Matrix-Assisted Laser Desorption-Ionization Time-of-Flight mass spectrometry using a custom-made database, biomarker assignment or mathematical classifiers does not differentiate *Shigella spp.* and *Escherichia coli*

Authors: Maaike J.C. van den Beld[1,2](orcid:0000-0001-8720-8434), John W.A. Rossen[2](orcid: 0000-0002-7167-8623), Arie Evers[1], A.M.D. (Mirjam) Kooistra-Smid[2,3](orcid:0000-0002-8856-581X), Frans A.G. Reubsaet[1]

Affiliations:

[1]Infectious Disease Research, Diagnostics and laboratory Surveillance, Centre for Infectious Disease Control National Institute for Public Health and the Environment (RIVM), , Bilthoven, the Netherlands

[2]Department of Medical Microbiology and Infection Prevention, University of Groningen, University Medical Center Groningen, Groningen, the Netherlands

[3]Department of Medical Microbiology, Certe, Groningen, the Netherlands

Corresponding author: Maaike van den Beld

Email: maaike.van.den.beld@rivm.nl

Abstract

Purpose: *Shigella spp.* and *E. coli* are closely related, and cannot be distinguished using Matrix-Assisted Laser Desorption-Ionization Time-of-Flight mass spectrometry (MALDI-TOF MS) with commercially available databases. Here, three alternative approaches using MALDI-TOF for identification and distinction of *Shigella spp*., *E. coli* and its pathotype EIEC were explored.

Methods: A custom-made database was developed, biomarkers were assigned and classification models using machine learning were designed and evaluated using spectra of 456 *Shigella spp*., 42 *E. coli* and 61 EIEC isolates, obtained by the direct smear method and the ethanol-formic acid extraction method.

Results: Identification with a custom-made database resulted in >94% *Shigella* identified at the genus level, and >91% *S. sonnei* and *S. flexneri* at the species level, but distinction of *S. dysenteriae*, *S. boydii* and *E. coli* was poor. Moreover, 10-15% of duplicates rendered discrepant results. With biomarker assignment, 98% *S. sonnei* isolates were correctly identified, although the *S. sonnei* biomarkers were not specific as other species were also identified as *S. sonnei*. Discriminating markers for *S. dysenteriae*, *S. boydii*, and *E. coli* were not assigned at all. Classifiers identified *Shigella* in 96% of isolates correctly, but most *E. coli* isolates were also assigned to *Shigella.*

Conclusion: None of the proposed alternative approaches is suitable for use in clinical diagnostics for the identification of *Shigella spp*., *E. coli* and EIEC because of their poor distinctive properties. We suggest the use of MALDI-TOF MS for identification of the *Shigella spp.*/*E. coli* complex, but other tests should be used for distinction.

Keywords: MALDI-TOF, *Shigella spp*., *Escherichia coli*, custom-made database, biomarker assignment, machine-learning classifiers

Introduction

The *E. coli* pathotype entero-invasive *E. coli* (EIEC) is thought to cause the same disease as *Shigella spp.* [1]. This pathotype consists of isolates that possess some of the *E. coli* phenotypic characteristics, but has also the invasive nature of *Shigella spp.* [2,3]. EIEC harbors the same virulence markers as *Shigella spp.* that are used in molecular diagnostics for detection of both *Shigella spp.* and EIEC, but are not suitable to distinguish them [4]. Differentiation of *E. coli* from *Shigella spp.* is historically performed using phenotypical tests, serotyping and the determination of virulence markers using PCR [5,6].

Currently, most clinical laboratories use Matrix-Assisted Laser-Desorption Ionization Time-of-Flight Mass Spectrometry (MALDI-TOF MS) for the identification of bacteria in a routine diagnostic setting. Commercially available databases, as MALDI biotyper® in combination with the MALDI Security-Relevant (SR) library® (Bruker Daltonik GmbH, Bremen, Germany) and VITEK® MS (BioMérieux, Marcy-l'Etoile, France) are able to distinguish *Shigella spp.* and *E. coli* from other *Enterobacteriaceae*. However, they are not able to distinguish between the different *Shigella* species and *E. coli*, including the EIEC pathotype [7].

The development of custom-made databases for identification of bacteria using MALDI-TOF MS as alternative to commercially available databases, proved successful for multiple species before [8-10]. Most notably, an earlier study developed a custom-made database for identification and distinction of *Shigella spp.* and *E. coli* specifically, however EIEC isolates were not included in their database [11]. Using a database approach, comparisons of unknown isolates to isolates in a database comprise the whole spectra for species identification. However, for closer related groups a more subtle approach can be essential in which variations within the spectra are examined in the form of presence or absence of specific peaks as biomarkers [12,13]. The biomarker approach was used to type *E. coli* isolates before,

with varying success rates [13]. These approaches mainly targeted the pathotype entero- hemorrhagic *E. coli* (EHEC) or the highly virulent ST131 clone [13], although two studies used biomarker typing specifically for *Shigella spp.* and *E. coli*, without EIEC isolates [14,15]. One of those studies identified biomarkers outside the mass range of 2,000-20,000 Da, which is used in routine applications [15], and the other did not specify in which species the biomarkers were present or absent [14]. Besides the determination of the presence or absence of single biomarkers, patterns of these biomarkers can be investigated and recognized with machine-learning algorithms [16]. These machine-learning based methods can be used to establish classifiers for identification of groups within species of bacteria [17,18]. Moreover, these classifiers were developed for the identification of *Shigella spp*. and *E. coli* before, although EIEC isolates were not included [14].

In this study, the ability of MALDI-TOF MS was assessed for the distinction of the four *Shigella* species, EIEC and non-invasive *E. coli* using alternatives for the commercially available databases. First, a custom-made database, including all species of *Shigella*, *E. coli*, including EIEC isolates, was developed and evaluated. Second, biomarkers were assigned and evaluated, and third, classifier models based on machine learning were defined, applied and evaluated.

Material and Methods

Bacterial isolates

A total of 559 isolates consisting of 36 *S. dysenteriae*, 156 *S. flexneri*, 32 *S. boydii*, 232 *S. sonnei*, 61 EIEC, and 42 other *E. coli* of human and animal origin comprising phylogroups A, B1, B2 and D [19] was used (Table 1). All isolates, except the references, were identified using a previously described culture-based identification algorithm [20]. They were divided into a set of training isolates (n=288) and a set of test isolates (n=271), both having similar species and serotype distributions. The training set was used to construct the custom-made database, assign biomarkers, and for defining and training of

machine-learning classifier models. The test set was used to test all of these algorithms in duplicate, with both direct smear and ethanol-formic acid extraction application methods.

MALDI-TOF preparation of isolates

All isolates were grown overnight on Columbia Sheep Agar (CSA, Biotrading, Mijdrecht, the Netherlands) at 37°C and were subsequently subjected to the direct smear method as well as to the ethanol- formic acid extraction with silica beads as previously described [21]. Colonies or 1 µl extract were applied onto a polished steel plate, air-dried and overlaid with 1 µl α-Cyano-4-hydroxycinnamic acid in 50% acetonitrile-2.5% trifluoroacetic acid (HCCA matrix). The samples were analyzed using a Bruker Microflex LT (Bruker Daltonik GmbH, Bremen, Germany) in a linear positive mode, with 30-40% laser power and within a mass-range of 2,000-20,000 Da.

The isolates from the training set were used to produce Main Spectrum Profiles (MSPs) according to manufacturer's instructions, using FlexControl V3.4 (Bruker Daltonik). The isolates from the training set were analyzed using Maldi Biotyper 3.0 Real Time Classification (RTC, Bruker Daltonik).

Database development

The MSPs produced from 288 isolates in the training set were used to build a custom-made database with Maldi Biotyper OC V3.1.66 (Bruker Daltonik). In addition, a dendrogram to assess the relatedness of these MSPs was inferred using default settings. The isolates in the test set were identified using this custom-made database. Additionally, the test isolates were also identified using the commercially available Bruker MALDI Biotyper database (V8.0.0.0) and the Bruker Security-Relevant Library (V1.0.0.0), and using a combination of the commercial and custom-made databases. Quality of the results was indicated by a log-score, calculated by Maldi Biotyper 3.0 RTC: a log-score of 2.000-2.300 corresponds to "secure genus identification, probable species identification" and a log-score of >2.300 corresponds to a "highly probable species identification". Both duplicate spots were analyzed, the

Table 1 Isolates used in this study, divided in training set and test set

Species and serotype/O-type	Training set		Test set	
	n	**Origin**	**n**	**Origin**
***S. dysenteriae* serotype 1**	2	CIP 57.28[T]; A1	1	1 ci[a]
***S. dysenteriae* serotype 2**	5	A2, 4 ci[a]	4	4 ci[a]
***S. dysenteriae* serotype 3**	5	AMC-43-G-93; 4 ci[a]	3	3 ci[a]
***S. dysenteriae* serotype 4**	2	AMC 43-G-86; 1 ci[a]	0	
***S. dysenteriae* serotype 5**	1	AMC 43-G-84	0	
***S. dysenteriae* serotype 6**	1	AMC 43-G-81	1	1 ci[a]
***S. dysenteriae* serotype 7**	1	AMC 43-G-76	1	1 ci[a]
***S. dysenteriae* serotype 9**	2	A58: 1646; 1 ci[a]	1	1 ci[a]
***S. dysenteriae* serotype 10**	1	A2050-52	0	
***S. dysenteriae* serotype 12**	2	2 ci[a]	1	1 ci[a]
***S. dysenteriae* serotype 14**	1	NCTC 11867	0	
***S. dysenteriae* serotype 15**	1	NCTC 11868	0	
Total number of *S. dysenteriae*	**24**		**12**	
***S. flexneri* serotype 1a**	3	B1A; 2 ci[a]	0	
***S. flexneri* serotype 1b**	5	B1B; 4 ci[a]	5	5 ci[a]
***S. flexneri* serotype 1c**	4	4 ci[a]	3	3 ci[a]
***S. flexneri* serotype 2a**	32	CIP 82.48[T]; B2A; 30 ci[a]	32	32 ci[a]
***S. flexneri* serotype 2b**	1	B2B	3	3 ci[a]
***S. flexneri* serotype 3a**	2	B3A; 1 ci[a]	14	14 ci[a]
***S. flexneri* serotype 3b**	2	B3B; B3C	3	3 ci[a]
***S. flexneri* serotype 4a**	1	B4A	4	4 ci[a]
***S. flexneri* serotype 4av**	4	5 ci[a]	0	
***S. flexneri* serotype 4b**	1	B4B	0	
***S. flexneri* serotype 4c**	3	3 ci[a]	0	
***S. flexneri* serotype 5b**	1	B5	1	1 ci[a]
***S. flexneri* serotype 6**	10	B6; 9 ci[a]	10	10 ci[a]
***S. flexneri* serotype X**	1			
***S. flexneri* serotype Y**	2	2 ci[a]	2	2 ci[a]
***S. flexneri* serotype Yv**	2	2 ci[a]	0	
***S. flexneri* provisional**	5	5 ci[a]	0	
Total number of *S. flexneri*	**79**		**77**	
***S. boydii* serotype 1**	2	AMC-43-G-58; 1 ci[a]	3	3 ci[a]
***S. boydii* serotype 2**	3	CIP 82.50[T]; P288; 1 ci[a]	4	4 ci[a]
***S. boydii* serotype 3**	1	D1	0	
***S. boydii* serotype 4**	2	AMC-43-G-63; 1 ci[a]	2	2 ci[a]
***S. boydii* serotype 5**	2	P143; 1 ci[a]	0	
***S. boydii* serotype 6**	1	CDC 9771 (D19)	0	
***S. boydii* serotype 7**	1	AMC 4006 (Lavington)	0	
***S. boydii* serotype 8**	0		1	1 ci[a]
***S. boydii* serotype 9**	1	1296/7	0	
***S. boydii* serotype 10**	1	430	1	1 ci[a]
***S. boydii* serotype 11**	1	34	0	
***S. boydii* serotype 12**	0		1	1 ci[a]
***S. boydii* serotype 13**	0		1	1 ci[a]
***S. boydii* serotype 14**	0		1	1 ci[a]
***S. boydii* serotype 15**	1	CDC C-703	0	
***S. boydii* serotype 18**	1	1 ci[a]	1	1 ci[a]
Total number of *S. boydii*	**17**		**15**	
S. sonnei	**117**	CIP 82.49[T]; 116 ci[a]	**115**	115 ci[a]
EIEC	**30**	DSM 9027; DSM 9028; CCUG 11335; CCUG 38080; CCUG 38092; CCUG 38093; EW227; 1624-56; 1184-68; 145/46; L119B-10; 19 ci[a]	**31**	31 ci[a]
Other *E. coli* pathotypes (human)	**11**	7 STEC ci[a], 4 EPEC ci[a]	**11**	8 STEC ci[a]; 3 EPEC ci[a]
Other *E. coli* pathotypes (animal)[b]	**10**	5 mussel, 3 pigeon, 2 turkey	**10**	4 mussel, 3 pigeon, 2 turkey, 1 oyster

[a]ci = clinical isolate. [b]isolated from animals, all other numbers are reference isolates.

highest log-score of at least 2.000 was considered as the definitive MALDI-TOF identification, as is done in a routine workflow. If an isolate had a log-score < 2.000, it was disregarded from further analysis. Isolates were then assigned to different discrimination levels “genus”, “pathotype”, “group” and “species” as displayed in Figure 1.

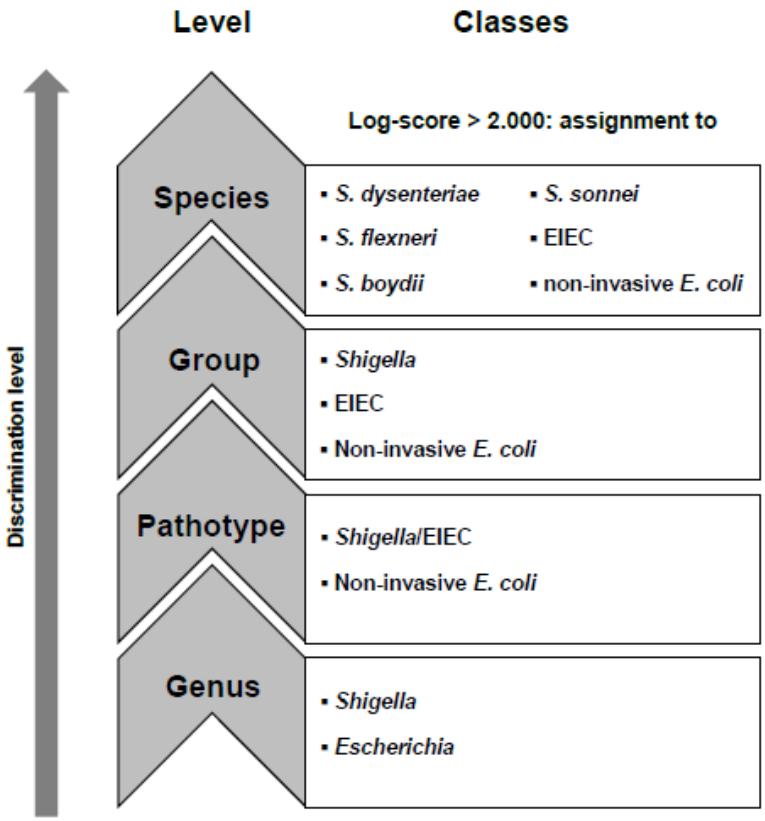

Figure 1 The classes in the different discrimination levels to which isolates were assigned

For accurate identification, only matches with database MSPs from the same species within a log-score range of 2.000-2.300 or >2.300 should be expected in one spot. To assess this, the ten MSPs from the database that produced the highest scores within a log-score of 2.000-2.300 or >2.300 per spot were determined. For each species identified with the culture dependent identification algorithm, the median number of species and quartile ranges per spot that had a log-score of 2.000-2.300 or >2.300 were calculated and visualized using SPSS 24.0.0.1.

Biomarker assignment and principal component analysis

Spectra files from MSPs of 288 isolates in the training set were exported as mzXML files using Compassxport CXP3.0.5. (Bruker Daltonik) or exported via a batch process in Flexanalysis (Bruker Daltonik). A new database was created in Bionumerics v7.6.3 (Applied Maths NV, http://www.applied-maths.com/) according to the manufacturers' instructions. All raw spectra were imported into the Bionumerics database with x-axis trimming to a minimum of 2000 m/z. Baseline subtraction, noise computing, smoothing, baseline detection and peak detection were performed with default settings. Spectra summarizing, peak matching and peak assignment was performed according to instructions from Bionumerics [22]. In short, all raw spectra were summarized into isolate spectra. Peak matching was performed on isolate spectra using a constant tolerance of 1.9, a linear tolerance of 550 and a peak detection rate of 10%. Binary peak matching tables were exported to summarize the presence of peak classes on all discrimination levels as depicted in Figure 1. For the levels genus, pathotype and groups, decision diagrams were produced (Online Resource 1a, 1b and 1c). The spectra files of isolates from the test set were imported and preprocessed in Bionumerics, using the same methods and settings as with spectra from the isolates in the training set. Peak matching with a constant tolerance of 1.9 and a linear tolerance of 550 on spectra of test isolates was performed using the option "existing peak classes only" to compare the presence of peaks in the test isolates with the presence of peaks in the isolates from the training set. Decision diagrams (Online Resource 1) and the presence or absence of peak masses as depicted in Table 2 were applied to assign unknown isolates from the test set according to the different levels as shown in Figure 1.

By assigning biomarkers, only presence and absence of peaks was investigated. To assess also quantitative peak data such as peak intensity and peak area, a principal component analysis (PCA) was performed on all isolates in the training set to visualize the position of isolates in three dimensions.

Table 2 Discrimination scheme of biomarkers, percentage of isolates in the training set with presence of specific biomarkers

Biomarkers (m/z)		**2691**	**2877**	**3129**	**3636**	**3647**	**3930**	**3939**	**4163**	**4189**	**4368**	**4501**	**4769**	**4775**
S. dysenteriae	(n=24)	92	4	100	0	0	0	100	0	100	100	100	0	100
S. flexneri	(n= 46)	100	0	100	0	63	53	18	1	94	97	99	22	97
S. boydii	(n=17)	88	0	100	18	0	0	94	0	100	88	100	18	100
S. sonnei	(n=117)	56	49	59	22	0	1	56	17	89	68	56	23	98
EIEC	(n=31)	100	0	100	3	6	0	97	0	97	100	94	26	97
Other *E. coli*	(n=21)	52	24	52	71	0	5	62	38	90	67	57	67	100

Biomarkers (m/z)	**4784**	**5156**	**5239**	**5386**	**5415**	**6262**	**6322**	**6412**	**6488**	**7275**	**7295**	**7715**	**7868**	**787**
S. dysenteriae	100	100	8	92	52	100	100	0	0	0	0	0	0	100
S. flexneri	76	97	55	99	45	99	99	4	42	1	83	0	41	23
S. boydii	76	94	18	88	59	100	88	18	6	18	0	6	12	88
S. sonnei	69	86	27	73	0	74	62	13	1	17	0	28	18	75
EIEC	71	100	39	100	42	94	97	19	0	23	6	3	16	84
Other *E. coli*	19	86	0	67	0	57	67	48	10	52	0	24	43	33

Biomarkers (m/z)	**8326**	**8370**	**8379**	**9002**	**9227**	**9535**	**9546**	**9563**	**9739**	**10300**	**10310**	**10488**	**10934**
S. dysenteriae	0	0	100	100	0	0	100	100	0	0	100	0	0
S. flexneri	5	4	90	94	4	17	82	92	8	9	86	0	0
S. boydii	0	18	82	88	12	18	82	88	6	18	82	0	12
S. sonnei	15	15	82	38	12	16	85	54	15	13	70	34	36
EIEC	6	16	77	87	13	23	77	81	6	19	77	0	0
Other *E. coli*	43	38	38	24	48	48	62	38	43	48	48	0	5

Presence of biomarkers identified in previous studies

All isolates in the training set as well as in the test set were examined for the unique masses (± 500 ppm) that were found in biomarker assignment to *Shigella spp*. and *E. coli* in previous studies [15,14]. Additionally, because peaks in our study were assigned at m/z values instead of masses only and masses could be potentially charged with two electrons, this is corrected for by also examining the previously published masses divided by two (±500 ppm) [14,15].

Classifier models based on machine learning

Peak data of the summarized isolate spectra of the 288 isolates in the training set was used to define and train machine learning based classifiers using Bionumerics v7.6.3 according to the manufacturers' instructions. In short, peak matching with a constant tolerance of 1.9 and a linear tolerance of 550 was performed on isolate spectra on the different levels: genus, pathotype, group and species. Classifiers were created at all levels using character values. Support Vector Machine (linear) learning was used as

a scoring method in which p-values were used as a rank score. The classifiers were trained and cross-validated to check their performance for identification. Subsequently, the classifier models were used to classify the unknown isolates in the test set at the different discrimination levels to evaluate their performance.

Results

Database development

All MSPs of 288 training isolates were added to a custom-made database, the relatedness of these MSPs is shown in a dendrogram (Figure 2). The Maldi Biotyper OC software recognized three large clusters of MSPs that are not species-specific within this custom database. This did not change if clusters were assigned manually with a lower distance level at 50-100 relative units, indicating that similarity in spectrum profiles is distributed over the species level (Figure 2).

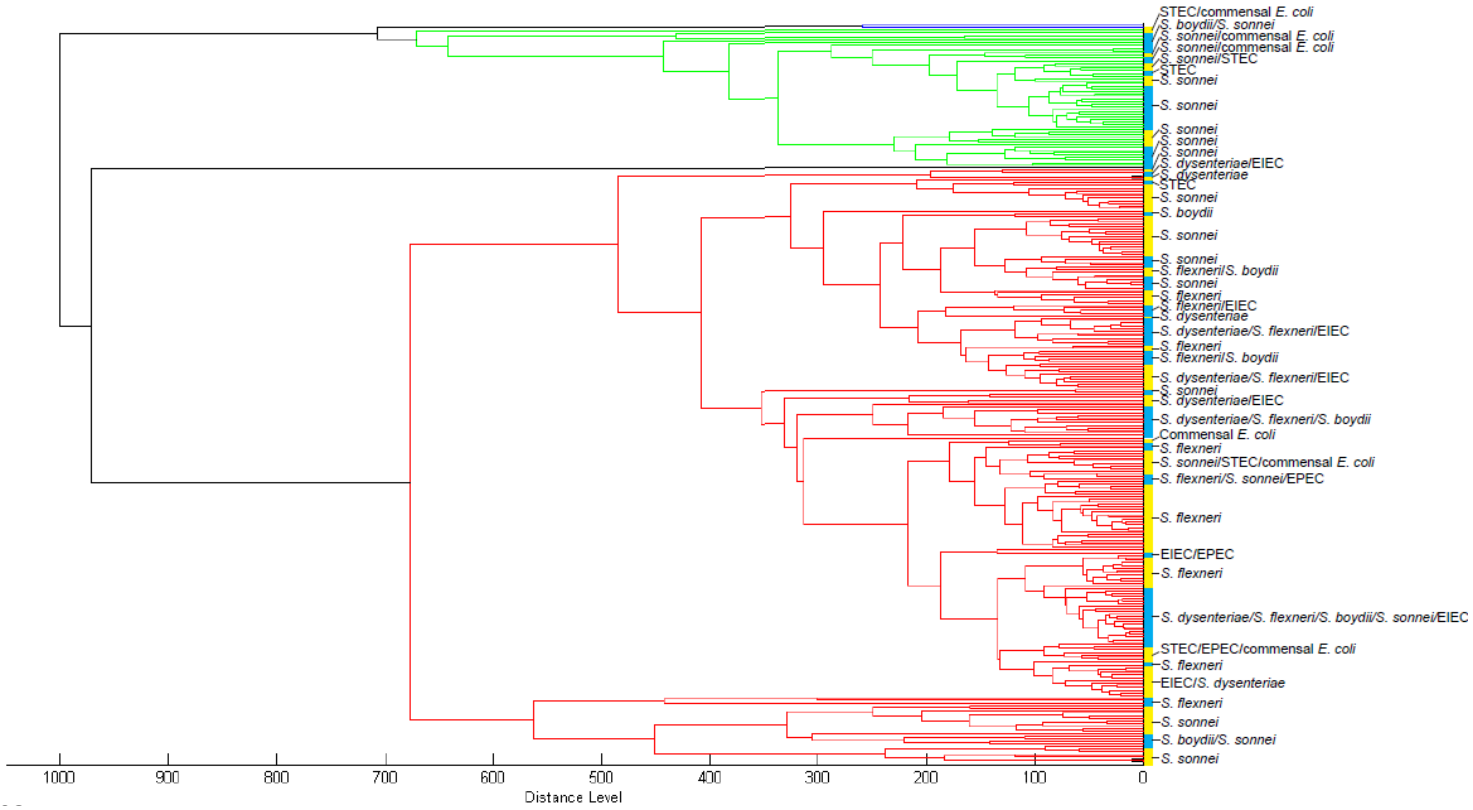

Figure 2 Dendrogram of MSPs of training isolates

Blue = cluster 1; green = cluster 2; red = cluster 3. Yellow/blue vertical band = manual cluster distinction at distance level 50-100 relative units with species designation using the culture-based identification algorithm.

Additionally, the duplicate spots of test isolates using either the direct smear or extraction method resulted in a different species designation in 10-15% of the samples. Furthermore, with an accurate distinction of species one would not expect assignment to multiple species above the threshold of log-score 2.000. However, with both application methods, most isolates were assigned to several species with a log-score of 2.000-2.300 or >2.300 per spot (Figure 3).

One isolate from the test set (*S. boydii* serotype 13) had a low-quality spectrum (log score 1.574-1.930) and one isolate *(S. dysenteriae* serotype 1) was originally incorrectly stored, as this isolate was identified as *Corynebacterium diphtheria*e using the Bruker databases. Both these isolates were ignored in further analyses. All other isolates had log-scores higher than 2.000, and percentages of

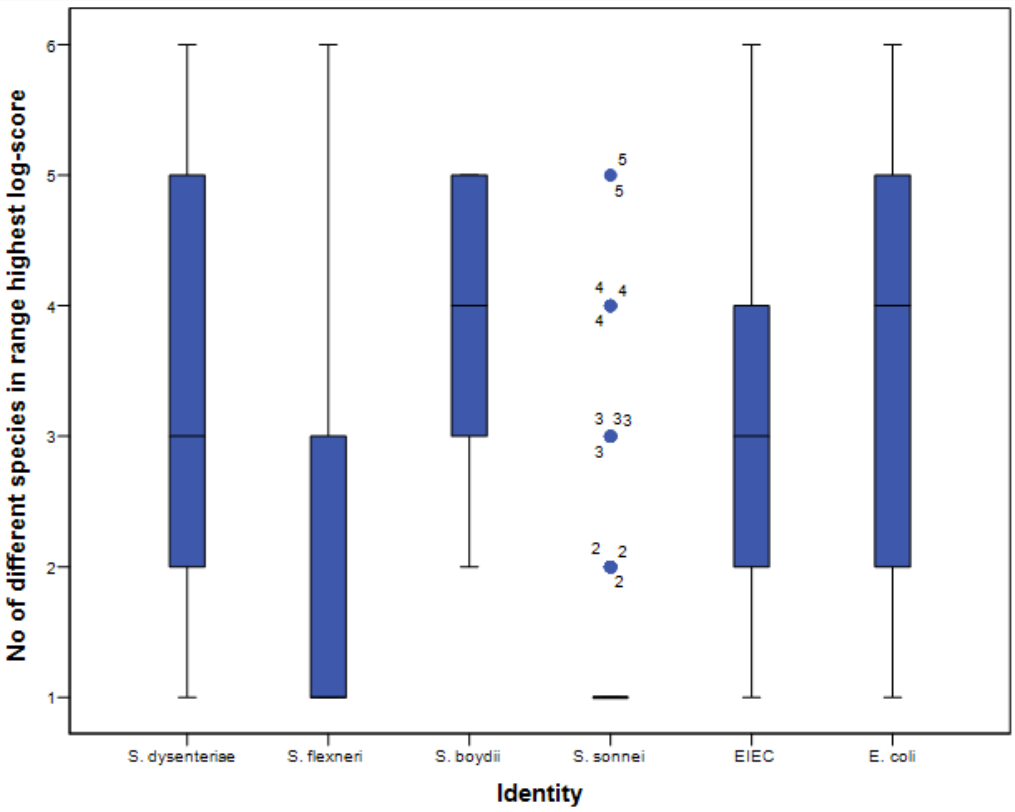

Figure 3 Number of different species in the first 10 matches per spot with the direct smear method
Identity (x-axis) was assigned using the culture-based identification algorithm. Black horizontal bars represent the median number of species, the 25%-75% interquartile ranges are indicated by the blue vertical bars, and 5%-95% intervals by the black vertical lines. Outliers are indicated with blue dots.

MALDI-TOF identification that were concordant with the original identification on all discrimination levels were as displayed in Table 3. With the Bruker databases only, percentages of correctly identified *Shigella spp.* on all discrimination levels is low, ranging from 6% to 45% correct designations, both for the direct smear method as for the extraction method (Table 3). In contrast, 90%-100% of *E. coli* isolates were correctly identified. When identification was based on the custom-made database with or without the Bruker databases, the percentage of correctly identified *E. coli* isolates decreased from 29%-71%, while *Shigella spp.* were correctly identified in 94%-99% of cases on the genus, pathotype and group levels. In addition, 91%-97% of *S. flexneri* and *S. sonnei* were correctly identified at the species level, in contrast to *S. dysenteriae* and *S. boydii*, for which the percentages of correct identification were poor (Table 3).

Biomarker assignment and principal component analysis

The decision diagrams based on biomarkers assigned to the isolates in the training set were used for identification of unknown isolates in the test set. Distinctive peaks on the species levels were summarized in Table 2. High percentages for correct identification of *S. sonnei* isolates were achieved at the species level using both the direct smear as the extraction method. However, the biomarkers are not specific for *S. sonnei* as other species also contain them. For other species the identified biomarkers correctly identified isolates below 38%. Specific biomarkers were not detected for all the classes at the different discrimination levels as depicted in Figure 1. Consequently, it was not possible to identify *S. dysenteriae, S. boydii,* and *E. coli* isolates at all, because of the absence of discriminating peaks for these species (Table 3).

In the PCA of the detected peaks in the isolates of the training set, one large cluster was formed, with at both ends a few outliers (Figure 4). If the isolates were colored according to their identity based on the culture-based identification method, in none of the discrimination levels separate groups of isolates were seen (Figure 4a-d).

Table 3 Correct identification results of isolates from test set

	Correct identification with MALDI-TOF, direct smear:					Correct identification with MALDI-TOF, ethanol-formic acid extraction:				
	Bruker databases[a], n (%)	**Custom-made database, n (%)**	**Bruker databases[a] + custom-made, n (%)**	**Biomarker assignment, n (%)**	**Classifier models, n (%)**	**Bruker databases[a], n (%)**	**Custom-made database, n (%)**	**Bruker databases[a]+ custom-made, n (%)**	**Biomarker assignment, n (%)**	**Classifier models, n (%)**
Genus										
Shigella (n =217)	19 (9)	205 (94)	205 (94)	10 (5)	209 (96)	12 (6)	207 (95)	205 (94)	15 (7)	217 (100)
E. coli (n = 52)	49 (94)	26 (50)	29 (56)	NA	11 (21)	47 (90)	35 (67)	37 (71)	NA	4 (8)
Unassigned[b]	2 (1)	1 (0.4)	3 (1)	257 (96)	0 (0)	1 (0.4)	0 (0)	3 (1)	250 (93)	0 (0)
Pathotype										
Shigella/EIEC (n = 248)	NA	233 (94)	241 (97)	217 (88)	145 (58)	NA	245 (99)	242 (98)	225 (92)	147 (59)
Other *E. coli* (n =21)	21 (100)	6 (29)	10 (48)	NA	14 (67)	21 (100)	11 (52)	13 (62)	NA	6 (29)
Unassigned[b]	2 (1)	1 (0.4)	3 (1)	46 (17)	0 (0)	0 (0)	0 (0)	3 (1)	27 (10)	0 (0)
Group										
Shigella (n = 217)	19 (9)	205 (94)	205 (94)	193 (89)	131 (60)	12 (6)	207 (95)	205 (94)	195 (90)	134 (62)
EIEC (n = 31)	NA	9 (29)	8 (26)	NA	2 (6)	NA	19 (61)	19 (61)	NA	0 (0)
Other *E. coli* (n =21)	21 (100)	6 (29)	10 (48)	NA	13 (62)	21 (100)	11 (52)	13 (62)	NA	7 (33)
Unassigned[b]	2 (1)	1 (0.4)	3 (1)	49 (23)	0 (0)	1 (0.4)	0 (0)	3 (1)	36 (13)	0 (0)
Species										
S. dysenteriae (n =11)	5 (45)	5 (45)	5 (45)	0 (0)	0 (0)	4 (36)	7 (64)	6 (55)	0 (0)	0 (0)
S. flexneri (n =77)	NA	70 (91)	70 (91)	24 (31)	6 (8)	NA	73 (95)	73 (95)	30 (39)	3 (4)
S. boydii (n =14)	NA	1 (7)	0 (0)	0 (0)	0 (0)	NA	0 (0)	0 (0)	0 (0)	0 (0)
S. sonnei (n = 115)	NA	110 (96)	110 (96)	113 (98)	92 (80)	NA	112 (97)	112 (97)	108 (94)	101 (88)
EIEC (n = 31)	NA	9 (29)	8 (26)	0 (0)	1 (3)	NA	19 (61)	19 (61)	1 (3)	3 (10)
Other *E. coli (n =21)*	21 (100)	6 (29)	10 (48)	0 (0)	12 (57)	21 (100)	11 (52)	13 (62)	0 (0)	4 (19)
Unassigned[b]	2 (1)	1 (0.4)	3 (1)	85 (32)	0 (0)	1 (0.4)	0 (0)	3 (1)	97 (36)	0 (0)

NA = not applicable, as no discriminating peaks were assigned to these classes. [a]Bruker MALDI Biotyper database (V8.0.0.0) and the Bruker Security-Relevant Library (V1.0.0.0). [b]Number of isolates that could not be assigned to a class.

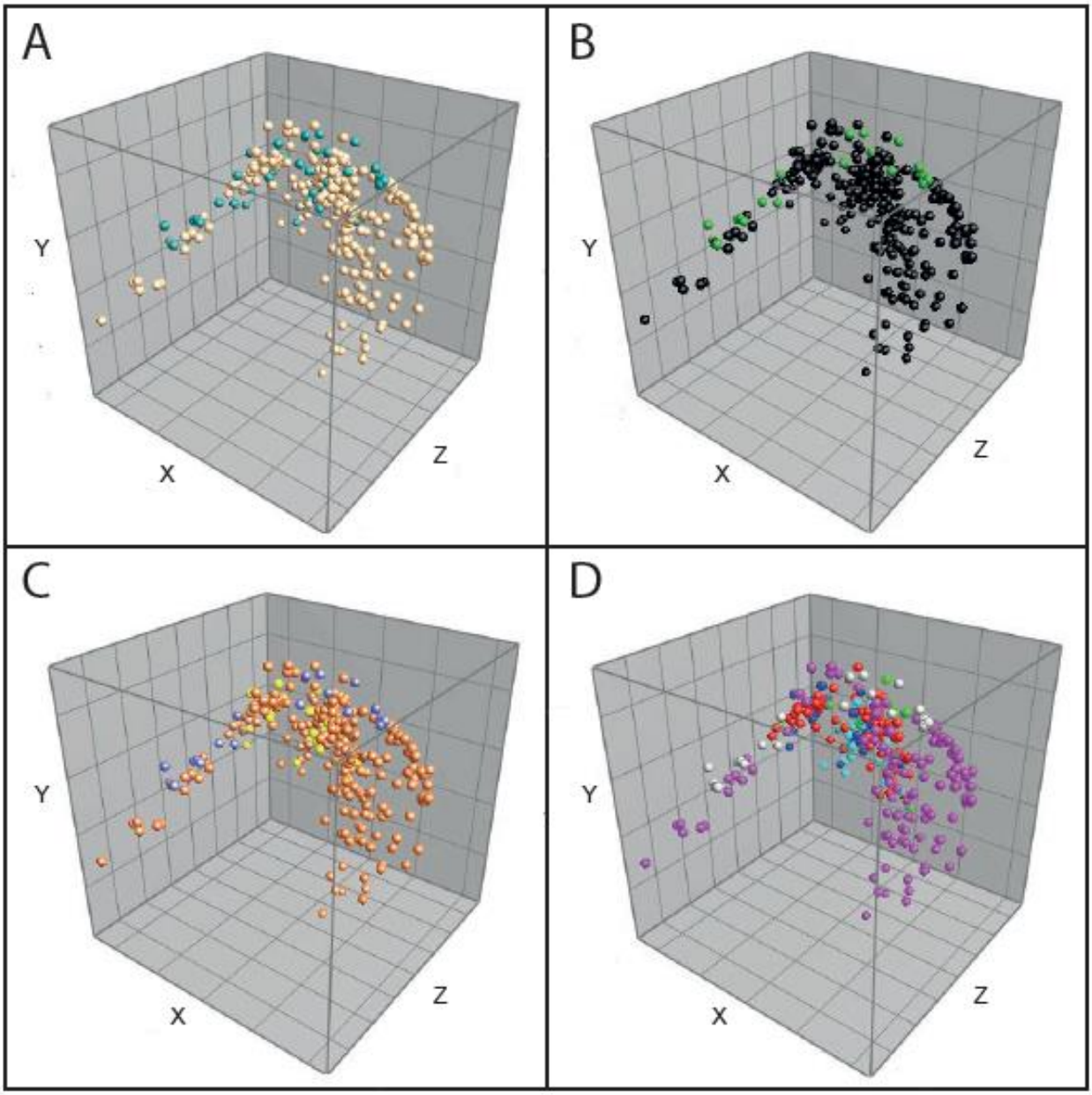

Figure 4 PCA of isolates in the training set

A. Colored at genus level: beige = Shigella; teal = Escherichia. B. Colored at pathotype level: black = Shigella/EIEC; green = E. coli (other than EIEC). C. Colored at group level: orange = Shigella spp.; yellow = EIEC; purple = E. coli (other than EIEC). D. Colored at species level: light blue = S. dysenteriae; red = S. flexneri; green = S. boydii; pink = S. sonnei; blue = EIEC; light grey = Other E. coli.

Presence of biomarkers identified in previous studies

The specific biomarkers for *S. flexneri*, *S. sonnei* and *E. coli* assigned by Everley *et al.* [15] were not present in any of the 559 isolates in this study when using an error limit of ± 500 ppm. Even if they were

corrected for a charge with 2 electrons, they were not present. A few biomarkers for *Shigella spp*. and *E. coli* described by Khot and Fisher [14] were present within a range of 500 ppm in isolates used in this study, i.e., 4163 Da, 7157 Da, 8326 Da, and 9227 Da, and corrected for a charge of 2 electrons, 5096 Da and 5752 Da.

Classifier models based on machine learning

Using the internal cross-validation of the classifiers at all discrimination levels, all but one class had an accuracy of more than 87.5%. The only class with a lower accuracy (77%) was the class "*Escherichia*" at the genus discrimination level.

When using machine learning based classifiers for identification, 96% of *Shigella spp.* isolates and 21% of the *E. coli* isolates from the test set were correctly identified at the genus level, using the direct smear application method and respectively 100% and 8% using the ethanol-formic acid extraction method (Table 3). Correct identification percentages for the pathotype, group and the species level were displayed in Table 3. Although more than 80% of *S. sonnei* isolates were correctly identified with the species classifier, specificity is low, as more than 70% of *S. flexneri* isolates were also identified as *S. sonnei.*

Discussion

Current commercially available MALDI-TOF databases cannot distinguish between *Shigella spp.* and *E. coli*. Therefore, three different alternatives were explored in this study. A custom-made database was developed, biomarkers were identified and classification models using machine learning were designed. All methods were tested using spectra of *Shigella spp*., *E. coli* and EIEC isolates obtained by the direct smear method and the ethanol-formic acid extraction method. The latter method resulted in slightly more correctly identified isolates at all discrimination levels. However, in routine clinical

diagnostics, identification relies on the direct smear method, as the extraction method is more time-consuming and laborious, especially for large sample quantities.

With the custom-made database with or without combination with the Bruker databases, >91% of *S. flexneri* and *S. sonnei* were correctly identified at the species level. For *S. boydii* and *S. dysenteriae*, our custom-made database performed poor as has been shown before [11]. However, compared to a previous study our custom-made database assigned less *E. coli* isolates correctly [11]. This indicates that the inclusion of EIEC isolates in the custom-made database as well as in the test set complicates the identification. Half of the EIEC isolates were assigned to one of the *Shigella* species, thereby decreasing the percentage of correctly identified *E. coli*. Additionally, the poor performance of identifying *E. coli* with our custom-made database can be the result of an overrepresentation of *S. flexneri* and *S. sonnei*. To investigate this hypothesis, a second custom-made database was developed, based on 17 isolates of each species, representing the diversity in serotypes. This database did not perform better or worse than the custom-made database that contained 288 MSPs (see Online Resource 2), indicating that a more evenly distribution of species in the database does not improve identification of *E. coli*. Although percentages of correct species assignments to *S. flexneri* and *S. sonnei* were high, other species were falsely assigned to them, both in our study as in a previous study [11]. In the latter study, correct species identification was based on the majority rule that three out of four spots should indicate the same species. Besides the fact that the interpretation of four spots per isolate is not feasible in clinical diagnostics, this indicates that distinction of spectra is poor and assignment of species is based on probabilities rather than actual variations in spectra. Our study confirms this phenomenon, because multiple species identifications within the same log-score range were made per spot. Moreover, 10-15% of duplicate spots resulted in different species assignments using the commercially available databases as well as the custom-made database. Another designation for poor distinction was shown in the dendrogram that was inferred from the MSPs in the custom-made database. Because the MSPs of the same species were not clustering together, one could expect on

forehand that the resulting database would not be capable to identify the isolates from the test set correctly.

Another alternative approach for the use of commercially available databases is the detection of discriminating biomarkers. However, in our study many isolates had an inconclusive identification, as specific biomarkers were not detected for most classes. Although more than 90% of *S. sonnei* isolates were identified at the species level, other species are also frequently falsely identified as *S. sonnei*, indicating poor distinctive properties of the assigned biomarkers.

Moreover, when analyzing also peak intensity and peak area rather than just peak presence, the PCA showed that *Shigella spp.* and *E. coli* did not represent separated groups based on their biomarkers. One large cluster was formed, containing *Shigella spp*., *E. coli* and EIEC, with a few outliers consisting only of *S. sonnei* isolates. Furthermore, only six out of 40 (15%) of the unique masses described for *Shigella spp./E. coli* in a former study were also present in our biomarker list, indicating that these biomarkers are stably present in the *E. coli* and *Shigella* population [14]. The absence of 85% of the masses that were assigned as biomarkers in this former study in our isolates, indicates that the detected biomarkers vary amongst isolate sets tested, and that a stable variation per species is not observed. Consequently, we anticipate that assignment of biomarkers based on yet another additional set of isolates will lead to even more diversity in biomarkers, indicating their unsuitability for distinct identification of *Shigella spp*., *E. coli* and EIEC.

The use of classifier models based on machine learning resulted in comparable percentages of correctly identified *Shigella* on the genus level, i.e. ≥ 94%, as reported in other studies [14]. Our classifier models also achieved a comparable percentage of correctly identified *E. coli* as earlier published semi-automated classifiers without EIEC isolates. In our classifier model designed on the pathotype level, EIEC isolates were not incorporated in the class *E. coli* and correct identification was 67%, compared to 56% in a previous study [14]. Nonetheless, the other remaining *E. coli* isolates were

falsely classified as *Shigella* both with our classifiers and with previously published ones [14]. Therefore, this decreases the specificity for the distinction of *Shigella,* making the classifiers unsuitable to use in clinical diagnostics. At the group and species level, classifiers performed even less, and most species could not be identified at all. It is possible that the poor performance of the classifier models was caused by an overrepresentation of *S. flexneri* and *S. sonnei,* as discussed for the custom-made database in our study. Therefore, the selection of 17 isolates of each species was used again and alternative classifiers were designed. These classifiers did not perform better or worse than the classifiers designed using all 288 isolates in the training set, indicating that an absence of an evenly distribution of species was not the cause for poor identification with classifiers (Online Resource 2).

Compared to previous studies, we used a substantial larger set of isolates and included the *E. coli* pathotype EIEC. Another strength was that multiple alternative approaches for the identification of *Shigella spp*. and *E. coli* using MALDI-TOF MS were explored. Although *S. sonnei* and *S. flexneri* isolates were overrepresented in both the training set and the test set, this distribution is representative for high-resource settings.

In conclusion, none of our explored alternative approaches for identification of *Shigella spp.*, *E. coli* and EIEC with MALDI-TOF MS was suitable to use in clinical diagnostics as all rendered a poor distinction based on spectra or biomarkers. We anticipate that with the use of an even larger and more diverse set of isolates, distinction of *Shigella spp*., *E. coli* and EIEC based on spectra using MALDI-TOF MS will remain challenging. Therefore, we propose an identification algorithm in which MALDI-TOF is used for the identification and differentiation of *Shigella/E. coli* as a group from other *Enterobacteriaceae*, followed by tests other than MALDI-TOF MS to distinguish between the different *Shigella species*, *E. coli*, and specific *E. coli* pathotypes including EIEC.

Author contributions:

MB, JR, MK, and FR conceived and designed the project. MB and AE performed the analyses. MB, JR, MK, and FR interpreted results. MB wrote the manuscript. All authors read, reviewed and approved the final manuscript.

Potential conflict of interest

The authors declare that they have no conflict of interest.

Data availability statement

The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

Electronic Supplementary Material

ESM_1.pdf	Decision diagrams of assigned biomarkers
ESM_2.pdf	Comparison of performance of a custom-made database and classifiers based on all 288 isolates or based on an evenly distribution of 17 isolates for each species

References

1. DuPont HL, Formal SB, Hornick RB, Snyder MJ, Libonati JP, Sheahan DG, LaBrec EH, Kalas JP (1971) Pathogenesis of *Escherichia coli* diarrhea. N Engl J Med 285 (1):1-9. doi:10.1056/NEJM197107012850101

2. Lan R, Alles MC, Donohoe K, Martinez MB, Reeves PR (2004) Molecular evolutionary relationships of enteroinvasive *Escherichia coli* and *Shigella spp*. Infect Immun 72 (9):5080-5088. doi:10.1128/IAI.72.9.5080-5088.2004 72/9/5080 [pii]

3. Pettengill EA, Pettengill JB, Binet R (2015) Phylogenetic analyses of *Shigella* and enteroinvasive *Escherichia coli* for the identification of molecular epidemiological markers: whole-genome comparative analysis does not support distinct genera designation. Front Microbiol 6:1573. doi:10.3389/fmicb.2015.01573

4. Kaper JB, Nataro JP, Mobley HL (2004) Pathogenic *Escherichia coli*. Nat Rev Microbiol 2 (2):123-140. doi:10.1038/nrmicro818

5. Bopp CAB, F.W.; Fields, P.I.; Wells, J.G.; Strockbine, N.A. (2003) *Escherichia*, *Shigella* and *Salmonella*. In: Murray PR (ed) Manual of Clinical Microbiology, vol 1. 8th edn. ASM Press, Washington D.C., pp 654-671

6. van den Beld MJ, Reubsaet FA (2012) Differentiation between *Shigella*, enteroinvasive *Escherichia coli* (EIEC) and noninvasive *Escherichia coli*. Eur J Clin Microbiol Infect Dis 31 (6):899-904. doi:10.1007/s10096-011-1395-7

7. Martiny D, Busson L, Wybo I, El Haj RA, Dediste A, Vandenberg O (2012) Comparison of the Microflex LT and Vitek MS systems for routine identification of bacteria by matrix-assisted laser desorption ionization-time of flight mass spectrometry. J Clin Microbiol 50 (4):1313-1325. doi:10.1128/JCM.05971-11

8. Cameron M, Perry J, Middleton JR, Chaffer M, Lewis J, Keefe GP (2018) Short communication: Evaluation of MALDI-TOF mass spectrometry and a custom reference spectra expanded database for the identification of bovine-associated coagulase-negative staphylococci. J Dairy Sci 101 (1):590-595. doi:10.3168/jds.2017-13226

9. Seuylemezian A, Aronson HS, Tan J, Lin M, Schubert W, Vaishampayan P (2018) Development of a custom MALDI-TOF MS database for species-level identification of bacterial isolates collected from spacecraft and associated surfaces. Front Microbiol 9:780. doi:10.3389/fmicb.2018.00780

10. Forero Morales MP, Hopprich R, Wise R, Shephard L, Weldhagen GF (2018) Cost-effective implementation of a custom MALDI-TOF library for the identification of South Australian *Nocardia* isolates. Pathology 50 (7):753-757. doi:10.1016/j.pathol.2018.08.006

11. Paauw A, Jonker D, Roeselers G, Heng JM, Mars-Groenendijk RH, Trip H, Molhoek EM, Jansen HJ, van der Plas J, de Jong AL, Majchrzykiewicz-Koehorst JA, Speksnijder AG (2015) Rapid and reliable discrimination between *Shigella species* and *Escherichia coli* using MALDI-TOF mass spectrometry. Int J Med Microbiol 305 (4-5):446-452. doi:10.1016/j.ijmm.2015.04.001

12. Josten M, Reif M, Szekat C, Al-Sabti N, Roemer T, Sparbier K, Kostrzewa M, Rohde H, Sahl HG, Bierbaum G (2013) Analysis of the matrix-assisted laser desorption ionization-time of flight mass spectrum of *Staphylococcus aureus* identifies mutations that allow differentiation of the main clonal lineages. J Clin Microbiol 51 (6):1809-1817. doi:10.1128/JCM.00518-13

13. Sauget M, Valot B, Bertrand X, Hocquet D (2017) Can MALDI-TOF Mass Spectrometry Reasonably Type Bacteria? Trends Microbiol 25 (6):447-455. doi:10.1016/j.tim.2016.12.006

14. Khot PD, Fisher MA (2013) Novel approach for differentiating *Shigella species* and *Escherichia coli* by matrix-assisted laser desorption ionization-time of flight mass spectrometry. J Clin Microbiol 51 (11):3711-3716. doi:10.1128/JCM.01526-13

15. Everley RA, Mott TM, Wyatt SA, Toney DM, Croley TR (2008) Liquid chromatography/mass spectrometry characterization of *Escherichia coli* and *Shigella species*. J Am Soc Mass Spectrom 19 (11):1621-1628. doi:S1044-0305(08)00578-3 [pii]10.1016/j.jasms.2008.07.003

16. De Bruyne K, Slabbinck B, Waegeman W, Vauterin P, De Baets B, Vandamme P (2011) Bacterial species identification from MALDI-TOF mass spectra through data analysis and machine learning. Syst Appl Microbiol 34 (1):20-29. doi:https://doi.org/10.1016/j.syapm.2010.11.003

17. Mather CA, Werth BJ, Sivagnanam S, SenGupta DJ, Butler-Wu SM (2016) Rapid Detection of Vancomycin-Intermediate *Staphylococcus aureus* by Matrix-Assisted Laser Desorption Ionization-Time of Flight Mass Spectrometry. J Clin Microbiol 54 (4):883-890. doi:10.1128/JCM.02428-15

18. Ho PL, Yau CY, Ho LY, Chen JHK, Lai ELY, Lo SWU, Tse CWS, Chow KH (2017) Rapid detection of cfiA metallo-beta-lactamase-producing *Bacteroides fragilis* by the combination of MALDI-TOF MS and CarbaNP. J Clin Pathol 70 (10):868-873. doi:10.1136/jclinpath-2017-204335

19. Clermont O, Bonacorsi S, Bingen E (2000) Rapid and simple determination of the *Escherichia coli* phylogenetic group. Appl Environ Microbiol 66 (10):4555-4558. doi:10.1128/aem.66.10.4555-4558.2000

20. van den Beld MJC, de Boer RF, Reubsaet FAG, Rossen JWA, Zhou K, Kuiling S, Friedrich AW, Kooistra-Smid M (2018) Evaluation of a culture dependent algorithm and a molecular algorithm for identification of *Shigella spp*., *Escherichia coli,* and enteroinvasive *E. coli* (EIEC). J Clin Microbiol 56:e00510-00518. doi:10.1128/JCM.00510-18

21. Saleeb PG, Drake SK, Murray PR, Zelazny AM (2011) Identification of mycobacteria in solid-culture media by matrix-assisted laser desorption ionization-time of flight mass spectrometry. J Clin Microbiol 49 (5):1790-1794. doi:10.1128/JCM.02135-10

22. Maths A Bionumerics: seven tutorials, spectra. http://www.applied-maths.com/tutorials#Spectra. Accessed 16-07-2019