *wzy*, apart from O28ab/ac and O112ab/ac. One *wzx* and *wzy* variant, obtained from the same O45 isolate, did not cluster with the remaining O45, and for *wzy*, the two O22 variants did not cluster.

Similarity among the *wzx*, *wzy*, *wzm*, and *wzt* database variants. Most full-length BLAST hits with high similarity were between variants of the same O type represented by more gene variants. This was the case for both *wzx* and *wzy* of the O types O6, O7, O18, O45, O55, O103, O111, O125, O126, O127, O128, O145, and O157, for *wzx* of O63, O86, O104, O123 and O148, and for *wzy* of O102, O117, O121, and O141. Within O groups, the identity on *wzx* was 98.99 to 99.93% ID, while *wzy* variants were 98.07 to 99.93% identical. Within groups O8 and O9, *wzm* and *wzt* variants had 97.20 to 99.25% ID and 91.20 to 99.18% ID, respectively. High similarities were found between O types O89, O101, and O162 (*wzm*, 95.38 to 99.87% ID; *wzt*, 98.41 to 99.87% ID). O groups O153/O178 and O42 and one O28ac had identical gene variants for both *wzx* and *wzy*, as shown in Table 2. For *wzx*, O2/O50, O17/O77, O118/O151, O169/O183, and O141ab/O141ac were identical, and for *wzy*, O13/O135, O17/O44, and O123/O186 were identical. These identical variants are represented with both O groups in the database.

All gene variants in the database are shown in Table S2 in the supplemental material in a pairwise BLAST hits similarity matrix.

High similarities were found between variants belonging to different O types and are listed in Table 2. O types which were similar to each other on both *wzx* and *wzy* variants were O90 and O127, O107 and O117, O20 and O137, and O13, O135, and O129. O68/O62 were 99.92% identical with regard to *wzx* while not having the same-size *wzy*. O36/O134 and O124/O164 were 99.82% identical with regard to *wzy* within both pairs, while the *wzx* genes were of different sizes.

O17, O44, O73, O77, and O106 were very closely related, with identities ranging from 99.68 to 100% and 99.62 to 100% for *wzx* and *wzy*, respectively. O17 and O77 were identical for *wzx*, and O17 and O44 were identical for *wzy*. O73 produced a full-length BLAST hit to the others on *wzy* only.

Apart from O20/O137 and O153/O187, all the O types that were highly similar with regard to gene variants were also known to cross-react phenotypically either one way or two way (Table 2).

SerotypeFinder H-typing database. A flagellin gene database was constructed for WGS-based H typing. The database contained 102 flagellin gene variants and covered all 53 known H types. The database consisted primarily of *fliC* gene variants, but for the H types determined by other flagellin genes (*flkA* for H3, H35, H36, H47, and H53, *fllA* for H44 and H55, *flmA* for H54, and *flnA* for H17), the variants of these were included. The gene content of the H database is listed in Table 1. The flagellin gene variants were H type specific, with a large proportion of H types being represented by more *fliC* variants in the database. In total, the database contained 93 *fliC* variants, five *flkA* variants, two *fllA* variants, and one variant each of the genes *flmA* and *flnA*. The relationship among all the flagellin gene variants is illustrated in Fig. 2. The flagellin genes clustered according to their H types.

The flagellin gene variants were highly variable in size, with the *fliC* variants ranging from 1,050 bp to 2,013 bp: *flkA* variants (each H type of a different size) ranged from 1,107 bp to 1,671 bp, *fllA* variants were 1,725 bp and 1,869 bp, and the *flmA* and *flnA* genes were 1,551 bp and 1,524 bp, respectively.

The flagellin gene variants within each H type were generally conserved in size and showed high sequence similarity to variants within the same H type, with the lowest similarity being 97.18% between the two H31 *fliC* variants. Some H types, however, harbored *fliC* genes with a gene size that was also present in other H

TABLE 1 Gene content of the SerotypeFinder O and H databases[a]

Antigen type	Gene(s)	No. of variants	Size (bp)[n]	% ID	Accession no. or reference strain[b]
O1	*wzx*, *wzy*	1, 1			GU299791
O2	*wzx*[c], *wzy*	1, 1			EU549863
O3	*wzx*, *wzy*	1, 1			EU694097
O4	*wzx*, *wzy*	1, 1			AY568960
O5	*wzx*, *wzy*	1, 2			AB811596,* (AB811596*, CAPL01000042)
O6	*wzx*, *wzy*	3, 3			(AJ426045, CP002185, AB811597*), (AJ426045, CP002185, AJ426423)
O7	*wzx*, *wzy*	3, 3			CP003034, AB490074, AF125322
O8	*wzx*, *wzy*	1, 1			AF013583
	wzm, *wzt*	2, 2			AB010150, AB811598*
O9 (O9a)	*wzm*, *wzt*	3, 3			AB010294, D43637, AB010293
O10	*wzx*, *wzy*	1, 1			AB811599*
O11	*wzx*, *wzy*	1, 1			HQ388393
O12	*wzx*, *wzy*	1, 1			AB811600*
O13	*wzx*, *wzy*[d]	1, 1			EU296422, EU296422/EU296423
O15	*wzx*, *wzy*	1, 1			AY647261
O16	*wzx*, *wzy*	1, 1			AB811601*
O17	*wzx*[e], *wzy*[f]	1, 1			AB812084*/AB972416*, DQ000314
O18(ab/ac)	*wzx*, *wzy*	2, 3			(AB811602*, GU299793), (AB811602*, GU299793, AB811603*)
O19	*wzx*, *wzy*	1, 1			AB811604*
O20	*wzx*, *wzy*	1, 1			AB811605*
O21	*wzx*, *wzy*	1, 1			EU694098
O22	*wzx*, *wzy*	1, 2			DQ851855, (DQ851855, AB811606)
O23	*wzx*, *wzy*	1, 1			AB811607*
O24	*wzx*, *wzy*	1, 1			DQ220292
O25	*wzx*, *wzy*	1, 1			GU014554
O26	*wzx*, *wzy*	1, 1			AF529080, (AF529080)
O27	*wzx*, *wzy*	1, 1			GU014555
O28(ab/ac[g])	*wzx*, *wzy*	3, 2			(AB811608*, AB811609*, DQ462205/FJ539194), (AB811608*, DQ462205/FJ539194)
O29	*wzx*, *wzy*	1, 1			EU294173
O30	*wzx*, *wzy*	1, 1			AB811610*
O32	*wzx*, *wzy*	1, 1			EU296410
O33	*wzx*, *wzy*	1, 1			AB811611*
O34	*wzx*, *wzy*	1, 1			AB811612*
O35	*wzx*, *wzy*	1, 1			FJ940774
O36	*wzx*, *wzy*	1, 1			AB811613*
O37	*wzx*, *wzy*	1, 1			AB811614*
O38	*wzx*, *wzy*	1, 1			AB811615*
O39	*wzx*, *wzy*	1, 1			AB811616*
O40	*wzx*, *wzy*	1, 1			EU296417
O41	*wzx*, *wzy*	1, 1			AB811617*
O42[g]	*wzx*, *wzy*	1, 1			DQ462205/FJ539194
O43	*wzx*, *wzy*	1, 1			AB811619*
O44	*wzx*, *wzy*[f]	1, 1			AB811620*, DQ000314
O45	*wzx*, *wzy*	3, 5			(JN859200, JN859208, CU463050), (AIGX01000028, JN859208, CU463050, JN859205, AY771223)
O46	*wzx*, *wzy*	1, 1			AB811621*
O48	*wzx*, *wzy*	1, 1			AB811622*
O49	*wzx*, *wzy*	1, 1			AB811623*
O50	*wzx*[c], *wzy*	1, 1			EU549863, AB811624*
O51	*wzx*, *wzy*	1, 1			AB812020*
O52	*wzm*, *wzt*	1, 1			AY528413
O53	*wzx*, *wzy*	1, 1			EU289392
O54	*wzx*, *wzy*	1, 1			AB812085*
O55	*wzx*, *wzy*	3, 2			(AB353132, AB353133, JH958641), (AB353132, CP003109)
O56	*wzx*, *wzy*	1, 1			DQ220293
O58	*wzx*, *wzy*	1, 1			EU294175
O59	*wzx*, *wzy*	1, 1			AY654590
O60	*wzm*, *wzt*	1, 1			AB812022*
O61	*wzx*, *wzy*	1, 1			GU220362
O62	*wzx*, *wzy*	1, 1			AB812023*
O63	*wzx*, *wzy*	2, 1			(EU549862, FJ539195), EU549862

(Continued on following page)

TABLE 1 (Continued)

Antigen type	Gene(s)	No. of variants	Size (bp)[n]	% ID	Accession no. or reference strain[b]
O64	*wzx*, *wzy*	1, 1			AB812025*
O65	*wzx*, *wzy*	1, 1			AB812026*
O66	*wzx*, *wzy*	1, 1			DQ069297
O68	*wzx*, *wzy*	1, 1			AB812027*
O69	*wzx*, *wzy*	1, 1			AB812028*
O70	*wzx*, *wzy*	1, 1			FN995094
O71	*wzx*, *wzy*	1, 1			GU445927
O73	*wzx*, *wzy*	1, 1			DQ000313
O74	*wzx*, *wzy*	1, 1			AB812030*
O75	*wzx*, *wzy*	1, 1			GU299795
O76	*wzx*, *wzy*	1, 1			AB812031*
O77	*wzx*[e], *wzy*	1, 1			(AB812084*/AB972416*), AB972416*
O78	*wzx*, *wzy*	1, 1			FJ940775
O79	*wzx*, *wzy*	1, 1			EU294162
O80	*wzx*, *wzy*	1, 1			AB812032*
O81	*wzx*, *wzy*	1, 1			CU928162
O82	*wzx*, *wzy*	1, 1			AB812034*
O83	*wzx*, *wzy*	2, 1			(AB812035, 289152760), AB812035*
O84	*wzx*, *wzy*	1, 1			AB812036*
O85	*wzx*, *wzy*	1, 1			GU299798
O86	*wzx*, *wzy*	3, 1			(AY220982, AY667408, AY670704), AY220982
O87	*wzx*, *wzy*	1, 1			EU294177
O88	*wzx*, *wzy*	1, 1			AB812037*
O89	*wzm*, *wzt*	1, 1			AB812038*
O90	*wzx*, *wzy*	1, 1			AB812039*
O91	*wzx*, *wzy*	1, 1			AY035396
O92	*wzm*, *wzt*	1, 1			AB812040*
O93	*wzx*, *wzy*	1, 1			AB812041*
O95	*wzm*, *wzt*	1, 1			AB812042*
O96	*wzx*, *wzy*	1, 1			AB812043*
O97	*wzm*, *wzt*	1, 1			AB812042*
O98	*wzx*, *wzy*	1, 1			DQ180602
O99	*wzm*, *wzt*	1, 1			FJ940773
O100	*wzx*, *wzy*	1, 1			AB812045*
O101	*wzm*, *wzt*	1, 1			GQ499340, AB812046*
O102	*wzx*, *wzy*	1, 2			JX087966, (AB812047*, JX087966)
O103	*wzx*, *wzy*	3, 7			(NC_013353, AB704860, EF027106), (EF027120, EF027119, EF027116, AP010958, EF027115, EF027117, EF027106)
O104	*wzx*, *wzy*	5, 1			(AF361371, KB021482, CP003301, AFPS01000083, AFOB02000091), AF361371
O105	*wzx*, *wzy*	1, 1			EU294171
O106	*wzx*, *wzy*	1, 1			DQ000315
O107	*wzx*, *wzy*	1, 1			EU694095
O108	*wzx*, *wzy*	1, 1			AB812048*
O109	*wzx*, *wzy*	1, 1			HM485572
O110	*wzx*, *wzy*	1, 1			AB812049*
O111	*wzx*, *wzy*	2, 2			(AP010960, JN887675), (NC_013364, JN887675)
O112(ab/ac)	*wzx*, *wzy*	2, 2			EU296413, EU296405
O113	*wzx*, *wzy*	1, 1			AF172324
O114	*wzx*, *wzy*	1, 1			AY573377
O115	*wzx*, *wzy*	1, 1			GU068041
O116	*wzx*, *wzy*	1, 1			AB812051*
O117	*wzx*, *wzy*	1, 2			EU694096, (EU694096, DQ465247)
O118	*wzx*[h], *wzy*	1, 1			HM204927/HM204926, HM204927
O119	*wzx*, *wzy*	1, 1			GQ499368
O120	*wzx*, *wzy*	1, 1			AB812052*
O121	*wzx*, *wzy*	1, 4			JN859209, (JN859209, JN859212, AY208937, JN859216)
O123	*wzx*, *wzy*[i]	2, 1			(DQ676934, DQ676933), DQ676934
O124	*wzx*, *wzy*	1, 1			EU296419
O125(ab/ac)	*wzx*, *wzy*	2, 2			2745-53Canioni*, 2129-54Ewing*
O126	*wzx*, *wzy*	2, 2			DQ465248, GU068042, DQ465248, GU068042
O127	*wzx*, *wzy*	3, 2			(FM180568, AY493508, AB812054*), (FM180568, AY493508)

(Continued on following page)

TABLE 1 (Continued)

Antigen type	Gene(s)	No. of variants	Size (bp)[n]	% ID	Accession no. or reference strain[b]
O128(ab/ac)	*wzx*, *wzy*	3, 2			(56-54Cigleris*, 5564-64*, AY217096), (56-54Cigleris*, 5564-64*)
O129	*wzx*, *wzy*	1, 2			EU296424, (EU296424, AB972421*)
O130	*wzx*, *wzy*	1, 1			EU296421
O131	*wzx*, *wzy*	1, 1			AB812055*
O132	*wzx*, *wzy*	1, 1			AB812056*
O133	*wzx*, *wzy*	1, 1			AB812057*
O134	*wzx*, *wzy*	1, 1			AB812058*
O135	*wzx*, *wzy*[d]	1, 1			EU296423, EU296422/EU296423
O136	*wzx*, *wzy*	1, 1			AB812059*
O137	*wzx*, *wzy*	1, 1			AB972423*
O138	*wzx*, *wzy*	1, 1			DQ109551
O139	*wzx*, *wzy*	1, 1			DQ109552
O140	*wzx*, *wzy*	1, 1			AB812060*
O141(ab/ac)	*wzx*[j], *wzy*	1, 3			E68*/RVC2907*, (DQ868765, E68*, RVC2907*)
O142	*wzx*, *wzy*	1, 1			AB812061*
O143	*wzx*, *wzy*	1, 1			EU294164
O144	*wzx*, *wzy*	1, 1			AB812062*
O145	*wzx*, *wzy*	2, 3			(CP006262, JN850039), (CP006262, AY863412, JN850039)
O146	*wzx*, *wzy*	1, 1			DQ465249
O147	*wzx*, *wzy*	1, 1			DQ868766
O148	*wzx*, *wzy*	2, 1			(DQ167407, AAJT02000037), DQ167407
O149	*wzx*, *wzy*	1, 1			DQ091854, DQ868764
O150	*wzx*, *wzy*	1, 1			EU294168
O151	*wzx*[h], *wzy*	1, 1			HM204927/HM204926, HM204926
O152	*wzx*, *wzy*	1, 1			EU294170
O153[k]	*wzx*, *wzy*	1, 1			AB812063*/E54071*
O154	*wzx*, *wzy*	1, 1			AB812064*
O155	*wzx*, *wzy*	1, 1			AY657020
O156	*wzx*, *wzy*	1, 1			AB812065*
O157	*wzx*, *wzy*	7, 9			(AB602252, AB602249, AB602253, AKMA01000036, AKLI01000048, JH959508, AB602250), (JH964427, AB602249, JH970567, AB602253, AMUN01000113, AF061251, AKLQ01000042, AKLV01000031, JH953200)
O158	*wzx*, *wzy*	1, 1			GU068044
O159	*wzx*, *wzy*	1, 1			EU294176
O160	*wzx*, *wzy*	1, 1			AB812066*
O161	*wzx*, *wzy*	1, 1			GU220361
O162	*wzm*, *wzt*	1, 1			AB812067*
O163	*wzx*, *wzy*	1, 1			AB812068*
O164	*wzx*, *wzy*	1, 1			EU296420
O165	*wzx*, *wzy*	1, 1			GU068045
O166	*wzx*, *wzy*	1, 1			GU299794
O167	*wzx*, *wzy*	1, 1			EU296408
O168	*wzx*, *wzy*	1, 1			EU296403
O169	*wzx*, *wzy*[l]	1, 1			AB812069*, AB812069*/AB627352
O170	*wzx*, *wzy*	1, 1			AB812070*
O171	*wzx*, *wzy*	1, 1			AB812071*
O172	*wzx*, *wzy*	1, 1			AY545992
O173	*wzx*, *wzy*	1, 1			GU068046
O174	*wzx*, *wzy*	1, 1			DQ008592
O175	*wzx*, *wzy*	1, 1			AB812073*
O176	*wzx*, *wzy*	1, 1			AB812074*
O177	*wzx*, *wzy*	1, 1			DQ008593
O178[k]	*wzx*, *wzy*	1, 1			AB812063*/AB812075*
O179	*wzx*, *wzy*	1, 1			AB812076*
O180	*wzx*, *wzy*	1, 1			JQ751058
O181	*wzx*, *wzy*	1, 1			AB812078*
O182	*wzx*, *wzy*	1, 1			AB812079*
O183	*wzx*, *wzy*[l]	1, 1			AB627352, AB812069*/AB627352
O184	*wzx*, *wzy*	1, 1			AB812080*
O185	*wzx*, *wzy*	1, 1			AB812081*
O186	*wzx*, *wzy*[i]	1, 1			AB812082*, DQ676934

(Continued on following page)

TABLE 1 (Continued)

Antigen type	Gene(s)	No. of variants	Size (bp)[n]	% ID	Accession no. or reference strain[b]
O187	*wzx*, *wzy*	1, 1			AB812083*
H1	*fliC*	2	1,788	99.83	L07387*, AB028471*
H2	*fliC*	2	1,494 (1)	99.60	AIHA01000023
			1,497 (1)		AF543692
H3	*flkA*	1	1,590		AB128916*
H4	*fliC*	3	1,050	98.46–98.86	AJ605764*, AJ605765, AJ536600
H5	*fliC*	2	1,311	98.32	AY249990*, AY337469
H6	*fliC*	3	1,647 (2)	98.97–99.64	AIEY01000041, AY249991*
			1,653 (1)		AIFA01000047
H7	*fliC*	16	1,758	97.50–99.94	AY337468, JH694260, KB000721, KB007180, AOES01000098, KB006714, JH946604, AF228487, AF228496, AF228495, AB334575, AB334574, AF228494, AF228491, AF228492, AB028474*
H8	*fliC*	2	1,479	99.12	AJ865465, AJ884569
H9	*fliC*	1	2,013		AY249994*
H10	*fliC*	2	1,263	99.76	AY337482, AY249995*
H11	*fliC*	2	1,467	99.86	AY337465, AY337472
H12	*fliC*	4	1,788	99.38–99.89	AB028475*, AY337474, AIFX01000055, AY337471
H14	*fliC*	1	1,653		AY249998*
H15	*fliC*	1	1,689		AY249999*
H16	*fliC*	6	1,575 (3)	99.75–99.94	AB128919, JH954529, JH953794
			1,578 (3)		AY337476, AY337477, AY337475
H17	*fliC*	2	1,050	99.24	AJ515904, AJ605766
	flnA	1	1,524		CP002291*
H18	*fliC*	1	1,665		AY250001*
H19	*fliC*	2	1,833 (1)	99.24	AY337479
			1,842 (1)		AY250002*
H20	*fliC*	1	1,731		AY250003
H21	*fliC*	1	1,476		AIHL01000060
H23	*fliC*	1	1,767		AB028476*
H24[m]	*fliC*	1	1,482		K72 (H25w)*
H25	*fliC*	1	1,332		AGSG01000116
H26	*fliC*	2	1,674	99.82	AY250008, AY337483
H27	*fliC*	1	1,461		AM231154
H28	*fliC*	2	1,740	98.79	AAJT02000052, AY250010
H29	*fliC*	3	1,323 (2)	98.57–99.92	JH965342, AY337485
			1,332 (1)		AY250012
H30	*fliC*	1	1,713		AY250011
H31	*fliC*	2	1,668	97.18	CP000247, AY250013
H32	*fliC*	1	1,713		AY250014
H33	*fliC*	1	1,287		AY250015
H34	*fliC*	2	1,638	99.82	AF345850, AY250016*
H35	*flkA*	1	1,509		EF392692
H36	*flkA*	1	1,671		EF392693
H37	*fliC*	1	1,686		AY250017*
H38	*fliC*	3	1,344	99.63–99.93	AY337466, AY337473, AY250018*
H39	*fliC*	1	1,299		AY250019*
H40	*fliC*	1	1,479		AJ884568*
H41	*fliC*	1	1,680		AY250020*
H42	*fliC*	2	1,281	99.84	AY337470, AY250021*
H43	*fliC*	1	1,506		AIGA01000038
H44	*fllA*	1	1,725		AB269770*
H45	*fliC*	1	1,707		AY250023*
H46	*fliC*	2	1,713 (1)	99.65	AB028478*
			1,719 (1)		AY250024*
H47	*flkA*	1	1,107		EF392694
H48	*fliC*	1	1,497		AY250025*
H49	*fliC*	2	1,695 (1)	99.82	AY250026*
			1,698 (1)		AB028480
H51	*fliC*	2	1,818	99.94	AY250027*, AB028481*
H52	*fliC*	2	1,344	99.93	AY250028*, AVRH01000047

(Continued on following page)

TABLE 1 (Continued)

Antigen type	Gene(s)	No. of variants	Size (bp)[n]	% ID	Accession no. or reference strain[b]
H53	*flkA*	1	1,272		AB128917*
H54	*flmA*	1	1,551		AB128918*
H55	*fllA*	1	1,869		AB269771*
H56	*fliC*	1	1,311		AY250029

[a] Gene sizes and similarities within genes are only shown for the flagellar genes *fliC*, *flkA*, *flmA*, *flnA*, and *fllA*.
[b] Variant sequences obtained by sequencing of reference strains (see http://www.ssi.dk/English/HealthdataandICT/National%20Reference%20Laboratories/Bacteria/~/media/49802860CB5E44D6A373E6116ABBDC0D.ashx) are marked with asterisks. Parentheses indicate the respective sources of the unique *wzx* and *wzy* sequences. For example, for O6, two *wzx* and *wzy* sequences were obtained from AJ426045 and CP002185, respectively, one *wzx* sequence was obtained from AB811597, and one *wzy* sequence was obtained from AJ426423.
[c] O2/O50 *wzx* variant.
[d] O13/O135 *wzy* variant.
[e] O17/O77 *wzx* variant.
[f] O17/O44 *wzy* variant.
[g] One variant of O28ac *wzx* and *wzy* is shared with O42.
[h] O118/O151 *wzx*.
[i] O123/O186 *wzy* variant.
[j] O141ab/O141ac *wzx* variant.
[k] O153/O178 *wzx* and *wzy* variants.
[l] O169/O183 *wzy* variant.
[m] The H24 strain K72 (H25w) was sequenced in this study, and the *fliC* variant was extracted for inclusion in the H database.
[n] Numbers in parentheses are the numbers of sequences of the given length.

types, specifically, H5 and H56 (1,311 bp), H25 and H29 (1,332 bp), H38 and H52 (1,344 bp), H8 and H40 (1,479 bp), H2 and H48 (1,497 bp), H6 and H14 (1,653 bp), H1 and H12 (1,788 bp), H4 and H17 (1,050 bp), and H30, H32, and H46 (1,713 bp). When subjected to a BLAST search against the other variants of the same gene size, most variants belonging to different H types produced alignments of less than 40% of the gene size (see Table S2 in the supplemental material).

Among the variants producing full-length BLAST hits to other H-type genes were H5 and H56 variants, which showed 91.97% ID and 92.65% ID between the H56 variant and the two H5 variants. H8 and H40 variants showed 94.46% ID and 95.20% ID between the different H type variants, and H1 and H12 variants showed 97.43 to 97.71% ID. The H30 and H32 variants had 93.99% ID, whereas one H4 *fliC* (AJ605764) and the two H17 variants showed very high sequence similarity to each other, ranging from 97.98% to 99.90%.

Evaluation of *in silico* serotyping using SerotypeFinder. For most *E. coli* isolates obtained for validation, both conventional O and H types were available, but for some isolates, a comparison could be made on only one of the antigens, since conventional serotype information was not available for both. Of the 682 *E. coli* genome sequences available for validation, 669 were assigned an O type by conventional methods, whereas 538 were assigned an H type. Since some of the reference strains available as genome sequences had been used for database construction, they were not included for validation of the specific database, and thus the O database and H database of SerotypeFinder were validated on a total of 601 and 509 isolates, respectively. Information on all validation isolates is included in Table S1 in the supplemental material.

Of the 601 isolates employed in validation of O prediction by SerotypeFinder, both genes (*wzx/wzy* or *wzm/wzt*) were found and used for O-type prediction for 548, whereas 21 validation assemblies of O groups O3, O55, O69, O111, O128, O145, and O157 were assigned to the correct O groups based on only *wzx* or *wzy*. In 51 isolates, the genes were detected by reference mapping. No O-processing genes were found in 32 of the isolates, and they could not be included for further analysis. In total, 560 of 569 isolates with detected genes were predicted with consistency with the conventional typing. An H type was predicted with consistency with the conventional typing by SerotypeFinder for 504 of 508 isolates with detected flagellin genes. Flagellin genes were not detected in only one of the 509 isolates employed in validation of H prediction.

The overall evaluation of the serotype prediction performed by SerotypeFinder can be seen from Table 3, while Table 4 shows the O and H groups and the number of available genomes that are currently covered by the validation.

Discrepancies between SerotypeFinder predictions and reported conventional results. For the 569 genomes used for validation of the O typing performed by SerotypeFinder, nine were predicted to be a different O type than what was determined by conventional methods. Discrepancies were also observed for two isolates [F10524-41 and 178-54(1986-54)], which were not part of the validation, as they were reference strains represented by variants in the database. All discrepancies are listed and annotated in Table 5.

Of the 509 genomes used for validation of the H typing, five were predicted to be different from what was reported by the conventional typing. Two of these isolates were replicate sequences of the same reference strain, W27 (test O115).

Discrepancy between O types predicted by the *wzx* and *wzy* (or *wzm* and *wzt*). For 37 isolates included in the validation of the O prediction, SerotypeFinder detected *wzx* and *wzy* (or *wzm* and *wzt*) genes belonging to different O types within the isolate, and prediction was made from just one of these variants. The observed differences are listed in Table 6.

The one O50 and three O2 isolates in the data set were identical for *wzx* but could be distinguished by *wzy*, which was specific to the O type. Likewise, for the two O169 isolates and the O183 isolate, which had identical *wzy* variants but specific *wzx* variants. Similarly, the two O123 and one O186 isolates had O123/O186 *wzy* but could be distinguished by *wzx*. The *wzy* genes of O17 and

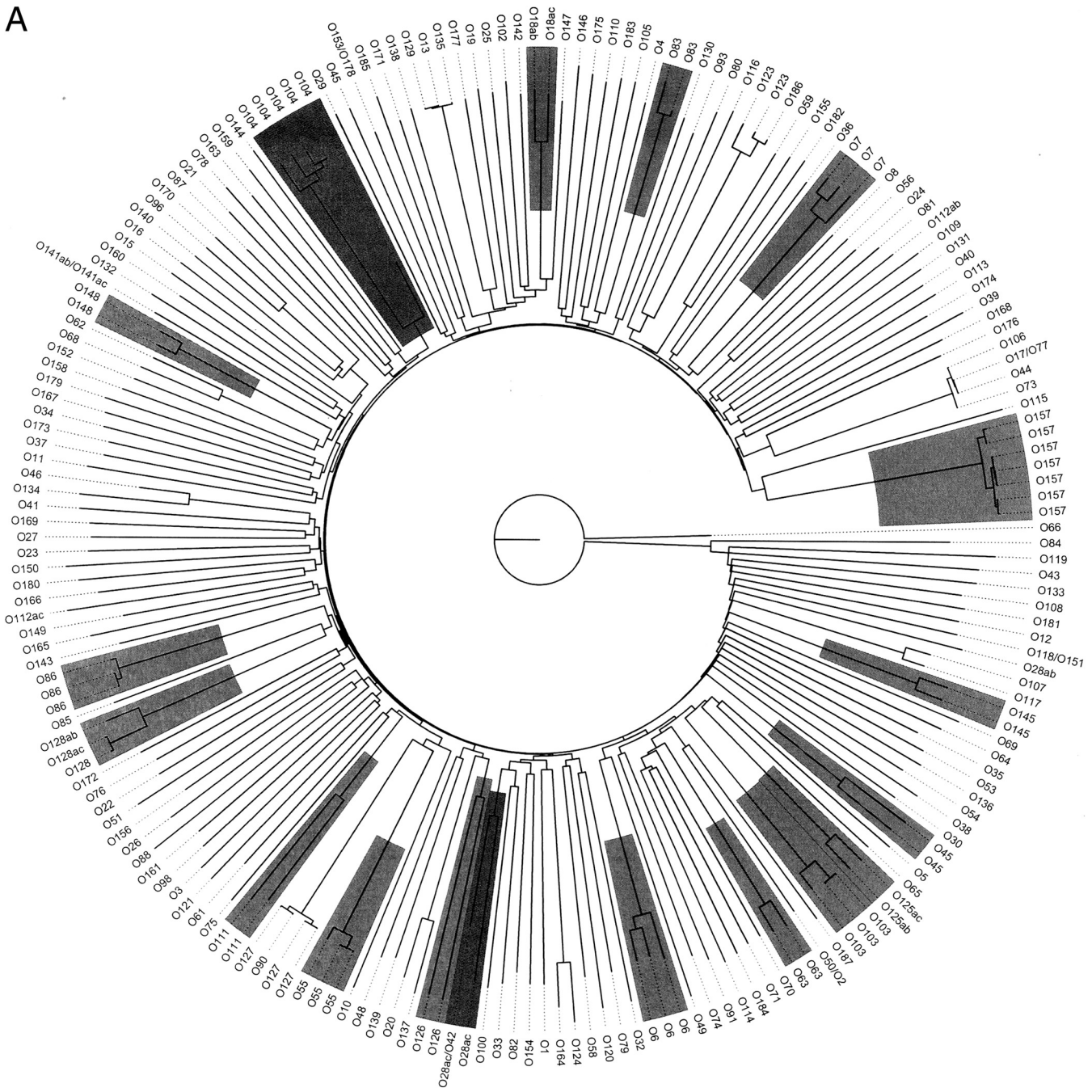

FIG 1 Relationship of the SerotypeFinder database *wzx* and *wzy* gene variants. The key genes used in the *in silico* O typing (*wzx* [A] and *wzy* [B]) were visualized by neighbor-joining trees, constructed using percent ID. O types represented by more variants are highlighted. The *wzx* and *wzy* variants clustered according to O types. One O45 antigen gene cluster sequence is known to differ from that of other O45 gene cluster sequences and has been proposed to have been acquired from other *Enterobacteriaceae* (36).

O44 were identical; however, these O types could be distinguished by their *wzx* variants, which for O44 was specific to the type, and for O17 was shared with O77.

For the six O118 isolates in the validation set, the *wzx* variant was shared (specific to O118 and O151), whereas five of the isolates had *wzy* specific for O151 and only one isolate had *wzy* specific for O118. In addition, the validation set included one O151, and this isolate also had O118/O151 *wzx*, while harboring the O151 *wzy*. Thus, the O118 and O151 could not be distinguished from each other. The same was true for the O135 isolate, where O13 *wzx* and O13/O135 *wzy* were detected.

The only O164 isolate had O164 *wzx* while having O124 *wzy*. One of the four O127 isolates in the validation set possessed a complete O90 *wzx* variant while having an O127 *wzy*. For one of the two O134 isolates, O134 *wzx* was detected while an incomplete *wzy* was found.

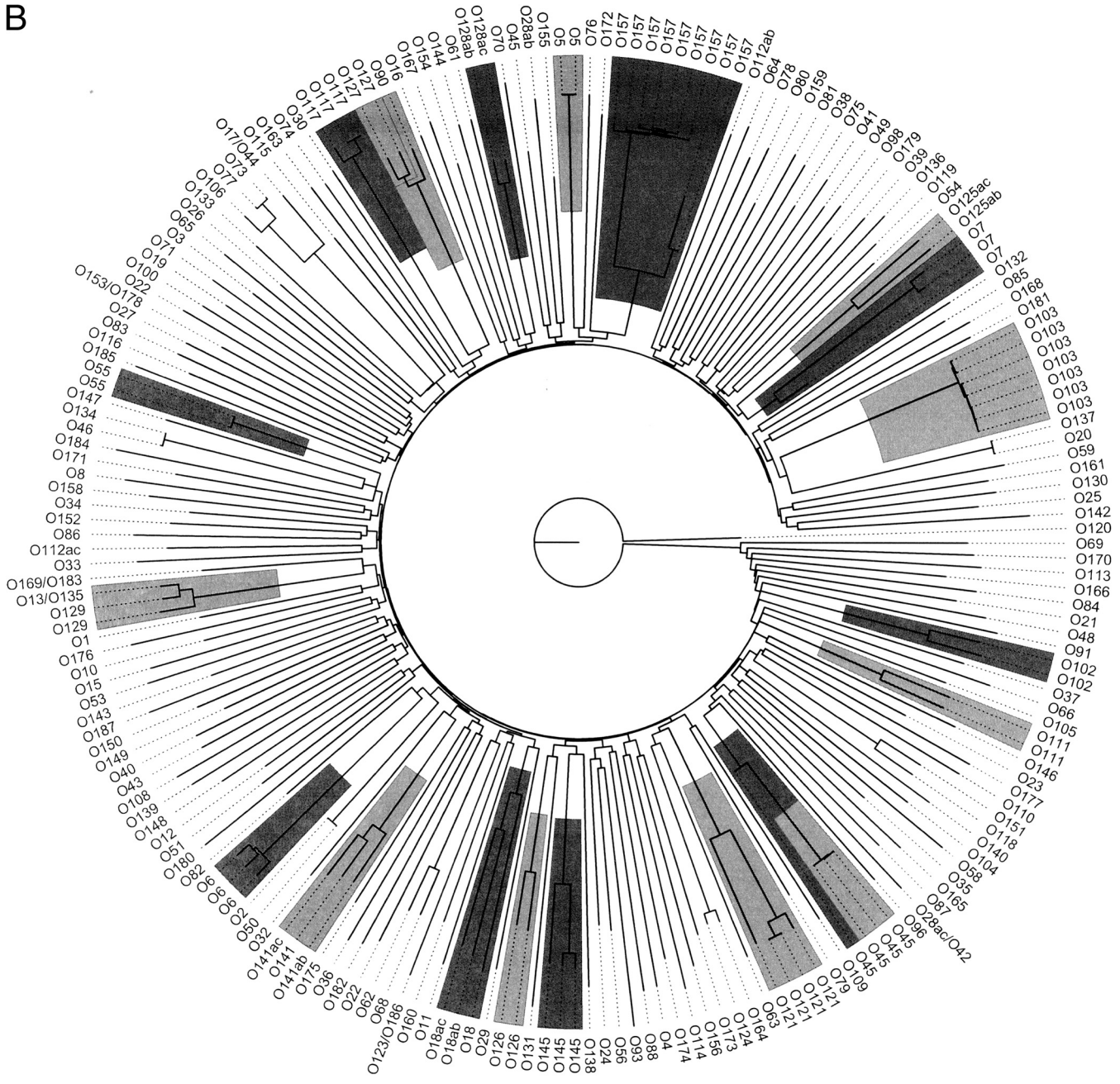

FIG 1 continued

For one of the two O101 isolates, a complete O162 *wzm* gene was found together with an O101 *wzt*. The other O101 isolate in the validation set harbored O101 *wzm* and *wzt*, and the O162 isolate had O162 *wzm* and *wzt*. Furthermore, for all ab/ac variations of O128, no difference was observed; all harbored the O128ab *wzx* and the O128ac *wzy*.

Typing of nonmotile, rough, and *Shigella* isolates by SerotypeFinder. SerotypeFinder was able to assign H types to all but one of the *E. coli* isolates in the validation set which had not been assigned H types by conventional methods. In total, this referred to 144 of the isolates, representing around a fifth of the total validation set. For 75 of these isolates, the H type had not been classified as not known (NK), not done (ND), or nontypeable (HN), while 69 of the isolates were nonmotile (H−) by conventional typing.

The validation set included 13 isolates with no conventionally assigned O types, seven of which were rough strains. SerotypeFinder could assign O types for six of these—BL21 (O7), BW2952 (O16), C895-12 (O104), KT114 (O36), SSI_AA058 (O39), and SSI_AA024 (O156)—whereas for the last isolate, C850-12, no O-processing genes (*wzx*, *wzy*, *wzm*, and *wzt*) were detected.

Of the remaining six isolates with no conventionally assigned O type, SerotypeFinder assigned O types for four of them but not for SSI_AA063 (O?K?H35) and SMS-3-5 (NK). Furthermore,

TABLE 2 Nucleotide similarities of O-processing genes *wzx* and *wzy* and the corresponding known cross-reactions using rabbit antisera

Category and O groups	% similarity		Cross-reaction using rabbit antisera[a]	
	wzx	*wzy*	Known	Additional
Identity in either *wzx* and/or *wzy* gene(s)				
O153/O178	100	100	No	O153 * O54
O28ac/O42[b]	100	100	Yes	O28ab * O28ac
O2/O50	100	99.90	Yes	Four with O2; 11 with O50[c]
O17/O77	100	99.92	Yes	—[d]
O118/O151	100	99.91	O118-O151[e]	O174-O151
O169/O183	NA[f]	100	Yes	O30 * O169; O105 * O183
O141ab/O141ac	100	99.82	Yes	
O13/O135	99.68	100	O135-O13[e]	12 with O13; Seven with O135[g]
O17/O44	99.92	100	Yes	—[d]
O123/O186	99.86–99.93	100	Yes	O12 * O123; O116 * O123; O4 * O186; O116 * O186; O124 * O186
High similaritiy on one gene but different sizes on the other				
O62/O68	99.92	Diff. sizes	Yes	O62 * O73; O62 * O106; O62 * O125ab/ac; O68 * O73; O68 * O125ab[h]
O36/O134	Diff. sizes	99.82	No	O36 * O43; O134 * O17; O134 * O171
O124/O164	Diff. sizes	99.82	Yes	O124 * O186; O4 * O164; O25 * O164[i]
High similarities on both genes				
O90/O127	99.76–99.92	99.14–99.23	Yes	O86 * O90; O127 * O128ab
O107/O117	99.75	99.77–99.85	Yes	Ten with O107, nine with O117[j]
O20/O137	99.86	99.84	No	
O13/O135/O129	99.60–99.92	98.96–100	Yes	O50 * O129
			O135-O13, only one way	O129 * O133

[a] Two-way cross-reactions are shown as "Yes" or with an asterisk, and one-way cross reactions are shown with a hyphen, e.g., "O118-O151." These are cross-reactions as known by the WHOCC as of 24 March 2015.
[b] Only one of the O28ac variants is 100% identical.
[c] O2 also cross-reacts with O53, O74, O117, and O138; O50 also cross-reacts with O1, O13, O19, O53, O107, O117, O129, O133, O135, O147, and O149.
[d] —, O17, O44, O73, O77, and O106 form a special O complex, which is very closely related both phenotypically and genotypically. All five O groups show phenotypic two way cross-reactions except for O77 which does not cross-react with O73 or O106. Outside this complex, two-way cross-reactions between O17, O44, O73, and O77 and five, four, seven and one other O groups are presently recorded at the WHOCC.
[e] Only the one way cross-reaction is known.
[f] NA, not applicable. Similarity on *wzx* genes is very low, sizes are different, and the pairwise BLAST results in a match only over 17 bp (see Table S2 in the supplemental material).
[g] O13 also cross-reacts with O4, O18ac, O19, O23, O25, O44, O50, O62, O73, O125ab, O129 and O147; O135 also cross-reacts with O4, O16, O17, O18ac, O44, O50 and O129.
[h] Fifteen O antigens react with O68 antiserum: O3, O4, O7, O13, O18ab, O19, O23, O36, O44, O102, O138, O141, O142, O147 and O148.
[i] O164 also reacts in O18ac.
[j] O107 also cross-reacts with O1, O14, O50, O92, O102, O108, O116, O139, O159, and O185; O117 also cross-reacts with O2, O26, O50, O53, O101, O112ac, O120, O149, and O185.

analysis of 4 *Shigella* isolates representing *Shigella boydii*, *S. dysenteriae*, *S. flexneri*, and *S. sonnei* by SerotypeFinder resulting in the predicted H types H45, H18, H14, and H16, respectively, and O-type prediction for *S. dysenteriae* and *S. flexneri* of O148 and O13/O135, respectively.

Coexistence of flagellin genes and sets of O-processing genes. For some isolates SerotypeFinder detected more than one flagellin gene (Table 7). This was observed for the non-*fliC*-encoded H types (H3, H35, H36, H44, H47, H53, H54, and H55) with the exception of H17. In these isolates with more than one flagellin gene, prediction of H type was based on the non-*fliC* gene.

More than one set of O-processing genes were detected by SerotypeFinder in the former test strain, H308b (O8:K84:H−), where a complete O93 *wzx/wzy* set and an O8 *wzm/wzt* set were detected, and in the O9:K9:H12 reference strain, Bi316-42, where a complete O104 *wzx/wzy* set and an O9 *wzm/wzt* set were detected.

DISCUSSION

The objective of this study was to construct and validate an *in silico* tool, SerotypeFinder, as part of the CGE Web tools to enable rapid and accurate WGS-based routine (O:H) serotyping of *Escherichia coli*.

There is an increasing need for development of user-friendly tools to enable extraction of relevant information for identification and typing from WGS data. This can provide clinical health personnel without bioinformatic skills with results that can be readily interpreted. Many different straightforward tools have already emerged that can be used for extraction of relevant data for identification and typing of different pathogens from WGS (28).

In this study, we demonstrate that serotype specific information can be extracted from WGS data and analyzed for (O:H) serotype prediction by SerotypeFinder. The use of SerotypeFinder enabled accurate detection of the O-processing genes *wzx*, *wzy*, *wzm*, and *wzt* and the flagellin genes *fliC*, *flkA*, *fllA*, *flmA*, and *flnA*

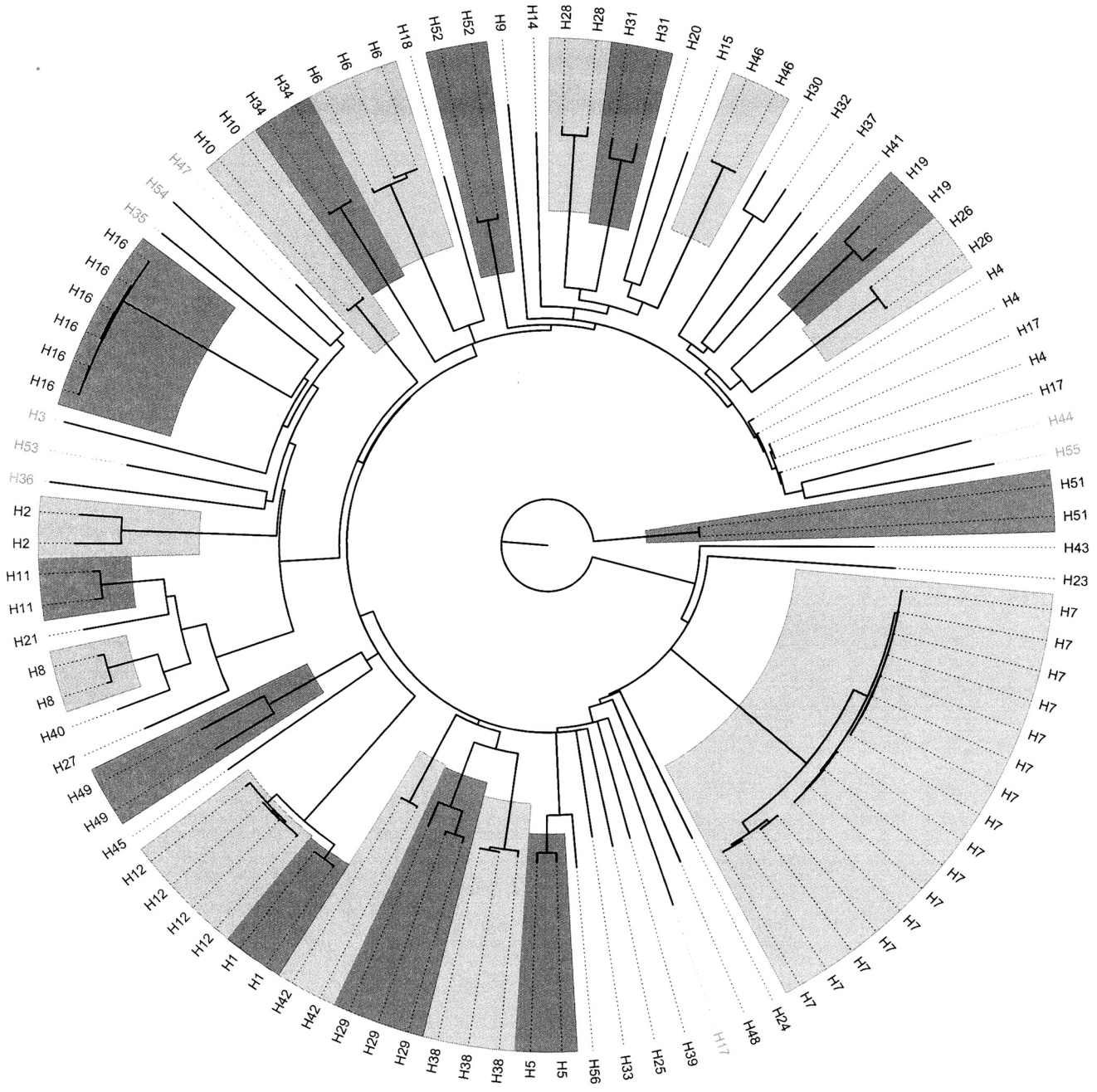

FIG 2 Relationship of the SerotypeFinder flagellin gene variants. The flagellin genes employed for *in silico* H typing (*fliC*, *flkA*, *fllA*, *flmA* and *flnA*) were visualized by a Neighbor joining tree, constructed using percent ID. H types in black are *fliC*-encoded, and gray are non-*fliC* flagellin genes. H types represented by more variants are highlighted. The *fliC* genes clustered according to H type.

TABLE 3 Evaluation of serotype prediction by SerotypeFinder

Typing	No. (%) of genomes: For validation	With detected genes	With consistent WGS and conventional results
O	601	569[a] (~95)	560[b] (~98)
H	509	508 (~100)	504 (~99)

[a] In 51 genomes, genes were found by reference mapping, and in 21 genomes, only one gene was used for prediction.
[b] Eleven predictions were ambiguous between the two O-processing genes [O118/O151(7), O164/O124, O134/O46, O90/O127, and O162/O101].

and serotype prediction based on these. However, SerotypeFinder was not successful in detecting the O-processing genes in 32 isolates and detected no flagellin genes in one isolate. This was most likely a feature of poor sequence data and genome assemblies for most of these isolates but could also indicate the existence of hitherto-unrecognized genetic O or H variation. However, as O14, O57, and O188 were not represented in the database for SerotypeFinder, isolates of these O types were not expected to be assigned an O type, as confirmed by the two O14 isolates included in the validation. O14 is considered O rough by the WHOCC (F. Scheutz, unpublished data).

TABLE 4 O groups and H types currently covered in the validation of SerotypeFinder

No. of genomes in agreement with conventional results (O total; H total)	O groups	H types
1 (71; 0)	O4, O10, O18, O21, O22, O23, O24, O29, O30, O33, O34, O35, O37, O38, O42, O43, O44, O48, O49, O50, O53, O54, O58, O65, O73, O76, O82, O83, O84, O87, O89, O95, O102, O105, O108, O109, O110, O116, O120, O130, O135, O136, O137, O138, O139, O140, O143, O144, O147, O149, O150, O151, O154, O155, O159, O162, O163, O164, O167, O168, O170, O172, O173, O175, O176, O177, O178, O179, O182, O183, O186	
2 (82; 0)	O1, O5, O20, O25, O28, O32, O39, O40, O46, O60, O70, O78, O79, O80, O81, O88, O92, O96, O98, O99, O101, O114, O123, O126, O131, O132, O134, O141, O153, O156, O158, O160, O161, O165, O166, O171, O177, O181, O184, O185, O187	
3 (33; 63)	O27, O75, O107, O112, O113, O125, O133, O142, O148, O169, O180	H3, H5, H12, H14, H15, H17, H20, H26, H27, H31, H36, H37, H38, H41, H45, H46, H48, H49, H51, H52, H53
4 (48; 44)	O2, O3, O9, O16, O17, O36, O51, O69, O104, O115, O127, O174	H23, H24, H29, H32, H35, H39, H42, H44, H47, H54, H55
5 (25; 25)	O11, O91, O117, O119, O146	H1, H9, H40, H43, H56
6 (30; 12)	O6, O7, O15, O86, O118	H18, H30
7 (7; 14)	O8	H28, H34
8 (8; 0)	O45	
≥10 (324; 345)	O103 (17), O55 (22), O128 (18), O145 (24), O121 (30), O26 (43), O111 (48), O157 (122)	H33 (9), H10 (10), H16 (12), H25 (11), H21 (13), H4 (14), H8 (14), H6 (17), H2 (34), H19 (39), H11 (45), H7 (127)

For some O types, WGS-based typing offered better resolution than was obtained by the conventional typing. For instance, four of the five rough strains could be typed by SerotypeFinder, and also the isolate SSI_AA029, which by conventional typing was either O36 or O68, was confirmed as O36 by SerotypeFinder. In addition, the isolate F8858-41, which according to the serological methods was O26, O50, or O133, was finally confirmed as O133 by SerotypeFinder.

TABLE 5 Discrepancies between SerotypeFinder predictions and the reported conventional results

Isolate	Serotype	SerotypeFinder ID	Comment
RN587/1	O157:**H8**	H45	Difficult to type according to authors (29)
W27[a]	O115:K−:**H18**	H19	Original transcription error suspected; now the strain is nonmotile and cannot be confirmed; three of 116 motile O115 strains in the WHOCC database are H19, and none are H18
U19-41	O51:K−:**H24**	H49	Original transcription error suspected; the isolate agglutinates only the H pool containing H49, and not the H pool containing H24
F7902-41	O15:K14:**H4**	H17	Prediction based on *fliC*; H4 and H17 *fliC* are very similar in sequence; unresolved
53638	**O144**	O124:H30	Suspected transcription error or typing mistake
178-54(1986-54)[b]	**O129**:K−:H11	O13/O135	Re-tested and confirmed O129
C2139-99	**O111**:H7	O96	Unresolved
F10524-41[b]	**O62**:K−:H30	O13/O135	Isolate is still subjected to O-antigen testing, as the antigen is rough
F8858-41	(**O19a**)[c]	O133	Suspected mix-up; the isolate was subjected to retyping of O antigen, and many cross-reactions occurred (O26, O50 and O133); final O-antigen determination was not possible
HW35	**O11**:Kne:H33	O99	Suspected original transcription error; the isolate was confirmed as O99, with no cross-reaction with the O11 antisera
SE11	**O152**:H28	O173	Unresolved
SSI_AA017	**O41**:H48	O60	Unresolved; poor sequence data and assembly
SSI_AA037	**O20**:K67:H17	O128ac	The O128 detected is suspected to be due to K67
SSI_AA099	**O43**:H46	O131	Retesting confirmed O131
SSI_AA095	**O44**:K53,93:H46	O17	Cross-reactions between O17, O44, O73, and O77 make it difficult to determine these O groups phenotypically; retesting confirmed O44

[a] The O115 reference strain was sequenced both at CDC and SSI with the same results in SerotypeFinder.
[b] Variant sequences for the reference strains for O129 [178-54(1986-54)] and O62 (F10524-41) were represented in the O database, and therefore, resequencing of these strains was not considered as part of the validation set.
[c] The parentheses indicate that final O-antigen determination was not possible.

TABLE 6 Discrepancies between O types reported by the *wzx* and *wzy* genes (or *wzm* and *wzt*)

Conventional O type (no. of isolates)	SerotypeFinder ID[a]			
	wzx	*wzy*	*wzm*	*wzt*
O50 (1)	O50/O2	**O50**		
O2 (3)	O50/O2	**O2**		
O169 (2)	**O169**	O169/O183		
O183 (1)	**O183**	O169/O183		
O123 (2)	**O123**	O123/O186		
O186 (1)	**O186**	O123/O186		
O17 (2)	**O17**/O77	**O17**/O44		
O44 (1)	**O44**	O17/O44		
O118 (6)	O118/O151	O151 (5), O118 (1)		
O151 (1)	O118/O151	O151		
O135 (1)	O13	O13/O135		
O164 (1)	**O164**	O124		
O127 (1)	O90	**O127**		
O134 (1)	O134	O46 (incomplete gene)		
O128 (4), O128a (1), O128ab (3), O128abc (4)	O128ab	O128ac		
O101 (1)			O162	**O101**

[a] Boldface indicates the gene variant that should be used for prediction of the specific O type. For 37 isolates included in the validation, SerotypeFinder detected *wzx* and *wzy* (or *wzm* and *wzt*) genes belonging to different O types within the isolate.

Some gene variants in SerotypeFinder are by nature identical. This was found for O153/O178 and O42 and some O28ac types. We also observed that some of the O types that we thought were identical with regard to only one of the O-processing genes were indistinguishable for both genes. Isolates of O118 and O151 were impossible to differentiate, and the only O135 isolate available for validation was predicted to be both O13 and O135. Further validation is thus necessary.

The members of the pairs O2/O50, O169/O183, O123/O186, and O17/O44 had one of their O-processing genes in common but were distinguishable by the other variant. However, one isolate (SSI_AA095), which was O44 according to the conventional typing, was predicted to be O17 by SerotypeFinder. As cross-reactions are known to occur between O17, O44, O73, and O77, it is difficult to phenotypically determine these O groups, and more isolates of these groups need to be evaluated in order to determine if they can be properly predicted by SerotypeFinder. The possibility that other genetic elements such as prophages, plasmids, or genomic islands could contain genes involved in processing of the O-antigen determinants should be examined in more detail.

For one O164, one O127, and one O101 isolate, which were each represented in the validation set by only a single isolate, one of the O-processing gene variants detected was specific to another O type, namely, O124 (*wzy*), O90 (*wzx*), and O162 (*wzm*), respectively. Therefore, we suggest that O164, O127, and O101 should be

TABLE 7 Coexistence of *fliC* and non-*fliC* flagellin genes detected by SerotypeFinder

Gene[a]					
fliC	Non-*fliC*	Serotype[b]	Isolate[b]	Pathotype/source	Country
fliC2	*flkA35*	O128abc:H35	SSI_AA057	Diarrhea	Guinea Bissau
		O?:K?:H35	SSI_AA063	Bacteremia	Denmark
		O128abc:H35	2012-60-2763-9	Pig	Denmark
fliC2	*flkA47*	OX185:H47	SSI_AA061	VTEC (*vtx2*)/Diarrhea	Denmark
		O128abc:H47	SSI_AA012	Diarrhea	England
fliC4	*fllA44*	O3:H44	SSI_AA030		USA
		O20:H44	KTE154		Denmark
fliC4	*fllA55*	O21:H55	SSI_AA026	Bacteremia	USA
fliC11	*flkA3*	Orough[c]:K−:H3	SSI_AA024	Calf	Turkey
fliC11[c]	*flkA47*	**O156**:H47	**E1585-68**	Diarrhea	Sharjah, Arabian Gulf
fliC11	*flmA54*	(O5):K−:H54	SSI_AA098 (ECOR-18)	Healthy *Celebes* ape	USA
fliC16	*flkA53*	O11:H53	SSI_AA027	Children	Liberia
fliC21	*flkA36*	O86:H36	SSI_AA066	Diarrhea	Guinea Bissau
fliC21	*flkA47**	O86:**H47**	**1755-58**	Diarrhea	
fliC21	*flmA54**	O161:**H54**	**E223-69**	Diarrhea	Sharjah, Arabian Gulf
fliC25	*fllA55*	O21:H55	SSI_AA026	Bacteremia	USA
fliC27	*flkA3*	O8:H3	SSI_AA028		Germany
fliC27	*flkA36*	O26:H36	SSI_AA044	Diarrhea	Denmark
		O51:H36	SSI_AA075	EAEC/healthy control	Mali
fliC27	*flmA54*	O117:H54	SSI_AA013	Urinary tract infection	Denmark
		O107:H54	SSI_AA088		England
		(O107):H54	SSI_AA090		Denmark
fliC38	*fllA44*	O16:K19:H44	SSI_AA021	Diarrhea	Denmark
fliC38	*fllA55**	O75:**H55**	**E2987-73**		England
		O75:H55	SSI_AA084	Diarrhea	Denmark
		O75:H55	SSI_AA045		Denmark
fliC40[d]	*flkA53*	O148:H53	SSI_AA046		Thailand

[a] Asterisks denote previously described combinations.
[b] Reference strains and their test H antigens are in boldface.
[c] Published as H21, but H11 and H21 cross-react serologically and it is not clear if absorbed antisera were used in the confirmation of H antigens (16).
[d] A presumed intact *fliC53* gene is highly homologous (98%) to *fliC40* (15).
[e] Predicted to be O156 according to SerotypeFinder.

predicted from *wzx*, *wzy*, and *wzt*, respectively, which represent their conventional O types. Further isolates should be validated to confirm this. In particular, strains of O124/O164, *Shigella boydii* 3, O13/O129/O135, and *Shigella flexneri* 2a, as well as strains of O169/O183 and *Shigella boydii* 6 and 10, should be examined because of the very high similarities between these O-processing gene clusters (23) and their conventional serological cross-reactions. One isolate, SSI_AA003 (O134:H52), contained an O46 *wzy* variant along with its O134 *wzx*. However, as the O46 *wzy* was incomplete and another O134 isolate in the data set, 4370-53, contained O134 variants of both *wzx* and *wzy*, and because both O46 isolates in the data set SSI_AA018 and SSI_AA031 contained O46 variants of both genes, we did not expect O134 to be a problem. This illustrates the importance of determining whether the detected genes are of a reasonable size for a trustworthy prediction.

The 12 O128ab/ac variants were indistinguishable by SerotypeFinder, which predicted O group O128 only. This was not surprising, as we had also observed that the O128ab and ac variants clustered closely together for both *wzx* and *wzy*. For the other O groups with ab/ac variations, we could not estimate if SerotypeFinder could distinguish between O18ab/O18ac and O141ab/O141ac, as the isolates available for these O types were either used for constructing the database or were not designated ab/ac. However, within both groups, the ab/ac variants were highly identical, and we were unable to identify the genetic determinants for ab/ac variations. For O125ab/ac, for which the gene variants were also very similar, the one isolate (C418-89) available that was not part of the database was predicted correctly as O125ac from both *wzx* and *wzy* gene variants. For these four O groups (O18ab/ac, O125ab/ac, O128ab/ac, and O141ab/ac), conventional typing at the WHOCC has not indicated significant clinical or epidemiological differences. In combination with the observed similarities between the *wzx* and *wzy* sequences, we see no reason to maintain the distinction between these ab/ac variants.

For O28ab/ac and O112ab/ac, for which the ab/ac *wzx* and *wzy* genes were clearly different, the correct variation was assigned for the one available isolate of each O type, and the distinction between these ab/ac variants should be maintained.

Regarding the H-type prediction, there were five disagreements among the 508 isolates used in the validation. The reasons for the differences were suspected to be original transcription errors (W27, U19-41, and RN587/1), as a result of either typing mistakes (U19-41), original difficulties in typing (RN587/1) (29), or nonmotility (W27). One isolate, F7902-41, was predicted to be H17 based on its *fliC* variant (AJ605766), while according to conventional typing, it was H4. Three H17 isolates included in the validation were correctly predicted to be H17 by SerotypeFinder but with regard to the H17 *fliC* variant (AJ515904) and not the *flnA17* gene. We thus speculate that the H17 *fliC* variant, AJ605766, should be designated H4. To confirm this, more H17 strains need to be examined. In contrast to previous descriptions, where H17 is determined by *flnA17*, we did not detect the *flnA17* gene in any of the three H17 isolates. This is in agreement with findings by Beutin et al., where the *flnA* was not detected in the H17 reference strain, P12b (30).

For the discrepancies in O types, we observed 11 in total. For one isolate, F8858-41 (O19a), we suspected a mix-up in the sequence data we had received, since this isolate was predicted to be O133 by SerotypeFinder. The mix-up was confirmed by retyping of the isolate, which was negative for O19a but positive for O26, O50, and O133. However, the final O type was not determined because of the absence of absorbed antisera. Another discrepancy was observed for the reference isolate HW35 (O11:Kne:H33), which was predicted to be O99 by SerotypeFinder. For this isolate, we suspected an original typing or transcription error, as retyping confirmed the O99 type detected by SerotypeFinder.

The isolate SSI_AA037 (O20:K67:H17) was predicted to be O128 by SerotypeFinder. However, we suspect that the now-obsolete K67 (B) antigen, which was originally described in the reference strain (Cigleris) for O128ab:(K67):H2, may be processed by the *wzx* and *wzy* genes, resulting in a neutral capsule-like polysaccharide. K67 was originally described as positive by agglutination of live culture by use of an OK O128ab antiserum and no agglutination of the live culture in O128ab antiserum. However, we chose not to reexamine the strain due to the lack of reproducibility of this historic serotyping procedure.

For the isolate, 53638 (O144), which was predicted to be O124:H30 by SerotypeFinder, we suspected that the conventional type assigned was incorrect, as O144 has previously been described as harboring only H25, while O124 has been known to carry H7, H30, or H32 (31). However, as the genome sequence and the serotype information were collected from the NCBI, we were not able to check for possible of incorrect entries or to confirm all isolates by conventional serotyping of the isolate.

For the isolate 178-54(1986-54), which was O129:K-:H11 according to the conventional serotyping, SerotypeFinder predicted the O type to be O13/O135. After reserotyping, the conventional O129 serotype was confirmed. Since O13/O135 and O129 are also almost identical with regard to both *wzx* and *wzy*, we conclude that they cannot be distinguished by this method, and the prediction should thus be reported as O13/O129/O135 by SerotypeFinder.

The isolate SSI_AA099 was O43:H46 according to conventional serotyping and was predicted to be O131 by SerotypeFinder. Serological retesting of this isolate confirmed it as O131.

SSI_AA095 was predicted to be O17 by SerotypeFinder, although according to conventional typing, it was O44, which was confirmed by retesting.

Three additional discrepancies detected in isolates C2139-99, SE11, and SSI_AA017 remain unresolved, while one isolate, F10524-41, is being retyped serologically and is still under investigation, as the antigen is rough.

Several of the reference strains for the H antigens have been described as having two flagellar genes, usually *fliC* together with a non-*fliC* gene. Our SerotypeFinder tool confirmed previously described combinations in reference strains, such as *fliC21* with *flkA47*, *fliC38* with *fllA55*, and *fliC21* with *flmA54*. In addition, SerotypeFinder identified 15 previously nonobserved combinations with coexisting flagellar genes in 22 of the strains in the validation set (Table 7). Of interest is the observation that in all the strains examined, the non-*fliC* genes were the ones that were phenotypically expressed. These findings confirm the diversity of the flagellin gene pool in the *E. coli* population and indicate that DNA methods that are limited with regard to known targets may either misidentify or completely fail to identify the flagellin-encoding gene. This is evident for restriction fragment length polymorphism (RFLP) typing of *fliC*, where H3 strains are typed as *fliC16* in both the systems of Fields et al. (32), *fliC*(F), and Machado et al. (33), *fliC*(M); H4 is typed as *fliC4*/*fliC44*/negative by *fliC*(F); H4

and H44 are typed as *fliC17* by *fliC*(M); H17, H53, and H54 are typed as negative by *fliC*(F); and finally, *fliC*(F) has identical RFLP patterns for H38 and H43 strains (34).

RFLP is one of the most successful molecular typing tools for predicting the H type (32, 33) and especially combining the two RFLP typing methods (34). However, with *in silico* serotyping by SerotypeFinder, all the known flagellin genes can be detected, and the H type can be correctly predicted.

From this study, it has become evident that serotype information can be extracted from WGS data and lead to rapid serotype prediction comparable to that obtained with conventional serotyping. The tool is now available for complete WGS-based routine typing and surveillance of *E. coli*. It is less costly and more rapid than the conventional serotyping and is thus an advantageous alternative to the current routine.

We previously presented VirulenceFinder for detection of *E. coli* virulence genes from WGS data and demonstrated that real-time routine typing and surveillance of VTEC infections can be performed as an advantageous alternative to conventional typing (18). For further improvement of the subtyping of *stx* performed by the VirulenceFinder, we have now provided an additional database for subtyping of *stx* genes based on the entire *stx* holotoxin sequences according to the nomenclature described by Scheutz et al. (35).

With this study, we have shown that O and H typing can be performed from WGS data. We have covered all known H types with our validation, each represented by at least three isolates and also validated on a large number of different O types. At the present stage, we do not feel that we have the definitive basis and understanding of the relatively few challenges posed by the molecular typing of some of the O types. Further validation of the molecular basis for O typing as well as the development of more definitive principles for establishment of new O and H types will be continued as part of an ongoing international collaboration.

Our validation demonstrates that the results obtained from *in silico* serotyping using WGS are already superior to the conventional methods in terms of reproducibility and discriminatory and definitive power.

ACKNOWLEDGMENTS

We are very grateful to Susanne Jespersen, Pia Møller Hansen, and Christian Vråby Pedersen for superb technical assistance. We extend our sincere thanks to Rebecca Lindsey and Peter Gerner-Smidt from the Centers for Disease Control and Prevention (CDC), Atlanta, GA, USA, for sharing WGS data and serotype information with us and for taking an interest in our tool. We thank Karen Leth Nielsen, SSI, Denmark, and Yvonne Agersø, National Food Institute, DTU, Denmark, for supplying isolates for validation of SerotypeFinder.

We have no conflicts of interest to declare.

REFERENCES

1. **Kaper JB.** 2005. Pathogenic *Escherichia coli*. Int J Med Microbiol **295:**355–356. http://dx.doi.org/10.1016/j.ijmm.2005.06.008.
2. **Kauffmann F.** 1947. The serology of the *E. coli* group. J Immunol **57:**71–100.
3. **Ørskov F, Ørskov I.** 1984. Serotyping of *Escherichia coli*. Methods Microbiol **14:**43–112.
4. **Ørskov I, Ørskov F, Jann B, Jann K.** 1977. Serology, chemistry, and genetics of O and K antigens of *Escherichia coli*. Bacteriol Rev **41:**667–710.
5. **Scheutz F, Cheasty T, Woodward D, Smith HR.** 2004. Designation of O174 and O175 to temporary O groups OX3 and OX7, and six new *E. coli* O groups that include verocytotoxin-producing *E. coli* (VTEC): O176, O177, O178, O179, O180 and O181. APMIS **112:**569–584. http://dx.doi.org/10.1111/j.1600-0463.2004.apm1120903.x.
6. **Ørskov I, Ørskov F, Bettelheim KA, Chandler M.** 1975. Two new *Escherichia coli* O antigens, O162 and O163, and one new H antigen, H56. Withdrawal of H antigen H50. Acta Pathol Microbiol Scand B **83:**121–124.
7. **Whitfield C, Roberts IS.** 1999. Structure, assembly and regulation of expression of capsules in *Escherichia coli*. Mol Microbiol **31:**1307–1319. http://dx.doi.org/10.1046/j.1365-2958.1999.01276.x.
8. **Whitfield C, Keenleyside WJ, Clarke BR.** 1994. Structure, function and synthesis of surface polysaccharides in *Escherichia coli*, p 437–494. *In* Gyles CL (ed), *Escherichia coli* in domestic animals and humans. CAB International, Wallingford, United Kingdom.
9. **Wang L, Wang Q, Reeves PR.** 2010. The variation of O antigens in Gram-negative bacteria, p 123–152. *In* Wang X, Quin PJ (ed), Endotoxins: structure, function and recognition. Subcellular biochemistry, vol. 53. Springer Science+Business Media, Dordrecht, the Netherlands.
10. **Joys TM.** 1988. The flagellar filament protein. Can J Microbiol **34:**452–458. http://dx.doi.org/10.1139/m88-078.
11. **Macnab RM.** 1987. Flagella, p 70–83. *In* Neidhardt FC, Ingraham JL, Low KB, Magasanik B, Schaecter M, Umbarger HE (ed), *Escherichia coli* and *Salmonella typhimurium*: cellular and molecular biology, vol 1. American Society for Microbiology, Washington, DC.
12. **Feng L, Liu B, Liu Y, Ratiner YA, Hu B, Li D, Zong X, Xiong W, Wang L.** 2008. A genomic islet mediates flagellar phase variation in *Escherichia coli* strains carrying the flagellin-specifying locus *flk*. J Bacteriol **190:**4470–4477. http://dx.doi.org/10.1128/JB.01937-07.
13. **Ratiner YA.** 1998. New flagellin-specifying genes in some *Escherichia coli* strains. J Bacteriol **180:**979–984.
14. **Ratiner YA, Sihvonen LM, Liu Y, Wang L, Siitonen A.** 2010. Alteration of flagellar phenotype of *Escherichia coli* strain P12b, the standard type strain for flagellar antigen H17, possessing a new non-*fliC* flagellin gene *flnA*, and possible loss of original flagellar phenotype and genotype in the course of subculturing through semisolid media. Arch Microbiol **192:**267–278. http://dx.doi.org/10.1007/s00203-010-0556-x.
15. **Tominaga A.** 2004. Characterization of six flagellin genes in the H3, H53 and H54 standard strains of *Escherichia coli*. Genes Genet Syst **79:**1–8. http://dx.doi.org/10.1266/ggs.79.1.
16. **Wang L, Rothemund D, Curd H, Reeves PR.** 2003. Species-wide variation in the *Escherichia coli* flagellin (H-antigen) gene. J Bacteriol **185:**2936–2943. http://dx.doi.org/10.1128/JB.185.9.2936-2943.2003.
17. **Ratiner YA.** 1983. Presence of 2 structural genes determining antigenically different phase-specific flagellins in some *Escherichia coli* strains. FEMS Microbiol Lett **19:**37–41. http://dx.doi.org/10.1111/j.1574-6968.1983.tb00506.x.
18. **Joensen KG, Scheutz F, Lund O, Hasman H, Kaas RS, Nielsen EM, Aarestrup FM.** 2014. Evaluation of real-time WGS for routine typing, surveillance and outbreak detection of verotoxigenic *Escherichia coli*. J Clin Microbiol **52:**1501–1510. http://dx.doi.org/10.1128/JCM.03617-13.
19. **Hazen TH, Sahl JW, Fraser CM, Donnenberg MS, Scheutz F, Rasko DA.** 2013. Refining the pathovar paradigm via phylogenomics of the attaching and effacing *Escherichia coli*. Proc Natl Acad Sci U S A **110:**12810–12815. http://dx.doi.org/10.1073/pnas.1306836110.
20. **Trees E, Strockbine N, Changayil S, Ranganathan S, Zhao K, Weil R, MacCannell D, Sabol A, Schmidtke A, Martin H, Stripling D, Ribot EM, Gerner-Smidt P.** 2014. Genome sequences of 228 Shiga toxin-producing *Escherichia coli* isolates and 12 isolates representing other diarrheagenic *E. coli* pathotypes. Genome Announc **2:**e00718-14. http://dx.doi.org/10.1128/genomeA.00718-14.
21. **Martin M.** 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet J **17:**10–12.
22. **Zerbino DR, Birney E.** 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res **18:**821–829. http://dx.doi.org/10.1101/gr.074492.107.
23. **Iguchi A, Iyoda S, Kikuchi T, Ogura Y, Katsura K, Ohnishi M, Hayashi T, Thomson NR.** 2015. A complete view of the genetic diversity of the *Escherichia coli* O-antigen biosynthesis gene cluster. DNA Res **22:**101–107. http://dx.doi.org/10.1093/dnares/dsu043.
24. **Edgar RC.** 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res **32:**1792–1797. http://dx.doi.org/10.1093/nar/gkh340.
25. **Saitou N, Nei M.** 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol **4:**406–425.
26. **Larsen MV, Cosentino S, Rasmussen S, Friis C, Hasman H, Marvig RL,**

Jelsbak L, Sicheritz-Ponten T, Ussery DW, Aarestrup FM, Lund O. 2012. Multilocus sequence typing of total-genome-sequenced bacteria. J Clin Microbiol **50:**1355–1361. http://dx.doi.org/10.1128/JCM.06094-11.
27. **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. http://dx.doi.org/10.1093/jac/dks261.
28. **Wyres KL, Conway TC, Garg S, Queiroz C, Reumann M, Holt K, Rusu LI.** 2014. WGS analysis and interpretation in clinical and public health microbiology laboratories: what are the requirements and how do existing tools compare? Pathogens **3:**437–458. http://dx.doi.org/10.3390/pathogens3020437.
29. **Blank TE, Lacher DW, Scaletsky ICA, Zhong H, Whittam TS, Donnenberg MS.** 2003. Enteropathogenic *Escherichia coli* O157 strains from Brazil. Emerg Infect Dis **9:**113–115. http://dx.doi.org/10.3201/eid0901.020072.
30. **Beutin L, Strauch E, Zimmermann S, Kaulfuss S, Schaudinn C, Mannel A, Gelderblom HR.** 2005. Genetical and functional investigation of *fliC* genes encoding flagellar serotype H4 in wildtype strains of *Escherichia coli* and in a laboratory *E. coli* K-12 strain expressing flagellar antigen type H48. BMC Microbiol **5:**4. http://dx.doi.org/10.1186/1471-2180-5-4.
31. **Tozzoli R, Scheutz F.** 2014. Diarrhoeagenic *Escherichia coli* infections in humans, p 1–18. *In* Morabito S (ed), Pathogenic *Escherichia coli*. Caister Academic Press, Norfolk, United Kingdom.
32. **Fields PI, Bloom K, Hughes HJ, Helsel LO, Feng P, Swaminathan B.** 1997. Molecular characterization of the gene encoding H antigen in *Escherichia coli* and development of a PCR-restriction fragment length polymorphism test for identification of *E. coli* O157:H7 and O157:NM. J Clin Microbiol **35:**1066–1070.
33. **Machado J, Grimont F, Grimont PAD.** 2000. Identification of *Escherichia coli* flagellar types by restriction of the amplified *fli*C gene. ResMicrobiol **151:**535–546.
34. **Prager R, Strutz U, Fruth A, Tschape H.** 2003. Subtyping of pathogenic *Escherichia coli* strains using flagellar (H)-antigens: serotyping versus *fliC* polymorphisms. Int J Med Microbiol **292:**477–486. http://dx.doi.org/10.1078/1438-4221-00226.
35. **Scheutz F, Teel LD, Beutin L, Pierard D, Buvens G, Karch H, Mellmann A, Caprioli A, Tozzoli R, Morabito S, Strockbine NA, Melton-Celsa AR, Sanchez M, Persson S, O'Brien AD.** 2012. Multicenter evaluation of a sequence-based protocol for subtyping Shiga toxins and standardizing Stx nomenclature. J Clin Microbiol **50:**2951–2963. http://dx.doi.org/10.1128/JCM.00860-12.
36. **Plainvert C, Bidet P, Peigne C, Barbe V, Medigue C, Denamur E, Bingen E, Bonacorsi S.** 2007. A new O-antigen gene cluster has a key role in the virulence of the *Escherichia coli* meningitis clone O45:K1:H7. J Bacteriol **189:**8528–8536. http://dx.doi.org/10.1128/JB.01013-07.