JCM
Journals.ASM.org

Novel Approach for Differentiating *Shigella* Species and *Escherichia coli* by Matrix-Assisted Laser Desorption Ionization–Time of Flight Mass Spectrometry

Prasanna D. Khot,[a] Mark A. Fisher[a,b]

ARUP Laboratories, Salt Lake City, Utah, USA[a]; Department of Pathology, University of Utah, Salt Lake City, Utah, USA[b]

***Shigella* species are so closely related to *Escherichia coli* that routine matrix-assisted laser desorption/ionization–time of flight mass spectrometry (MALDI-TOF MS) cannot reliably differentiate them. Biochemical and serological methods are typically used to distinguish these species; however, "inactive" isolates of *E. coli* are biochemically very similar to *Shigella* species and thus pose a greater diagnostic challenge. We used ClinProTools (Bruker Daltonics) software to discover MALDI-TOF MS biomarker peaks and to generate classification models based on the genetic algorithm to differentiate between *Shigella* species and *E. coli*. Sixty-six *Shigella* spp. and 72 *E. coli* isolates were used to generate and test classification models, and the optimal models contained 15 biomarker peaks for genus-level classification and 12 peaks for species-level classification. We were able to identify 90% of *E. coli* and *Shigella* clinical isolates correctly to the species level. Only 3% of tested isolates were misidentified. This novel MALDI-TOF MS approach allows laboratories to streamline the identification of *E. coli* and *Shigella* species.**

Shigella species and *Escherichia coli* are very closely related Gram-negative bacteria belonging to the family *Enterobacteriaceae*. Phenotypically, *Shigella* species and *E. coli* species share many common characteristics; genotypically, they could be considered the same species (1–4). Due to this close relatedness, the differentiation of *Shigella* species from *E. coli* species can be difficult. In particular, variants of *E. coli* termed "inactive" (e.g., nonmotile, non-lactose-fermenting, or non-gas-producing isolates) are biochemically very similar to *Shigella* species and such isolates can pose a significant diagnostic challenge. Currently, methods based on biochemical tests and serotyping are preferred for identification of these species; however, these approaches may have suboptimal diagnostic performance. Unfortunately, advanced molecular methods such as sequencing the 16S rRNA gene and routine matrix-assisted laser desorption/ionization–time of flight mass spectrometry (MALDI-TOF MS) are unable to reliably differentiate between *Shigella* species and *E. coli* (5, 6).

There are four commonly recognized *Shigella* species (*S. sonnei*, *S. flexneri*, *S. boydii*, and *S. dysenteriae*), all of which may cause the well-characterized disease known as shigellosis (bacillary dysentery) (7). In contrast, *E. coli* strains in the human gut are typically commensal, although some can be pathogenic. Shigellosis is endemic throughout the world and is responsible for nearly 165 million cases of severe dysentery each year (7, 8). Since shigellosis is highly communicable ($<$100 viable cells can produce disease in healthy adults), it is a serious health concern at childcare centers and in developing countries with poor sanitation conditions. In the United States, approximately 14,000 cases of shigellosis occur each year, with *S. sonnei* and *S. flexneri* identified as the predominant pathogens (9). The Shiga-toxin-producing species *S. dysenteriae*, although infrequently isolated in the United States, may produce more-serious disease that can be fatal if left untreated.

We used a novel approach based on MALDI-TOF MS and ClinProTools (Bruker Daltonics) software to discover biomarker peaks that distinguish *Shigella* species from *E. coli* species. ClinProTools is a data-mining software program that helps identify potential biomarkers in complex mass spectra (10, 11). It also allows calculation of mathematical models based on biomarker peaks to develop "classifiers" of unknown isolates. Three strategies to identify biomarker peaks were evaluated: first, a semiautomated approach which relied on the presence or absence of peaks; second, a fully automated approach that relied on both differences in intensity and masses of biomarker peaks; and third, a hybrid approach which used a combination of biomarker peaks from the first two strategies.

MATERIALS AND METHODS

Bacterial strains. A total of 138 archived clinical isolates identified by a consensus approach of biochemical, serological, and genetic testing (12–14), including 66 *Shigella* species (35 *S. sonnei*, 23 *S. flexneri*, 4 *S. boydii*, and 4 *S. dysenteriae*) and 72 *E. coli* species (31 typical and 41 inactive), were chosen for analysis by MALDI-TOF MS. Of the 138 isolates, 131 were collected from 2006 to 2012 from diverse body sites (blood, tissue, genital, respiratory, stool, urine, and wound) and originated from at least 17 states across the United States. Five of the 138 isolates were ATCC reference strains of *E. coli* (25922), *S. sonnei* (25931), *S. flexneri* (12022), *S. boydii* (8700), and *S. boydii* (BAA-1247). Two of four *S. dysenteriae* isolates were genetically related strains lacking Shiga toxin expression: CVD 1254 (ΔstxAB) and CVD 1255 (ΔguaBA Δsen ΔstxAB; gifts from Eileen Barry, University of Maryland) (15). *E. coli* isolates were characterized as "inactive" if they displayed two or more of the following properties: lack of lactose fermentation, lack of motility, and lack of gas production (13). Within a species or biotype (e.g., normal or inactive *E. coli*), half of the available isolates were randomly assigned to groups used either to generate classification models or to test the models (16). In other words, there was no overlap between the isolates used to develop and test the classifi-

Received 14 June 2013 Returned for modification 1 July 2013
Accepted 23 August 2013
Published ahead of print 28 August 2013
Address correspondence to Mark A. Fisher, mark.fisher@aruplab.com.

doi:10.1128/JCM.01526-13

cation models. All three MALDI-TOF MS data analysis approaches used the same cohort of isolates for model generation and testing.

Biochemical methods. Isolates were identified by routine phenotypic and biochemical methods (13), including lactose fermentation characteristics on MacConkey agar, MIO (motility, indole, and ornithine decarboxylase), LIA (lysine iron agar), and TSI (triple sugar iron) agars (Hardy Diagnostics, Santa Maria, CA), as well as by the use of a BD Phoenix Automated Microbiology System (Phoenix NID panel; BD Diagnostics, Sparks, MD).

Serotyping. *Shigella* species and inactive *E. coli* were serotyped using antisera A to D (BD Diagnostics and Remel, Lenexa, KS). Isolates were cultivated in pure culture on Columbia sheep blood agar (Hardy Diagnostics) at 35°C, and a heavy suspension of organisms was made in 0.5 ml of 0.85% saline solution. Single drops of antisera for antigens A, B, C, and D were mixed with single drops of the organism suspension using wooden applicator sticks. After gentle rocking for 1 min, reactions were observed for strong agglutination which indicated the specific serotype. If agglutination was negative or weak for isolates that resembled *Shigella* species by other methods, then a 1-h boiling step was performed to remove any blocking envelope antigens prior to retesting, as recommended by the manufacturer.

Quantitative PCR (qPCR) assays. Two PCR assays were developed to target the *lacY* (β-galactoside permease) and *ipaH* (invasion plasmid antigen H) genes which were previously shown to distinguish *E. coli* from *Shigella* species (12, 14). Genomic DNA was extracted from pure cultures using a MagaZorb DNA Mini-Prep kit (Promega, Madison, WI) and quantified by spectrophotometry. Quantitative PCR was performed on a SmartCycler real-time PCR instrument (Cepheid, Sunnyvale, CA) using the double-stranded DNA (dsDNA) binding dye LCGreen Plus+ (BioFire Diagnostics, Salt Lake City, UT) as described below using genomic DNA of known *E. coli* and *S. sonnei* (*lacY*) or *S. flexneri* and *E. coli* (*ipaH*) clinical isolates as positive and negative controls, respectively, with every run.

***lacY* gene qPCR.** A 102-bp segment of the *lacY* gene was amplified using forward primer 5′-CTGCTTCTTTAAGCAACTGGCGA-3′ and reverse primer 5′-ACCAGACCCAGCACCAGATAAG-3′. Each 25-μl PCR mixture contained 1× Colorless GoTaq Flexi DNA polymerase (Promega), 3 mM $MgCl_2$, a 0.3 mM (each) deoxynucleoside triphosphate (dNTP) blend (Promega), 0.5 μM (each) forward and reverse primer, 0.5× LCGreen Plus+ (BioFire Diagnostics), and 20 ng genomic DNA. PCR cycling conditions consisted of a premelt at 95°C for 2 min and then 30 cycles of 95°C for 20 s, 58°C for 30 s, and 72°C for 20 s followed by final extension of 72°C for 5 min and a melt curve analysis step to confirm the PCR product. A test isolate was considered positive for the presence of the *lacY* gene when the threshold cycle (C_T) value was within 10-fold of the value for the positive control ($\Delta C_T \pm 3.32$) and had a characteristic melt peak (melting temperature [T_m] = 83.5°C) and was considered negative when the C_T value differed from the positive-control value by less than 0.001 ($\Delta C_T > 9.97$).

***ipaH* gene qPCR.** A 147-bp segment of the *ipaH* gene was amplified using forward primer 5′-TCGATAATGATACCGGCGCTC-3′ and reverse primer 5′-CTGCGAGCATGGTCTGGAA-3′. PCR and data interpretation conditions were identical to those for the *lacY* PCR except for use of 100 ng genomic DNA, a 55°C annealing temperature, and a characteristic melt peak of 85.7°C.

MALDI-TOF MS data acquisition. Isolates were cultivated in pure culture on MacConkey agar (Hardy Diagnostics) at 35°C. Organisms were harvested at 18 to 24 h. The formic acid-acetonitrile extraction method was employed on all isolates, and mass spectra were acquired as previously described on triplicate spots of each isolate extract (5). Data were collected between 2 K and 20 K *m/z* in linear positive-ionization mode (microflex; Bruker Daltonics, Billerica, MA). Each spectrum was a sum of 500 shots collected in increments of 100. When identification scores from the initial automated data collection were <1.9 for *E. coli* in the Biotyper analysis (Bruker Daltonics), new spectra were collected in manual acquisition mode. Spectra were further analyzed with FlexAnalysis 3.3 (Bruker Daltonics) and ClinProTools 2.2 (Bruker Daltonics) as described below. If spectra did not give satisfactory values for the default recalibration parameters in ClinProTools, isolates were regrown and extracted and new spectra collected as described above.

MALDI-TOF MS data analysis. Three approaches were used to generate biomarker-based classifiers (also called models). In all approaches, spectra from the model generation cohort were used to create a peak list to distinguish between classes of isolates (e.g., species) and test cohort spectra were then classified by the model to evaluate its performance. The classification algorithm in ClinProTools involved two steps. The first step distinguished between 2 classes (*Shigella* species and *E. coli*), and the second step distinguished among 5 classes (*S. sonnei*, *S. flexneri*, *S. boydii*, *S. dysenteriae*, and *E. coli*). If results from the 2-class (genus-level) and 5-class (species-level) models were consistent (e.g., if the results showed agreement with respect to genus identification), then the species-level identification was accepted. If results were inconsistent, the isolate was flagged for further workup, which in a typical laboratory would involve additional testing by serotyping, biochemical methods, and/or PCR. Accuracy was calculated with respect to agreement between MALDI-TOF MS identification and the reference identification based on serotyping, biochemical methods, and PCR.

Semiautomated approach. Biomarker peaks were identified by pairwise comparison of classes using the "Peak Statistic Table" function in ClinProTools followed by manual confirmation that peaks were distinguishable using FlexAnalysis. Mass lists for each spectrum were exported into Excel (Microsoft, Redmond, WA), and frequencies of biomarker peaks were calculated by class using custom code (available upon request) written in MATLAB (Mathworks, Natick, MA). The resulting "reference peak profiles" were compared with mass lists for each test isolate, and a Pearson's correlation coefficient was calculated. The profile that resulted in the highest correlation coefficient score was designated the identification of the unknown test isolate.

Automated approach. The automated approach was performed using three ClinProTools functions: data preparation, model generation, and spectra classification. Data preparation involved baseline subtraction (top hat; 10% minimal baseline width), normalization (total ion current), recalibration (1,000 ppm maximal peak shift and 30% match to calibrant peaks, with exclusion of spectra that could not be recalibrated), average spectrum calculation (resolution = 800), average peak list calculation (signal-to-noise threshold = 5), peak calculation in the individual spectra, and normalization of peak lists. Model generation using the genetic algorithm (17) was performed using the following settings: ≤15 peaks, automatic detection of initial number of peak combinations, ≤50 generations, 0.2 mutation rate, 0.5 crossover rate, no varying random seed, and 3 neighbors. Classification of unknown spectra was achieved by using the "Classify" function in ClinProTools. If ≥2 of 3 spectra per isolate were assigned to the same class, the identification was accepted.

During automated model generation, two parameters called "Cross Validation" and "Recognition Capability" were calculated by ClinProTools. Cross Validation is a measure of the model's reliability and may be used to predict its future performance. It is calculated by randomly splitting the model generation spectra into a model subset and a test subset. A model is generated and subsequently tested for its ability to correctly classify spectra in the test subset. This process is repeated multiple times to calculate a normalized Cross Validation value (18). Recognition Capability is a measure of the model's ability to correctly classify the spectra that were used to generate the model. It is calculated by testing each spectrum used to generate the model against the model itself and dividing the number of correctly classified spectra by the total. In other words, it is the percentage of model generation spectra that were correctly classified by the model.

Automated-hybrid approach. The data analysis used the same settings as the automated approach, except the "Force Peak into Model" command in ClinProTools was used to generate a hybrid model by inclusion of peaks from the semiautomated approach. Peaks were empirically

TABLE 1 Biomarker peaks used in the 3 MALDI-TOF MS approaches

Peak (*m/z*) determined by the semiautomated approach	Peak (*m/z*) determined by the automated approach		Peak (*m/z*) determined by the automated-hybrid approach	
	Genus-level model	Species-level model	Genus-level model	Species-level model
2,400	2,848	2,701	2,400^{a}	2,400^{a}
3,792	3,577	3,673	3,577^{b}	3,578^{b}
4,162	3,673	5,096	3,673^{b}	3,673^{b}
4,856	5,120	5,136	3,792^{a}	5,096^{a,b}
4,869	5,326	8,324	4,162^{a}	5,136^{b}
5,096	6,507	8,444	4,856^{a}	6,668^{b}
5,752	6,668	9,533	5,326^{b}	8,324^{b}
7,288	6,825	10,135	6,507^{b}	8,444^{b}
7,302	6,857	12,222	6,668^{b}	8,455^{a}
8,323	7,157	13,601	7,157^{b}	9,533^{b}
8,455	8,349	14,725	8,349^{b}	10135^{b}
9,711	9,223		9,223^{b}	13,601^{b}
9,736	9,264		9,448^{b}	
10,458	9,448		9,711^{a}	
	11,706		11,731	

[a] Peak from the semiautomated approach selected for inclusion in the ClinProTools model.

[b] Peak from the automated approach selected for inclusion in the ClinProTools model.

chosen for inclusion in the hybrid model if they improved the Cross Validation and Recognition Capability scores in comparison to those determined using the automated model.

Statistical analysis. Associations between categorical variables were analyzed by Fisher's exact test using statistical computing software R (v.2.15.0; http://www.R-project.org). *P* values < 0.05 were considered to represent statistical significance.

RESULTS

Accuracy of MALDI-TOF MS. The semiautomated approach was based upon statistical analysis of peaks by class (species) in ClinProTools followed by manual review to identify a set of biomarker peaks. This process resulted in 14 peaks that were potentially useful in distinguishing among the five species (*S. sonnei*, *S. flexneri*, *S. boydii*, *S. dysenteriae*, and *E. coli*, Table 1). The rationale for developing this model was to determine if the simple presence or absence of peaks could distinguish among the species. The accuracy of this approach among the 69 test cohort isolates was 94% (31 of 33) for detecting *Shigella* species but only 56% (20 of 36) for *E. coli*. Most (14 of 16) of the incorrect identifications were *E. coli* isolates identified as *S. sonnei*.

The automated approach, based on the genetic algorithm (17), resulted in 15 and 11 biomarker peaks for the genus- and species-level models, respectively (Table 1). In contrast to the semiautomated approach, peak selection using this approach considered differences in both intensity and mass. This model generation method improved the ability to distinguish isolates at both the genus and species levels compared to the semiautomated approach. Analysis of the test cohort resulted in 94% (65 of 69) accuracy with the genus model and 91% (63 of 69) accuracy with the species model. When the two-step testing algorithm that requires agreement between the models was implemented, 59 of the 69 test isolates (86%) were correctly identified. Of the remaining 10 isolates, nine were flagged for additional testing due to model disagreement; thus, only 1 of 69 (1.4%) isolates was misidentified (*S. flexneri* as *S. boydii*) at the species level.

ClinProTools allows peaks to be manually included in classification models. Eight distinguishing peaks from the semiautomated approach (five from the genus- and three from the species-level model) were selected for inclusion in the hybrid models (Table 1), which yielded improved performance relative to both the semiautomated and the automated approaches. The genus- and species-level hybrid models correctly classified 96% (66 of 69) and 91% (63 of 69) of the test cohort isolates, respectively, and the two-model testing algorithm yielded 90% (62 of 69) accuracy (Table 2). Five isolates were flagged for further workup, and only two isolates were misidentified (Table 3). One of the two misidentified isolates (isolate 102) was a typical lactose-fermenting *E. coli* isolate which, during the routine bacterial identification workflow, would likely not be tested with this specialized MALDI-TOF MS assay. Although the two-step classification algorithm resulted in slightly fewer correct identifications compared to the species-level model alone (Table 2), it resulted in fewer misidentifications (2 versus 4; Table 3).

TABLE 2 Accuracy of the hybrid MALDI-TOF MS assay, serotyping, and Phoenix with reference identification for the test isolate cohort

Organism (no. of isolates tested)	No. (%) correctly identified				
	MALDI-TOF MS				
	Genus-level model	Species-level model	Two-step classification^{a}	Serotyping	Phoenix
S. sonnei (18)	18 (100)	17 (94)	17 (94)	13 (72)	16 (89)
S. flexneri (11)	11 (100)	10 (91)	10 (91)	11 (100)	8 (73)
S. boydii (2)	2 (100)	2 (100)	2 (100)	2 (100)	2 (100)
S. dysenteriae (2)	1 (50)	1 (50)	1 (50)	2 (100)	1 (50)
Shigella species (33)^{b}	32 (97)	30 (91)	30 (91)	28 (85)	27 (82)
E. coli, typical (16)	15 (94)	15 (94)	15 (94)	ND^{c}	16 (100)
E. coli, inactive (20)	19 (95)	18 (90)	17 (85)	15 (75)	16 (80)
E. coli (36)^{d}	34 (94)	33 (92)	32 (89)	ND	32 (89)
Total (69)	66 (96)	63 (91)	62 (90)	43 (81)^{e}	59 (86)

[a] Final MALDI-TOF identification was accepted when results were consistent between the genus-level and species-level models.

[b] All *Shigella* species combined.

[c] ND, not determined.

[d] Typical and inactive *E. coli* combined.

[e] A total of 53 isolates were tested by serotyping.

TABLE 3 Discrepant results from the automated-hybrid MALDI-TOF MS approach

		ID based on MALDI			PCR amplification	
Isolate	Reference identification	Genus level	Species level	Result	*lacY* gene	*ipaH* gene
22[a]	*E. coli* (inactive)	*E. coli*	*S. sonnei*	Further workup	−	−
82	*S. sonnei*	*Shigella* spp.	*E. coli*	Further workup	−	+
102	*E. coli* (typical)	*Shigella* spp.	*S. sonnei*	*S. sonnei*	+	−
123	*E. coli* (inactive)	*Shigella* spp.	*E. coli*	Further workup	+	−
124	*E. coli* (inactive)	*E. coli*	Inconclusive	Further workup	+	−
129	*S. flexneri*	*Shigella* spp.	*S. boydii*	*S. boydii*	−	+
136	*S. dysenteriae*	*E. coli*	Inconclusive	Further workup	−	+

[a] Although the *lacY* gene for isolate 22 did not amplify, it was determined to be an *E. coli* gene based on *ipaH* gene PCR, serotyping, and biochemical tests.

During model generation, ClinProTools software calculates "Cross Validation" and "Recognition Capability" values, which are indicators of the model's performance and may be useful predictors of the model's ability to classify test isolates. The automated approach generated Cross Validation values, which reflect the model's ability to handle variability among test spectra, of 99.4% for the genus-level model and 90.3% for the species level model, whereas the hybrid models showed improved performance at 99.8% and 97.2%, respectively (Table 4). The Recognition Capability value, which reflects the model's ability to correctly identify its component spectra, was 100% for the genus-level model in both approaches and improved from 99.8% (automated) to 100% (hybrid) for the species-level model (Table 4). Due to its superior performance based on Cross Validation and Recognition Capability values, as well as on the test cohort, the automated-hybrid approach was used for comparison with other standard identification methods.

Comparison of MALDI-TOF MS with serotyping and automated biochemical identification. The performance of this MALDI-TOF MS assay was compared to the performance of the routine methods of serotyping and automated biochemical identification (BD Phoenix) for all *Shigella* ($n = 33$) and inactive *E. coli* ($n = 21$) test isolates. Because they are not routinely subjected to *Shigella* serotyping, typical *E. coli* isolates ($n = 20$) were compared using only the Phoenix method. The accuracies of MALDI-TOF MS, Phoenix, and serotyping compared to the reference identification are shown in Table 2. MALDI-TOF MS outperformed both serotyping and the Phoenix; however, the differences were not statistically significant ($P > 0.05$; Fisher's exact test). Note that, although the MALDI-TOF MS assay had a combined accuracy value of 90% (62 of 69), of the seven discrepant isolates, five isolates were simply flagged for further workup (e.g., the results were inconclusive) and only two isolates (2.9%) were misidentified (Table 3). In contrast, discrepant serotyping and Phoenix data would result in predominantly incorrect identifications (for serotyping, 9 of 10; for Phoenix, 9 of 10). Taken together, these data show that this novel MALDI-TOF approach is equivalent or superior to current phenotypic methods for distinguishing *E. coli* and *Shigella* species.

TABLE 4 Cross Validation and Recognition Capability values for automated approaches

	Value (%) determined by indicated approach			
	Automated		Automated-hybrid	
Parameter	Genus-level model	Species-level model	Genus-level model	Species-level model
Cross Validation	99.4	90.3	99.8	97.2
Recognition Capability	100	99.8	100	100

DISCUSSION

The differentiation of *Shigella* species and *E. coli* continues to pose a diagnostic challenge for clinical laboratories. Sequencing of the 16S rRNA gene and routine MALDI-TOF MS-based identification cannot distinguish between these species, and identification usually relies on a few distinct phenotypic and biochemical characteristics (5–7), which require additional time beyond primary isolation and may still not resolve all isolates. Since the discovery of the first species (*S. dysenteriae*) in 1898, *Shigella* species have been generally considered distinct from *E. coli* species from a clinical perspective (7, 19). Most *E. coli* species are commensals found as part of the normal gut flora, whereas *Shigella* species are generally considered pathogenic. Based on DNA hybridization, multilocus enzyme electrophoresis, and comparison of genomes and housekeeping genes, it would be reasonable to conclude that *E. coli* and *Shigella* species are part of the same phylogenetic continuum rather than clearly distinct species (1, 4, 20, 21). To further complicate diagnosis, our results indicate that inactive *E. coli* isolates may be misidentified as *Shigella* by commercial assays (e.g., Phoenix and serotyping), necessitating additional testing to reach a conclusive identification. Our newly developed MALDI-TOF MS-based assay using ClinProTools software (Bruker Daltonics) enables rapid distinction of *Shigella* species from *E. coli* species and could be adopted by clinical laboratories already using MALDI-TOF for routine bacterial identification.

ClinProTools software offers the ability to generate classification models from large numbers of spectra in a relatively rapid and flexible way. The aim of model generation is to determine a common signature among spectra of each of the model generation classes (e.g., different genera or species) in such a way that spectra of test isolates can be classified by the model. Among the three model generation algorithms available, models based on the genetic algorithm performed better for our study isolates than the other algorithms (Supervised Neural Network and QuickClassifier; data not shown). Two recent studies have used classification models generated using ClinProTools software for distinguishing between two *Staphylococcus aureus* strains (22) and between *Streptococcus pneumoniae* and *S. mitis* (23). The genetic algorithm was either the optimal algorithm for model generation or worked as well as the other options in both studies. Both of those studies developed single models that relied on only three distinguishing biomarker peaks, which was likely adequate because they were discriminating between only two classes (i.e., strains or species). For our study, models were developed to differentiate as many as five classes (4 *Shigella* species and *E. coli*). Two aspects seemed critical for this to be effective: first, utilization of a larger number of peaks than previously described (15 and 12 peaks; Table 1), and second, use of a hybrid-automated approach which combined manually validated peaks and those selected by the default ClinProTools algorithm. The flexibility of ClinProTools in allowing customization of parameters was indispensable in developing an assay that could distinguish between *Shigella* species and *E. coli*.

Despite issues with interpretation of agglutination, serotyping is arguably considered the gold standard for identification of *Shigella* species and is widely recommended in the clinical microbiology setting (13, 24). Overall, serotyping performed well in identifying most species of *Shigella* (Table 2); however, we were surprised to see that 28% (5 of 17) of the *S. sonnei* isolates in our test cohort were misidentified by this method. Repeat serotyping and PCR results of the *lacY* and *ipaH* genes helped resolve these discrepant results. Many laboratories rely on automated biochemical systems for the identification of enteric bacteria, and the overall accuracy of performance of the Phoenix system for identifying the *E. coli* in this cohort was 89% (32 of 36 isolates). However, this system had difficulties identifying *Shigella* and inactive *E. coli* isolates, with only 82% (27 of 33) of *Shigella* species and 80% (16 of 20) of inactive *E. coli* species (Table 2) identified correctly in our study. These data are consistent with those reported by Carroll et al., which showed that the Phoenix system misidentified approximately 17% of their *Shigella* isolates as *E. coli* (25). Unfortunately, that study did not specifically examine inactive *E. coli* isolates. As seen with the Phoenix, other automated systems may have difficulties correctly identifying some *Shigella* species and *E. coli* (26–29). Together, these data reinforce the idea that the phenotypic distinction between *Shigella* species and some *E. coli* isolates may be quite difficult or impossible to achieve using traditional methods. The specialized MALDI-TOF MS assay described here provides an alternative testing strategy that could improve upon currently available methods.

Our study had some limitations. First, the numbers of *S. dystenteriae* and *S. boydii* isolates included were low because these species are uncommon in the United States. Second, given that the ClinProTools models were generated using half of the available study isolates, it is possible that the diagnostic performance of this customized MALDI-TOF MS assay could improve if the models were generated with a larger set of isolates. Lastly, our study relied on a protein extraction method using cells grown on MacConkey agar. Spectra generated from isolates growing on sheep blood agar and the direct smear sample preparation method would have enabled a more streamlined workflow; however, preliminary manual analysis of a limited number of spectra (sheep blood agar versus MacConkey agar) suggested that data obtained from isolates growing on a selective medium such as MacConkey agar contained more discriminatory peaks. Further investigation will be required to determine if this strategy is feasible using other growth media.

The inability of routine MALDI-TOF MS to distinguish between *Shigella* species and *E. coli* is well recognized (5, 30). Furthermore, even though specialized mass spectrometry combined with liquid chromatography or affinity probes was shown to have the potential to differentiate between *S. flexneri* or *S. sonnei* and *E. coli* species, these techniques will require further development and validation to be applicable in routine clinical laboratory settings (31, 32). Our study demonstrated that MALDI-TOF MS, using a routine sample preparation combined with a specialized and yet automated data analysis approach, can overcome existing analysis limitations. The performance of this assay exceeded that of currently accepted methods such as serotyping and use of the Phoenix instrument. In addition, although serotyping is the recommended approach for identifying *Shigella* to the species level, this assay could enable species-level identification without the labor-intensive and subjective process of serotyping. This assay could be adopted by clinical laboratories to rapidly distinguish inactive and other non-lactose-fermenting *E. coli* species from *Shigella* species, although to replicate the analysis approach presented here may require laboratories to possess large cohorts of *E. coli* and *Shigella* isolates. However, the ability to transfer MALDI-TOF MS spectra between laboratories may promote wider use of such data analysis approaches to generate alternative classification tools. Overall, our study has demonstrated that MALDI-TOF MS is a powerful technology that is driving improvements in bacterial identification, and the currently observed limitations may simply be due to a lack of sufficient analysis tools rather than to inherent shortcomings of the method.

ACKNOWLEDGMENTS

Four *Shigella* isolates (2 *S. dysenteriae*, 2 *S. flexneri*) were a generous gift from Chad Campbell, Utah Department of Health, Salt Lake City, UT. Two *S. dysenteriae* strains were a generous gift from Eileen Barry, Center for Vaccine Development, University of Maryland School of Medicine, Baltimore, MD. We thank Ann Croft and members of the ARUP Bacteriology laboratory for assistance in completion of this study.

REFERENCES

1. **Brenner DJ, Fanning GR, Miklos GV, Steigerwalt AG.** 1973. Polynucleotide sequence relatedness among *Shigella* species. Int. J. Syst. Evol. Microbiol. **23:**1–7.
2. **Lan R, Reeves PR.** 2002. *Escherichia coli* in disguise: molecular origins of *Shigella.* Microbes Infect. **4:**1125–1132.
3. **Goris J, Konstantinidis KT, Klappenbach JA, Coenye T, Vandamme P, Tiedje JM.** 2007. DNA-DNA hybridization values and their relationship to whole-genome sequence similarities. Int. J. Syst. Evol. Microbiol. **57:** 81–91.
4. **Lukjancenko O, Wassenaar TM, Ussery DW.** 2010. Comparison of 61 sequenced *Escherichia coli* genomes. Microb. Ecol. **60:**708–720.
5. **Khot PD, Couturier MR, Wilson A, Croft A, Fisher MA.** 2012. Optimization of matrix-assisted laser desorption ionization–time of flight mass spectrometry analysis for bacterial identification. J. Clin. Microbiol. **50:**3845–3852.
6. **CLSI.** 2008. Interpretive criteria for identification of bacteria and fungi by DNA target sequencing; approved guideline, CLSI document MM18-A. Clinical and Laboratory Standards Institute, Wayne, PA.
7. **Niyogi SK.** 2005. Shigellosis. J. Microbiol. **43:**133–143.
8. **Kotloff KL, Winickoff JP, Ivanoff B, Clemens JD, Swerdlow DL, Sansonetti PJ, Adak GK, Levine MM.** 1999. Global burden of *Shigella* infections: implications for vaccine development and implementation of control strategies. Bull. World Health Organ. **77:**651–666.
9. **CDC.** 2013. Shigellosis. Centers for Disease Control and Prevention, Washington, DC. http://www.cdc.gov/nczved/divisions/dfbmd/diseases/shigellosis/.
10. **Elssner T, Kostrzewa M.** 2005. CLINPROT—a MALDI-TOF MS based system for biomarker discovery and analysis, p 167–178. *In* Conrad K, Bachmann M, Lehmann W, Sack U (ed), Methods, possibilities and perspectives of pre-symptomatic tumor diagnostics, vol 1. Pabst Science Publishers, Lengerich, Germany.
11. **Ketterlinus R, Hsieh SY, Teng SH, Lee H, Pusch W.** 2005. Fishing for biomarkers: analyzing mass spectrometry data with the new ClinProTools software. Biotechniques **2005**(Suppl):37–40.

12. **Pavlovic M, Luze A, Konrad R, Berger A, Sing A, Busch U, Huber I.** 2011. Development of a duplex real-time PCR for differentiation between *E. coli* and *Shigella* spp. J. Appl. Microbiol. **110:**1245–1251.
13. **Versalovic J (ed).** 2011. Manual of clinical microbiology, 10th ed. ASM Press, Washington, DC.
14. **Vu DT, Sethabutr O, Von Seidlein L, Tran VT, Do GC, Bui TC, Le HT, Lee H, Houng HS, Hale TL, Clemens JD, Mason C, Dang DT.** 2004. Detection of *Shigella* by a PCR assay targeting the *ipaH* gene suggests increased prevalence of shigellosis in Nha Trang, Vietnam. J. Clin. Microbiol. **42:**2031–2035.
15. **Wu T, Grassel C, Levine MM, Barry EM.** 2011. Live attenuated *Shigella dysenteriae* type 1 vaccine strains overexpressing Shiga toxin B subunit. Infect. Immun. **79:**4912–4922.
16. **Haahr M.** 1998. Random integer set generator. http://www.random.org/integer-sets/.
17. **Holland JH.** 1992. Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence, 1st ed. MIT Press, Cambridge, MA.
18. **Kearns M, Mansour Y, Ng AY, Ron D.** 1997. An experimental and theoretical comparison of model selection methods. Mach. Learn. **27:** 7–50.
19. **Johnson JR.** 2000. *Shigella* and *Escherichia coli* at the crossroads: Machiavellian masqueraders or taxonomic treachery? J. Med. Microbiol. **49:** 583–585.
20. **Pupo GM, Lan R, Reeves PR.** 2000. Multiple independent origins of *Shigella* clones of *Escherichia coli* and convergent evolution of many of their characteristics. Proc. Natl. Acad. Sci. U. S. A. **97:**10567–10572.
21. **Pupo GM, Karaolis DK, Lan R, Reeves PR.** 1997. Evolutionary relationships among pathogenic and nonpathogenic *Escherichia coli* strains inferred from multilocus enzyme electrophoresis and mdh sequence studies. Infect. Immun. **65:**2685–2692.
22. **Boggs SR, Cazares LH, Drake R.** 2012. Characterization of a *Staphylococcus aureus* USA300 protein signature using matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. J. Med. Microbiol. **61:**640–644.
23. **Ikryannikova LN, Filimonova AV, Malakhova MV, Savinova T, Filimonova O, Ilina EN, Dubovickaya VA, Sidorenko SV, Govorun VM.** 27 November 2012. Discrimination between *Streptococcus pneumoniae* and *Streptococcus mitis* based on sorting of their MALDI mass spectra. Clin. Microbiol. Infect. [Epub ahead of print.] doi:10.1111/1469-0691.
24. **Garcia LS (ed).** 2010. Clinical microbiology procedures handbook, 3rd ed. ASM Press, Washington, DC.
25. **Carroll KC, Glanz BD, Borek AP, Burger C, Bhally HS, Henciak S, Flayhart D.** 2006. Evaluation of the BD Phoenix automated microbiology system for identification and antimicrobial susceptibility testing of *Enterobacteriaceae.* J. Clin. Microbiol. **44:**3506–3509.
26. **Gavin PJ, Warren JR, Obias AA, Collins SM, Peterson LR.** 2002. Evaluation of the Vitek 2 system for rapid identification of clinical isolates of gram-negative bacilli and members of the family *Streptococcaceae.* Eur. J. Clin. Microbiol. Infect. Dis. **21:**869–874.
27. **Gupta S, Aruna C, Muralidharan S.** 2011. Misidentification of a commensal inactive *Escherichia coli* as *Shigella sonnei* by an automated system in a critically ill patient. Clin. Lab. **57:**767–769.
28. **Nadarajah R, Leonard ST, Brooks GF.** 2004. Comparison of BD Phoenix automated microbiology system with the MicroScan Rapid Neg ID plus Neg MIC Panel Type 30 for identification and susceptibility testing of Gram-negative bacilli. Technical Center, white papers, identification/susceptibility. Becton, Dickinson and Company, Frankin Lakes, NJ.
29. **O'Hara CM, Miller JM.** 2000. Evaluation of the MicroScan rapid neg ID3 panel for identification of *Enterobacteriaceae* and some common gram-negative nonfermenters. J. Clin. Microbiol. **38:**3577–3580.
30. **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:**1313–1325.
31. **Chen WJ, Tsai PJ, Chen YC.** 2008. Functional nanoparticle-based proteomic strategies for characterization of pathogenic bacteria. Anal. Chem. **80:**9612–9621.
32. **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:**1621–1628.