Charakterisierung klinisch-relevanter Bakterien mittels Proteotypisierung

Dissertation
zur Erlangung des mathematisch-naturwissenschaftlichen Doktorgrades
"Doctor rerum naturalium"
der Georg-August-Universität Göttingen

im Promotionsprogramm Biologie
der Georg-August-University School of Science (GAUSS)

vorgelegt von
Matthias Frederik Emele
aus Stuttgart

Göttingen, 2019

Betreuungsausschuss:

Prof. Dr. med. Uwe Groß
(Institut für Medizinische Mikrobiologie, Abteilung Medizinische Mikrobiologie, Universitätsmedizin Göttingen)

Prof. Dr. rer. nat. Fabian Commichau
(Institut für Mikrobiologie und Genetik, Abteilung Allgemeine Mikrobiologie, Georg-August-Universität Göttingen)

PD Dr. med. Andreas E. Zautner
(Institut für Medizinische Mikrobiologie, Abteilung Medizinische Mikrobiologie, Universitätsmedizin Göttingen)

Mitglieder der Prüfungskomission:

Referent: Prof. Dr. med. Uwe Groß
(Institut für Medizinische Mikrobiologie, Abteilung Medizinische Mikrobiologie, Universitätsmedizin Göttingen)

Korreferent: Prof. Dr. rer. nat. Fabian Commichau
(Institut für Mikrobiologie und Genetik, Abteilung Allgemeine Mikrobiologie, Georg-August-Universität Göttingen)

Weitere Mitglieder der Prüfungskommission:

Prof. Dr. rer. nat. Markus Bohnsack
Institut für Molekularbiologie, Abteilung Molekularbiologie, Universitätsmedizin Göttingen

Prof. Dr. rer. nat. Rolf Daniel
(Institut für Mikrobiologie und Genetik, Abteilung Genomische und Angewandte Mikrobiologie, Georg-August-Universität Göttingen)

Prof. Dr. rer. nat. Carsten Lüder
(Institut für Medizinische Mikrobiologie, Abteilung Medizinische Mikrobiologie, Universitätsmedizin Göttingen)

Prof. Dr. rer. nat. Jörg Stülke
(Institut für Mikrobiologie und Genetik, Abteilung Allgemeine Mikrobiologie, Georg-August-Universität Göttingen)

Tag der mündlichen Prüfung:
30.04.2019

Meiner Familie

Danksagung

Da das Ende meiner Promotionszeit nun erreicht ist, möchte ich gerne einigen Leuten danken, deren Unterstützung maßgeblich zu einem erfolgreichen Abschluss beigetragen hat.
Zunächst möchte ich meinen Betreuern Herrn PD Dr. Andreas E. Zautner, Herrn Prof. Dr. Uwe Groß und Herrn Prof. Dr. Fabian Commichau danken, die mich während der gesamten Zeit mit ihrer langjährigen Erfahrung und Expertise unterstützt haben. Auch während komplizierter Projektphasen hatten sie stets neue und kreative Ansätze parat.
Außerdem möchte ich den Kooperationspartnern der Projekte, Herrn Dr. Thomas Riedel und Frau Dr. Schneiderhan-Marra danken. Die Zusammenarbeit war stets reibungslos und gewinnbringend. Außerdem konnte ich viele neue Methoden erlernen und wertvolle Erfahrungen sammeln.

Während des gesamten Zeitraums haben mich außerdem viele Leute innerhalb und außerhalb der Universitätsmedizin Göttingen unterstützt. Sei es im Labor oder während Gastaufenhalten. Insbesondere möchte ich gerne folgenden Mitarbeitern danken: Dr. Oliver Bader, Felix Joppe, Yvonne Laukat, Agnieska Goretzki, Emilia Gomez, Ines Oehmig, Pia Sternisek, Sabrina Minatelli, Julian Schwanbeck, Roswitha Nast, Dr. Raimond Lugert, Anna Günther, Jens Gruber, Angela Aurin-Gerke, Ruth Rosenhagen, Maik Tomm, Marco Köhler und Frank Zickenrott. Vielen Dank für die wertvolle Hilfe in jeglicher Form und alles Gute für die Zukunft.

Zuletzt möchte ich mich von Herzen bei meiner Familie für die unglaubliche Unterstützung in den letzten Jahren bedanken, die es mir ermöglicht hat, diesen Weg zu gehen.

Inhaltsverzeichnis

Abkürzungsverzeichnis

AFPL	Amplified Fragment Length Polymorphism
aspA	Aspartase A (Gen)
Da	Dalton (Atomare Masseneinheit)
DNA	deoxyribonucleic acid (dt.: Desoxyribonukleinsäure)
et al.	lat. *et alii* – und andere
glnA	Glutaminsynthase (Gen)
gltA	Zitratsynthase (Gen)
glyA	Serinhydroxymethyltransferase (Gen)
gyrA	Gyrase-Untereinheit-A (Gen)
ICMS	eng. intact cell mass spectrometry
m/z	Masse-zu-Ladung-Verhältnis [kg/C] – die Einheit wird i.d.R. nicht angegeben, bei einfach geladenen Biomarkermassen entspricht der *m/z*-Wert dem Molekulargewicht+1H+
MALDI-TOF MS	Matrix-unterstützte Laser-Desorption/ Ionisation-basierte – eng. matrix-assisted laser desorption/ ionization (MALDI), Massenspektrometrie (MS) mit Flugzeitanalyse – eng. time of flight (TOF)
MLST	eng. multilocus sequence typing
MRSA	Methicillin-resistenter Staphylococcus aureus
MSPP	Massenspektrometriebasierte Phyloproteomik
NGS	eng. next-generation-sequencing
PCA	eng. Principal component analysis – dt. Hauptkomponentenanalyse
PCR	eng. polymerase chain reaction – dt. Polymerase-Kettenreaktion
PFGE	Pulsfeldgelelektrophorese
pgm	Phosphoglukomutase (Gen)
porA	Major-outer-membrane-protein-Gen
PTM	posttranslationale Modifikationen
rMLST	eng. ribosomal multilocus sequence typing
rp	Ribosomales Protein (Gen)
SCC*mec*	eng. staphylococcal cassette chromosome mec

ST	Sequenztyp
UPGMA	eng. Unweighted Pair Group Method with Arithmetic mean
VISA	Vancomycin-intermediärer Staphylococcus aureus
VSSA	Vancomycin-suszeptibler Staphylococcus aureus
wgMLST	eng. whole-genome multilocus sequence typing

Abbildungsverzeichnis

Zusammenfassung

Im Fall von Ausbruchsgeschehen ist die schnelle Identifikation der Infektionsquelle entscheidend. Folglich ist die Typisierung mikrobieller Spezies auch unterhalb der Spezies- und Subspeziesebene eine wichtige Aufgabe klinisch-mikrobiologischer Labore. Im Rahmen nosokomialer Infektionen ist insbesondere die Differenzierung mikrobieller Subtypen wichtig, die spezifische antimikrobielle Resistenzen oder eine erhöhte Humanvirulenz aufweisen.

Über die Jahre wurden diverse Subtypisierungsmethoden entwickelt, wobei DNA-Sequenz-basierte Methoden wie die Multilocus-Sequenztypisierung (MLST) der Goldstandard sind. Nachteile dieser Methoden sind, dass sie entweder zeit- und arbeitsintensiv (Sanger Sequenzierung) oder aber kostenintensiv (Next Generation Sequencing) sind. Aus diesem Grund sind sie für die klinische Routinediagnostik nicht optimal, da täglich eine hohe Anzahl von Proben analysiert werden muss.

Neben den sequenzbasierten Methoden hat sich die MALDI-TOF-Massenspektrometrie zu der Standardmethode für die Genus- und Speziesidentifizierung entwickelt, welche auf der Massenbestimmung niedermolekularer, hauptsächlich ribosomaler Proteine beruht. Allel-Isoformen dieser ribosomalen Proteine (Biomarkerproteine), die sich in ihrer Masse unterscheiden, bilden die Grundlage eines neuen Typisierungsverfahrens, das initial als Massenspektrometrie-basierte Phyloproteomik (MSPP) bezeichnet, mittlerweile aber in Proteotypisierung umbenannt wurde. Grundlage des Verfahrens ist eine Aminosäuresequenzliste alleler Isoformen, die aufgrund nicht-synonymer Mutationen in den Genen der Biomarkerproteine auftreten und als Massenverschiebungen bei der Überlagerung kalibrierter MALDI-TOF-Spektren sichtbar sind.

Auf Basis der detektierbaren Biomarkermassen und Massenverschiebungen kann eine isolatspezifische Kombination von Aminosäuresequenzen abgeleitet werden. Mittels hierarchischer Clusteralgorithmen können, analog zu sequenzbasierten Methoden, phyloproteomische Dendrogramme errechnet werden.

Die Proteotypisierung besitzt ein großes Potenzial für die Typisierung von Bakterien unterhalb der Spezies- und Subspeziesebene, was in dieser Arbeit anhand der Bakterien *Campylobacter coli*, *Campylobacter fetus* und *Clostridioides difficile* demonstriert wird.

Summary

In case of a disease outbreak fast and reliable detection of the source of infection is crucial. For this reason, typing of microbial species, also at the below-species level is an important task of clinical microbiological laboratories. In the context of nosocomial infections especially differentiation of microbial subtypes with specific antimicrobial resistances and increased human virulence is of importance.

Through the years, a wide range of subtyping methods has been developed, whereby DNA sequence-based methods like multilocus sequence typing (MLST) are the current gold standard. The disadvantage of these methods is, that they are either time-consuming and labour-intensive (sanger sequencing) or costly (Next Generation Sequencing). For this reason, these methods are not ideal for clinical routine diagnostics, where a large number of samples has to be analyzed every day.

Besides sequence-based methods MALDI-TOF-mass spectrometry has become a standard method for genus and species identification. This method is based on the detection of low-molecular, primarily ribosomal proteins. Allelic isoforms of these proteins (biomarker proteins) differing in mass are the cornerstone of a new typing procedure referred to as proteotyping. Basis of this procedure is an amino acid sequence list of allelic isoforms that occur due to non-synonymous mutations in genes coding for biomarker proteins and appear as mass shifts when overlaying calibrated MALDI-TOF spectra.

Based on the detectable biomarker masses and mass shifts it is possible to deduce an isolate-specific combination of amino acids. Using hierarchical clustering algorithms, it is possible to calculate phyloproteomic dendrograms analogous to sequence-based methods.

Proteotyping offers a great potential for subspecies differentiation, which will be demonstrated by means of the bacteria *Campylobacter coli*, *Campylobacter fetus* and *Clostridioides difficile.*

1 Einleitung

1.1 Anwendungsgebiete für die Subtypisierung von Mikroorganismen und Anforderungen an assoziierte Verfahren

Die Identifikation einer Infektionsquelle kann äußerst kompliziert sein, ist jedoch für die Vorbeugung und Kontrolle von Ausbruchsgeschehen essentiell. Während bei einer Vielzahl von Pathogenen eine Transmission von Mensch zu Mensch stattfindet, können bestimmte Infektionen auch über die Nahrung, Tiere, Insekten oder Umweltquellen wie Gewässer erworben werden (Sandora *et al.*, 2014). Insbesondere im Fall zoonotischer Erkrankungen besteht aufgrund der diversen ökologischen Nischen der verschiedenen Wirte, aber auch in unterschiedlichen anatomischen Regionen des Menschen eine ausgeprägte biologische Diversität innerhalb einer mikrobiellen Spezies (Rosef *et al.*, 1983; Waldenström *et al.*, 2002; Sheppard *et al.*, 2009a; Sheppard *et al.*, 2009b; Griekspoor *et al.*, 2015).

Da die hohe phänotypische- und genotypische Variabilität von Mikroorganismen hinreichend bekannt ist, hat deren Charakterisierung auch in der medizinischen Mikrobiologie eine zentrale Rolle eingenommen - sowohl unterhalb der Spezies- als auch der Subspeziesebene (Conway *et al.*, 2001; Wolters *et al.*, 2011; Lartigue, 2013; Zautner *et al.*, 2015). Die Subtypisierung von Bakterien bestimmt die Ähnlichkeit zweier unterschiedlicher Isolate derselben Spezies oder Subspezies. Weisen zwei Isolate denselben Subtyp auf, ist es wahrscheinlicher, dass sie miteinander in Verbindung stehen, als wenn sie unterschiedliche Subtypen aufweisen. Teilen also zwei Patienten ein Zimmer und der aus Patient 1 isolierte Erreger weist den gleichen Subtyp auf, wie der aus Patient 2 isolierte Erreger, könnte der Erreger direkt oder indirekt von dem einen auf den anderen Patienten übertragen worden sein. Dass die Infektion unterschiedlichen Ursprungs ist, ist jedoch unwahrscheinlich (Sandora *et al.*, 2014).

Die Applikation antimikrobieller Substanzen (Antiinfektiva) im Rahmen der ärztlichen Behandlung resultiert in einer Einengung dieser mikrobiellen Diversität sowie in der Zerstörung der mikrobiotavermittelten Kolonisationsresistenz (Rolfe 1984; Kinnebrew *et al.*, 2010; Kachrimanidou und Malisiovas, 2011; Arias und Murray, 2012). Auf diese Weise werden mikrobielle Subtypen selektiert, die Resistenzen gegenüber Antibiotika aufweisen oder multiresistent sind, die besser in der Umwelt persistieren können oder gar eine höhere Virulenz gegenüber dem Menschen zeigen (Berendonk *et al.*, 2015; Knapp *et al.*, 2010; Amador *et al.*, 2015; Gibreel und Taylor, 2006; Alfredson und Korolik, 2007; Bolton, 2015). Diese Subtypen haben folglich Selektionsvorteile in der klinischen Umgebung, was ihre nosokomiale Ausbreitung begünstigt (Khan *et al.*, 2015).

Um eine hinreichende Überwachung besagter Ausbruchsgeschehen nosokomialer Infektionen sowie der Migration arzneimittelresistenter Erregerstämme bezüglich ihrer phylogenetischen Verwandtschaft gewährleisten zu können, ist die Entwicklung schneller, kostengünstiger und gut standardisierter Verfahren unerlässlich (Barbut *et al.*, 2014; Pfaller und Castanheira, 2015).

Das Differenzierungsvermögen entsprechender Verfahren sollte ausreichend sein, um hochvirulente von niedrig- oder avirulenten Stämmen unterscheiden zu können. Darüber hinaus sollten Stämme, die spezifische Resistenzen innehaben, von suszeptiblen Stämmen unterschieden werden können. Die Tatsache, dass die Akzession von Resistenzmechanismen und Virulenzfaktoren mitunter durch horizontalen Gentransfer über Plasmide erfolgt (Dodd, 2012; Tang *et al.*, 2017), schmälert die Aussagekraft gängiger Typisierungsverfahren teilweise erheblich. Aufgrund der phylogenetischen Verwandtschaft können zwar Aussagen über die Präsenz chromosomal-kodierter Resistenzmechanismen oder Virulenzfaktoren getroffen werden, was die Wahrscheinlichkeit einer erfolgreichen Therapie signifikant erhöht, durch das Plasmid kodierte Resistenzmechanismen und Virulenzfaktoren werden phylogenetisch jedoch nur bedingt erfasst (Maiden *et al.*, 1998; Leekitcharoenphon *et al.*, 2012).

1.2 Übersicht über die Entwicklung wichtiger mikrobiologischer Subtypisierungsverfahren

Das Ziel der Erregersubtypisierung bei Ausbrüchen ist wie eingangs erwähnt das Erkennen von Infektionsclustern durch die Identifikation klonaler Zusammenhänge bei Erregerisolaten. Hierfür wurde über die Jahre ein breites Spektrum an Methoden entwickelt. Die im Kontext dieser Arbeit relevanten Methoden sind jedoch allesamt phylogenetischer Natur.

Häufig verwendet werden Methoden, die auf sogenannten Restriktionsfragment-Längenpolymorphismen (RFLPs) beruhen. RFLPs sind vererbbare, lokal auftretende DNA-Sequenzveränderungen, die bei Verdau dieser DNA mit

Restriktionsenzymen zu Modifikationen im ursprünglichen Restriktionsfragment führen können. Potenzielle Änderungen im Spaltungsmuster werden anschließend via Pulsfeld-Gelelektrophorese analysiert und daraus resultierend die Phylogenie abgeleitet (Schwartz und Cantor, 1984; Maslow, 1993). Für die in dieser Arbeit untersuchten Spezies *Campylobacter coli*, *Campylobacter fetus* sowie *Clostridioides difficile* existieren jeweils spezifische und etablierte RFLP Protokolle (Yan *et al*., 1991; Bowman *et al*., 1991). Eine ähnliche Funktionsweise weist die sogenannte *Amplified Fragment-Length Polymorphism* (AFLP)-Methode auf (Vos *et al*., 1995). Im Rahmen dieser Methode wird ein genetischer Fingerabdruck erstellt, indem die DNA zunächst durch zwei Restriktionsenzyme fragmentiert wird, gefolgt von der Amplifikation einiger Fragmente mittels PCR. Durch die Varianz in der Anzahl der Restriktionsstellen entstehen unterschiedlich lange Fragmente, welche wiederum unterschiedliche Muster auf einem Elektrophorese-Gel ergeben. Die unterschiedlichen Muster dienen der Speziesunterscheidung sowie der Ableitung der Phylogenie (Vos *et al*., 1995; Lindstedt *et al*., 2000; Velappan *et al*., 2001).

Mittlerweile haben sich sequenzbasierte Methoden weitestgehend durchgesetzt, da sie neben der Möglichkeit einer dezentralen Analyse und Archivierung auch eine bessere Reproduzierbarkeit aufweisen. Für viele mikrobielle Spezies ist die Multilokus Sequenz Typisierung (MLST) zum Goldstandard bei der Ermittlung der Phylogenie geworden (Maiden *et al*., 1998; Perez-Losada *et al*., 2013; Dingle *et al*., 2005; Lemee *et al*., 2004; Griffiths *et al*., 2010; van Bergen *et al*., 2005). Bei der MLST werden Allele von üblicherweise sechs bis zehn *Housekeeping*-Genen auf Punktmutationen untersucht. Für jedes Allel ergibt sich eine Nummer und somit ein Zahlencode (Allelprofil), der jeweils einem Sequenztyp zugeordnet ist.

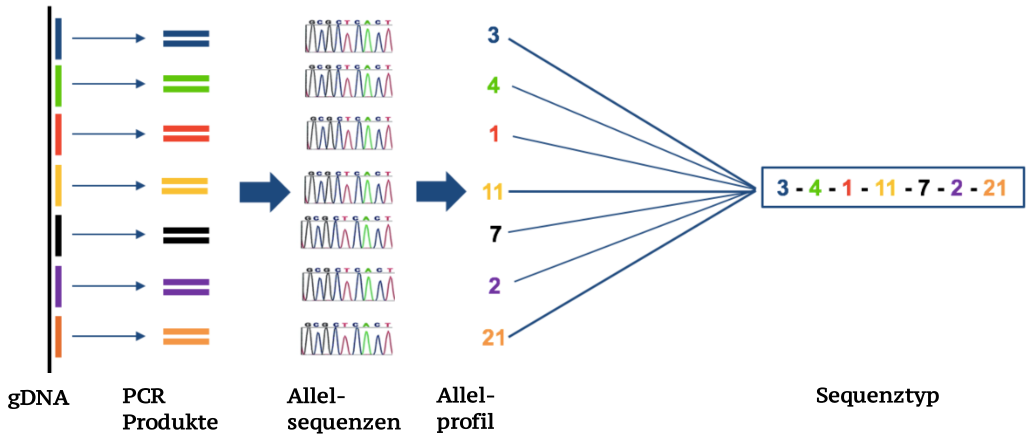


Abbildung 1 Multilocus-Sequenztypisierung (MLST) Die MLST ist der momentane Goldstandard bei der Spezies- und Subspeziesidentifikation. Üblicherweise werden Genfragmente von sieben bis neun *Housekeeping*-Genen sequenzanalysiert. Nach erfolgter PCR und anschließender Sanger-Sequenzierung kann ein Allelprofil abgeleitet werden, welches letztendlich wiederum einen spezifischen Sequenztypen ergibt.

Für *Campylobacter coli* und *Campylobacter fetus* werden dieselben Genloci analysiert: *aspA* (Aspartase A), *glnA* (Glutaminsynthase), *gltA* (Zitratsynthase), *glyA* (Serinhydroxymethyltransferase), *pgm* (Phosphoglukomutase), *tkt* (Transketolase) und *uncA* (ATP-Synthase alpha-Untereinheit) (Dingle *et al*., 2001; Dingle *et al*., 2005). Nach Ragimbeau *et al*. (2014) kann das Schema um den Genlokus *gyrA* (Gyrase-Untereinheit-A) erweitert und mit *porA* (Variables äußeres Membranprotein) und *flaA* (Flagellin kodierender Lokus A) kombiniert werden, um einerseits die Quellenzuordnung und andererseits die Detektion temporärer humaner Cluster zu optimieren (Ragimbeau *et al*., 2014).

Die folgenden Loci werden hingegen für die MLS-Typisierung von *C. difficile* verwendet: *aroE* (Shikimat Dehydrogenase), *ddl* (D-Alanin: D-Alanin Ligase), *dutA* (dUTP Pyrophosphatase), *gmk* (Guanylat Kinase), *recA* (Rekombinase), *sodA* (Superoxid Dismutase) und *tpi* (Triosephosphat Isomerase) (Lemee *et al*., 2004). Wie aus der Beschreibung der Funktionsweise der Methode hervorgeht, müssen die jeweiligen Genloci mit nicht unerheblichem Arbeitsaufwand per PCR amplifiziert und anschließend sequenziert (Sanger Sequenzierung) werden (Maiden, 2006). Die Entwicklung des Next Generation Sequencing (NGS) trägt hier zu einer erheblichen Verbesserung der MLS-Typisierung bei. NGS liegt die Idee einer massiven parallelen Sequenzierung von mehreren tausend bis hin zu Millionen DNA-Fragmenten im Rahmen eines einzigen Sequenzierlaufs zugrunde. Die Technologie hat sich bereits bei der Generierung von Sequenzdaten als nützlich erwiesen, wenn nur wenige Informationen über den Zielorganismus vorlagen und zwar durch die Bereitstellung von Rohmaterial für die Ermittlung von MLST-Schemata (Pérez-Losada *et al.*, 2013).

Traditionelle MLST-Schemata benötigen Referenzgenome, um geeignete Marker zu entwickeln. Durch die Analysegeschwindigkeit des NGS nimmt die Anzahl solcher Referenzgenome erheblich zu und das MLST-Schema kann signifikant erweitert werden. Im Rahmen des *whole genome* MLST (wgMLST) werden so sämtliche innerhalb einer mikrobiellen Spezies ubiquitären Genloci berücksichtigt (Boers *et al.*, 2012; Cody *et al.*, 2013; Carrillo *et al.*, 2012). Ein MLST-Schema sollte grundsätzlich so gestaltet sein, dass die berücksichtigten Genloci eine suffiziente Variabilität aufweisen, um einerseits die Phylogenie abzuleiten und andererseits Fragen bezüglich der Epidemiologie beantworten zu können. Welche und wie viele Genloci ein solches MLST-Schema beinhaltet, hängt in erster Linie von der epidemiologischen Fragestellung und der zu untersuchenden Spezies ab. In der klinischen Diagnostik ist beispielsweise häufig eine Unterscheidung zwischen hoch- und niedrigvirulenten Subspezies von Interesse.

Grundsätzlich lässt sich sagen, dass phylogenetische und epidemiologische Beziehungen umso besser abgebildet werden, je höher die Variabilität der im Typisierungsschema enthaltenen *Housekeeping*-Gene ist. Hypervariable und transposable Genelemente sollten hingegen ausgeschlossen werden (Leekitcharoenphon *et al.*, 2012).

Das mitunter am besten etablierte NGS-basierte MLST-Schema ist das ribosomale MLST- (rMLST) Schema, welches alle 53 Gene beinhaltet, die für die Untereinheiten des bakteriellen Ribosoms kodieren (*rps* Gene). Die *rps* Loci eignen sich hervorragend als universelles Charakterisierungsschema, da sie zum einen in allen Bakterien vorhanden sind und zum anderen über das gesamte Chromosom verteilt und hochkonserviert sind (Jolley *et al.*, 2012).

1.3 MALDI-TOF Massenspektrometrie zur Speziestypisierung

Trotz des umfassenden Potenzials der beschriebenen sequenzbasierten Methoden, bringen diese gewisse Nachteile mit sich, insbesondere hinsichtlich der Eignung für die klinische Routinediagnostik. PCR-Analysen gefolgt von anschließender Sangersequenzierung bedeuten einen relativ hohen Zeitaufwand, das NGS ist mit einem nicht unerheblichen Kosten- sowie Schulungsaufwand bei der Etablierung der Methode verbunden.

Neben den bisher beschriebenen Methoden hat sich eine weitere Methode als Standardmethode in klinisch-mikrobiologischen Laboren zur Gattungs- und Speziesidentifikation etabliert: Die Ganzzell-Massenspektrometrie (engl. *intact cell mass spectrometry* – ICMS) (Seng *et al*., 2010; Croxatto *et al*., 2012; Opota *et al*., 2017). Bei dieser Methode werden Massenspektren aus Zelllysaten im Massenbereich zwischen 2 und 20 kDa gemessen. Hierbei steht nicht die Charakterisierung einzelner Proteine im Fokus, sondern das gesamte, bei der Messung erzeugte spektrale Muster ist relevant. In diesem Zusammenhang wird auch von einem „Proteinfingerabdruck" des Bakteriums gesprochen. Durch den Abgleich des gemessenen Spektrums mit großen Datensätzen bekannter Bakterienkulturen kann der untersuchte Erreger sehr wahrscheinlich identifiziert werden. Darüber hinaus ist bekannt, dass MALDI-TOF MS die Klassifikation unbekannter Erreger erleichtert, indem Übereinstimmungen in den Massenspektren dieser bakteriellen Erreger mit Proteinbiomarkern in vorhandenen Datenbanken abgeglichen werden (Conway *et al*., 2001).

Auf Massenspektrometrie basierende Typisierung, auch Proteotypisierung genannt, wird seit nunmehr fast 20 Jahren für die Charakterisierung von mikrobiellen Gemeinschaften, einzelnen Proteinen und Geweben, aber auch Viren und Bakterien verwendet (Karlsson *et al*., 2015; Hugo *et al*., 2012; Rodriguez *et al*., 2006; Shillingford *et al*., 2003; Schwahn *et al*., 2010). Beispielweise wurde das Verfahren bereits erfolgreich im Rahmen der Subtypisierung Shiga toxinbildender *Escherichia coli*-Stämme, methicillinresistenter *Staphylococcus aureus*-Abstammungslinien und *Listeria monocytogenes*-Abstammungslinien verwendet (Christner *et al*., 2014; Wolters *et al*., 2011; Ojima-Kato *et al*., 2016).

Die Biomarker-Ionen, die im angesprochenen Massenbereich zwischen 2 und 20 kDa detektiert werden, sind in erster Linie hochkonservierte ribosomale Proteine, die ein spezifisches Massenprofil aufweisen. Für die Analyse müssen die zu untersuchenden Bakterien in Reinkultur vorliegen, wobei eine stecknadelkopfgroße Menge ausreichend ist. Der bei den sequenzbasierten Methoden angesprochene hohe Zeit- beziehungsweise Kostenaufwand ist bei dieser Methode nicht gegeben: Die Materialkosten belaufen sich auf wenige Cent pro Einzelanalyse, der Schulungsaufwand ist aufgrund der Endbenutzerfreundlichkeit der Methode vernachlässigbar. Außerdem ist ein MALDI-TOF Microflex-Gerät üblicherweise in klinisch-mikrobiologischen Laboren verfügbar. Die Methode erlaubt darüber hinaus, mit einer Messvarianz <1 Da, eine äußerst präzise Massenbestimmung und ermöglicht so auch die Differenzierung von Isolaten unterhalb der Spezies- und Subspeziesebene (Lartigue, 2013; Zautner *et al*., 2015; Emele *et al*., 2019). Es existieren simple, jedoch relativ unpräzise mathematische Algorithmen, die Unterschiede in den per MALDI-TOF gemessenen Massenspektren identifizieren und daraus phyloproteomische Verwandtschaftsbeziehungen ableiten können. Für eine erfolgreiche Applikation dieser Algorithmen ist die Identifikation der in dieser Arbeit berücksichtigten Biomarker nicht von Nöten.

Ein solcher mathematischer Algorithmus ist in der MALDI Biotyper Software (Bruker Daltonics, Bremen) integriert. Genauer handelt es sich um eine sogenannte PCA-basierte hierarchische Clusterung.

In Machbarkeitsstudien offenbarte sich jedoch ein Problem, das mit der PCA-basierten Clusteranalyse einhergeht: Sowohl die Kulturbedingungen, als auch der Zeitpunkt der Messung haben einen erheblichen Einfluss auf das Ergebnis. Im ersten Versuch hatte sich noch gezeigt, dass die Unterscheidung zwischen zwei Clustern, die nahezu ausschließlich aus *Salmonella Typhi*-Isolaten bestanden, von einem größeren Cluster, das ausschließlich aus nicht-*S. Typhi*-Isolaten bestand, möglich ist (Kuhns *et al*., 2012). Im zweiten Versuch wurden die Salmonellen-Isolate an verschiedenen Tagen und auf verschiedenen Agarplattenchargen kultiviert und gemessen. Nun war festzustellen, dass die unter vergleichbaren Kulturbedingungen angezüchteten Isolate ein gemeinsames Cluster bildeten, wodurch die im ersten Versuch gelungene Differenzierung zwischen Serovar Typhi-/ Nicht-Serovar Typhi-Isolaten nicht mehr möglich war (Kuhns *et al*., 2012).

Der Grund für die unterschiedlichen Ergebnisse ist, dass bei der PCA-basierten Clusterung neben Biomarkermassen auch die Intensität der lokalen Maxima berücksichtigt wird. Die Intensität der lokalen Maxima hängt davon ab, wie stark ein Protein exprimiert wird, was maßgeblich durch die Kulturbedingungen beeinflusst wird. Folglich ist es im Rahmen dieser Analyse essenziell, die Isolate zeitgleich und mit derselben Nährbodencharge anzuzüchten sowie im selben Zeitrahmen zu messen. In einer weiteren Studie wurden diese Bedingungen berücksichtigt, wodurch klinisch relevante *C. jejuni*-Stämme unterschieden werden konnten. Die

Beherzigung der Prozessstandardisierung erlaubte auch die Reproduzierbarkeit der Messungen (Zautner *et al*., 2013).

1.3.1 Proteotypisierung

Die angesprochene Berücksichtigung der Präsenz und Absenz von Biomarkerionen und der stark variablen Intensitäten lokaler Maxima im Rahmen von Gesamtspektrum-Cluster-Algorithmen zur Ableitung phylogenetischer beziehungsweise phyloproteomischer Verwandtschaftsbeziehungen resultiert in fehlerhaften und schwer reproduzierbaren Ergebnissen.

Diese Tatsache veranlasste Zautner *et al*. (2015) zu der Entwicklung einer Methode, welche ausschließlich Veränderungen der Biomarkermasse berücksichtigt, die auf spezifische allele Isoformen desselben Proteins zurückzuführen sind. Diese Methode wurde von unserer Arbeitsgruppe zunächst als Massenspektrometrie-basierte Phyloproteomik (MSPP) bezeichnet, mittlerweile aber in Proteotypisierung umbenannt.

Das Proteotypisierungsschema ist wie folgt aufgebaut: Initial wird das Massenspektrum eines genomsequenzierten Referenzstammes via MALDI-TOF MS aufgezeichnet. Anschließend werden die messbaren lokalen Maxima (Biomarker-Ionen) auf Grundlage der kalkulierten Masse mit den proteinkodierenden Genen korreliert. Mit Hilfe der online verfügbaren Genomdatenbank (NCBI) wird eine eigene Datenbank erstellt, die die Aminosäuresequenzen sämtlicher alleler Isoformen des jeweiligen Proteinbiomarkers enthält. Für jede Isoform wird anschließend die molekulare Masse berechnet. Ein wesentlicher Aspekt dieser Methode ist, dass mögliche posttranslationale Modifikationen (PTM), wie die Abspaltung des *N*-terminalen Methionins, berücksichtigt werden. Erwähnenswert ist, dass Fagerquist *et al*. (2006) zeigen konnten, dass die Massenverschiebungen der Biomarker nicht auf Unterschieden in den PTM, sondern ausschließlich auf Aminosäuresubstitutionen beruhen.

Ist die Isoformenliste erstellt, werden sämtliche zu typisierende Isolate kultiviert. Die Isolate werden anschließend auf zweierlei Weise für die Messung präpariert: Zum einen werden Extraktspektren gemessen, wofür die Proben mit Ameisensäure/ Azetonitril behandelt werden, zum anderen wurden Schmierspektren gemessen. Anschließend erfolgt die Aufzeichnung der Spektren sämtlicher Isolate. Die gemessenen „Roh-Spektren“ werden durch die zugehörige Evaluations-

Software (FlexAnalysis) geglättet und kalibriert. Im nächsten Schritt wird das Spektrum des genomsequenzierten Referenzstammes mit den Spektren der anderen Stämme vergleichend analysiert. Durch Abgleich der Biomarkermassen im Spektrum mit denen der Isoformenliste kann die jeweilige allele Isoform identifiziert werden.

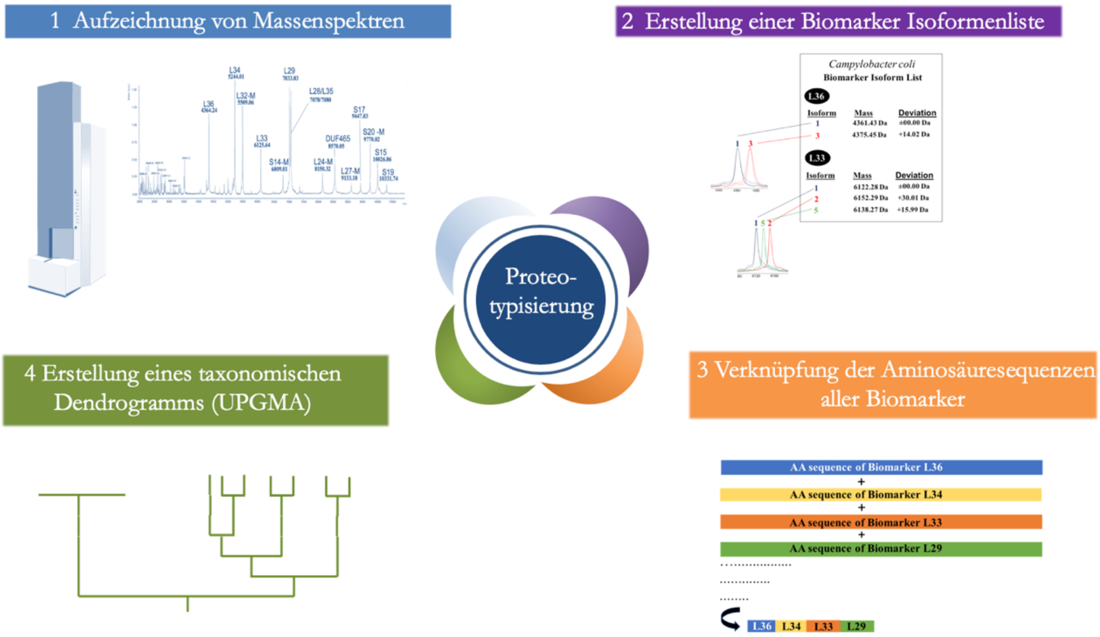


Abbildung 2 Proteotypisierung Grafische Darstellung der wesentlichen Schritte. 1) Aufzeichnung der Massenspektren eines genomsequenzierten Referenzstammes sowie sämtlicher in der Testkohorte enthaltenen Stämme. 2) Erstellung einer allelischen Isoformenliste basierend auf Genomsequenzen aus wgMLST und rMLST Datenbanken (NCBI) gefolgt von einer Analyse der Spektren basierend auf der durch die Isoformenliste vorhergesagten Massen. 3) Verknüpfung der AS-Sequenzen aller Biomarkerionen für jedes Isolat zu einer Sequenz. 4) Ableitung der Phylogenie via UPGMA-Methode.

Für jedes Isolat lässt sich so eine spezifische Isoform für jeden im Typisierungsschema enthaltenen Biomarker ableiten. Anschließend werden für jedes Isolat die spezifischen Biomarker Aminosäuresequenzen fusioniert. Entsprechend der Vorgehensweise bei der MLS-Typisierung kann so für jedes Isolat ein proteotypisierungsbasierter Sequenztyp abgeleitet und final ein phyloproteomisches UPGMA Dendrogramm berechnet werden.

Untersuchungen, die auf der Kombination genetischer Informationen (Genomik) und der Beobachtung lokaler Maxima in MALDI-TOF MS Spektren (Proteomik) beruhen, haben bereits vielversprechende Ergebnisse hervorgebracht:

Unter anderem gelang Ojima-Kato *et al*. (2016) die Differenzierung zwischen *Listeria monocytogenes* und anderen *Listeria*-Spezies (*L. ivanovii, L. grayi, L. innocua*, *L. welshimeri*, *L. seeligeri*, *L. rocourtiae*) basierend auf acht Biomarkern (ribosomalen Proteinen) (Ojima-Kato *et al*., 2016).

In einer anderen Studie konnten auf diese Weise *C. difficile*-Klade 4-Stämme von anderen *C. difficile*-Stämmen abgegrenzt werden (Cheng *et al*., 2018). Eine weitere Studie zur Subtypisierung von *C. difficile* zeigte die Unterscheidbarkeit von *C. difficile*-MLST-Typ 1 Stämmen von anderen MLST-Typen (Corver *et al*., 2018). Darüber hinaus existieren viele weitere, vielversprechende Studien zur Subtypisierung unterschiedlicher Spezies mit dieser Methodik, die im Rahmen der Einleitung nicht alle explizit erläutert werden können (z.B. Suarez *et al*., 2013; Durighello *et al*., 2014; Rizzardi *et al*., 2015; Ortega *et al*., 2018).

Eine Vielzahl der Publikationen zur Erregerdiagnostik via MALDI-TOF MS beruht jedoch auf einer statistischen Analyse der Spektren. Die Proteine, die sich hinter den Biomarkern verbergen, sind jedoch nicht bekannt.

Auf Grundlage des von Zautner *et al*. (2015) entwickelten Proteotypisierungsschemas ist unserer Arbeitsgruppe bereits die Subtypisierung von *C. jejuni jejuni*- (Zautner *et al*., 2015) sowie *C. jejuni doylei* (Zautner *et al*., 2016) Isolaten gelungen. Da bei dieser Methode Genprodukte mit Biomarkerionen assoziiert werden können, ist sie potenziell eine verlässliche Alternative zu den zuvor beschriebenen massenspektrometrischen Untersuchungen. Aufgrund der vielversprechenden Ergebnisse bisheriger Studien wurde in dieser Arbeit untersucht, ob sich die Proteotypisierung auch für die Subtypisierung weiterer klinisch-relevanter Isolate eignet.

Im Folgenden werden nun die im Rahmen dieser Dissertation durchgeführten Arbeiten präsentiert.

2 Ergebnisse

Der Ergebnisteil besteht aus Publikationen beziehungsweise Manuskripten zum Thema Erregercharakterisierung mittels Proteotypisierung, die im Rahmen der Dissertation entstanden sind. Da die Publikationen wie üblich in englischer Sprache verfasst sind, ist den Publikationen jeweils eine Kurzbeschreibung des Inhaltes in deutscher Sprache vorangestellt. Außerdem erfolgt eine Beschreibung

- der Autoren und deren Beitrag zur praktischen Arbeit sowie
- des Status des Manuskripts.

2.1 Differenzierung von *Campylobacter coli*-Subspezies mittels Proteotypisierung

Campylobacter coli ist neben *Campylobacter jejuni* weltweit der häufigste Erreger der bakteriellen Enteritis. Es sind drei Kladen von *C. coli* bekannt, die jeweils mit der Probenquelle assoziiert sind. Während Klade 2 und Klade 3 Isolate in erster Linie in Gewässern und der Umwelt vorkommen, werden Stämme der Klade 1 mit akuter Diarrhö beim Menschen in Verbindung gebracht.

Die phylogenetische Klassifikation von Isolaten erfolgt typischerweise mit Hilfe der relativ aufwendigen Multilokus Sequenz Typisierung (MLST). Ziel dieser Studie war es, ein Typisierungsschema für *C. coli* basierend auf der Proteotypisierungsmethode zu entwickeln und so eine Alternative zu sequenzbasierten Methoden schaffen.

Insgesamt wurden hierfür 97 *C. coli*-Isolate*,* welche die etablierten Kladen der Spezies abdeckten, mittels MALDI-TOF MS analysiert und darauf aufbauend ein *C. coli*-Proteotypisierungsschema entwickelt. Die MLST diente als Referenzmethode.

Verschiedene Isoformen identifizierter Biomarker (ribosomale Proteine) wurden jeweils mit ihren Aminosäuresequenzen assoziiert und in das *C. coli*-Proteotypisierungsschema aufgenommen.

Insgesamt wurden 16 Biomarker identifiziert, die die Unterscheidung der drei Kladen sowie der drei Subkladen der *C. coli*-Klade 1 ermöglichen.

Letztendlich konnte in dieser Studie die Proteotypisierungsmethode erfolgreich für *C. coli* adaptiert werden, was die Unterscheidung der drei etablierten *C. coli*-Kladen und Klade 1 Subkladen ermöglicht. Das wesentliche Ergebnis der Studie ist, dass die einzige klinisch-relevante Klade, Klade 1, von den anderen Kladen abgegrenzt werden kann.

Autoren: **Matthias Frederik Emele**, Sonja Smole Možina, Raimond Lugert, Wolfgang Bohne, Wycliffe Omurwa Masanta, Thomas Riedel, Uwe Groß, Oliver Bader, Andreas Erich Zautner

Beitrag der Autoren zur praktischen Arbeit:

Matthias Frederik Emele:
Dateninterpretation, Bioinformatik, Erstellung von Abbildungen und Grafiken, Anfertigung des Manuskripts.
Sonja Smole Možina:
Sammlung bakterieller Isolate, Dateninterpretation, Korrektur des Manuskripts.
Raimond Lugert:
Bakteriologie, Dateninterpretation, Korrektur des Manuskripts.
Wolfgang Bohne:
Bioinformatik, Korrektur des Manuskripts
Wycliffe Omurwa Masanta:
Bakteriologie, Probenvorbereitung, MLST, Korrektur des Manuskripts
Oliver Bader:
Massenspektrometrie, Studiendesign, Korrektur des Manuskripts
Thomas Riedel:
Genomsequenzierung, Core Genome Alignment, Hinterlegung der Biomarkersequenzen bei GenBank
Andreas Erich Zautner:
Studiendesign, Dateninterpretation, Bioinformatik, Korrektur des Manuskripts

Status des Manuskripts:
Publiziert; Journal: ScientificReports (Nature)

SCIENTIFIC REPORTS

OPEN

Proteotyping as alternate typing method to differentiate *Campylobacter coli* clades

Received: 11 April 2018
Accepted: 25 February 2019
Published online: 12 March 2019

Matthias Frederik Emele[1], Sonja Smole Možina[2], Raimond Lugert[1], Wolfgang Bohne[1], Wycliffe Omurwa Masanta[1,3], Thomas Riedel[4,5], Uwe Groß[1], Oliver Bader[1] & Andreas Erich Zautner[1]

Besides *Campylobacter jejuni*, *Campylobacter coli* is the most common bacterial cause of gastroenteritis worldwide. *C. coli* is subdivided into three clades, which are associated with sample source. Clade 1 isolates are associated with acute diarrhea in humans whereas clade 2 and 3 isolates are more commonly obtained from environmental waters. The phylogenetic classification of an isolate is commonly done using laborious multilocus sequence typing (MLST). The aim of this study was to establish a proteotyping scheme using MALDI-TOF MS to offer an alternative to sequence-based methods. A total of 97 clade-representative *C. coli isolates* were analyzed by MALDI-TOF-based intact cell mass spectrometry (ICMS) and evaluated to establish a *C. coli* proteotyping scheme. MLST was used as reference method. Different isoforms of the detectable biomarkers, resulting in biomarker mass shifts, were associated with their amino acid sequences and included into the *C. coli* proteotyping scheme. In total, we identified 16 biomarkers to differentiate *C. coli* into the three clades and three additional sub-clades of clade 1. In this study, proteotyping has been successfully adapted to *C. coli*. The established *C. coli* clades and sub-clades can be discriminated using this method. Especially the clinically relevant clade 1 isolates can be differentiated clearly.

Intact cell mass spectrometry (ICMS) emerged as the standard method for the identification of microbial species in clinical microbiological laboratories[1–3]. In this method, species identification is not based on the analysis of individual biomarkers or mass spectrometric fingerprints, but on a comparison of the mass spectrum with a microbial spectra database[4] or a database of ribosomal protein sequences taking into account *N*-terminal methionine cleavage[5]. Besides species identification, ICMS allows distinction of subspecies by accurate discrimination based on strain specific biomarkers[6]. It has also been demonstrated that MALDI-TOF MS facilitates the classification of unknown bacterial isolates, based on similarities in the mass spectra of these bacterial isolates with protein biomarker databases, also known as phyloproteomics[7]. Mass spectrometry-based typing methods, generally referred to as proteotyping[8], have been used for about two decades for the characterization of tissues[9], individual proteins[10], microbial communities[11], viruses[12] and, as already mentioned, bacteria. Among others, mass spectrometry (MS) fingerprinting has already been successfully used for subtyping of methicillin-resistant *Staphylococcus aureus* lineages[13], *Clostridioides difficile* PCR ribotypes[14], Shiga-toxigenic *Escherichia coli* strains[15], *Listeria monocytogenes* lineages[16], and *Salmonella* serotypes[17]. In previous studies we have, for example, shown that it is possible to discriminate *Salmonella enterica* ssp. *enterica* serovar Typhi from non-typhi serovars which cause less severe gastrointestinal infections[18]. Also we have shown that it is possible to discriminate different sequence types of *Campylobacter jejuni* ssp. *jejuni* by analyzing isoforms of L32-M[19]. These strain-specific characteristics form the basis for the development of a novel microbial typing method that we initially named Mass Spectrometry-based PhyloProteomics (MSPP)[20,21], which we will, in accordance with the terminology now used in the scientific community[8], refer to as proteotyping, as our method refers to a limited number of biomarkers and

[1]Institut für Medizinische Mikrobiologie, Universitätsmedizin Göttingen, Kreuzbergring 57, 37075, Göttingen, Germany. [2]Department of Food Science and Technology, Biotechnical Faculty, University of Ljubljana, Jamnikarjeva 101, 1000, Ljubljana, Slovenia. [3]Present address: Department of Medical Microbiology, Maseno University Medical School, Private Bag, Maseno, Kenya. [4]Leibniz-Institut DSMZ-Deutsche Sammlung von Mikroorganismen und Zellkulturen, Braunschweig, Germany. [5]Deutsches Zentrum für Infektionsforschung (DZIF), Standort Hannover-Braunschweig, Braunschweig, Germany. Correspondence and requests for materials should be addressed to A.E.Z. (email: azautne@gwdg.de)

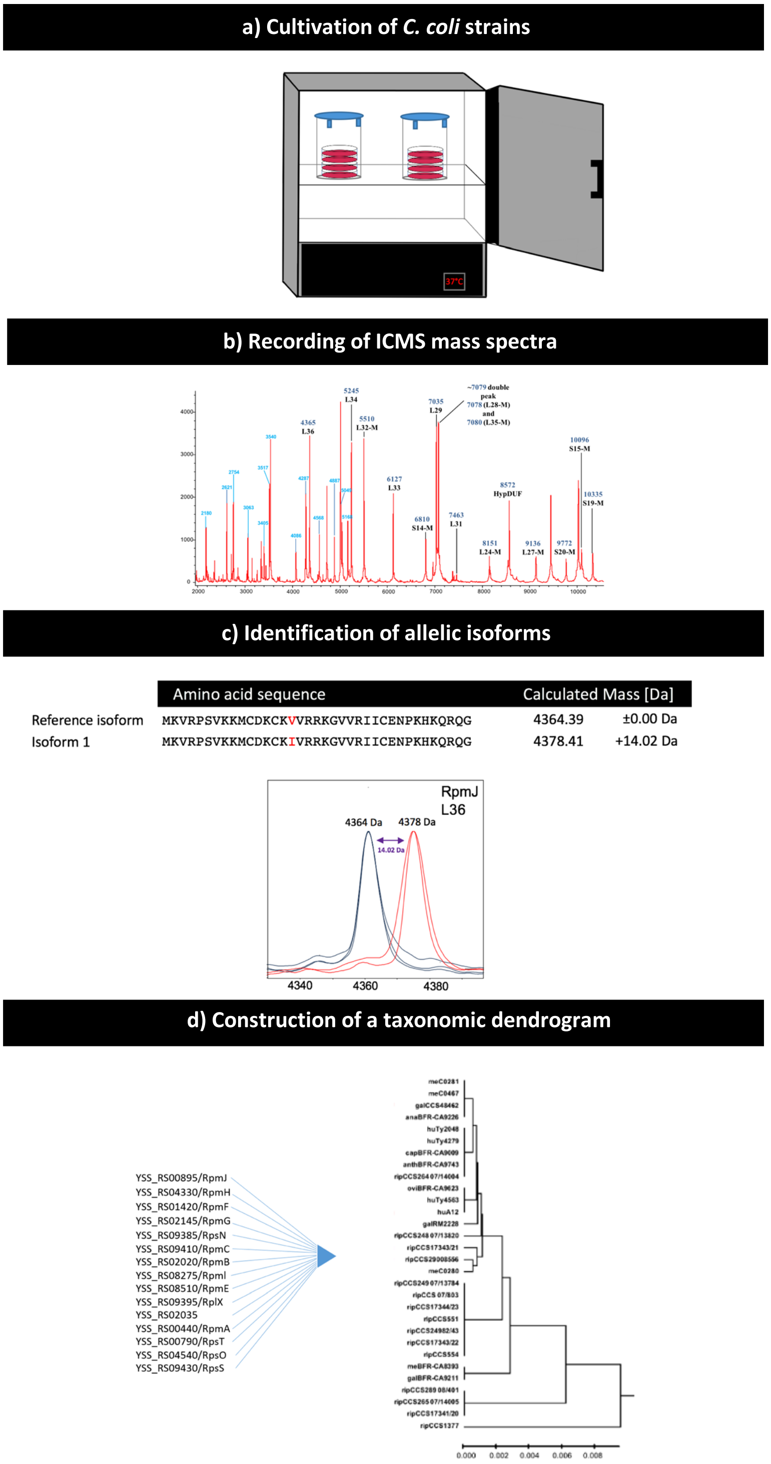
not to all the proteins present in the sample. At the core of the method of proteotyping is an amino acid sequence list of all isoforms that have evolved through non-synonymous mutations in the biomarker genes. These isoforms can be recognized as mass shifts in a superposition of calibrated MALDI-TOF spectra. For each bacterial isolate to be typed, the proteotyping scheme can be used to derive a combination of amino acid sequences from the detected biomarker masses. The functionality of this approach was proven by comparison of proteotyping to the current gold standard multilocus sequence typing (MLST)[22]. The advantage of proteotyping over whole spectrum clustering approaches is that only mass changes associated with a particular set of allelic isoforms of the same protein are considered for phylogeny derivation. Other methods take into account the presence or absence of individual masses as well as peak intensity, what delivers less accurate results[20]. Proteotyping provides further advantages in comparison to common subtyping methods like MLST, ribosomal MLST (rMLST) or whole-genome MLST (wgMLST). MLST has the problem of combining sufficiently variable genes into a typing scheme in order to map phylogenetic relationships[23]. Another disadvantage is that it only considers sufficiently variable core genes, whereas hypervariable, transposable gene sites and the entire genome sequence are not considered[24]. Even well-established whole genome sequencing-based MLST schemes are very expensive and time-consuming[25–27]. Therefore, these methods are not used in everyday clinical routine diagnostics and subtyping of microorganisms is currently restricted to a limited cohort, mostly in epidemiological surveys. In the light of the above, a fast and precise subtyping method like proteotyping enables the conduction of numerous experiments that involve the determination of phylogenetic relatedness.

Besides *C. jejuni*, *C. coli* is the most common bacterial cause of gastroenteritis worldwide[28,29]. The housekeeping genes of *C. jejuni* and *C. coli* exhibit 86.5% sequence identity[30], similar to that observed between the enteric bacteria *E. coli* and *S. enterica*, which are well studied and thought to have diverged 120 million years ago[31]. *C. coli* can be subdivided into three genetic clades, which differ in various ways. Clade 1 isolates of *C. coli* are most frequently isolated from farm animals and clinical stool samples of humans suffering from acute diarrhea, whereas clade 2 and clade 3 strains, which are more closely related to each other, are mainly found in environmental waters and samples from waterfowl[32–35]. In a previous study, Sheppard and coworkers showed, that all of the examined cases of human *C. coli* infection were caused by lineages belonging to clade 1[33]. Clade 1 is further subdivided into two clonal complexes: ST-828, which makes up 70.5% of the *C. coli* isolates, and ST-1150, which makes up 4.5% of *C. coli* isolates, whereas clades 2 and 3 do not exhibit a clonal complex substructure[33]. An examination of the divergence in *C. jejuni* estimated the speciation of *C. jejuni* and *C. coli* to have occurred 6580 years ago and clonal complex sub-structuring even more recently[36]. For the maintenance of the three *C. coli* clades, gene pools of these clades have to be kept separate. A simple explanation for how these gene pools are kept separate would be through a general reduction in the overall level of recombination by recombinational barriers, but as previously mentioned, there is frequent recombination within each clade[33]. In principle, three kinds of recombinational barriers can be described. The first kind of recombinational barrier that enables the maintenance of the *C. coli* clade system are mechanistic barriers, which are imposed by the homology dependence of recombination[37] or other factors, like modification and restriction systems[38]. The second kind of recombinational barriers are ecological barriers, meaning a physical separation of bacterial populations in distinct niches. The third are adaptive barriers, describing a selection against hybrid genotypes[39]. Subtypes belonging to *C. coli* clade 1 numerically dominate in clinical samples. It is possible that there are genomic differences affecting the pathogenicity of *C. coli* clade 1 isolates but these differences are not required to explain the overrepresentation of this clade in human samples as isolates of this clade plainly dominate in disease reservoirs and food chain sources[33]. Comparative analysis of *C. coli* clades suggests that potential virulence factors and resistance mechanisms are not restricted to a single clade. Genes encoding proteins involved in chemotaxis and capsule formation were observed in different clades of *C. coli*[40]. The clustered regularly interspaced short palindromic repeat (CRISPR) locus, which is considered to serve as prokaryotic immune system and protection against invasion of alien genetic elements is also present in all *C. coli* clades, although its genomic location differs[41,42]. Also, the cytolethal distending toxin (*cdt)* genes are reported to be ubiquitous in all *C. coli* strains[43–46]. The *cdt* genes are well conserved in *C. coli*, although size and sequences of the respective genes do vary between strains[47].

In this study, we have established a proteotyping scheme for subtyping of *C. coli* isolates. *C. coli* isolates from different sources were MLST-typed and therewith it was shown that our test cohort included isolates of all three established clades and subclades. These isolates were typed by ICMS/proteotyping and their phyloproteomic relatedness was deduced. Comparison of the obtained phyloproteomic proteotyping-based unweighted pair group method with arithmetic mean (UPGMA) tree with the corresponding MLST-based UPGMA dendrogram demonstrated that proteotyping is able to differentiate the clinically relevant clade 1 isolates from clade 2 and 3 isolates.

Results and Discussion

Previously, we have established a standard workflow for setting up a new proteotyping (MSPP) scheme and a proteotyping procedure[20]. Following this workflow for *C. coli*, (i) we recorded a mass spectrum of the genome sequenced *C. coli* reference strain RM2228 (ATCC BAA-1061) and assigned ICMS spectrum masses to open reading frames; (ii) we have compiled a collection of allelic isoforms of the assignable spectrum masses by analyzing the total 1,565 *C. coli* sequence datasets deposited in the wgMLST and rMLST databases. Accordingly, we were able to calculate a frequency distribution of the individual allelic isoforms based on these 1,565 *C. coli* genomes (Supplementary Table 2). According to the proteotyping scheme (Fig. 1), the spectra of the 97 cultured *C. coli* isolates were recorded, following pre-processing and calibration. Mass shifts in comparison to the *C. coli* reference strain RM2228 were estimated and the allelic isoforms were assigned by matching of the measured biomarker mass with the calculated masses from the isoform database set. A phyloproteomic proteotyping-based UPGMA-tree was calculated after fusing the amino acid sequences of all biomarker ions included in the *C. coli* proteotyping scheme for each tested isolate.



	Amino acid sequence	Calculated Mass [Da]	
Reference isoform	MKVRPSVKKMCDKCKVVRRKGVVRIICENPKHKQRQG	4364.39	±0.00 Da
Isoform 1	MKVRPSVKKMCDKCKIVRRKGVVRIICENPKHKQRQG	4378.41	+14.02 Da

Figure 1. Proteotyping workflow (**a**) Culturing *C. coli* strains under microaerophilic conditions. (**b**) Recording of MALDI-TOF mass spectra. (**c**) Designation of allelic isoforms by comparison of mass spectra of all measured *C. coli* strains with the allelic isoform list established on the basis of sequence data available in the wgMLST and rMLST databases. (**d**) Concatenation of the amino acid sequences of the identified isoforms into a single continuous sequence and calculation of a taxonomic dendrogram (UPGMA).

Identification of biomarker ions.

With reference to the genome sequence of the *C. coli* strain RM2228, 16 single charged biomarker masses, in the range of 4,000 and 10,500 *m/z*, were associated to a specific gene (Figs 2 and 3). The standard deviation for a measurement representing a sum of 6 recordings was less than 0.8 Da and the difference between measured mass and calculated average mass was at maximum 1.35 Da (Supplementary Table 3). The identified biomarkers were RpmJ (L36; 4365 Da), RpmH (L34; 5245 Da), RpmF (L32-M; 5510 Da), RpmG (L33; 6127 Da), RpsN (S14-M; 6810 Da), RpmC (L29; 7035 Da), RpmB (L28-M; 7078 Da), RpmI (L35-M;

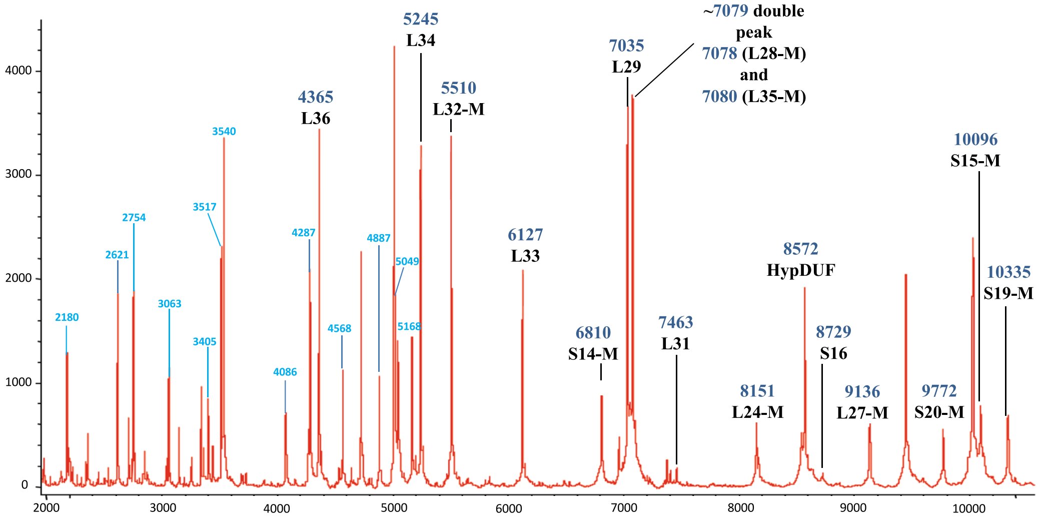


Figure 2. Mass spectrum of the genome sequenced *C. coli* reference strain RM2228. Singularly charged biomarker ions identified by comparison of measured molecular masses with calculated masses based on the reference genome are marked in black, doubly/multiply charged ions are labeled in blue, and two so far not identified biomarker ions are labeled with a question mark "?". The peak at $m/z \approx 7{,}079$ corresponds to a fused double peak of biomarkers L28-M ($m/z = 7{,}078$) and L35-M ($m/z = 7{,}080$). In *C. coli* isolates of the MLST-Clade 3, there is an allelic isoform for L28-M, which has a molecular weight 16 Da lower than the mass of L35-M and therefore two single peaks for L28-M and L35-M can be registered instead of the double peak (see Fig. 3).

7080 Da), RpmE (L31; 7463 Da), RplX (L24-M; 8151 Da), hypothetical protein DUF465 (Cj0449c homologue; 8572 Da), RpsP (S16; 8729 Da), RpmA (L27-M; 9136 Da), RpsT (S20-M; 9743), RpsO (S15-M; 10096 Da), and RpsS (S19-M; 10335 Da). The genes of the 16 biomarker proteins included in the *C. coli* proteotyping scheme are distributed across the entire genome of strain RM2228, similar to the seven established MLST markers, and are therefore suitable for the derivation of phylogeny.

These 16 biomarkers are generally identical to those in the proteotyping scheme of *C. jejuni* ssp. *jejuni* and *C. jejuni* ssp. *doylei*[20,21]. Differences were that in case of RpsU (S21; 9140.9 Da), RpsQ (S17; 9591.5 Da), and RplW (L23; 10554.3 Da), as well as in case of their de-methioninated isoforms, no visible peak could be detected in any of the examined *C. coli* strains. Therefore, these three biomarkers were not included in the current *C. coli* proteo-typing scheme.

In contrast to *C. jejuni* ssp. *doylei*, the biomarker L22-M could be detected in the *C. coli* mass spectrum and therefore included in the scheme. L22-M was de-methioninated as in the mass spectrum of *C. jejuni* ssp. *jejuni*.

The *N*-terminal methionines of the biomarkers S14-M, S20-M, L24-M, and L32-M were cleaved off in *C. coli* as well as in *C. jejuni* ssp. *jejuni* and *C. jejuni* ssp. *doylei*.

However, five differences were found with respect to the posttranslational modification of the biomarkers by proteolytic removal of the *N*-terminal methionine: In comparison to *C. jejuni* ssp. *jejuni*, the *N*-terminal methionine of the biomarker ions S15, S19, L28, and L35 is removed in *C. coli*, which is also the case with *C. jejuni* ssp. *doylei*[21].

As with *C. jejuni* ssp. *jejuni*, but in contrast to *C. jejuni* ssp. *doylei*, the *N*-terminal methionine of L27 remains attached in *C. coli*.

Since all five differences were observed in each case for all isolates of the different *Campylobacter* species or sub-species, this confirms the findings of Fagerquist and coworkers that the post-translational modifications are species- and sub-species-specific but not isolate-specific[48]. Accordingly, one can distinguish the three *Campylobacter* species or sub-species solely on the basis of the presence or absence of the *N*-terminal methionine of L27 and S15, S19, L28, or L35.

Establishment of an allelic isoform list. In the next step, we compiled a collection of allelic isoforms of each of the 16 biomarkers of the *C. coli* proteotyping scheme. For this purpose, we used the 1,565 *C. coli* genome sequences available in the wgMLST and rMLST databases.

The gene sequence deposited for each biomarker isoform was translated into an amino acid sequence and aligned. Subsequently, the molecular mass for each individual isoform was calculated. Between 3 and 9 isoforms for each biomarker ion could be identified within the data received from the rMLST and wgMLST databases. The frequency of occurrence of isoforms varied from >99% to a single occurrence of the isoform, where in cases of single occurrences in the rMLST and wgMLST databases, a sequencing error must also be considered. For each of the 16 biomarkers, at least two isoforms with a relative increased frequency were found in the database, which means that these masses can serve as phylogenetic discriminators (Supplementary Table 2).

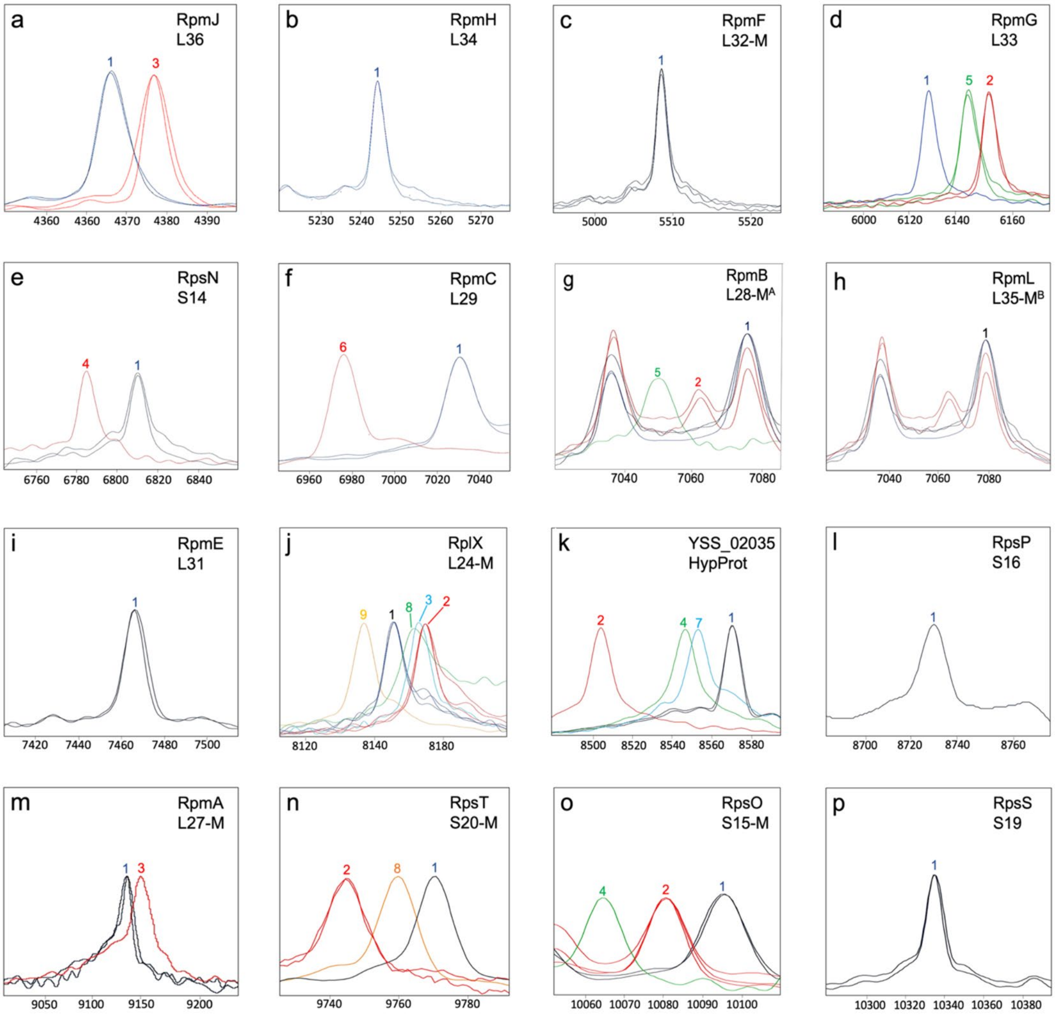


Figure 3. *C. coli*-specific proteotyping biomarkers (**a**–**o**). Spectra of representative *C. coli* strains were superimposed to illustrate the mass differences between allelic isoforms detected in our *C. coli* collection. X-Axis: mass [Da] charge-1 ratio, scale 200 Da. Y-Axis: intensity [10x arbitrary units], spectra were individually adjusted to similar noise level for better visualization of low-intensity peaks. Color codes: the isoform of *C. coli* reference strain RM4661 is illustrated in blue; red, light green, dark green, purple and orange are further isoforms. Isoforms lacking *N*-terminal methionine are appended with "-M". ^{A}(**g**) The peak at $m/z \approx 7{,}079$ is a superposition of the biomarker ion masses L28-M ($m/z = 7{,}078$) and L35-M ($m/z = 7{,}080$). In contrast, the allelic isoforms 2 and 3 (−14 Da and −28 Da, respectively) are mere L28-M peaks. ^{B}(**h**) For the biomarker L35-M we could only detect one allelic isoform in our test cohort, which is superimposed by the biomarker mass L28-M in the spectrum of *C. coli* RM2228. In order to show the not superimposed L35-M peak in h an additional spectrum of a clade 3 *C. coli* isolate was added, in which the L28-M peak is shifted by −14 Da and therefore L35-M is not superimposed.

MLST Typing of a microbial isolate collection. To validate the *C. coli* proteotyping scheme a cohort of 101 isolates (*C. coli* reference strain RM2228, 96 *C. coli* Isolates, and 4 *C. jejuni* isolates) was typed by both MLST and proteotyping. The isolates were chosen in such a way that all clades and sub-clades were represented. According to the MLST results, 83 isolates belonged to clade 1. Out of these clade 1 isolates, six belonged to the sub-clade 1B and two further belong to sub-clade 1 C, while the remaining 75 isolates formed sub-clade 1A (ST828). These clade 1 isolates were mainly isolated from human faeces (19), and food-associated samples like chicken meat (21), waterfowl (7), turkey meat (6), swine meat (6), and cattle (5). But only four isolates originated from environmental water. Seven isolates, originating from environmental water, belonged to clade 2, and three isolates also originating from environmental water belonged to clade 3. Additionally, we included four isolates outside the defined MLST clades, but also identified as *C. coli* by conventional MALDI-TOF MS. MLST results of three of these four isolates meC0280 (ST6994), mecC0281 (ST6992), and meC0467 (ST6993) originating from turkey cloacal swabs suggested a closer relationship to *C. jejuni* and the fourth isolate CCS1377 (ST7908), an environmental water isolate, formed a separate clade in between clade 2 and clade 3 (Supplementary Fig. 1).

ORF No. (RM4661)	Gene product	Forward primer (5′ → 3′)	Reverse primer (5′ → 3′)	Amplicon length [bp]
YSS_RS00895	RpmJ/L36	AGCTGCTGCTTCATCTTCACT	AGCCTTGATAAAGGGCGTATC	490
YSS_RS04330	RpmH/L34	AAATGCTCGGGCAAATTGATTA	GCCATCGCAATACCACTTTT	512
YSS_RS01420	RpmF/L32	TGCACCACTATGTCCTGCTG	TGCCACAATGCAAGGTTTTGT	728
YSS_RS02145	RpmG/L33	AGCTGATGGCGTTGAAATGG	ACCCCCAACCATCGGATTTG	430
YSS_RS09385	RpsN/S14	ACACGACGACCTGGTTTAGA	TCGGTCTTGATGAGCAGTTGA	611
YSS_RS09410	RpmC/L29	GGTCTGCATTCAACCGCTAC	GCCAAATTGAAGCAGCTCGT	668
YSS_RS02020	RpmB/L28	CGTCAAGTTCATTATGGCGCT	TGGAACAAAATGCCCGTCCA	742
YSS_RS08275	RpmI/L35	GCAAGCAGCATTGATACGCA	GCTTGGCTATTTTGCAAAGGATT	715
YSS_RS08510	RpmE/L31	GCAAGGTTTTTCCTGATGCTGT	TGGCATACCCGCATCACTC	756
YSS_RS09395	RplX/L24	TCGGAACTCGTATCTTTGGGC	CAGGAAAACCTTCACGCACT	578
YSS_RS02035	DUF465	GCTGCTGGGTAAGATTTTGGT	TCGTGTAACCCTAGAAGATGGC	584
YSS_RS00440	RpmA/L27	AGTTAGCGTTGGCGATGAGTT	AACGAAGATGATATCCCCGCC	783
YSS_RS00790	RpsT/S20	GCTCTTCTTCGAGTTTGGGTT	GGTGGATTGGGTGTTATGCT	765
YSS_RS04540	RpsO/S15	ATATCGGATACAACCGCGCA	GCATACTCGCTAGCTTTGGT	636
YSS_RS09430	RpsS/S19	AGCACCAGCATCTACACGAC	ATGGCAAGTATCGGCGAAGT	782

Table 1. Oligonucleotide primers used for sequencing of the *C. coli* biomarker genes included in the proteotyping scheme.

Identification of allelic isoforms. Measurements of the isolates of the study cohort were performed in the same way as for the reference strain *C. coli* RM2228. Allelic isoforms were identified by comparison of the masses of candidate allelic isoforms to the reference spectrum of *C. coli* RM2228 and by matching the mass differences with the isoform list. For isoforms with the same mass difference to the reference in RM2228, or more precisely, with the same amino acid substitutions, but at different positions in the amino acid sequence, additional DNA sequencing was done using the primers listed in Table 1.

Within this study population, we detected five isoforms for RplX (L24-M) and four isoforms for protein DUF465. Three isoforms each were detected for RpmG (L33), RpmB (L28-M), RpsT (S20-M), RpsO (S15-M) and two isoforms each for RpmJ (L36), RpsN (S14-M), RpmC (L29) and RpmA (L27-M). For RpmH (L34), RpmF (L32-M), RpmI (L35-M), RpsP (S16) and RpsS (S19-M) only one isoform was detected (Fig. 3, Supplementary Table 2).

Computing of a phyloproteomic UPGMA-dendrogram. The amino acid sequences of the 16 identified biomarker isoforms were concatenated to one continuous sequence for each isolate, which was in turn used to compute a phyloproteomic tree by conventional clustering algorithms (UPGMA).

Within our test cohort, the combined amino acid sequences in our collection yielded 12 (plus two for *C. jejuni*) different proteotyping-based sequence types. For an evaluation of the constructed proteotyping-based UPGMA-tree, an MLST-based UPGMA-tree was computed for comparison. This was done with 30 *C. coli* isolates and 4 *C. jejuni* isolates representative of all MLST clades and sub-clades as well as all 12 proteotyping-derived types. For clarity, the complete test cohort was reduced from 101 isolates to 34 representative isolates. The UPGMA-tree deduced from the concatenated biomarker protein sequences was generally concordant with MLST results (Fig. 4).

The *C. coli* proteotyping scheme was clearly able to distinguish *C. jejuni* and *C. coli* isolates. Since the three biomarkers RpsU/S21, RpsQ/S17, and RplW/L23 were not detectable in the *C. coli* mass spectrum, the *C. coli* proteotyping scheme had to be reduced by these three biomarkers, which nevertheless still allows sufficient differentiation between the two microbial biospecies. As already stated above, it is feasible to distinguish both microbial species solely on the basis of the presence or absence of the *N*-terminal methionine of the biomarkers of L27 and S15, S19, L28, or L35. In addition, there are allelic isoforms of the biomarkers, which are characteristic for each of the biospecies e.g.: L32-M − T48N; L31 − T23V + A29S + N38S; and S20-M − N41K + G42N (using *C. jejuni* NCTC 11168 as reference strain).

Furthermore, the *C. coli*-specific proteotyping scheme precisely discriminated isolates belonging to different clades, illustrated by the absence of crossing connection lines of different colors in Fig. 4. All isolates of sub-clade 1A, and of the sub-clades 1B and 1C as well as of clade 2 and 3 form individual clusters. However, only the sub-clades 1A and 1B form neighboring clusters, while the isolates of sub-clade 1C are to be found between the clades 2 and 3.

Besides the isolates representing the well-established clades and sub-clades of *C. coli*, four isolates not belonging to either of these clades were included in our study: CCS1377, meC0280, mecC0281, and meC0467.

Isolate CCS1377 is, in both the MLST-based and the proteotyping-based dendrograms, a single isolate placed outside the *C. coli* clades, which is evolutionarily more closely related to *C. jejuni*.

In contrast, the three isolates meC0280, mecC0281, and meC0467, which form a separate clade in the MLST-based neighbor-joining tree branching off at the basis of the *C. jejuni* branch (Supplementary Table 1), did not form a common cluster in the proteotyping-based tree. The isolates mecC0281 and meC0467 clustered together with the clade 1 A isolates, in contrast meC0280 clustered together with the isolates of sub-clade 1B. Using a whole genome neighbor-joining parsnp algorithm as reference we could demonstrate that the isolates

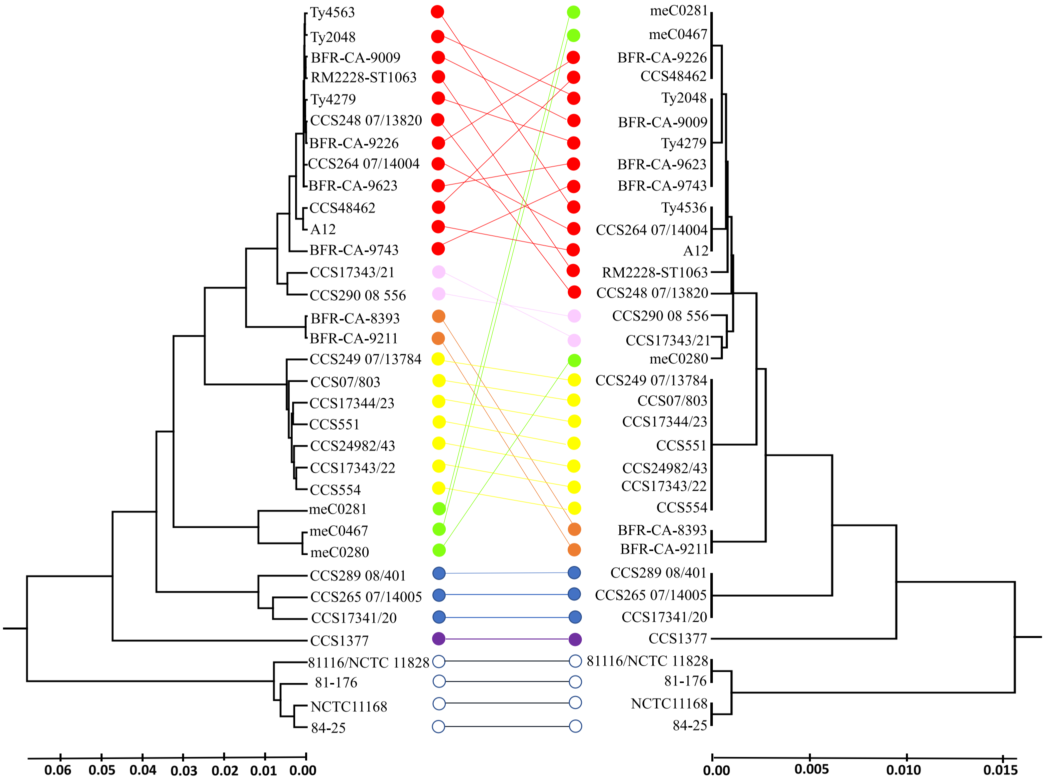


Figure 4. Comparison of MLST-based and proteotyping-based UPGMA dendrograms. The MLST-based phylogenetic tree (left) as well as the proteotyping-based dendrogram (right) were constructed by UPGMA. The MLST dendrogram resulted from 7 loci, the proteotyping-based dendrogram from the amino acid sequences of 16 identified biomarker ions. The different *C. coli* clades and sub-clades are represented by different colors. In addition, four *C. jejuni* isolates have been included in the illustration, which form their own *C. jejuni* clade. Color codes: clade 1A – red, clade 1B – pink, clade 1C – orange, clade 2 – yellow, clade 3 – blue, isolate CCS1377 – purple, isolates meC0280, mecC0281, and meC0467 – green, *C. jejuni* isolates – white. Lines connect the corresponding isolates in the different trees. As it can be seen, there are only crossings of connecting lines within one clade (corresponding to one color), whereas different colors (clades) do not cross each other. This demonstrates that proteotyping can be used to distinguish the clades clearly from each other. The only exceptions are the three isolates meC0280, mecC0281, and meC0467 labeled in green. These form their own clade in the MLST-based tree (Supplementary Fig. 1), but in the core genome alignment (Supplementary Fig. 2) they cluster with *C. coli* clade 1. This means that for isolates of this group the proteotyping-based tree is similar to a core genome alignment, while MLST is less suitable.

meC0280, mecC0281, and meC0467 integrate into the cluster of clade 1 *C. coli* isolates (Supplementary Fig. 2). Therefore, the clustering in the proteotyping-based UPGMA-tree corresponds more closely to the clustering of the whole genome neighbor-joining parsnp-tree. Here proteotyping proves to be a sufficient differentiation tool that seems superior to 7-gene MLST-based phylogeny.

In summary, our proteotyping scheme clearly differentiates the clinically relevant clade 1 isolates from the other clades. If this scheme would be integrated into a subtyping module of the mass spectrometry evaluation software, we would be able to determine the clade and the clinical relevance of an isolate as early as in the mass spectrometric species determination phase.

Materials and Methods

Campylobacter coli and *Campylobacter jejuni* isolates.

A total of 101 *Campylobacter* isolates were included in the presented study. Of these were 97 *C. coli* Isolates including 21 isolates from chicken, 19 from human feces (clinical isolates of patients with campylobacteriosis), 15 from environmental water, 9 from turkey, 7 from water fowl, 6 from swine, 5 from cattle, 3 from wild bird, 3 from sheep, 2 from goat feces, 2 from ape feces, 2 from wild boar, and one from deer, bivalves and Eurasian otter. Twenty four of these isolates (including all 15 riparian and 9 chicken isolates) were provided by the Department of Food Science and Technology, at University of Ljubljana, Slovenia; 54 isolates (animal isolates) were provided by the German *Campylobacter* Reference Center of the Bundesinstitut für Risikobewertung (Federal Institute for Risk Assessment) in Berlin, Germany; 19 isolates (human isolates) originated from stool samples of suspected campylobacteriosis patients treated at the University Medical Center Göttingen, Germany. The genome-sequenced *C. coli* reference strain, RM2228, as well as the four *C. jejuni* reference strains NCTC 11168, 81–176, 84–25, and 81116/NCTC 11828 were obtained from the National

Collection of Type Cultures (NCTC), Salisbury, UK, Manassas, Virginia, USA. The isolates, especially the subset for Fig. 4, were picked so that the test collection represented a high genetic diversity. Initial species identification was performed using the MALDI Biotyper system (Bruker Daltonics, Bremen, Germany). Results with MALDI Biotyper identification score values ≥ 2.000 were assessed as correct. Additionally, the well-established multiplex polymerase chain reaction of Vandamme and coworkers was used to distinguish between *C. jejuni* and *C. coli*[49].

Bacterial culture. *C. coli* and *C. jejuni* strains were stored for long-term storage in Cryobank tubes at −80 °C (Mast Diagnostica, Reinfeld, Germany). For the experiments, they were incubated as one batch overnight under microaerophilic conditions (5% O_2, 10% CO_2, 85% N_2) on Columbia agar base (Merck, Darmstadt, Germany) supplemented with 5% sheep blood (Oxoid Deutschland GmbH, Wesel, Germany). Experiments were carried out under biosafety level 2 conditions.

The preparation of a matrix solution containing human insulin. To prepare the matrix solution, the matrix substance, purified with α-cyano-4-hydroxy-cinnamic acid (HCCA; Bruker Daltonics, Bremen, Germany), was dissolved in the standard solvent consisting of acetonitrile 50%, water 47.5% and trifluoroacetic acid 2.5%. The resulting concentration was 10 mg HCCA/mL. Recombinant human insulin (Sigma-Aldrich, Taufkirchen, Germany) in HCCA solution was added to serve as an internal calibrant for spectrum evaluation. The final concentration of human insulin in 50% aqueous acetonitrile was 10 pg/μL. The exact determination of the insulin peak mass was carried out experimentally by mixing with the Bruker Test Standard and consecutive recording of mass spectra. The insulin peak was detected at an $m/z = 5,806.1$. The insulin peak functioned as an internal calibrant for all *C. coli* mass spectra. Insulin proved to be particularly suited, because its mass did not coincide with other recorded biomarker masses. The use of an internal calibrant significantly increases precision in the determination of biomarker mass changes. With this approach, we were able to detect mass differences with a standard deviation of less than 1 Da.

Recording MALDI-TOF mass spectra. The preparation of the samples used in MALDI-TOF MS was carried out in two variants: by smear preparation and extraction. Five colonies of an overnight agar plate culture were harvested for the preparation of the extract samples and then given into 300 μL double distilled water. The colonies were suspended by rigorous mixing. Subsequently, 900 μL absolute ethanol was added and the suspension was thoroughly mixed by repeated up-and-down pipetting. After complete suspension of the bacterial cells, the suspensions were centrifuged for 1 minute at 13,000 × g. Subsequently, the supernatant was discarded and the pellets dried at room temperature for 10 minutes. By vortexing during the drying process, the pellet was thoroughly resuspended in 50 μL of 70% formic acid. In the next step, 50 μL acetonitrile was added to each tube and mixed with the pipette, followed by centrifugation of the mixture at 13,000 × g for 2 min. After centrifugation, 1 μL of the supernatant was pipetted onto a sample position on a polished steel MALDI target plate, and was left to dry for about 5 minutes at room temperature. Subsequently, each sample position was coated with 1 μL of the HCCA matrix containing the internal calibrant, human insulin. Again, the matrix-coated target was left to dry at room temperature. Once the matrix had dried, the samples were ready for mass spectrometric measurement[50]. Recording of the mass spectra was performed according to the standard recommendations for the MALDI Biotyper System (Bruker Daltonics, Bremen, Germany). Six hundred spectra in a mass range of 2–20 kDa were recorded in 100-shot steps on an Autoflex III system and summed up. Only if the MALDI Biotyper identification score values were ≥2,000 they were judged to be valid.

Assignment of specific allelic isoforms to biomarker ions in mass spectra. Analysis of mass spectra was performed using FlexAnalysis and the algorithms implemented therein (Bruker Daltonics, Bremen, Germany). First, the spectra were calibrated internally to the set insulin peak ($m/z = 5,806.1$), followed by subsequent pre-processing, baseline subtraction and smoothing. The theoretical average molecular weight of the proteins that correspond to any open reading frame (see Supplementary Table 3) was derived from the amino acid sequence using the molecular weight calculator in the ExPASy Bioinformatics Resource Portal (http://web.expasy.org/compute_pi/). It is important to note that posttranslational modifications occasionally occur in ribosomal proteins of *Enterobacteriaceae*. For this reason, further optional molecular weights had to be considered for each open reading frame[51]. Plausible post-translational modifications are proteolytic removal of *N*-terminal initiator methionine (iMet) which was considered to result in a mass difference of −131 Da, *N*-terminal acetylation[52,53] and the presence or absence of disulfide bonds for example in the calibrant human insulin.

The unambiguous naming of biomarker masses, i.e. the assignment of a biomarker peak to a specific allelic isoform, was done by comparing the measured masses with the calculated masses from the reference *C. coli* RM2228 genome. If there was no clear correspondence between a biomarker mass in the recorded mass spectrum of a specific isolate in the test cohort to the mass calculated from the *C. coli* RM2228 reference genome, the biomarker mass identification was done by matching the measured biomarker mass to calculated masses in entries of the ribosomal MLST (rMLST) database or the whole genome MLST (wgMLST) database, respectively. If still no unambiguous matching was found for the biomarker mass, the mass spectrum was examined for peaks with biomarker masses that correspond to possible mass shifts due to mutations in the original biomarker resulting in amino acid exchanges (Supplementary Table 2).

Every recorded allelic isoform in the test cohort was reconfirmed by amplification via PCR and consecutive Sanger sequencing of the obtained amplicons (Seqlab, Göttingen, Germany). All primers used in the experiment are listed in Table 1. The parameters of the PCR reactions were set as follows: pre-denaturation at 94 °C for 300 sec; denaturation at 94 °C for 30 sec; annealing at 55 °C for 30 sec; elongation at 72 °C for 30 sec; repetition for 30 amplification cycles; post-elongation at 72 °C for 600 sec. In each of the isolates studied, the predicted mutations were found in the genes encoding the corresponding biomarker protein, which proved the identities of the peaks.

Both, nucleotide and amino acid sequences of the allelic isoforms of biomarkers newly described during the study, have been deposited at the Genbank. The accession numbers of all biomarkers (nucleotide and amino acid sequences) are listed in Supplementary Table 4. MLST sequence types of all isolates analyzed in the study have been deposited at the *Campylobacter* MLST database (https://pubmlst.org/campylobacter/).

Calculation of phylogenetic and phyloproteomic dendrograms. The Molecular Biology and NGS Analysis Tool Geneious V11.1.2 (http://www.geneious.com) was used to translate and align the protein sequences taken from the rMLST and wgMLST databases. Additionally, Geneious was used to trim and align sequences from confirmatory sanger sequencing[54].

Calculation of the MLST- and proteotyping-based UPGMA-dendrogram was done with the help of the MEGA7 software[55]. For the assignment of MLST sequence types and clonal complexes, the *C. coli*/*C. jejuni* MLST website (https://pubmlst.org/campylobacter/) was consulted[56]. The evolutionary history was inferred using the Neighbor-joining method[57]. The evolutionary distances were computed using the Maximum Composite Likelihood method[58]. Core-genome alignments were computed using Parsnp and FastTree2[59] was used to calculate the maximum-likelihood (ML) phylogenetic tree. Parsnp and FastTree2 are both implemented in the Harvest package[60].

Ethical Approval. Ethical approval for the study was obtained from Ethics Commission of the University Medical Center Göttingen, Germany. No humans, animals, or personalized data were used for this study.

References

1. Seng, P. *et al*. MALDI-TOF-mass spectrometry applications in clinical microbiology. *Future Microbiol* **5**, 1733–54 (2010).
2. Bader, O. MALDI-TOF-MS-based species identification and typing approaches in medical mycology. *Proteomics* **13**, 788–99 (2013).
3. Zingue, D., Flaudrops, C. & Drancourt, M. Direct matrix-assisted laser desorption ionisation time-of-flight mass spectrometry identification of mycobacteria from colonies. *Eur. J. Clin. Microbiol. Infect. Dis. Off. Publ. Eur. Soc. Clin. Microbiol.* **35**, 1983–1987 (2016).
4. Seng, P. *et al*. Ongoing Revolution in Bacteriology: Routine Identification of Bacteria by Matrix-Assisted Laser Desorption Ionization Time-of-Flight Mass Spectrometry. *Clin Infect Dis* **49**, 543–51 (2009).
5. Pineda, F. J. *et al*. Microorganism identification by matrix-assisted laser/desorption ionization mass spectrometry and model-derived ribosomal protein biomarkers. *Anal. Chem.* **75**, 3817–3822 (2003).
6. Lartigue, M. F. Matrix-assisted laser desorption ionization time-of-flight mass spectrometry for bacterial strain characterization. *Infect Genet Evol* **13**, 230–5 (2013).
7. Conway, G. C., Smole, S. C., Sarracino, D. A., Arbeit, R. D. & Leopold, P. E. Phyloproteomics: species identification of Enterobacteriaceae using matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. *J Mol Microbiol Biotechnol* **3**, 103–12 (2001).
8. Karlsson, R. *et al*. Proteotyping: Proteomic characterization, classification and identification of microorganisms–A prospectus. *Syst. Appl. Microbiol.* **38**, 246–257 (2015).
9. Shillingford, J. M. *et al*. Proteotyping of mammary tissue from transgenic and gene knockout mice with immunohistochemical markers: a tool to define developmental lesions. *J. Histochem. Cytochem. Off. J. Histochem. Soc.* **51**, 555–565 (2003).
10. Rodriguez, C. *et al*. Proteotyping of human haptoglobin by MALDI-TOF profiling: Phenotype distribution in a population of toxic oil syndrome patients. *Proteomics* **6**(Suppl 1), S272–281 (2006).
11. Hugo, A. *et al*. Proteotyping of microbial communities by optimization of tandem mass spectrometry data interpretation. *Pac. Symp. Biocomput. Pac. Symp. Biocomput.* 225–234 (2012).
12. Schwahn, A. B., Wong, J. W. H. & Downard, K. M. Rapid differentiation of seasonal and pandemic H1N1 influenza through proteotyping of viral neuraminidase with mass spectrometry. *Anal. Chem.* **82**, 4584–4590 (2010).
13. Wolters, M. *et al*. MALDI-TOF MS fingerprinting allows for discrimination of major methicillin-resistant *Staphylococcus aureus* lineages. *Int J Med Microbiol* **301**, 64–8 (2011).
14. Reil, M. *et al*. Recognition of *Clostridium difficile* PCR-ribotypes 001, 027 and 126/078 using an extended MALDI-TOF MS system. *Eur J Clin Microbiol Infect Dis* **30**, 1431–6 (2011).
15. Christner, M. *et al*. Rapid MALDI-TOF Mass Spectrometry Strain Typing during a Large Outbreak of Shiga-Toxigenic *Escherichia coli*. *PLoS One* **9**, e101924 (2014).
16. Ojima-Kato, T., Yamamoto, N., Takahashi, H. & Tamura, H. Matrix-assisted Laser Desorption Ionization-Time of Flight Mass Spectrometry (MALDI-TOF MS) Can Precisely Discriminate the Lineages of *Listeria monocytogenes* and Species of *Listeria*. *PloS One* **11**, e0159730 (2016).
17. Ojima-Kato, T. *et al*. Application of proteotyping Strain Solution^{TM} ver. 2 software and theoretically calculated mass database in MALDI-TOF MS typing of *Salmonella* serotype. *Appl. Microbiol. Biotechnol.* **101**, 8557–8569 (2017).
18. Kuhns, M. *et al*. Rapid discrimination of *Salmonella enterica* serovar Typhi from other serovars by MALDI-TOF mass spectrometry. *PLoS One* **7**, e40004 (2012).
19. Zautner, A. E. *et al*. Discrimination of multilocus sequence typing-based *Campylobacter jejuni* subgroups by MALDI-TOF mass spectrometry. *BMC Microbiol* **13**, 247 (2013).
20. Zautner, A. E., Masanta, W. O., Weig, M., Groß, U. & Bader, O. Mass Spectrometry-based PhyloProteomics (MSPP): A novel microbial typing Method. *Sci. Rep.* **5**, 13431 (2015).
21. Zautner, A. E. *et al*. Subtyping of *Campylobacter jejuni* ssp. *doylei* Isolates Using Mass Spectrometry-based PhyloProteomics (MSPP). *JoVE J. Vis. Exp*. e54165–e54165 (2016).
22. Larsen, M. V. *et al*. Multilocus sequence typing of total-genome-sequenced bacteria. *J. Clin. Microbiol.* **50**, 1355–1361 (2012).
23. Maiden, M. C. *et al*. Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Proc Natl Acad Sci U A* **95**, 3140–5 (1998).
24. Leekitcharoenphon, P., Lukjancenko, O., Friis, C., Aarestrup, F. M. & Ussery, D. W. Genomic variation in *Salmonella enterica* core genes for epidemiological typing. *BMC Genomics* **13**, 88 (2012).
25. Jolley, K. A. *et al*. Ribosomal multilocus sequence typing: universal characterization of bacteria from domain to strain. *Microbiology* **158**, 1005–15 (2012).
26. Bennett, J. S. *et al*. A genomic approach to bacterial taxonomy: an examination and proposed reclassification of species within the genus *Neisseria*. *Microbiology* **158**, 1570–80 (2012).
27. Cody, A. J. *et al*. Real-time genomic epidemiological evaluation of human *Campylobacter* isolates by use of whole-genome multilocus sequence typing. *J Clin Microbiol* **51**, 2526–34 (2013).
28. Coker, A. O., Isokpehi, R. D., Thomas, B. N., Amisu, K. O. & Obi, C. L. Human campylobacteriosis in developing countries. *Emerg. Infect. Dis.* **8**, 237–244 (2002).

29. Samuel, M. C. *et al*. Epidemiology of Sporadic *Campylobacter* Infection in the United States and Declining Trend in Incidence, FoodNet 1996–1999. *Clin. Infect. Dis.* **38**, S165–S174 (2004).
30. Dingle, K. E., Colles, F. M. & Falush, D. & Maiden, M. C. Sequence typing and comparison of population biology of *Campylobacter coli* and *Campylobacter jejuni*. *J. Clin. Microbiol.* **43**, 340–347 (2005).
31. Groisman, E. A. & Ochman, H. How *Salmonella* became a pathogen. *Trends Microbiol.* **5**, 343–349 (1997).
32. Sheppard, S. K., McCarthy, N. D. & Falush, D. & Maiden, M. C. Convergence of *Campylobacter* species: implications for bacterial evolution. *Science* **320**, 237–9 (2008).
33. Sheppard, S. K. *et al*. Evolution of an agriculture-associated disease causing *Campylobacter coli* clade: evidence from national surveillance data in Scotland. *PLoS One* **5**, e15708 (2010).
34. Colles, F. M., Ali, J. S., Sheppard, S. K., McCarthy, N. D. & Maiden, M. C. J. *Campylobacter* populations in wild and domesticated Mallard ducks (Anas platyrhynchos). *Environ. Microbiol. Rep.* **3**, 574–580 (2011).
35. Sheppard, S. K. *et al*. Progressive genome-wide introgression in agricultural *Campylobacter coli*. *Mol Ecol* **22**, 1051–64 (2013).
36. Wilson, D. J. *et al*. Rapid evolution and the importance of recombination to the gastroenteric pathogen *Campylobacter jejuni*. *Mol. Biol. Evol.* **26**, 385–397 (2009).
37. Fraser, C., Hanage, W. P. & Spratt, B. G. Recombination and the nature of bacterial speciation. *Science* **315**, 476–480 (2007).
38. Eggleston, A. K. & West, S. C. Recombination initiation: Easy as A, B, C, D… X? *Curr. Biol. CB* **7**, R745–749 (1997).
39. Zhu, P. *et al*. Fit genotypes and escape variants of subgroup III *Neisseria meningitidis* during three pandemics of epidemic meningitis. *Proc. Natl. Acad. Sci. USA* **98**, 5234–5239 (2001).
40. Skarp-de Haan, C. P. *et al*. Comparative genomics of unintrogressed *Campylobacter coli* clades 2 and 3. *BMC Genomics* **15**, 129 (2014).
41. Gardner, S. P. & Olson, J. W. Barriers to Horizontal Gene Transfer in *Campylobacter jejuni*. *Adv. Appl. Microbiol.* **79**, 19–42 (2012).
42. Dugar, G. *et al*. High-resolution transcriptome maps reveal strain-specific regulatory features of multiple *Campylobacter jejuni* isolates. *PLoS Genet.* **9**, e1003495 (2013).
43. Pickett, C. L. *et al*. Prevalence of cytolethal distending toxin production in *Campylobacter jejuni* and relatedness of *Campylobacter* sp. *cdtB* gene. *Infect Immun* **64**, 2070–8 (1996).
44. Eyigor, A., Dawson, K. A., Langlois, B. E. & Pickett, C. L. Detection of cytolethal distending toxin activity and *cdt* genes in *Campylobacter* spp. isolated from chicken carcasses. *Appl. Environ. Microbiol.* **65**, 1501–1505 (1999).
45. Eyigor, A., Dawson, K. A., Langlois, B. E. & Pickett, C. L. Cytolethal distending toxin genes in *Campylobacter jejuni* and *Campylobacter coli* isolates: detection and analysis by PCR. *J. Clin. Microbiol.* **37**, 1646–1650 (1999).
46. Dassanayake, R. P. *et al*. Characterization of cytolethal distending toxin of *Campylobacter* species isolated from captive macaque monkeys. *J. Clin. Microbiol.* **43**, 641–649 (2005).
47. Asakura, M. *et al*. Comparative analysis of cytolethal distending toxin (*cdt*) genes among *Campylobacter jejuni*, *C. coli* and *C. fetus* strains. *Microb. Pathog.* **42**, 174–183 (2007).
48. Fagerquist, C. K. *et al*. Sub-speciating *Campylobacter jejuni* by proteomic analysis of its protein biomarkers and their post-translational modifications. *J Proteome Res* **5**, 2527–38 (2006).
49. Vandamme, P. *et al*. *Campylobacter hyoilei* Alderton *et al*. 1995 and *Campylobacter coli* Veron and Chatelain 1973 are subjective synonyms. *Int J Syst Bacteriol* **47**, 1055–60 (1997).
50. Mellmann, A. *et al*. Evaluation of matrix-assisted laser desorption ionization-time-of-flight mass spectrometry in comparison to 16S rRNA gene sequencing for species identification of nonfermenting bacteria. *J Clin Microbiol* **46**, 1946–54 (2008).
51. Gonzales, T. & Robert-Baudouy, J. Bacterial aminopeptidases: properties and functions. *FEMS Microbiol Rev* **18**, 319–44 (1996).
52. Ouidir, T., Jarnier, F., Cosette, P., Jouenne, T. & Hardouin, J. Characterization of *N*-terminal protein modifications in *Pseudomonas aeruginosa* PA14. *J. Proteomics* **114**, 214–225 (2015).
53. Kentache, T., Jouenne, T., Dé, E. & Hardouin, J. Proteomic characterization of $N\alpha$- and $N\varepsilon$-acetylation in *Acinetobacter baumannii*. *J. Proteomics* **144**, 148–158 (2016).
54. Kearse, M. *et al*. Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* **28**, 1647–1649 (2012).
55. Kumar, S., Stecher, G. & Tamura, K. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Mol. Biol. Evol.* msw054 (2016).
56. Jolley, K. A. & Maiden, M. C. BIGSdb: Scalable analysis of bacterial genome variation at the population level. *BMC Bioinformatics* **11**, 595 (2010).
57. Saitou, N. & Nei, M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4**, 406–425 (1987).
58. Tamura, K., Nei, M. & Kumar, S. Prospects for inferring very large phylogenies by using the neighbor-joining method. *Proc. Natl. Acad. Sci. USA* **101**, 11030–11035 (2004).
59. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2–approximately maximum-likelihood trees for large alignments. *PloS One* **5**, e9490 (2010).
60. Treangen, T. J., Ondov, B. D., Koren, S. & Phillippy, A. M. The Harvest suite for rapid core-genome alignment and visualization of thousands of intraspecific microbial genomes. *Genome Biol.* **15** (2014).

Acknowledgements

We are grateful to Iris Iben and Agnieszka Goretzki (Göttingen) for excellent technical support, and Kerstin Stingl (BfR, Berlin, Germany) for providing several *C. coli* Isolates included in this study. We thank Anastasia Lübke for proofreading our text and correcting our English. This paper was funded by the Open Access support program of the Deutsche Forschungsgemeinscha and the publication fund of the Georg August Universität Göttingen. SSM is thankful to ARRS for financial support of the projects (P4-0116, J4-7608), which enabled isolation of many bacterial strains included in study.

Author Contributions

All listed coauthors contributed significantly to the study. M.F.E.: data interpretation, bioinformatics, wrote manuscript, figures; S.S.M.: collection of bacterial isolates, data interpretation, correction of manuscript; R.L.: bacteriology, data interpretation, correction of manuscript; W.B.: bioinformatics, correction of manuscript; W.O.M.: bacteriology, sample preparation, M.L.S.T., correction of manuscript; U.G.: study design, correction of manuscript; O.B.: mass spectrometry, study design, wrote manuscript; T.R.: genomic sequencing, core genome alignment, GenBank deposit; A.E.Z.: study design, data interpretation, bioinformatics, wrote manuscript.

Additional Information

Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-019-40842-w.

Competing Interests: The authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

2.2 Differenzierung von *Campylobacter fetus*-Subspezies mit Hilfe der Proteotypisierung

Campylobacter fetus ist in erster Linie ein Erreger intestinaler Erkrankungen, vereinzelt verursacht er jedoch auch systemische Erkrankungen und Meningitis. Gegenwärtig sind drei Subspezies von *C. fetus* bekannt: *C. fetus* Subspezies *fetus* (*Cff*), *C. fetus* Subspezies *venerealis* (*Cfv*) und *C. fetus* Subspezies *testudinum* (*Cft*). *Cff* und *Cfv* sind in erster Linie mit Säugetieren assoziiert, während *Cft* am häufigsten aus Reptilien isoliert wird.

Die phylogenetische Klassifikation der Spezies erfolgt üblicherweise mittels Multilokus Sequenz Typisierung (MLST) und PCR-Ribotypisierung, welche relativ zeitaufwändige und teure Methoden sind. Um diese arbeitsintensiven DNA-Sequenz basierten Methoden zu ersetzen, war das Ziel dieser Studie, ein Typisierungsschema basierend auf der Proteotypisierungsmethode zu entwickeln.

Insgesamt wurde 41 *C. fetus*-Stämme, die die drei bekannten Subspezies abdeckten, via ICMS analysiert und mit den entsprechenden MLST-Ergebnissen verglichen. Die im Massenspektrum des *C. fetus*-Referenzstammes LMG 6442 (NCTC 10842) identifizierten Biomarker sowie korrespondierende Isoforme wurden mit den zugehörigen Aminosäuresequenzen in Verbindung gebracht und in das *C. fetus*-Proteotypisierungsschema aufgenommen.

In Kombination erlauben die neun identifizierten Biomarker die Unterscheidung der *Cft*-Stämme von *Cff*- und *Cfv*-Stämmen. Biomarker, welche die Unterscheidung zwischen *Cff* und *Cfv* ermöglichen, wurden nicht identifiziert. Die Ergebnisse der Studie belegen die Stabilität und Eignung der Proteotypisierung als Intraspezies-Typisierungsmethode, zeigen aber auch deren Grenzen auf.

Autoren: **Matthias F. Emele**, Matti Karg, Helmut Hotzel, Linda van der Graaf-van Bloois, Uwe Groß, Oliver Bader, Andreas E. Zautner

Beitrag der Autoren zur praktischen Arbeit:

Matthias F. Emele:

Anfertigung des Manuskripts, Erstellung von Grafiken und Tabellen, Erstellung einer Isoformendatenbank.

Matti Karg:

Massenspektrometrische Analysen, MLST, Bioinformatik.

Helmut Hotzel:

Sammlung von Isolaten, Korrektur des Manuskripts

Linda van der Graaf-van Bloois:

Sammlung von Isolaten, Korrektur des Manuskripts

Uwe Groß:

Studiendesign, Korrektur des Manuskripts

Oliver Bader:

Studiendesign, Korrektur des Manuskripts

Andreas E. Zautner:

Studiendesign, Sammlung von Isolaten, Korrektur des Manuskripts

Status des Manuskripts:

Under Review; European Journal of Microbiology and Immunology

Differentiation of *Campylobacter fetus* subspecies by Proteotyping

Matthias F. Emele[1], Matti Karg[1], Helmut Hotzel[2], Linda van der Graaf-van Bloois[3,4], Uwe Groß[1], Oliver Bader[1], and Andreas E. Zautner[1,§]

[1] Institut für Medizinische Mikrobiologie, Universitätsmedizin Göttingen, Göttingen, Germany
[2]Institut für bakterielle Infektionen und Zoonosen, Friedrich-Loeffler-Institut Bundesforschungsinstitut für Tiergesundheit, Jena, Germany
[3]Department of Infectious Diseases and Immunology, Faculty of Veterinary Medicine, Utrecht University, The Netherlands
[4]WHO Collaborating Center for Campylobacter/OIE Reference Laboratory for Campylobacteriosis, Utrecht, The Netherlands

§ Corresponding author: Andreas E. Zautner
Universitätsmedizin Göttingen
Institut für Medizinische Mikrobiologie
Kreuzbergring 57
D-37075 Göttingen, Germany
Phone: +49-551-398549
FAX: +49-551-395861
E-Mail: azautne@gwdg.de

Short running title: Proteotyping of *Campylobacter fetus* subspecies

Keywords: MALDI-TOF MS; *Campylobacter fetus*; below species differentiation; ICMS; proteotyping; MLST

Abstract

Campylobacter fetus is a causative agent of intestinal illness and, sometimes, severe systemic infections and meningitis. *C. fetus* currently comprises three subspecies: *C. fetus* subspecies *fetus* (*Cff*), *C. fetus* subspecies *venerealis* (*Cfv*) and *C. fetus* subspecies *testudinum* (*Cft*). *Cff* and *Cfv* are primarily associated with mammals whereas *Cft* is associated with reptiles.

To offer an alternative to laborious sequence-based techniques such as multilocus sequence typing (MLST) and PCR-ribotyping for this species, the purpose of the study was to develop a typing scheme based on proteotyping.

In total, 41 representative *C. fetus* strains were analyzed by intact cell mass spectrometry and compared to MLST results. Biomarkers detected in the mass spectrum of *C. fetus* subsp. *fetus* reference strain LMG 6442 (NCTC 10842) as well as corresponding isoforms were associated with the respective amino acid sequences and added to the *C. fetus* proteotyping scheme.

In combination, the 9 identified biomarkers allow the differentiation of *Cft* subspecies strains from *Cff* and *Cfv* subspecies strains. Biomarkers to distinguish between *Cff* and *Cfv* were not found. The results of the study show the potential of proteotyping to differentiate different subspecies, but also the limitations of the method.

Introduction

Campylobacter spp. can cause gastrointestinal and extra-intestinal infections [1]. Although the majority of cases (>90%) of intestinal campylobacteriosis are caused by *Campylobacter jejuni* and *Campylobacter coli* a small number is also caused by *C. fetus* [2–5]. Among these, *C. fetus* is the most common cause of *Campylobacter* bacteremia. The frequency of detection in blood cultures varies between 19% and 53% [6–8] of all Campylobacterioses. The reported case fatality rate of invasive *C. fetus* infections is at 14% [9]. Due to the high incidence rate of campylobacteriosis worldwide, this shows that *C. fetus* infections occur frequently and have the potential to become a significant public health issue. However, not much is known about the source of infection and the people at risk, so far. Most reported *C. fetus* infections were observed in AIDS patients and other immunocompromised individuals [1, 10].

C. fetus is a Gram-negative, microaerophilic bacterium, growing between 25°C and 37°C. Clinical symptoms of human *C. fetus* infection vary from acute diarrhea to systemic illness [11, 12] and presentation of the symptoms depends on where the disseminated pathogen is localized. Septicemia with fever, but without apparent localized infection, for example, is reported in 24% to 41% of cases [7, 9]. Other manifestations can be the result of neurological infections (i.e. meningoencephalitis, meningitis or brain abscesses), arthritis, lung abscesses, osteomyelitis, and perinatal infections (i.e. abortion, infection in utero or placentitis) [12]. Furthermore *C. fetus* infections may also cause vascular pathology (i.e. endocarditis, pericarditis, vasculitis, mycotic aneurysms) [13].

Currently, three subspecies of *C. fetus* are known. These are *Cff*, *Cfv* and *Cft*. For *Cfv* also the biovar intermedius (*Cfvi*) has been identified in previous studies [14, 15]. Subspecies *Cff* and *Cfv* are primarily associated with mammals [13, 14] whereas the third subspecies *Cft* is linked with reptiles [15, 16]. *Cff* and *Cfv* are genetically very closely related [17, 18] but differ in host adaption. *Cff* can cause sporadic infections in humans, abortion in sheep and cattle and can be isolated from different sites in different hosts [19]. Occurrence of *Cfv* is restricted to the genital tract of cattle and is furthermore responsible for bovine genital campylobacteriosis (BGC). This syndrome is causes fertility problems in cattle [20]. Previous studies have demonstrated a large genetic divergence between strains of mammal and reptile origin [21, 22] and molecular and phenotypic characterization of human cases and 3 reptiles identified a new subspecies and proposed the name *C. fetus* subsp. *testudinum* subsp. nov. [15, 23].

In recent years intact cell MALDI-TOF mass spectrometry (ICMS) became a standard method for microbial species identification in clinical diagnostic laboratories [24, 25]. MALDI-TOF MS also offers the opportunity to classify unknown bacterial isolates by identifying similarities in mass spectra of unknown bacteria and biomarkers in existing databases, a procedure referred to as phyloproteomics [26]. Typing methods, which are based on mass spectrometric analysis are generally known as proteotyping [27] and have previously been used for characterization of microbial communities, tissues, individual proteins, viruses and bacteria for several years now [28–31]. Among clinically relevant bacteria *Salmonella* serotypes, *Clostridiodes difficile* PCR ribotypes, and methicillin-resistant *Staphylococcus aureus* lineages have been shown to be detectable by proteotyping, to name just a few [32–34].

Previous studies of our working group demonstrated the potential of bacterial subtyping on *Campylobacter* species in the clinical context, as it was possible to differentiate clinically relevant from clinically less relevant subgroups (Figure 1) [35–39]. At the heart of our approach is a list of allelic isoforms that result from due to non-synonymous mutations and posttranslational modifications in biomarker gene sequences, which are detectable as mass shifts in MALDI-TOF spectra. In this way a combination of amino-acid sequences specific for each of the isolates to be typed can be derived, in a similar manner as for MLST. By using proteotyping only changes in mass associated with a certain set of allelic isoforms of the same protein are taken into account for the derivation of phylogeny whereas visibility or absence of particular masses as well as their intensity are not considered. This improves the measurement accuracy wherefore ICMS is a very promising subtyping approach and a realistic alternative to currently used sequence-based techniques [37].

The goal of this study was to complete the set of typing schemes for clinically relevant *Campylobacter* species by developing a *C. fetus*-specific proteotyping scheme. A set of 41 *C. fetus* isolates covering all currently known subspecies of *C. fetus* was used. All isolates were characterized by proteotyping and MLST followed by the deduction of the phylogenetic relations.

Materials and methods

C. fetus isolates

The test cohort was compiled in way that all subspecies of the bacterial species were represented. In total, 41 *C. fetus* isolates were included in our study: 20 *Cff*, 11 *Cfv*, 7 *Cft*, and 3 *Cfvi* isolates (Table 1). The isolates were of different biological origins, namely preputial washing of cattle (4 *Cff*, 7 *Cfv*), vaginal mucus of cattle (2 *Cfv*), foetuses of cattle (2 *Cfv*), cattle (not further specified, 3 *Cfv*), bovine sperm (1 *Cff)*, bull genitals (1 *Cff*), calf foetus (2 *Cff*), intestinal content of a calf (1 *Cff*), intestinal content of a pig (1 *Cff*), foetus brain of a sheep (1*Cff*), reptile cloak swab (3 *Cft*), human blood culture (7 *Cff*, 4 *Cft*), and 2 *Cff* strains of unknown origin. Animal isolates were provided by the Friedrich-Loeffler-Institut Bundesforschungsinstitut für Tiergesundheit, Jena, Germany. The following strains were received from the Belgian co-ordinated collections of micro-organisms (BCCM; http://bccm.belspo.be/about-us/bccm-lmg): LMG6443 (*Cfv*), LMG6442 (*Cff*), LMG6570 (*Cfv*), LMG27499 (*Cft*), LMG06569 (*Cff*), LMG06571 (*Cff*), LMG06727 (*Cff*). Human blood-culture isolates were provided by the routine diagnostic laboratory of the University Medical Center, Göttingen, Germany (Table 1).

Bacterial culture conditions

C. fetus isolates used in the experiments were kept as cryobank stocks (Mast Diagnostica, Reinfeld, Germany) at -80°C. For the subsequent MALDI-TOF MS analysis the isolates were incubated under microaerophilic conditions (5% O_2, 10% CO_2, 85% N_2) in Mueller-Hinton agar supplemented with horse blood at 37°C for 2-3 days.

Preparation of matrix solution

As part of the measurement preparation α-cyano-4-hydroxy-cinnamic acid (HCCA) purified matrix substance (Bruker Daltonics, Bremen, Germany) was dissolved in standard solvent (acetonitrile 50%, trifluoroacetic acid 2.5% in ddH_2O) to 10 mg HCCA/mL. Purified recombinant human insulin (Sigma-Aldrich, Taufkirchen, Germany) was added to the HCCA solution as an internal calibrant to a final concentration of 10 pg/µL. The exact mass of the internal calibrant was experimentally determined (m/z = 5806.1) with reference to the Bruker Test

Standard (BTS). The calibrant did not overlap with any of the biomarker masses of interest and allowed a very precise internal mass calibration of the spectra.

MALDI-TOF mass spectrometry

To prepare samples for the measurements, two different variants were used: smear preparation and formic acid/ acetonitrile extraction. Smear preparation by experience yields clearer peaks in the *m/z* range <10,000 Da, whereas the extraction variant allows more precise analysis in the field >10,000 Da [39].

The samples for the measurements were prepared as described before [37, 39] In the measurement process, 600 spectra (mass range 2 to 20 kDa) were obtained in 100-shots steps on an Autoflex III system and summed up. If the MALDI Biotyper (Database release 2016) identification score values were ≥ 2.00 they were considered correct.

Identification of biomarkers in ICMS spectra

The obtained mass spectra were analyzed by standard algorithms of FlexAnalysis (Bruker Daltonics, Bremen, Germany). Initially, spectra were internally calibrated to the spiked human insulin peak. Subsequently, the baseline was subtracted, and the spectra were smoothened (standard MBT method).

For determination of the theoretical average weight of the amino acid sequences corresponding to the respective open reading frames of ribosomal proteins, the amino acid sequences were uploaded one by one to the ExPASy Bioinformatics Resource Portal (https://web.expasy.org/compute_pi/) where a molecular weight calculator tool is provided.

Proteins used for previous proteotyping schemes sometimes underwent post-translational modifications [40, 41]; therefore further molecular weights were calculated for each biomarker, taking into account potential proteolytic removal of the *N*-terminal methionine (-131.04 Da), acetylation, phosphorylation, formylation, and methylation (Table 2).

Biomarker masses observed in the reference genome of reference strain LMG 6442 (NCTC 10842) (Figure 2) were matched to the calculated masses. In contrast, biomarker masses observed in the spectrum of clinical isolates which could not be assigned to the calculated masses from the *C. fetus* reference genome the spectra were considered as novel isoforms of the particular biomarker.

For each isolate of the *C. fetus* test cohort all biomarker genes were amplified by PCR using primers listed in Table 3 and the amplicon was sequenced (Microsynth Seqlab, Göttingen, Germany). To confirm the respective allelic isoforms the gene sequences obtained from the amplicons were translated *in silico* and the amino acid sequences were subsequently aligned.

Multilocus sequence typing (MLST)

For MLST a procedure modified from the original typing schemes was used [18, 23]. In brief, the annealing temperature of the PCR was decreased from 48°C to 47°C and the *glyA2* oligonucleotide primer for the amplification of the *glyA* locus was replaced with the primer *glyS4* [18]. After concatenating of the MLST gene sequences for each strain the software MEGA X was also used to construct an MLST-based UPGMA dendrogram [42].

Phylogenetic and phyloproteomic analyses

A list of amino acid sequences of all allelic isoforms of the 9 identified biomarkers was compiled (Table 4). GenBank accession numbers for the biomarker sequences observed in this study are listed in Table 5.

To analyze the biomarkers' protein sequences translated from the NCBI nucleotide database (Geneious V10.1.3) they were concatenated for each strain and a UPGMA dendrogram (MEGA X) was constructed [42].

Results and discussion

In 2015 our working group set up a new proteotyping workflow for the proteotyping of microorganisms (Figure 1) [37]. Now the established procedure was used to develop a *C. fetus*-specific proteotyping scheme. According to the standard workflow masses emerging in the mass spectrum of the genome sequenced *Cff* reference strain LMG 6442 (NCTC 10842) were analyzed and MS biomarker ions were related with gene products consistent with the observed mass. By evaluating the 67 *C. fetus* nucleotide sequences available in the NCBI database a collection of allelic isoforms for all biomarkers observed in the reference spectrum was set up (Table 4). In accordance with the established proteotyping procedure, mass spectra of all strains included in the test cohort were recorded. Subsequently spectra were edited (baseline subtraction and smoothing) and overlaid with the spectrum of *Cff* reference strain LMG 6442 (NCTC 10842). Recorded biomarker masses were matched with the calculated average protein masses and mass shifts in relation to the masses of the references strain were analyzed. After concatenation of amino acid sequences of the biomarkers included in the *C. fetus* typing scheme a UPGMA tree based on these strain specific proteotyping-based types was calculated.

Identification of biomarker ions

In total, the analysis based on the genome of *Cff* reference strain LMG 6442 (NCTC 10842) yielded nine, single charged biomarker masses between *m/z* = 4,300 and 10,300 which were presumptively correlated with a specific gene-product. To provide reliable statements on reproducibility of our measurements the standard deviation was calculated on the basis of six measurements. The highest standard deviation (0.959) was observed for isoform 1 of biomarker S20-M whereas the lowest standard deviation (0.271) was observed for isoform 5 of biomarker L33-M (Table 6). The following biomarkers were identified: L36 (4,331.35 Da), L34 (4,217.26 Da), L32-M (5,530.47 Da), L33-M (6,205.31 Da), S14-M (6,728.11 Da), L29 (6,893.22 Da), L24-M (8,026.59 Da), S20-M (9,741.33 Da), S19-M (10,277.10 Da). De-methionation was observed for biomarkers L32-M, L33-M, S14-M, L24-M, S19-M, and S20-M (Table 2, Figure 2 & 3). In case of MLST the established markers are distributed over the whole genome of the reference strain. As the biomarkers identified in this study show a comparable distribution they were suitable for the deduction of phylogenetic relations.

Comparing *C. fetus* proteotyping biomarkers to biomarkers identified within the context of *C. jejuni* subsp. *jejuni*, *C. jejuni* subsp. *doylei* and *C. coli* proteotyping [37–39] several differences can be noted: In the case of *C. jejuni* subsp. *jejuni* 19 biomarkers were identified and associated with the respective peak in the ICMS spectrum whereas less than half (9) were found for *C. fetus*. Furthermore, biomarker L33 lacked *N*-terminal methionine in case of *C. fetus* (L33-M) but it was present in *C. jejuni* subsp. *jejuni*, *C. jejuni* subsp. *doylei* and *C. coli*. These observations confirm the results published by Fagerquist and colleagues according to which posttranslational modification patterns are microbial species-specific. Within the isolate collection biomarker mass shifts were observed in seven out of nine biomarkers [43].

Establishment of an allelic isoform database

Following the identification of biomarker ions, an amino acid sequence isoform list for each of the biomarkers identified in the previous step was compiled. In this context, we analyzed the 67 *C. fetus* genome sequences that can be found on NCBI. The number of identified isoforms for the respective biomarker varied: The highest number was six, whereas one biomarker showed just a single isoform. Differences were also observed regarding frequency of occurrence; whereas some isoforms occurred in >99% of the cases, other isoforms were only found once. Regarding single occurrence of isoforms, a sequencing error is possible. Except of biomarker L36, all identified biomarkers showed at least 3 different isoforms, demonstrating their suitability in the *C. fetus* subtyping context.

The amino acid sequences of all biomarker isoform are listed in Table 4. Variations of the amino acid sequences obtained by alignment of the sequences are indicted in red, additionally the computed average protein mass for each isoform is listed. It should be noted that due to some draft genomes in GenBank the number of available sequences may vary, as there were no contigs with the sequences coding for each biomarker in all genomes.

MLST and proteotyping of the isolate collection

To proof functionality of the *C. fetus* proteotying scheme the test cohort (41 *C. fetus* strains) was typed by MLST as well as proteotyping. The composition of the test cohort was such that all known subspecies of the species were covered. The isolate collection comprised the following 14 MLST sequence types: ST2 (3 isolates), ST3 (7 isolates), ST4 (14 isolates), ST5 (2 isolates), ST6 (4 isolates), ST11 (1 isolate) , ST15 (2 isolates), ST16 (1 isolate), ST20 (2 isolates), ST27 (1 isolate), ST30 (1 isolate), ST31 (1 isolate), ST66 (1 isolate) and ST68 (1 isolate, Table 1).

The concatenated amino acid sequences of the different biomarkers yielded four proteotyping-derived types (Figure 4, right dendrogram). Proteotyping-derived type A comprised most of the *Cff* and *Cfv* isolates (31/41). More precisely it comprised 3 MLST-ST2 isolates, 7 MLST-ST3 isolates, 14 MLST-ST4 isolates, 4 MLST-ST6 isolates, one MLST-ST11 isolate and one MLST-ST68 isolate.

Proteotyping-derived type B consisted of two *Cff* MLST-ST 20 isolates, while proteotyping-derived type C consisted of one MLST-ST5 isolate (*Cff*).

The most interesting finding was that proteotyping-derived type D consisted only of *Cft* isolates. Regarding MLST sequence types it comprised particularly one isolate of ST16, 2 isolates of ST15, one isolate of ST27, one isolate of ST30, one isolate of sequence type 31 and one isolate of ST61.

Identification of allelic isoforms

The test cohort was measured in exactly the same manner as it was done for the reference strain LMG 6442 (NCTC 10842). The evaluation of the measurements of mass spectra of the strains was done based on the comparison with the spectrum of this reference strain. Observed mass shifts were compared to the sequence list of amino acid isoforms whereby a particular allelic isoform could be identified.

If two different isoforms with the same mutation at different positions were observed which though did not differ regarding mass difference to the reference isoform, the variants were further examined by DNA sequencing. In the test cohort, 3 allelic isoforms for biomarker L33-M (RpmG), and two for biomarkers L34 (RpmH), L32-M (RpmF), L29 (RpmC), L24-M (RplX), S20-M (RpsT), S19-M (RpsS) were detected. For biomarkers L36 (RpmJ) and S14-M (RpsN) only one allelic isoform was identified (Table 2, Figure 3).

Construction of an UPGMA-dendrogram

To deduce the phylogenetic relationships of the species, amino acid sequences of the 9 identified proteotyping biomarkers were fused into a single sequence. The concatenated sequence was then further processed with the MEGA X software to calculate a phyloproteomic tree (UPGMA). The 9 identified biomarkers allowed a clear differentiation of a group of *Cff* and *Cfv* strains from a group of utterly *Cft* strains. In order to assess the quality of the proteotyping results, another UPGMA tree was calculated based on MLST data (Figure 4). Comparative analysis of the trees revealed some differences between the two resulting phylogenies: While the test cohort was differentiated into 14 MLST sequence types, the proteotyping-based analysis led to a division into only 4 different groups. The most interesting finding was that proteotyping-based type D comprised all of the

Cft isolates, showing that here our approach is comparable to the quality of the current gold standard MLST.

Unfortunately, the MLST-ST4 corresponding to the subspecies *Cfv* could not be differentiated by means of proteotying. Here, proteotying proves to be inferior to MLSTyping in its discriminatory resolution.

A previous study by Fitzgerald and coworkers showed that it is possible to distinguish *Cft* from other *C. fetus* subspecies. Based on multiple unidentified biomarker peaks, a dendrogram was calculated using Pearson correlation [15]. A factor, which is reducing the informative value of these results, was the lack of knowledge about the proteins responsible for each of the discriminating peaks.

In contrast to this study, we were able to identify at least 9 defined ribosomal proteins as biomarkers. As *Cft* strains exhibited different biomarker isoforms compared to the other two *C. fetus* subspecies they could be clearly differentiated. PCR and subsequent Sanger sequencing of the respective biomarkers further confirmed these differences.

Regarding the limitations of proteotyping the number of sequence data available is decisive for the quality of the typing scheme. In case of *C. fetus* much less sequences (67) were available as compared to *C. jejuni* subsp. *jejuni* (more than 3000) [37]. Another factor affecting the quality of the typing scheme is the number of biomarkers it comprises. Further studies should therefor focus on the identification of additional reliable biomarkers that can be included in the existing scheme.

The prerequisite for the application of the technique is the visibility of all biomarkers of the typing scheme. If this is not the case it is advisable to use sequence-based techniques.

Conclusion

As the results obtained so far demonstrate, proteotyping is a promising tool for microbial typing at the species, subspecies, and even below subspecies level. A smart bioinformatics solution and the development of an easy-to-handle user interface would allow the application of the technique in daily diagnostic routine as the corresponding equipment for proteotyping is available in modern clinical laboratories anyway. The rapidly growing sequence databases due to NGS are opening up a wide range of opportunities for the development of further proteotyping schemes that possibly allow a rapid detection in case of a disease outbreak.

Ethical approval

Ethical approval for the study was obtained from Ethics Commission of the University Medical Center Göttingen, Germany. No humans, animals, or personalized data were used for this study.

Authors' contributions

MFE and MK contributed equally to the work. MFE, OB and AEZ wrote the manuscript and established the biomarker isoform database *in silico.* MK performed MALDI measurements and confirmatory PCRs. HH, LvdG and AEZ collected bacterial isolates, data interpretation, bioinformatics, correction of manuscript. OB, UG and AEZ designed the experiments and evaluated the data.

Conflicts of interest

There are no conflicts of interest.

References

1. Forbes BA, Sahm DF, Weissfeld AS (2002): Diagnostic microbiology. In: Bailey & Scott's Diagnostic Microbiology, pp. 11–14

2. Gillespie IA, O'Brien SJ, Frost JA, Adak GK, Horby P, Swan AV, Painter MJ, Neal KR: A Case-Case Comparison of *Campylobacter coli* and *Campylobacter jejuni* Infection: A Tool for Generating Hypotheses. Emerg Infect Dis 8, 937–942 (2002)

3. Nielsen H, Hansen KK, Gradel KO, Kristensen B, Ejlertsen T, Østergaard C, Schønheyder HC: Bacteraemia as a result of *Campylobacter* species: a population-based study of epidemiology and clinical risk factors. Clin Microbiol Infect 16, 57–61 (2010)

4. Havelaar AH, Kirk MD, Torgerson PR, Gibb HJ, Hald T, Lake RJ, Praet N, Bellinger DC, Silva NR de, Gargouri N, Speybroeck N, Cawthorne A, Mathers C, Stein C, Angulo FJ, Devleesschauwer B, Group on behalf of WHOFDBER: World Health Organization Global Estimates and Regional Comparisons of the Burden of Foodborne Disease in 2010. PLOS Med 12, e1001923 (2015)

5. World Health Organization (2015): WHO estimates of the global burden of foodborne diseases: foodborne disease burden epidemiology reference group 2007-2015.

6. Guerrant RL, Lahita RG, Winn WC, Roberts RB: Campylobacteriosis in man: Pathogenic mechanisms and review of 91 bloodstream infections. Am J Med 65, 584–592 (1978)

7. Pacanowski J, Lalande V, Lacombe K, Boudraa C, Lesprit P, Legrand P, Trystram D, Kassis N, Arlet G, Mainardi J-L, Doucet-Populaire F, Girard P-M, Meynard J-L: *Campylobacter* Bacteremia: Clinical Features and Factors Associated with Fatal Outcome. Clin Infect Dis 47, 790–796 (2008)

8. Fernández-Cruz A, Muñoz P, Mohedano R, Valerio M, Marín M, Alcalá L, Rodriguez-Créixems M, Cercenado E, Bouza E: *Campylobacter* Bacteremia: Clinical Characteristics, Incidence, and Outcome Over 23 Years. Medicine (Baltimore) 89, 319 (2010)

9. Gazaigne L, Legrand P, Renaud B, Bourra B, Taillandier E, Brun-Buisson C, Lesprit P: *Campylobacter fetus* bloodstream infection: risk factors and clinical features. Eur J Clin Microbiol Infect Dis Off Publ Eur Soc Clin Microbiol 27, 185–

189 (2008)

10. Elshafie SS, Asim M, Ashour A, Elhiday AH, Mohsen T, Doiphode S: *Campylobacter* peritonitis complicating continuous ambulatory peritoneal dialysis: report of three cases and review of the literature. Perit Dial Int 30, 99–104 (2010)

11. Klein BS, Vergeront JM, Blaser MJ, Edmonds P, Brenner DJ, Janssen D, Davis JP: *Campylobacter* infection associated with raw milk. An outbreak of gastroenteritis due to *Campylobacter jejuni* and thermotolerant *Campylobacter fetus* subsp. *fetus*. JAMA 255, 361–364 (1986)

12. Man SM: The clinical importance of emerging *Campylobacter* species. Nat Rev Gastroenterol Hepatol 8, 669–685 (2011)

13. Wagenaar JA, van Bergen MAP, Blaser MJ, Tauxe RV, Newell DG, van Putten JPM: *Campylobacter fetus* infections in humans: exposure and disease. Clin Infect Dis Off Publ Infect Dis Soc Am 58, 1579–1586 (2014)

14. Véron M, Chatelain R: Taxonomic Study of the Genus *Campylobacter* Sebald and Véron and Designation of the Neotype Strain for the Type Species, *Campylobacter fetus* (Smith and Taylor) Sebald and Véron. Int J Syst Evol Microbiol 23, 122–134 (1973)

15. Fitzgerald C, Tu Z chao, Patrick M, Stiles T, Lawson AJ, Santovenia M, Gilbert MJ, van Bergen M, Joyce K, Pruckler J, Stroika S, Duim B, Miller WG, Loparev V, Sinnige JC, Fields PI, Tauxe RV, Blaser MJ, Wagenaar JA: *Campylobacter fetus* subsp. *testudinum* subsp. nov., isolated from humans and reptiles. Int J Syst Evol Microbiol 64, 2944–2948 (2014)

16. Gilbert MJ, Kik M, Timmerman AJ, Severs TT, Kusters JG, Duim B, Wagenaar JA: Occurrence, Diversity, and Host Association of Intestinal *Campylobacter*, *Arcobacter*, and *Helicobacter* in Reptiles. PLOS ONE 9, e101599 (2014)

17. On SLW, Harrington CS: Evaluation of numerical analysis of PFGE-DNA profiles for differentiating *Campylobacter fetus* subspecies by comparison with phenotypic, PCR and 16S rDNA sequencing methods. J Appl Microbiol 90, 285–293 (2001)

18. van Bergen M a. P, Linnane S, van Putten JPM, Wagenaar JA: Global detection and identification of *Campylobacter fetus* subsp. *venerealis*. Rev Sci Tech Int Off Epizoot 24, 1017–1026 (2005)

19. Thompson SA, Blaser MJ (2000): Pathogenesis of Campylobacter fetus infections. In: Campylobacter, American Society Microbiolgy Press, Washington,

DC, pp. 321–347

20. Dekeyser J (1984): Bovine genital campylobacteriosis. In: Campylobacter infection in man and animals, CRC Press, Boca Raton, pp. 181–191

21. Van der Graaf-van Bloois L, Miller WG, Yee E, Duim B, Wagenaar JA (2012): Whole genome sequencing of Campylobacter fetus subspecies. In: American Society for Microbiology General Meeting, San Francisco, CA, pp. 16–19

22. Tu Z-C, Zeitlin G, Gagner J-P, Keo T, Hanna BA, Blaser MJ: *Campylobacter fetus* of Reptile Origin as a Human Pathogen. J Clin Microbiol 42, 4405–4407 (2004)

23. Dingle KE, Blaser MJ, Tu Z-C, Pruckler J, Fitzgerald C, Bergen MAP van, Lawson AJ, Owen RJ, Wagenaar JA: Genetic Relationships among Reptilian and Mammalian *Campylobacter fetus* Strains Determined by Multilocus Sequence Typing. J Clin Microbiol 48, 977–980 (2010)

24. Seng P, Rolain JM, Fournier PE, La Scola B, Drancourt M, Raoult D: MALDI-TOF-mass spectrometry applications in clinical microbiology. Future Microbiol 5, 1733–54 (2010)

25. Zingue D, Flaudrops C, Drancourt M: Direct matrix-assisted laser desorption ionisation time-of-flight mass spectrometry identification of mycobacteria from colonies. Eur J Clin Microbiol Infect Dis Off Publ Eur Soc Clin Microbiol 35, 1983–1987 (2016)

26. Conway GC, Smole SC, Sarracino DA, Arbeit RD, Leopold PE: Phyloproteomics: species identification of *Enterobacteriaceae* using matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. J Mol Microbiol Biotechnol 3, 103–12 (2001)

27. Karlsson R, Gonzales-Siles L, Boulund F, Svensson-Stadler L, Skovbjerg S, Karlsson A, Davidson M, Hulth S, Kristiansson E, Moore ERB: Proteotyping: Proteomic characterization, classification and identification of microorganisms--A prospectus. Syst Appl Microbiol 38, 246–257 (2015)

28. Hugo A, Baxter DJ, Cannon WR, Kalyanaraman A, Kulkarni G, Callister SJ: Proteotyping of microbial communities by optimization of tandem mass spectrometry data interpretation. Pac Symp Biocomput Pac Symp Biocomput 225–234 (2012)

29. Shillingford JM, Miyoshi K, Robinson GW, Bierie B, Cao Y, Karin M,

Hennighausen L: Proteotyping of mammary tissue from transgenic and gene knockout mice with immunohistochemical markers: a tool to define developmental lesions. J Histochem Cytochem Off J Histochem Soc 51, 555–565 (2003)

30. Rodriguez C, Quero C, Dominguez A, Trigo M, Posada de la Paz M, Gelpi E, Abian J: Proteotyping of human haptoglobin by MALDI-TOF profiling: Phenotype distribution in a population of toxic oil syndrome patients. Proteomics 6 Suppl 1, S272-281 (2006)

31. Schwahn AB, Wong JWH, Downard KM: Rapid differentiation of seasonal and pandemic H1N1 influenza through proteotyping of viral neuraminidase with mass spectrometry. Anal Chem 82, 4584–4590 (2010)

32. Ojima-Kato T, Yamamoto N, Takahashi H, Tamura H: Matrix-assisted Laser Desorption Ionization-Time of Flight Mass Spectrometry (MALDI-TOF MS) Can Precisely Discriminate the Lineages of *Listeria monocytogenes* and Species of *Listeria*. PloS One 11, e0159730 (2016)

33. Reil M, Erhard M, Kuijper EJ, Kist M, Zaiss H, Witte W, Gruber H, Borgmann S: Recognition of *Clostridium difficile* PCR-ribotypes 001, 027 and 126/078 using an extended MALDI-TOF MS system. Eur J Clin Microbiol Infect Dis 30, 1431–6 (2011)

34. Wolters M, Rohde H, Maier T, Belmar-Campos C, Franke G, Scherpe S, Aepfelbacher M, Christner M: MALDI-TOF MS fingerprinting allows for discrimination of major methicillin-resistant *Staphylococcus aureus* lineages. Int J Med Microbiol 301, 64–8 (2011)

35. Kuhns M, Zautner AE, Rabsch W, Zimmermann O, Weig M, Bader O, Groß U: Rapid discrimination of *Salmonella enterica* serovar Typhi from other serovars by MALDI-TOF mass spectrometry. PLoS One 7, e40004 (2012)

36. Zautner AE, Masanta WO, Tareen AM, Weig M, Lugert R, Gross U, Bader O: Discrimination of multilocus sequence typing-based *Campylobacter jejuni* subgroups by MALDI-TOF mass spectrometry. BMC Microbiol 13, 247 (2013)

37. Zautner AE, Masanta WO, Weig M, Groß U, Bader O: Mass Spectrometry-based PhyloProteomics (MSPP): A novel microbial typing Method. Sci Rep 5, 13431 (2015)

38. Zautner AE, Lugert R, Masanta WO, Weig M, Groß U, Bader O: Subtyping of *Campylobacter jejuni* ssp. *doylei* Isolates Using Mass Spectrometry-based PhyloProteomics (MSPP). JoVE J Vis Exp e54165–e54165 (2016)

39. Emele MF, Možina SS, Lugert R, Bohne W, Masanta WO, Riedel T, Groß U, Bader O, Zautner AE: Proteotyping as alternate typing method to differentiate *Campylobacter coli* clades. Sci Rep 9, 4244 (2019)

40. Gonzales T, Robert-Baudouy J: Bacterial aminopeptidases: properties and functions. FEMS Microbiol Rev 18, 319–44 (1996)

41. Varland S, Osberg C, Arnesen T: *N*-terminal modifications of cellular proteins: The enzymes involved, their substrate specificities and biological effects. Proteomics 15, 2385–2401 (2015)

42. Kumar S, Stecher G, Li M, Knyaz C, Tamura K: MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. Mol Biol Evol 35, 1547–1549 (2018)

43. Fagerquist CK, Bates AH, Heath S, King BC, Garbus BR, Harden LA, Miller WG: Sub-speciating *Campylobacter jejuni* by proteomic analysis of its protein biomarkers and their post-translational modifications. J Proteome Res 5, 2527–38 (2006)

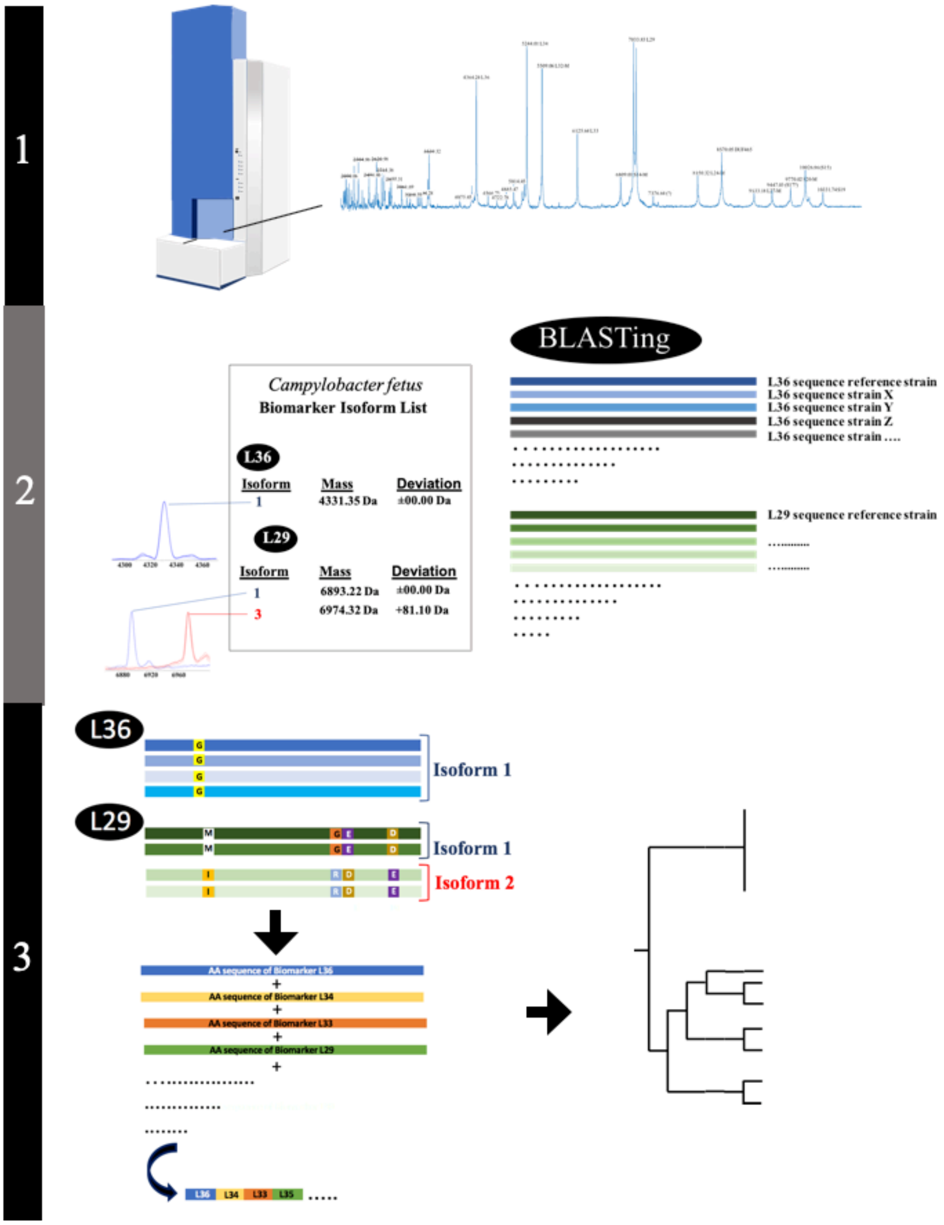


Fig. 1. Illustration of the different proteotying steps. 1) Recording of ICMS mass spectra of the *C. fetus* test cohort and reference strain *Cff* LMG 6442 (NCTC 10842). 2) Establishment of a *C. fetus*-specific allelic isoform list by blasting the genome sequences obtained from the NCBI database against the genome of the *C. fetus* reference strain. Subsequently, allelic isoforms in the test cohort are identified by comparing with the newly established allelic isoform list. 3) For each strain in the test cohort a specific set of biomarker isoforms is obtained. Subsequently the amino acid sequences of the biomarkers are fused into a single sequence what results in specific proteotyping-based sequence types for each of the strains and allows the calculation of a proteotying-derived taxonomic dendrogram.

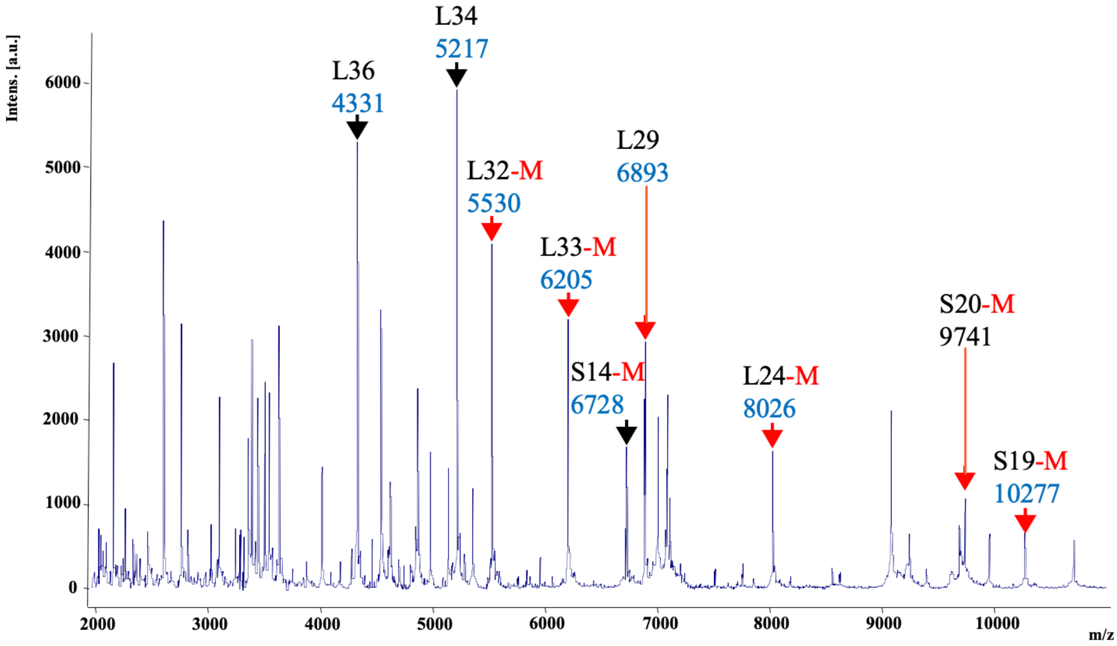


Fig. 2. ICMS spectrum of *C. fetus* subsp. *fetus* reference strain LMG 6442 (NCTC 10842). Singularly charged biomarkers that were part of the *C. fetus* proteotyping scheme labelled with a black arrow or, in case of an *N*-terminal methionine cleavage (posttranslational modification) with a red arrow. Multiple charged ions are not marked in this illustration.

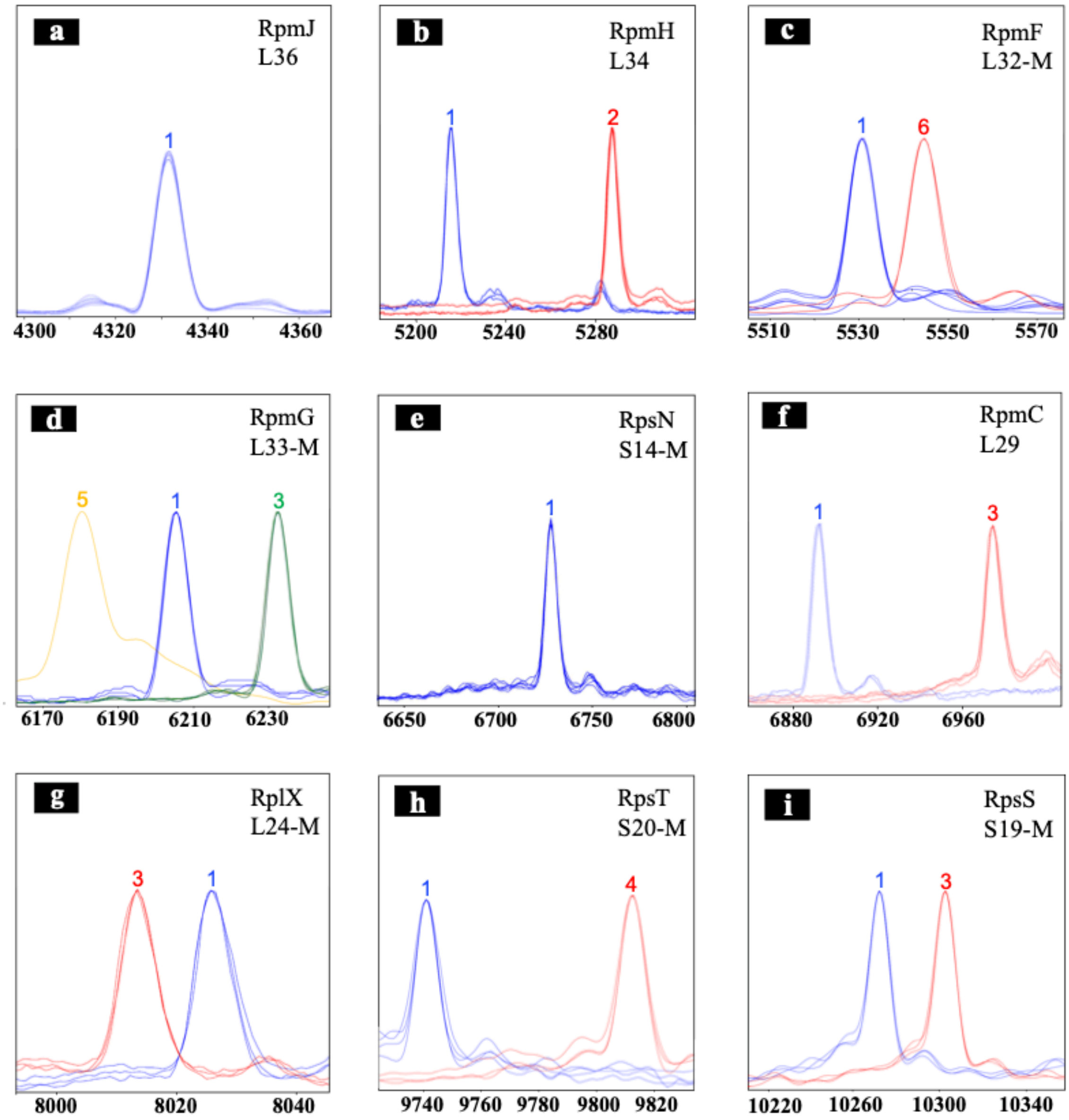


Fig. 3. Overview of *C. fetus* proteotyping biomarkers (a-i). To illustrate the observed mass differences of the allelic isoforms, spectra of different proteotyping-based sequence types were overlaid using the FlexAnalysis evaluation tool. X-Axis: mass [Da] charge-1 ratio, scale 200 Da. Y-Axis: intensity [10x arbitrary units]. For the graphical illustration, the peak intensity of high respectively low peaks was adjusted. Color code of the spectra: Spectra of strains with the isoform of *C. fetus* subsp. *fetus* reference strain LMG 6442 (NCTC 10842) are blue, whereas differing isoforms are colored red, green and yellow. If the *N*-terminal methionine of a ribosomal protein was cleaved the respective illustration is provided with an "-M".

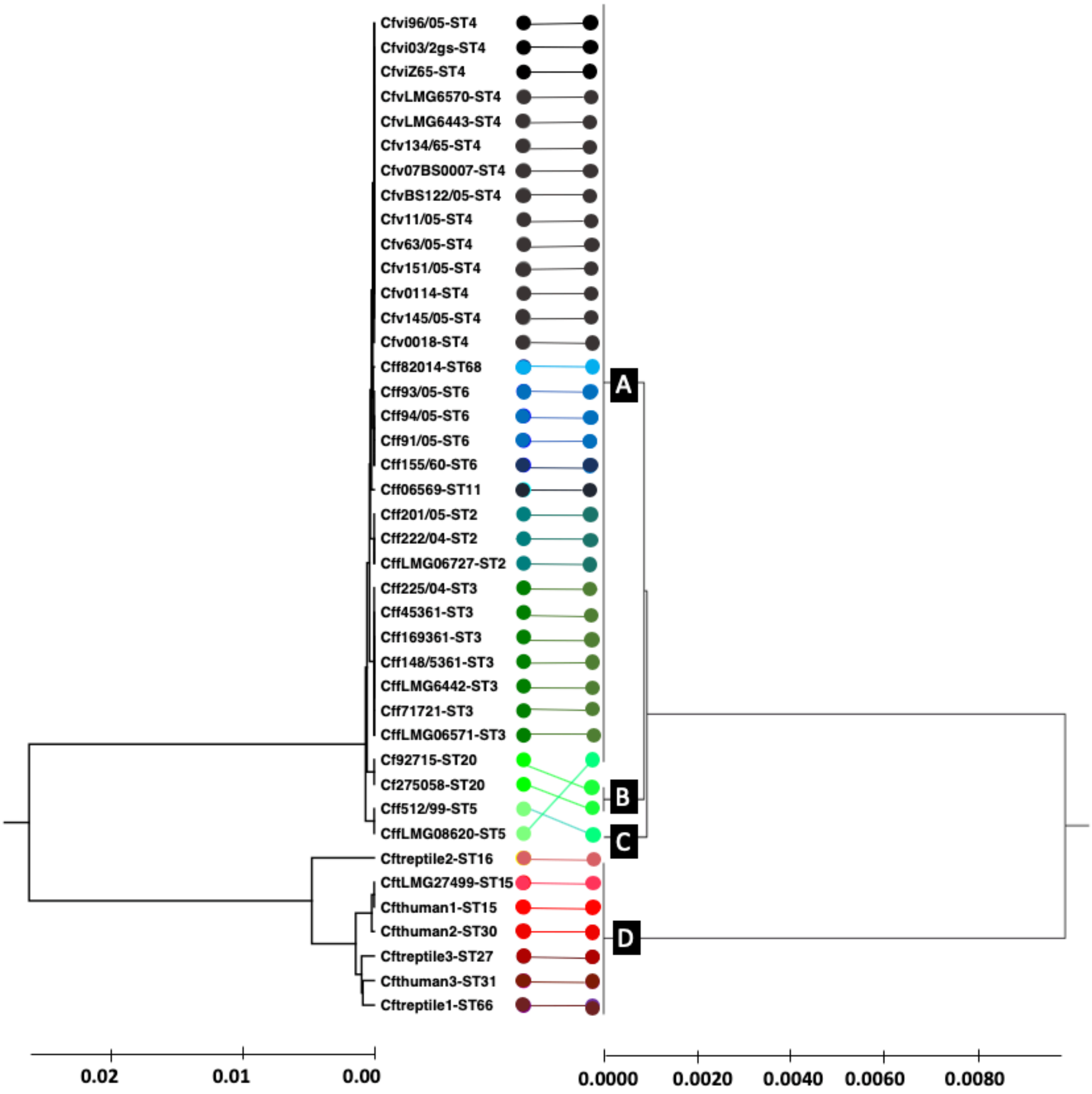


Fig. 4 Comparison of MLST- and proteotyping-derived phylogenies. On the left: Evolutionary tree calculated based on MLST by means of the maximum composite likelihood method (UPGMA). In total, 14 different MLST sequence types were identified which are illustrated in different colors.

On the right: Evolutionary tree based on proteotyping and calculated using UPGMA. Four different proteotyping-derived types were identified. Type A contains most of the *C. fetus* subsp. *fetus* and *C. fetus* subsp. *venerealis* strains. Type B and C contain two MLST ST 5 and one MLST ST 20 strain. The most interesting proteotyping-derived type is type D, which contains all *C. fetus* subsp. *testudinum* strains and thereby allows the differentiation of the subspecies from other *C. fetus* subspecies. The different proteotyping-based sequence types are marked at the branches of the evolutionary tree (A, B, C, D).

Table 1: List of *C. fetus* isolates used in the study.

isolate	origin	region	date	other strain designations	MLST-ST
*Cfv*0018	Preputial washing	Lower-Saxony	28.04.2009		4
*Cfv*145/05	Preputial washing	S-Bavaria	02.08.2005		4
*Cfv*0114	Vaginal sample cattle	Lower-Saxony	19.12.2006		4
*Cfv*151/05	Preputial washing	S-Bavaria	10.08.2005		4
*Cfv*93/05	Preputial washing	Thuringia	12.05.2005		6
*Cff*94/05	Preputial washing	Thuringia	12.05.2005		6
Cfvi 96/05	Preputial washing	Thuringia	12.05.2005		4
*Cff*225/04	Foetus calf	Thuringia	16.12.2004		3
*Cff*512/99	Calf intestinal con-	Thuringia	24.09.1999		5
*Cfv*63/05	Preputial washing	N-Bavaria	6.03.2005		4
*Cfv*11/05	Preputial washing	N-Bavaria	21.01.2005		4
*Cfv*BS122/05	Foetus, cattle	Baden-W.	14.06.2005		4
*Cfv*07BS000	Preputial washing	Baden-W.	26.09.2007		4
*Cfv*134/65	Foetus, cattle	S-Bavaria	12.07.2005		4
*Cff*201/05	-	Thuringia	23.11.2005		2
*Cff*91/05	Preputial washing	Thuringia	12.05.2005		6
*Cff*155/60s	Preputial washing	Baden-W.	07.09.2006		6
*Cff*222/04	Bovine sperm	Saxony	16.12.2004		2
*Cff*45361	Human blood culture	Germany			3
*Cff*169361	Human blood culture	Germany			3
*Cff*148/5361	Human blood culture	Germany			3
*Cfv*LMG6443	Cow, vaginal mucus	United Kingdom	1962	ATCC 19438; CCUG 538; CIP 68.29; JCM 2528;	4
*Cff*LMG6442	Sheep foetus brain	Sweden, Göteborg	1972	ATCC 27374; CCTM La3023; CCUG 6823A; CECT 564; CIP 53.96; JCM 2527; LMG 8849; NCTC	3
*Cfv*LMG6570	Cattle	Belgium	1985	CCUG 7477; CIP 53.105; Florent 483; NIDO 483	4
*Cff*71721	Human blood culture	Germany, Duderstadt	2016		3
*Cff*82014	Human blood culture	Germany, Herzberg am	2015		68
*Cff*LMG27499	Human blood culture	USA, New York	2003	ATCC BAA-2539; Blaser 03-427	15
*Cff*LMG06569	Calf foetus	Belgium	1985	CCUG 17693; CIP 68.8; Florent 7572; NIDO 7572	11
*Cff*LMG06571	Bull genitals	Belgium	1985	CCUG 17694; De Keyser 2125/4; NIDO 2125/4	3
*Cff*LMG0672		Belgium	1985	CCUG 17695A; LMG	2

Table 2: Theoretical biomarker masses predicted by the genome sequence of *C. fetus* reference strain LMG 6442 (NCTC 10842) under consideration of possible posttranslational modifications.

Biomarker	[-Met M+H+]	[-Met mM+H+]	[-Met +PO4 M+H+]	[M+H+]	[fM+H+]
L36	4197	4211	4277	**4332**	4360
L34	5083	5097	5163	**5218**	5246
L32-M	**5527**	5541	5607	5662	5690
L33-M	**6202**	6216	6282	6337	6365
S14-M	**6725**	6739	6805	6860	6888
L29	6759	6773	6839	**6894**	6922
L24-M	**8023**	8037	8103	8158	8186
S20-M	**9738**	9752	9818	9873	9901
S19-M	**10274**	10288	10354	10409	10437

Legend:

[-Met M+H+] = unmodified mass - demethioninated form

[-Met mM+H+] = methylated mass - demethioninated form

[-Met +PO4 M+H+] = phosphorylated mass - demethioninated form

[M+H+] = unmodified mass

[fM+H+] = formylated mass

Table 3: Primers used for sequencing of *C. fetus* genes coding for ribosomal proteins included in the proteotyping scheme.

Biomarker	Gene	Forward primer [5'→3']	Reverse primer [5'→3']	Amplicon length [bp]
L36	RpmJ	CGGGTGATCGCGTTAAAGTT	TACGAATCGCAGCAGCTTCA	522
L34	RpmH	AGTTATGCCGCAAACACCTAT	TTTTCAAGCCCTGCTTTTGCT	699
L32-M	RpmF	ACCACTATTGTGATAGATGCGGT	ACATCAGTAGCACTTTCTCCCA	596
L33-M	RpmG	CCCAGTTGCACTTGAAGAAGG	ACGATCGCTACAACAGCAAAT	539
S14-M	RpsN	AGGACTTCCGTGGTCTTCCA	ACGCTTCTACCACGTTCGTC	624
L29	RpmC	CGCCAGATAGAATCAGCTCGT	GCGGAAGCTTTTTCTAGCAC	701

L24-M	RplX	TTTGACGAAAATGCAGCCGT	ACTGGGAAGCCTTCACGAAC	621
S20-M	RpsT	TTCTCCGGCTCTGCCTCTAA	GCGAGTTCGCCTAGTTCTGG	736
S19	RpsS	GGGCAAACG-TAACTATCGGC	GAACAGGACCGGCATCTACT	752

Table 4: *C. fetus*-specific allelic isoform list.

Locus	Full name / product (ORF Locus tag in LMG 6442)	calc. Average mass [Da]		Frequency in database
RpmJ/ L36				
sequence	MKVRPSVKKMCDKCKIVKRKGIVHVICENPKHKQRQG (37aa)			
1 *	reference isoform LMG 6442 (NCTC 10842)	4331.35	±0.00	100.000% (67/67)

RpmH/ L34				
sequence	MKRTYQPHKTPKKRTHGFRGRMKTKNGRKVINARRAKGRKRLAA (44aa)			
1 *	reference isoform LMG 6442 (NCTC 10842)	5217.26	±0.00	47.761% (32/67)
2*	MKRTYQPHKTPKKRTHGFRERMKT-KNGRKVINARRAKGRKRLAA(44aa)	5289.32	+72.06	32.836% (22/67)
3	MKRTYQPHKTPKKRTHGFRERMRTKNGRKVL-NARRAKGRKRLAA(44aa)	5317.33	+100.07	19.403% (13/67)

RpmF/ L32-M				
sequence	(M)AVPKRRVSHTRAAKRRTHYKVTLPMPVKDKDGSWKMPHRINKTTGEY* (48aa)			
1 *	reference isoform LMG 6442 (NCTC 10842)	5530.47	±0.00	77.941% (53/68)
2	(M)AVPKRRVSHTRAAKRRTHYKVTLPMPVKDKDGS-WKMPHRMNKTTGEY (48aa)	5548.50	+18.03	17.647% (12/68)
3	(M)AVPKRRVSHTRAAKRRTHYKVTLPMPVKDKDGS-WKMPHRINKITGEY (48aa)	5542.52	+12.05	1.471% (1/68)
4	(M)AVPKRRVSHTRAAKCRTHYKVTLPMPVKDKDGS-WKMPHRINKTTGEY (48aa)	5477.42	-53.05	1.471% (1/68)
5	(M)AVPKRLVSHTRAAKRRTHYKVTLPMPVKNKDGS-WKMPHRINKTTGEY (48aa)	5486.45	-44.02	1.471% (1/68)
6*	(M)AVPKRRVSHTRAAKRRTHYKITLPMPVKDKDGS-WKMPHRINKTTGEY (48aa)	5544.49	+14.02	New sequence

RpmG/ L33-M				
sequence	(M)ASANRVKIGLKCAECNDINYTTTKNSKTTTEKLELKKY CPRLKKHTVHKEVKLK (55aa)			
1 *	reference isoform LMG 6442 (NCTC 10842)	6205.31	±0.00	46.970% (31/66)
2	(M)ASANRIKIGLKCVECGDINYTTTKNSKKTTEKLELKKYCPRLKKHTEHKEVKLK (55aa)	6247.39	+42.08	18.182% (12/66)
3 *	(M)ASANRVKIGLKCAECNDINYTTTKNSKKTTEKLELKKYCPRLKKHTVHKEVKLK (55aa)	6232.38	+27.07	33.333% (22/66)
4	(M)ASVNRIKIGLKCVECGDINYTTTKNSKKTTEKLELKKYCPRLKKHTEHKEVKLK (55aa)	6275.44	+70.13	1.515% (1/66)
5*	(M)ASANRVKIGLKCAECNDINYTTTKNSKTTTEKSELKKYCPRLKKHTVHKEVKLK (55aa)	6179.23	-26.08	New sequence

RpsN/ S14-M				
sequence	(M)AKKSMIAKAARKPKFSARGYTRCQICGRPHSVYKDFGICRV CLRKMANEGLIPGLKKASW (61aa)			
1 *	reference isoform LMG 6442 (NCTC 10842)	6728.11	±0.00	80.303% (53/66)
2	(M)AKKSMIAKAARKPKFSVRGYTRCQICGRPHSVYKDFGICRVCLRKMANEGLIPGLKKASW (61aa)	6756.16	+28.05	3.030% (2/66)
3	(M)AKKSMIAKAARAPKFSSRGYTRCQICGRPHSVYKDFGICRVCLRKMANEGLIPGLKKASW (61aa)	6687.01	+41.10	16.667% (11/66)

RpmC/ L29				
sequence	MKYIDISAKSMSELNALLKEKKVLLFTLRQKLKTMQLTNPNE IGETKKDIARINTAISAAK(61aa)			
1 *	reference isoform LMG 6442 (NCTC 10842)	6893.22	±0.00	47.761% (32/67)
2	MKYTEISAKSVSELTALLKEKKVLLFTLRQKLKTMQLTNPNEIRDTKKEIARINTAISAAK (61aa)	6949.27	+56.05	19.403% (13/67)
3*	MKYIDISAKSISELNALLKEKKVLLFTLRQKLKTMQLTNPNEIRDTKKEIARINTAISAAK(61aa)	6974.32	+81.10	32.836% (22/67)

RplX/L24-M				
sequence	(M)AVKYKIKKGDEVKVIAGDDKGKVAKVIAVLPKKGQVIVE GVKVAKKAVKPTEKNPNGGFISKEMPIDISNVAKVEG(77aa)			
1 *	reference isoform LMG 6442 (NCTC 10842)	8026.59	±0.00	47.751% (32/67)

RplX/L24-M				
2	(M)AIKYKIKKGDEVKVIAGDDKGKVAKVLAVLPKKGQVIVEGVKVAKKAVKPTDKNPNGGFVSKEMPIDISNVAKVEG (77aa)	8012.56	-14.03	17.910% (12/67)
3 *	(M)AVKYKIKKGDEVKVIAGDDKGKVAKVIAVLPKKGQVIVEGVKVAKKAVKPTDKNPNGGFISKEMPIDISNVAKVEG (77aa)	8012.56	-14.03	32.836% (22/67)
4	(M)AIKYKIKKGDEVKVIAGDDKGKVAKVLAVLPKKGQVIVEGIKVAKKAVKPTDKNPNGGFVSKEMPIDISNVSKVEG (77aa)	8042.59	+16.00	1.493% (1/67)

RpsT/ S20-M				
sequence	(M)ANHKSAEKRARQTIKRTERNRFYRTRLKNLTKAVRVAV ASGDKDAALVALKDANKNFHSFVSKGFLKKETASRKVSRLA KLVSTLAA (88aa)			
1*	reference isoform LMG 6442 (NCTC 10842)	9741.33	±0.00	48.529% (33/68)
2	(M)ANHKSAEKRARQTIKRTERNRFYRTRLKNLTKAVRVAVANGDKDAALLALKDVNKNFHSFVSKGFLKKETASRKVSRLAKLVSTLAA (88aa)	9810.43	+69.10	16.176% (11/68)
3	(M)ANHKSAEKRARQTIKRTERNRFYRTRLKNLTKAVRVAVANGDKDAALLALKDVNKNFHSFVSKGFLKKKTASRKVSRLAKLVSTLAA (88aa)	9809.49	+68.14	1.471% (1/68)
4*	(M)ANHKSAEKRARQTIKRTERNRFYRTRLKNLTKAVRVAVTSGDKDAALLALKDVNKNFHSFVSKGFLKKETASRKVSRLAKLVSTLAA (88aa)	9813.43	+72.10	32.353% (22/68)
5	(M)ANHKSAEKRARQTIKRTERNRFYRTRLKNLTKAVRVAVANGDKDAALLALKDVNKNFHSFVSKGFLKKETASRKVGRLAKLVSTLAA (88aa)	9780.41	+39.08	1.471% (1/68)

RpsS/ S19-M				
sequence	(M)ARSLKKGPFVDDHVMKKVLAAKAANDNKPIKTWSRRSM IIPEMIGLTFNVHNGKGFIPVYYVTENHIGYKLGEFAPTRTF KGHKGSVQKKIGK (93aa)			
1*	reference isoform LMG 6442 (NCTC 10842)	10277.10	±0.00	47.761% (32/67)
2	(M)ARSLKKGPFVDDHVMKKVLAAKAANDNKPIKTWSRRSTIIPEMIGLTFNVHNGKSFIPVYYVTENHIGYKLGEFAPTRTFKGHKGSVQKKIGK (93aa)	10277.04	-0.06	14.925% (10/67)

RpsS/ S19-M				
3*	(M)ARSLKKGPFVDDHVMKKVLAAKAANDNKPIKT-WSRRSMIIPEMIGLTFNVHNGKSFIPVYYV-TENHIGYKLGEFAPTRTFKGHKGSVQKKIGK (93aa)	10307.12	+30.02	32.836% (22/67)
4	(M)ARSLKKGPFVDDHVMEKVLAAKATNDNKPIKT-WSRRSTIIPEMIGLTFNVHNGKSFIPVYYV-TENHIGYKLGEFAPTRTFKGHKGSVQKKIGK (93aa)	10308.00	+30.90	4.478% (3/67)

Legend: * observed in test population AA numbering including start-methionine, if the mass spectrometric measurements indicate its absence it is written in brackets (M)

Table 5: Accession numbers of *C. fetus*-specific proteotyping biomarker isoforms.

Biomarker	Isoform	Gene Bank Accession	Locus Tag	Protein ID
L36	1	MK463617		
L34	1	CP000487.1:557520-557654	CFF8240_0551	ABK82017.1
L34	2	CP027287.1:608973-609107	C6B32_03095	AVK80859.1
L32-M	1	CP000487.1:210702-210848	CFF8240_0235	ABK81894.1
L32-M	6	MK463615		
L33-M	1	CP000487.1:1313847-1313949	CFF8240_1324	ABK82614.1
L33-M	3	CP027287.1:c1398913-1398746	C6B32_06940	AVK81560.1
L33-M	5	MK463616		
S14-M	1	CP000487.1:39526-39711	CFF8240_0047	ABK82398.1
L29	1	CP000487.1:37925-38110	CFF8240_0042	ABK82084.1
L29	3	CP027287.1:36898-37083	C6B32_00200	AVK80319.1
L24-M	1	CP000487.1:38746-38979	CFF8240_0045	ABK83333.1
L24-M	3	CP027287.1:37719-37952	C6B32_00215	AVK80322.1
S20-M	1	CP000487.1:1678191-1678457	CFF8240_1718	ABK82453.1
S20-M	4	CP027287.1:1762618-1762884	C6B32_08820	AVK81906.1
S19-M	1	CP000487.1:36187-36468	CFF8240_0038	ABK81869.1
S19-M	3	CP027287.1:35160-35441	C6B32_00180	AVK80315.1

Table 6: Measured and calculated biomarker masses.

Biomarker	Isoform	Measured Mass [Da]	Standard deviation	∆ Measured mass/ Average mass	Monoisotopic mass [Da]	Average Mass [Da]
L36	Isoform 1	4331	0.765	0.35	4328.40	4331.35
L34	Isoform 1	5217	0.425	0.26	5214.02	5217.26

L34	Isoform 2	5290	0.593	0.68	5286.04	5289.32
L32-M	Isoform 1	5530	0.478	0.47	5526.99	5530.47
L32-M	Isoform 6	5544	0.475	0.49	5541.01	5544.49
L33-M	Isoform 1	6205	0.867	0.31	6201.37	6205.31
L33-M	Isoform 3	6232	0.381	0.38	6228.42	6232.38
L33-M	Isoform 5	6179	0.271	0.23	6175.32	6179.23
S14-M	Isoform 1	6728	0.445	0.11	6854.63	6728.11
L29	Isoform 1	6893	0.321	0.22	6888.84	6893.22
L29	Isoform 3	6975	0.877	0.68	6969.96	6974.32
L24-M	Isoform 1	8026	0.928	0.59	8021.62	8026.59
L24-M	Isoform 3	8012	0.620	0.56	8007.61	8012.56
S20-M	Isoform 1	9741	0.959	0.33	9735.50	9741.33
S20-M	Isoform 4	9813	0.361	0.43	9807.56	9813.43
S19-M	Isoform 1	10277	0.499	0.10	10270.58	10277.10
S19-M	Isoform 3	10308	0.635	0.88	10300.59	10307.12

2.3 Identifizierung von *Clostridioides difficile* PCR Ribotyp 027 Stämmen via Proteotypisierung

Clostridioides difficile, ein Gram-positives sporenbildendes Bakterium, ist weltweit die Hauptursache nosokomialer Diarrhö und daher eine immense Belastung für das Gesundheitssystem. Während des letzten Jahrzehnts ist eine Population des hypervirulenten PCR-Ribotypen (RT) 027 auf der ganzen Welt sehr schnell emporgekommen. Stämme dieses Ribotyps werden mit einem schlimmeren Krankheitsverlauf und einer höheren Mortalitätsrate assoziiert, weshalb eine schnelle Identifikation dieser Stämme äußerst wichtig ist. Während gängige diagnostische Methoden wie MLST oder die Ribotypisierung via PCR zeitaufwändig sind, bietet die auf der MALDI-TOF Massenspektrometrie basierende Proteotypisierung eine schnelle, günstige und verlässliche Alternativlösung.

In dieser Studie wurde ein *C. difficile*-spezifisches Proteotypisierungsschema entwickelt. Insgesamt wurden 77 ribotypisierte Stämme, die alle fünf bekannten MLST-Kladen abdeckten, mittels Massenspektrometrie analysiert. Die MLST, basierend auf kompletten Genomsequenzen und PCR-Ribotypisierung wurden als Referenzmethoden verwendet. Isoforme der identifizierten Biomarkermassen, in der Regel ribosomale Proteine, wurden mit den zugehörigen Aminosäuresequenzen assoziiert und in das *C. difficile*-Proteotypisierungsschema aufgenommen.

Letztlich ist es gelungen, neun Biomarker mit den dafür kodierenden Genen in Verbindung zu bringen und diese in das *C. difficile*-Proteotypisierungschema aufzunehmen.

Die Diskriminierungsfähigkeit des entwickelten Schemas basierte im Wesentlichen auf Isoformen der Biomarker L28-M (zwei Haupt-Isoforme), L35-M (vier Haupt-Isoforme) und S20-M (zwei Haupt-Isoforme), die in insgesamt 16 verschiedenen Proteo-Typen resultierten.

In unserer Testkohorte konnten fünf der insgesamt 16 verschiedenen Proteo-Typen identifiziert werden. Diese fünf Typen stimmten nicht exakt mit den fünf MLST-basierten *C. difficile*-Kladen überein, die Tiefe der Subtypisierung war jedoch äquivalent. Die wesentliche Entdeckung war, dass Klade B ausschließlich Isolate des hypervirulenten Ribotyps 027 enthielt.

Die Proteotypisierung ist eine stabile und in der Anwendung einfache Methode für die Subtypisierung von Spezies und eine vielversprechende Alternative zu gegenwärtig verwendeten, molekularen Methoden. Da RT027 Isolate von nicht-RT027 Isolaten via Proteotypisierung unterschieden werden können, ist die Methode für den Einsatz im Rahmen der Routinediagnostik ausgesprochen interessant.

Autoren: **Matthias F. Emele**, Felix M. Joppe, Thomas Riedel, Jörg Overmann, Maja Rupnik, Paul Cooper, R. Lia Kusumawati, Friederike Laukien, Ortrud Zimmermann, Wolfgang Bohne, Uwe Gross, Oliver Bader, Andreas E. Zautner

Beitrag der Autoren zur praktischen Arbeit:

Matthias Frederik Emele:
Datenanalyse, PCR und Sangersequenzierung, Anfertigung des Manuskripts
Felix Manoel Joppe:
Erstellung der Isoformendatenbank, Korrektur des Manuskripts
Thomas Riedel, Jörg Overmann:
Whole genome sequencing, Korrektur des Manuskripts
Maja Rupnik:
Ribotypisierung, Korrektur des Manuskripts
Friederike Laukien:
Bakterienkulturen, Aufzeichnung von Massenspektren, Korrektur des Manuskripts
Ortrud Zimmermann, Paul Cooper, R. Lia Kusamawati:
Sammlung und Identifikation der *C. difficile*-Isolate, Korrektur des Manuskripts
Wolfgang Bohne:
WGS Analyse, Korrektur des Manuskripts
Uwe Groß, Oliver Bader, Andreas E. Zautner:
Planung der Studie, Datenanalyse, Manuskript und Bearbeitung von Grafiken, Korrektur des Manuskripts
Andreas Erich Zautner:
MLST Analyse und Berechnung der Dendrogramme

Status des Manuskripts:
In Re-Revision; Frontiers in Microbiology (Infectious diseases)

Proteotyping of *Clostridioides difficile* as alternate Typing Method to Ribotyping is able to differentiate the Ribotype 027

Matthias F. Emele[1], Felix M. Joppe[1], Thomas Riedel[2,3], Jörg Overmann[2,3], Maja Rupnik[4,5], Paul Cooper[6], R. Lia Kusumawati[7], Friederike Laukien[1], Ortrud Zimmermann[1], Wolfgang Bohne[1], Uwe Gross[1], Oliver Bader[1], and Andreas E. Zautner[1]

[1] Institut für Medizinische Mikrobiologie, Universitätsmedizin Göttingen, Göttingen, Germany

[2] Leibniz-Institut DSMZ- Deutsche Sammlung von Mikroorganismen und Zellkulturen, Braunschweig, Germany

[3]Deutsches Zentrum für Infektionsforschung (DZIF), Standort Hannover-Braunschweig, Braunschweig, Germany

[4] Institute of Public Health Maribor, Maribor, Slovenia

[5] Faculty of Medicine, University of Maribor, Maribor, Slovenia

[6] St. Martin de Porres Hospital, Eikwe, Ghana.

[7] Department of Microbiology, Faculty of Medicine, Universitas Sumatera Utara, Medan, Indonesia

Corresponding author: PD Dr. med. habil. Andreas E. Zautner, Institut für Medizinische Mikrobiologie, Universitätsmedizin Göttingen, Kreuzbergring 57, D-37075 Göttingen, Germany, Phone: +49(0)551 39-8549, Fax: +49(0)551 39-5861, E-Mail: azautne@gwdg.de

Running Title: Proteotyping of *Clostridioides difficile*

Keywords: MALDI-TOF MS, *Clostridioides difficile*, *Clostridium difficile*, below species differentiation, proteotyping

ABSTRACT

Clostridioides difficile, a Gram-positive spore-former, is the main cause of nosocomial diarrhea worldwide and therefore a substantial burden to the healthcare system. During the past decade, the hypervirulent PCR-ribotype (RT) 027 population emerged rapidly all over the world, associated with both, higher severity and mortality rates. It is thus of great importance to identify epidemic strains such as RT027 as fast as possible. While commonly used diagnostic methods, e.g. multi locus sequence typing (MLST) or PCR-ribotyping, are time-consuming, proteotyping offers a fast, inexpensive, and reliable alternative solution.

In this study, we established a MALDI-TOF-based typing scheme for *C. difficile.* A total of 77 ribotyped strains representative for five MLST clades were analyzed by mass spectrometry. MLST, based on whole genome sequences, and PCR-ribotyping were used as reference methods. Isoforms of MS-detectable biomarkers, typically ribosomal proteins, were related with the deduced amino acid sequences and added to the *C. difficile* proteotyping scheme. In total, we were able to associate nine biomarkers with their encoding genes and include them in our proteotyping scheme. The discriminatory capacity of the *C. difficile* proteotyping scheme was mainly based on isoforms of L28-M (2 main isoforms), L35-M (4 main isoforms), and S20-M (2 main isoforms) giving rise to at least 16 proteotyping-derived types.

In our test population, five of these 16 proteotyping-derived types were detected. These five proteotyping-derived types did not correspond exactly to the included five MLST-based *C. difficile* clades, nevertheless the subtyping depth of both methods was equivalent. Most importantly, Proteotyping-derived clade B contained only isolates of the hypervirulent RT027.

Proteotyping is a stable and easy-to-perform intraspecies typing method and a promising alternative to currently used molecular techniques. It is possible to distinguish RT027 isolates from non-RT027 isolates using proteotyping, providing a valuable diagnostic tool.

INTRODUCTION

Clostridioides difficile (Lawson et al., 2016) is a Gram-positive anaerobic spore former and the most frequent cause of antibiotic-associated diarrhea (Leffler and Lamont, 2015; Lo Vecchio and Zacur, 2012; Martin et al., 2016; Smits et al., 2016). Current research revealed that this pathogen is responsible for more than 152,000 reported healthcare-associated *C. difficile* infections and more than 8,300 associated deaths every year (Cassini et al., 2016). The incidence rate observed in the United states was consistent with the European one (Martin et al., 2016). The symptoms of a *C. difficile* infection (CDI) appear in various manifestations: The spectrum comprises rather weak symptoms like mild diarrhea but also serious forms like toxic megacolon, pseudomembranous colitis (PMC) or perforation of the colon (Nanwa et al., 2015). Although the potential for severe disease is high, most of the colonized individuals do not show any symptoms (Donskey et al., 2015; Elliott et al., 2017). Despite the fact that the involvement of the small intestine has been observed, characteristic PMC lesions are normally restricted to the colon (Jacobs et al., 2001; Keel and Songer, 2006). Infections outside of the intestine only occur very rarely (Byl et al., 1996).

Over the last decade, different hypervirulent *C. difficile* strains emerged. The most prominent of these hypervirulent strains has been categorized as RT027, which has emerged especially in Canada, North America, and various European countries (Brazier et al., 2008; Indra et al., 2008; Loo et al., 2005; McDonald et al., 2005; Pépin et al., 2004; Valiente et al., 2014). Outbreak studies from these and other countries all over the world revealed that RT027 is associated with an intensification of the worldwide epidemic of nosocomial *C. difficile* infections, resulting in repeated occurrence and high mortality rates (Hubert et al., 2007; Loo et al., 2005; Mooney, 2007; Redelings et al., 2007). Furthermore, these studies point out the rapid spreading of RT027 strains, and, despite of the currently declining cases in certain geographical areas, a persisting worldwide dissemination can be observed (Arvand et al., 2014; Arvand and Bettge-Weller, 2016; He et al., 2013).

Besides adherence- and motility factors virulence of *C. difficile* mainly depends on toxins encoded by the pathogenicity locus (PaLoc) (Carter et al., 2015). Most pathogenic *C. difficile* strains own two homologous toxins (TcdA and TcdB) and three proteins presumably responsible for production and secretion of these toxins are PaLoc-encoded (Awad et al., 2014; Monot et al., 2015; Smits, 2013). These toxins are produced due to insufficient supply of nutrients, primarily harm epithelial cells of the intestine and are then taken up by endocytosis. Once in the cytosol they are activated what results in necrosis of epithelial cells. In this way intestinal membrane integrity is lost and the host is exposed to microorganisms in the intestine what finally initiates the host inflammatory response (Abt et al., 2016). Although most *C. difficile* strains exhibit an identical localization of the PaLoc, some studies identified strains with an atypical localization (Hunt and Ballard, 2013).

Additionally, the PaLoc encodes the proteins TcdC, TcdE and TcdR. The alternative sigma factor TcdR enables binding of the RNA polymerase to the *tcdA* and *tcdB* promoters and to its own promoter by a positive feedback loop (Abt et al., 2016; Mani and Dupuy, 2001). When the stationary growth phase of *C. difficile* is reached TcdR drives the transcription of *tcdA* and *tcdB* (Mani and Dupuy, 2001). In the phase of exponential growth, TcdC is expressed at a higher level by *C. difficile*, what possibly serves as anti-sigma factor and thereby as a suppressor for the *tcdA* and *tcdB* transcription (Matamouros et al., 2007). To inhibit transcription, typical anti-sigma factors normally bind to sigma factors. In *C. difficile* transcription of *tcdA* and *tcdB* is possibly inhibited by direct binding of TcdC to single-stranded DNA (van Leeuwen et al., 2013). The fact that hypervirulence of strains was associated with deletions in the sequence of *tcdC* is another indicator for the possible involvement of TcdC in limiting *C. difficile* toxin expression (Spigaglia and Mastrantonio, 2002). The assumption that TcdC is an important repressor of the toxin production is also supported by the observation that the absence of a functional TcdC due to a frame shift mutation (D117 bp) in the *tcdC* gene is associated with an increased production of toxins in at least some (hyper-)virulent strains, like RT027 (Bakker et al., 2012; Curry et al., 2007; Warny et al., 2005).

In contrast to this, other studies revealed that the level of toxin production *in vitro* was neither affected by a restorated *tcdC* expression in a RT027 strain nor in genetically modified *tcdC* mutants (Bakker et al., 2012; Cartman et al., 2012).

Another gene in the PaLoc is *tcdE*, which encodes a holin-like protein. Recent studies point to the fact that this protein facilitates secretion of TcdA and TcdB, as these are proteins without conventional secretion signal sequences. Research findings on the influence of TcdE on secretion also diverge: As some studies underline its importance for toxin secretion, other studies show secretion in its absence (Govind and Dupuy, 2012; Olling et al., 2012). However, recently published work showed the involvement of TcdE in TcdA and TcdB secretion of high-toxin producing *C. difficile* strains (Govind et al., 2015; Mehner-Breitfeld et al., 2018).

Another toxin expressed by some *C. difficile* strains like hypervirulent RT027 which is not encoded by the PaLoc is the binary toxin or *C. difficile transferase* (CDT) (Stubbs et al., 2000; Sundriyal et al., 2010) Although this toxin is supposed to enhance virulence and some studies show a correlation between its presence and an increased mortality rate, its exact role is unknown so far (Barbut et al., 2005; Gerding et al., 2014; McEllistrem et al., 2005). CDT consists of two proteins: CdtA, which is a ADP-ribosyl transferase, responsible for ribosylation of actin in cells of eukaryotes. The second protein is CdtB, which is responsible for the formation of pores in acidified endosomes. Furthermore, it contributes to the transport of CdtA into the cytosol. Papatheodorou and coworkers identified the lipolysis-stimulated lipoprotein receptor as the membrane receptor responsible for CDT uptake by target cells (Papatheodorou et al., 2011).

The interference of ribosylation and actin polymerization leads to microtubule-caused cellular protrusion and an enhanced delivery of fibronectin to the surface of the cell. This mechanism finally improves target cell adhesion of *C. difficile* (Schwan et al., 2014).

There are several possible reasons for the spreading of RT027 strains: One reason is, that these strains show a higher resistance to fluoroquinolones compared to other strains (Dannheim et al., 2017; Drudy et al., 2007; Sebaihia et al., 2006). Comparing RT027 strains with historic, pre-epidemic strains revealed that each of the epidemic lineages exhibits a *gyrA* gene mutation responsible for an increased resistance to fluoroquinolones. As it is almost certain, that fluoroquinolone resistance was crucial for the spreading of RT027 strains the resistance was also shown in non-epidemic *C. difficile* strains, what suggests that additional factors contributed to its emergence (Spigaglia et al., 2008, 2010).

Pangenome studies concerning antimicrobial resistance genes (ARG) of *C. difficile* revealed that the most common ARG is the chromosome-encoded *cdeA*, which is a well-known multidrug transporter. Other frequently observed ARGs were those encoding resistance to tetracycline, aminoglycosides, and erythromycin (Knetsch et al., 2018).

Another factor that possibly contributed to the spreading is the implementation of trehalose as a food additive, which came into the market shortly before the rise of virulent strains RT078 and RT027. RT027 strains exhibit a single point mutation in the repressor of trehalose what leads to a more than 500-fold increase of sensitivity to trehalose. Trehalose also increased the virulence of RT027 strains in mouse models of *C. difficile* infections (Collins et al., 2018; Robinson et al., 2014).

Moreover, it was proposed that, after an antibiotic treatment, the re-colonization of the gut by commensals is inhibited by a phenol derivate, *p*-cresol, produced by *C. difficile* (Dawson et al., 2008). Also the ability to form spores has been proposed to contribute to the difference in virulence between RT027 and other *C. difficile* strains (Burns et al., 2010; Lanis et al., 2010).

There is a wide range of diagnostic methods available to investigate on the phylogeny of *C. difficile*, including PCR-ribotyping and multi locus sequence typing (MLST) (Knetsch et al., 2013). The most common method in Europe is PCR-ribotyping what meanwhile also applies to the United States (Fawley et al., 2015; Janezic and Rupnik, 2010; Waslawski et al., 2013). This approach was first described by Gürtler (Gürtler, 1993) and makes use of length differences (200-600 bp) of the intergenic spacer region (ISR) between 16S and 23S rRNA genes. Furthermore, different *C. difficile* strains also exhibit different numbers of alleles in the ribosomal operon. By combining ISR length- and allele number variation, a specific banding pattern can be obtained for the respective ribotype by PCR amplification with a single primer pair (Janezic, 2016).

In contrast, MLST discriminates isolates using nucleotide sequences of housekeeping gene fragments (Maiden et al., 1998), where a sequence type (ST) number is assigned to each unique combination of alleles. The MLST technique is also scalable to high-throughput robotic systems (Pavón and Maiden, 2009).

To respond immediately in case of a disease outbreak, fast, accurate and inexpensive diagnostic methods are indispensable. Since PCR-ribotyping and MLST are relatively expensive and time-consuming, matrix-assisted laser desorption/ionization mass spectrometry (MALDI-TOF MS) represents a promising alternative (Lavigne et al., 2013; Patel, 2015). This technique has become the current standard for species identification in many areas of clinical microbiological laboratories (Bader, 2013; Seng et al., 2010). Beside species identification, MALDI-TOF MS allows distinction of subspecies by accurate discrimination of strain-specific biomarkers (Durighello et al., 2014; Lartigue, 2013; Suarez et al., 2013). Previous studies have shown the possibility to differentiate *Salmonella enterica ssp. Enterica serovar Typhi* from *Salmonella enterica* ssp. *enterica* serotypes, which are of minor clinical relevance (Kuhns et al., 2012). Moreover, it was shown that it is even possible to discriminate different MLST sequence types (STs) of *Campylobacter jejuni* ssp. *jejuni* using a single biomarker ion (Zautner et al., 2013). Cheng and colleagues recently discovered that it is possible to differentiate Clade 4 strains of *C. difficile* from other *C. difficile* strains by MALDI-TOF MS on the basis of 5 markers (Cheng et al., 2018). In another recent study Corver and coworkers identified two peptide markers (*m/z* = 4927.81 and *m/z* = 5001.84) that enable the identification of *C. difficile* MLST types 1 and 11 by MALDI-MS (Corver et al., 2018). Another MALDI-TOF MS-based subtyping approach was published by Ortega and coworkers: They used a technique called high molecular weight (HMW) typing where a protein profile within the mass range of 30 to 50 kDa is analyzed (Ortega et al., 2018; Rizzardi and Åkerlund, 2015). More precisely this method groups *C. difficile* strains according to proteins of their surface layers. Within the study, they identified different HMW profiles. One of those profiles only harbors RT027 strains, what makes it an interesting tool for rapid subtyping (Ortega et al., 2018).

The main problem of clustering-based MALDI-TOF MS-typing methods is the lack of knowledge about the protein that corresponds to the respective peak in the mass spectrum. This problem can be solved to a certain degree using proteotyping. This microbial typing method that we initially named Mass Spectrometry-based PhyloProteomics (MSPP), but which we will, in accordance with the terminology now used in the scientific community (Karlsson et al., 2015), refer to as proteotyping, was previously successfully used for subtyping of *C. jejuni* ssp. *jejuni* and ssp. *doylei* isolates (Zautner et al., 2015, 2016). The essential characteristic of our proteotyping method is an amino acid sequence catalogue of isoforms of alleles. These isoforms are the result of non-synonymous mutations in genes coding for ribosomal proteins (biomarker genes). These mutations can be detected in the form of mass shifts within MALDI-TOF spectra. It is then possible to assign an isolate to a specific proteotyping-derived type by analyzing the scheme of recorded biomarker masses and deducing the respective amino acid sequence. The key advantage of proteotyping in comparison to whole mass spectrum clustering approaches is that only mass changes assigned to a specific set of allelic isoforms of the same protein are considered for deduction of phylogeny. Alternative methods that focus on presence or absence of single masses as well as peak intensity are leading to imprecise results (Matsumura et al., 2014; Novais et al., 2014; Suarez et al., 2013; Zautner et al., 2013).

For this study, we compiled a collection of 94 *C. difficile* strains to develop and test a *C. difficile*-specific proteotyping scheme.

MATERIALS AND METHODS

Clostridioides difficile isolates

In total, 94 *C. difficile* isolates were chosen in a way, that the test collection represented a high genetic diversity and the currently clinical relevant and most prevalent five out of six established clades of this species (Dingle et al., 2014; Knetsch et al., 2012; Knight et al., 2015; Riedel et al., 2017; Stabler et al., 2009). For the MALDI-TOF analyses 77 isolates were selected for which a complete genome was already sequenced (data not shown). To broaden the basis for the differentiation of RT027 isolates, 17 additional RT027 isolates were included in the study for which no genomic data was available. More precisely, 46 clade 1 strains, 24 clade 2 strains, 2 clade 3 strains, 17 clade 4 strains and 5 clade 5 strains were selected for the experiments (Supplementary Table 1). Isolates of clade C-I were not available and were not included in the study. The entire collection consisted of clinical isolates from four different countries: Germany, Great Britain, Ghana, and Indonesia (Seugendo et al., 2018).

Bacterial culture conditions

C. difficile isolates were kept in store in the form of cryobank stocks (Mast Diagnostica, Reinfeld, Germany) at -80 °C. Isolates were incubated for 48 h at 37 °C on Columbia agar (Merck, Darmstadt, Germany) supplemented with 5 % sheep blood (Oxoid, Wesel, Germany) under anaerobic condition using a COY anaerobic gas chamber (COY Laboratory Products, USA). The atmosphere used consisted of 85 % N_2, 10 % H_2, 5 % CO_2. All experiments were carried out under biosafety level 2 conditions.

Preparation of matrix solution

To prepare the matrix solution used for the experiments α-cyano-4-hydroxy-cinnamic acid (HCCA) purified matrix substance (Bruker Daltonics, Bremen, Germany) was dissolved in standard solvent consisting of 47.5 % MALDI-grade water, 50 % acetonitrile, and 2.5 % trifluoroacetic acid (all Sigma-Aldrich, Taufkirchen, Germany) by what the solution had a final concentration of 10 mg HCCA/mL. In order to have an internal calibrant for the measurements purified recombinant human insulin (Sigma-Aldrich, Taufkirchen, Germany) was added to HCCA. In the following step, human insulin was dissolved in 50 % aqueous acetonitrile to attain a final concentration of 10 pg/µL. The precise determination of the insulin peak mass was done experimentally by mixing with Biotyper Test Standard (BTS, Bruker Daltonics) and yielded an *m/z* of 5,806.1. The insulin peak was chosen as internal calibrant for all *C. difficile* mass spectra because it did not cover any biomarker masses of interest. An internal calibrant has a crucial effect on precision during determination of biomarker mass variations. This approach enabled us to detect mass difference with an accuracy of up to 1 Da.

MALDI-TOF mass spectrometry

Sample preparation for MALDI-TOF MS measurements was done using two different procedures following the manufacturer's instructions: (i) smear preparation, which, from experience, allows a better detection of peaks in the *m/z* range > 10,000 kDa and (ii) formic acid/acetonitrile extraction, facilitating more precise analysis in the *m/z* range < 10,000 kDa.

Briefly, to prepare extract samples, five colonies that were plated for 48 h on agar were thoroughly resuspended in 300 µL ddH$_2$O followed by the addition of 900 µL of absolute ethanol. The suspension was then mixed by pipetting up and down repeatedly. After complete suspension of the bacterial colonies the suspensions were centrifuged for 1 min (13,000 × g). The supernatant was discarded followed by drying of the pellets for approx. 10 min at room temperature. To resuspend the pellet in 50 µL of 70 % formic acid it was vortexed thoroughly. 50 µL of acetonitrile were then added to each sample and again mixed by pipetting as previously described, followed by centrifugation of the mixture for 2 min (13,000 × g) removing cellular debris. Subsequently, 1 µL of the supernatant was transferred into the designated field on a MALDI target plate, consisting of polished steel. It was left to dry at room temperature for approx. 5 min and subsequently overlaid with 1 µL of HCCA matrix containing the human insulin. After another drying step at room temperature samples were ready for MS-analysis.

MALDI-TOF MS measurements were performed according to the MALDI Biotyper standard procedures (Bruker Daltonics, Bremen, Germany). During analysis, 600 spectra in a mass range between 2 and 20 kDa were collected in 100-shots steps on an Autoflex III system and summed up. Results obtained with MALDI Biotyper (database release 2016) identification score values ≥ 2.000 were considered correct.

Identification of biomarkers in mass spectra

To analyze the received mass spectra, the software FlexAnalysis (Bruker Daltonics, Bremen, Germany) and its embedded standard algorithms were used. First, spectra were internally calibrated according to the known insulin peak (*m/z* = 5,806.1), followed by baseline subtraction (TopHat) and smoothing as implemented in the standard MBT method.

To determine the theoretical average molecular weight of the ribosomal proteins corresponding to the respective open reading frame of the different genomes (data not shown), the deduced amino acid sequences were uploaded separately to the molecular weight calculator tool at the ExPASy Bioinformatics Resource Portal (http://web.expasy.org/compute_pi/). Eukaryotic as well as ribosomal proteins of Enterobacteriaceae frequently undergo post-translational modifications (Gonzales and Robert-Baudouy, 1996; Varland et al., 2015). Consequently, further potential molecular weights needed to be calculated for each biomarker. In our context, the most relevant post-translational modification was proteolytic removal of *N*-terminal methionine, which was considered with a mass difference of -131.04 Da. In addition to the cleavage of the *N*-terminal methionine, further post-translational modifications may occur, e.g. acetylation, phosphorylation, formylation, and methylation (Kentache et al., 2016; Ouidir et al., 2015).

In order to identify biomarker masses, more precisely to assign a calculated biomarker mass to a certain allelic isoform, measured masses were checked against the calculated masses of the *C. difficile* 630 (DSM 27543) reference genome (Dannheim et al., 2017). If there was no clear correspondence between biomarker mass in the spectrum of a particular clinical isolate and the masses calculated from the *C. difficile* 630 (DSM 27543) reference genome (BioProject PRJNA275406, Genome sequence: CP010905.2), the spectrum was examined regarding peaks with a different molecular weight or more specifically amino acid substitutions that could be causal for the mass shift. Allelic isoforms in the test cohort were reconfirmed by *in silico* translation of the gene sequences taken from the complete bacterial genome and subsequent alignment of the resulting amino acid sequences. For each of the cases the predicted amino acid exchanges could be confirmed, which also served as additional argument in favor of the identity of the peak.

Phylogenetic analysis and Proteotyping

For handling of trace data, nucleotide sequences, and subsequent alignment of the deduced protein sequences, Geneious V 11.1.2, the Molecular Biology and NGS Analysis Tool was used (Biomatters Ltd., Auckland, New Zealand). For each biomarker (ribosomal protein encoding gene) identified in strain *C. difficile* 630 (DSM 27543), the sequences were screened against the respective genome sequence of the 77 isolates for which genome sequence data was available. Subsequently, an amino acid sequence list containing all allelic isoforms of the 9 biomarkers included in the proteotyping scheme was assembled. To construct the unweighted pair group method using average linkages (UPGMA)-tree, Molecular Evolutionary Genetics Analysis X (MEGA X) software was used (Kumar et al., 2018).

The respective PCR-ribotypes of the isolates were determined by agarose (isolates from Indonesia and Ghana) or capillary gel electrophoresis (isolates from Germany) following consensus protocols (ECDIS-Net, CDRN) described in previous publications (Berger et al., 2018; Fawley et al., 2016; Janezic and Rupnik, 2010; van Dorp et al., 2016).

For the 17 isolates without genome sequence data, the gene loci for the ribosomal proteins L28 and L35 were sequenced using the following primers: CdiffL28-F01: 5’-GTT-ATC-ATT-TTA-AGG-AGG-TGT-GCG-3’ and CdiffL28-R01: 5’-TGG-CTG-GAT-TTG-GTC-AGC-AC-3’; CdiffL35-F01: 5’-ACC-AAC-AAA-AGC-CCC-TGC-AT-3’ and CdiffL35-R01: 5’-TCT-TGC-CAT-CGT-TAT-GAC-CTC-C-3’. PCR-reactions were conducted with the following parameters: two denaturation steps at 95 °C for 30 s; annealing at 60 °C for 1 min; two elongation steps at 68

°C for 1- and 5-min. Sanger sequencing of the amplificates was performed by SeqLab-Microsynth (Göttingen, Germany).

RESULTS

The previously established proteotyping (MSPP) workflow (Zautner et al., 2015) was used to develop a *C. difficile*-specific proteotyping scheme as outlined in detail below (Fig. 1). In summary, the mass spectrum of the genome sequenced *C. difficile* reference strain 630 (DSM 27543) was recorded followed by the assignment of spectrum masses to protein-coding genes. Analysis of genome sequences received from the NCBI database enabled the establishment of an allelic isoforms list of the assignable spectrum masses. For all isolates included in the study observed mass shifts in comparison to the spectrum of *C. difficile* reference strain 630 (DSM 27543) were noted and the allelic isoforms assigned by comparing observed mass shifts with the established isoform list. A proteotyping-based phyloproteomic tree was calculated from concatenated biomarker amino acid sequences (as required by the MEGA X software) and compared to the respective MLST data constructed in an analogous fashion.

Identification of reference biomarker ions

The initial step of the proteotyping workflow was the measurement of *C. difficile* reference strain 630 (DSM 27543). The reproducibility of the MALDI-TOF mass spectra was sufficiently high. The standard deviation (based on six measurements) ranged from 0.231 (S21-M) to 0.931 (L36). The difference between the measured average mass and the calculated average mass ranged from 0.05 Da (L33) to 1.00 Da (S21-M) (Supplementary Table 3).

Subsequently, the different MS biomarker ions were ascribed to gene products deduced from the genome sequence corresponding to the measured mass taking into account potential post-translational modifications (Fig. 2, Supplementary Table 4). In total nine singly charged masses of biomarkers were observed in between *m/z* = 4,200 and 9,700 and matched to a specific gene with less than 1.0 Da mass tolerance. The following biomarkers have been identified: RpmJ (L36; 4,277 Da), RpmH (L34; 5,566 Da), RpmG (L33; 5,959 Da), RpmF (L32-M; 6,366 Da), RpmB (L28-M; 6,648 Da), RpmD (L30-M; 6,722 Da), RpsU (S21-M; 6,888 Da), RmpI (L35-M; 7,074 Da), and RpsT (S20; 9,651 Da). As indicated, a posttranslational cleavage of the *N*-terminal methionine has been observed in the case of RpmF/L32-M, RpmB/L28-M, RpmD/L30-M, RpsU/S21-M, RmpI/L35-M and RpsT/S20-M (Supplementary Table 4).

Establishment of an *in silico* allelic isoform database

With the help of 1,312 *C. difficile* sequences deposited in the NCBI database at the time of analysis (June 26th, 2018) we were able to compile a comprehensive list of allelic isoforms for all biomarker ions belonging to the *C. difficile*-specific proteotyping scheme.

Gene sequences deposited for the biomarker isoforms were translated into the respective amino acid sequence and aligned followed by calculation of the protein mass for each individual isoform. The maximum number of biomarker isoforms obtained from the database was 7 for L35-M and S20-M, the minimum number was 3 for L36 and L28-M. Occurrence frequency varied from >99% to a single occurrence of the isoform (Supplementary Table 2). In case of a single occurrence of an isoform, sequencing errors on the submitter's side cannot be ruled out. Ignoring all isoforms occurring only 1-3 times in the database, the *C. difficile* proteotyping scheme was mainly based on isoforms of L28-M (two main isoforms), L35-M (four main isoforms), and S20-M (two main isoforms) giving rise to at least 16 proteotyping-derived types. Potentially, there are significantly more proteotyping-derived types to be expected in the population.

Mass shifts and allelic isoforms in test isolate collection

Initially, the *C. difficile* reference strain 630 (DSM 27543) was analyzed by MALDI-TOF MS. In the study, all mass shift measurements were done with reference to this strain. To identify allelic isoforms the mass shift was compared with the list containing all amino acid sequences. For biomarker RmpI (L35-M) we detected 3 isoforms (7,074.6 Da; 7,090.6 Da; 7,047.5 Da) in the tested isolate cohort, and for biomarker RpmB (L28-M) two isoforms (6,647.8 Da; 6,705.8 Da). RpmJ (L36; 4,277.3 Da), RpmH (L34; 5,565.5 Da), RpmG (L33; 5,959.0 Da), RpmF (L32-M; 6,366.4 Da), RpmD (L30-M; 6,722.9 Da) RpsU (S21-M; 6,889.0 Da), and RpsT (S20-M; 9,651.3 Da) were invariable in the tested isolate cohort (Fig. 3), and the biomarker masses corresponded to the respective reference isoforms. Nucleotide and amino acid sequences of the allelic isoforms of biomarkers newly described during the study have been deposited at GenBank. The accession numbers of all biomarkers (nucleotide and amino acid sequences) are listed in Supplementary Table 5.

Phyloproteomic analysis

Following the principle of MLST to cluster DNA sequences, the biomarker amino acid sequences of each isolate were concatenated and used to deduce phylogeny by UPGMA method (conventional clustering algorithm). The combination of amino acid sequences resulted in five different proteotyping-derived types/clades (Fig. 4, right dendrogram), here the clades were designated with A-E to prevent confusion with MLST clades, which served as the main comparator (Fig. 4, left dendrogram).

The largest proteotyping-derived clade A contained the majority of isolates of MLST clades 1 and 4, while the second largest proteotyping-derived clade D combines isolates of MLST clades 1 and 2.

The smaller proteotyping-derived clades allowed discrimination of more distinctive isolate groups: clade C was formed by all tested *C. difficile* isolates of MLST clade 3 (corresponding to RTs 023 and 127) and MLST clade 5 (corresponding to RTs 078 and 126). Clade E exclusively contained a subgroup of MLST clade 4 isolates, namely RTs 243 and 254. The most interesting finding was that isolates of the highly pathogenic *C. difficile* RT027 formed a unique proteotyping-derived clade (clade B, indicated in red, Fig. 3). Since only three isolates belonged to RT027 in the initial test population, which consists of isolates for which a complete genome sequence was available (data not shown), we subsequently analyzed 17 further RT027 isolates. The proteotyping-derived type of the RT027 isolates (B) results from the biomarker RpmB/L28-M isoform no. 2 (6,705.8 Da), which corresponds to the amino acid substitution G9D when compared to the *C. difficile* 630 (DSM 27543) reference isoform, while the biomarker RpmI/L35-M isoform no. 1 (7,074.6 Da) is identical to the *C. difficile* 630 (DSM 27543) reference isoform. Sanger sequencing of the gene loci of RpmB/L28-M and RpmI/L35-M confirmed that all 17 additionally tested isolates also carried the constellation of the RT027 typical isoforms for L28-M and L35-M. This underpins the fact that proteotyping can be used to unambiguously discriminate isolates of RT027.

DISCUSSION

In this study, the proteotyping technique previously established for *C. jejuni* (Zautner et al., 2015, 2016) was successfully adapted to *C. difficile.*

The current *C. difficile* proteotyping scheme is based on nine biomarkers, which are exclusively ribosomal proteins. In contrast, the *C. jejuni* proteotyping scheme comprised 19 biomarkers, one being a non-ribosomal protein. The smaller number of detectable biomarkers might be explained by the fact that *C. difficile*, in contrast to *C. jejuni*, is a Gram-positive bacterium and that the Gram-positive cell wall makes it more difficult to release proteins from the cell.

Patterns of posttranslational modifications such as the cleavage of *N*-terminal methionine have been shown to be specific for a microbial species (Fagerquist et al., 2006). Six of 9 biomarkers in the *C. difficile* mass spectrum showed a cropped methionine, while only 6 of 19 biomarkers with a cropped methionine were detectable in *C. jejuni* (Zautner et al., 2015). This form of posttranslational modification appears thus to be more frequent in the detectable *C. difficile* proteotyping biomarkers than in *C. jejuni*. *N*-terminal methionine is cleaved by the ubiquitous and essential methionine aminopeptidase MAP (Frottin et al., 2006). The *N*-terminal methionine is often removed when the residue at the second position (P1′) in the primary sequence is small and uncharged, i.e. if at position P1' there is an alanine (A), cysteine (C), glycine (G), proline (P), serine (S), threonine (T), or valine (V). In accordance with this information the biomarkers L32-M (P1’ = A), L28-M (P1’ = A), L30-M (P1’ = A), S21-M (P1’ = S), L35-M (P1’ = P), and S20-M (P1’ = A) are de-methioninated, and the *N*-terminal methionine of L36 (P1’ = K) and L33 (P1’ = R) remains attached. An exception to the aforementioned is L34, which is not de-methioninated although there is a serine at position P1'. It should be noted that the L34 isoforms 3 and 4 have a lysine at position P1' due to a deletion at position 2.

In our isolate cohort only two biomarkers, L28-M and L35-M, showed mass shifts. The proteotyping-derived phyloproteomic tree (Fig. 4) is therefore deduced only from the combination of the two detectable isoforms for L28-M and the three detectable isoforms for L35-M. Of the six (2 x 3) possible combinations of these biomarker isoforms, five combinations or proteotyping-derived types, or clades, were present in the tested isolate cohort. According to our genome analysis, considerably more combinations can be expected in the *C. difficile* population.

Especially with the isoforms of the biomarker S20-M seen in the database analysis, 16 or more proteotyping-derived clades can be expected.

Our most relevant finding was the possibility to differentiate between the highly virulent RT027 *C. difficile* isolates and non-RT027 isolates using this proteotyping scheme. While some of our results on the identification of *C. difficile* RT027 strains by MALDI-TOF MS are comparable to those of other studies, most importantly to the one of Reil and coworkers (Reil et al., 2011), also significant differences were detected: We identified a biomarker L28-M isoform lacking *N*-terminal methionine with an average mass of 6,705.8 Da (L28-M isoform no. 2), whereas Reil *et al*. identified a mass signal at 6,707 Da to be specific for *C. difficile* RT027. The small mass difference (2 Da) of this mass signal in comparison to the one seen in our study may well be attributed due to a difference in calibration. However, Reil and coworkers did not perform further analysis on the gene encoding for the protein indicated by this mass signal, precluding the final confirmation of its identity. Another crucial difference in our study was that we could also demonstrate the corresponding L28-M isoform 1 (6,647.8 Da) to be present in all non-RT027 strains. This finally enabled us to securely differentiate these highly virulent strains from others.

For biomarker L35-M we observed an isoform with an average mass of 7,090.6 Da, likely also shifted by 2 Da in the study by Reil and coworkers (at *m/z* = 7,092 when analyzing RT027 strains). However, they did not consider it to be relevant for differentiation of RT027 from other ribotypes. In our study we found two more L35-M isoforms at 7,074.6 Da and 7,047.5 Da, for which there were no corresponding mass signals in the study of Reil and coworkers. Reil and coworkers also recognized specific markers not only for RT027 but also for the ribotypes RT001 and RT078/126.

Others have also shown the possibility to discriminate between MLST ST37 strains and non-ST37 strains by observing the distribution of two major mass signals at *m/z* = 3,242 and *m/z* = 3,286, respectively (Li et al., 2018). ST37, which mainly corresponds to ribotype 017, has been a dominant strain in adult *C. difficile* infections in China (Gu et al., 2015; Jin et al., 2016). Indeed, we were able to detect the mass with *m/z* = 3,242 (Fig. 2, the mass signal is indicated by “Li”), but unfortunately, we were not able to assign it to a gene, so this mass was excluded from the proteotyping scheme.

A further study demonstrated that, using a 3-peak pair cluster analysis (*m/z* = K1: 35,38.0/3,545.8; K2: 6,577.9/6,592.8; K3: 7,075.6/7,091.1 Da), it is feasible to detect binary toxin producers of *C. difficile* (Kuo et al., 2015). Mass signals for all three biomarkers were detectable in our recordings (Fig. 2). While K2 could not be assigned to any gene and was therefore not included in the proteotyping scheme, K3 corresponded to biomarker L35-M, isoforms no. 1 and 2. K1 likely represented the $M+2H^+$ form (doubly charged biomarker mass) of K3, and was not included. Since the *C. difficile* strain 630 (DSM 27543) is *tcdA*$^+$, *tcdB*$^+$, *cdtA*$^-$, and *cdtB*$^-$ (Dannheim et al., 2017; Sebaihia et al., 2006), this is in line with the findings of Kuo and his team (Kuo et al., 2015).

The work of Li *et al*. and Kuo *et al*. indicates that the potential of proteotyping may probably be even higher, if it were possible to assign further genes to masses of unknown identity.

Cheng and coworkers were able to distinguish *C. difficile* clade 4 isolates from other *C. difficile* isolates based on five different biomarkers (Cheng et al., 2018). A PCA-algorithm was established on the base of mass spectra of 135 isolates. Subsequently 25 isolates were used for the validation of the model. The isolates used in the study covered only clades 1,3 and 4 of the 8 known *C. difficile* clades.

In comparison to the approach of Cheng and coworkers our proteotyping approach is based on biomarkers of known origin. In our approach the phylogeny is deduced by UPGMA method, not by PCA. Previous studies have shown that PCA results depend on culture conditions as well as time of measurement as it also considers the intensity of peaks. As proteotyping results are not dependent on these factors, they are more reliable (Zautner et al., 2015).

Another relevant study in the context of our study was recently published by Corver and coworkers (Corver et al., 2018). They performed ultrahigh-resolution MALDI-FTICR-MS and identified two mass peaks (*m/z* = 4,927.81 and *m/z* = 5,001.84, a mass change of 74 Da, corresponding to a transition between a single Glycin and Methionine) that allow differentiation of MLST types 1 and 11. The sensitivity and specificity was determined based on the analysis of *C. difficile* sequences in the NCBI database. Both mass peaks could be assigned to two different isoforms of an uncharacterized protein. According to a BLAST-search the peptide could be a fragment of the protein CDIF630_01208. Thus, this protein is a potential candidate for the extension of the proteotyping scheme. However, the detection of this biomarker requires the ultrahigh-resolution MALDI-FTICR-MS, a technique not available in diagnostic-microbiological routine. Thus, this biomarker is currently not ready for integration in our proteotyping scheme.

Last, a method designated high molecular weight (HMW) typing has been demonstrated to allow *C. difficile* typing. In this method, a protein profile in the range between 30 and 50 kDa is analyzed. Although the method was less discriminatory than PCR-ribotyping, results have been obtained fast, simple and cost-effective (Ortega et al., 2018; Rizzardi and Åkerlund, 2015). For this method, too, it must be acknowledged that special mass spectrometric equipment must be available which goes beyond the current standards of routine diagnostics.

In comparison to sequence-based methods such as MLST or MLVA (Multiple loci variable-number tandem repeat analysis), the discriminatory depth of proteotyping is limited. Compared to whole genome sequence based methods, it is not feasible to show the clonality of isolates by proteotyping. Therefore, the practicability of proteotyping depends on the specific epidemiological question. To improve discriminatory capacity of our method, future studies should focus on the identification of additional biomarkers and the assignment to the respective gene loci. Furthermore, it should be aimed to extend the recordable mass spectrum. The development of a user-friendly bioinformatic solution and implementation into the standard software of the manufacturer could further facilitate the application of the technique in daily routine diagnostics.

CONCLUSION

The crucial difference between proteotyping and other MALDI-TOF MS-based techniques is that in case of proteotyping the protein isoform behind the peaks is known. Where other methods consider presence or absence of single masses as well as the peak intensity, in proteotyping differentiation is achieved on the basis of an exclusive combination of known, and ideally genetically verified, biomarker masses. This is achieved by genomic analysis of genes coding for ribosomal proteins of a species. Mutations resulting in changes of the amino acid sequences result in peak shifts in the MALDI-TOF spectra.

Our study shows that our formal *C. difficile*-specific 9 biomarker proteotyping scheme is sufficiently discriminatory to differentiate between RT027 strains and non-RT027 strains. While both methods are not congruent, the discriminatory depth of *C. difficile* proteotyping corresponds at least to the MLST clade classification but is potentially also higher. More genome sequences resulting in more isoforms and therewith more proteotyping-derived types would improve the discriminatory depth of the method significantly. In addition to the number of isoforms, more precise mass spectrometric methods can also be used to increase the number of biomarkers and thus the discriminatory depth.

Since immediate responses are highly important in case of disease outbreaks (mainly corresponding to RT027), our method offers a fast, accurate and inexpensive initial diagnostic tool that can provide indications of RT027 outbreaks.

AUTHOR CONTRIBUTIONS

MFE performed data analysis, PCR and Sanger sequencing, and wrote the manuscript, FMJ created the isoform database, TR and JO performed sequencing, MR performed ribotyping, FL performed bacterial culture and recorded the mass spectra, OZ, PC, and RLK isolated, collected and identified all *C. difficile* isolates, WB performed WGS sequence analysis, UG, OB, and AEZ conceived and designed the experiments, performed data analysis, and wrote the manuscript including figures. AEZ performed MLST analysis and calculated the taxonomic dendrograms. All authors have proofread the manuscript and agreed on publication.

FUNDING

This work was funded by the Federal State of Lower Saxony, Niedersächsisches Vorab (VWZN2889/3215/3266). The Open Access Support Program of the Deutsche Forschungsgemeinschaft and the publication fund of the Georg-August-Universität Göttingen funded publication of this paper.

ACKNOWLEDGEMENTS

We thank Boyke Bunk, Cathrin Spröer, Simone Severitt, Nicole Heyer and Carolin Pilke at the Leibniz Institute DSMZ for helpful advices and technical assistance.

ETHICAL APPROVAL

Ethical clearance for the analysis was obtained from Ethics Committee of the University Medical Center Göttingen, Germany. No humans, animals, or personalized data were used for this study.

References

Abt, M. C., McKenney, P. T., and Pamer, E. G. (2016). *Clostridium difficile* colitis: pathogenesis and host defence. *Nat. Rev. Microbiol.* 14, 609–620. doi:10.1038/nrmicro.2016.108.

Arvand, M., and Bettge-Weller, G. (2016). *Clostridium difficile* ribotype 027 is not evenly distributed in Hesse, Germany. *Anaerobe* 40, 1–4. doi:10.1016/j.anaerobe.2016.04.006.

Arvand, M., Vollandt, D., Bettge-Weller, G., Harmanus, C., Kuijper, E. J., and Clostridium difficile study group Hesse (2014). Increased incidence of *Clostridium difficile* PCR ribotype 027 in Hesse, Germany, 2011 to 2013. *Euro Surveill. Bull. Eur. Sur Mal. Transm. Eur. Commun. Dis. Bull.* 19.

Awad, M. M., Johanesen, P. A., Carter, G. P., Rose, E., and Lyras, D. (2014). *Clostridium difficile* virulence factors: insights into an anaerobic spore-forming pathogen. *Gut Microbes* 5, 579–593.

Bader, O. (2013). MALDI-TOF-MS-based species identification and typing approaches in medical mycology. *Proteomics* 13, 788–99. doi:10.1002/pmic.201200468.

Bakker, D., Smits, W. K., Kuijper, E. J., and Corver, J. (2012). TcdC does not significantly repress toxin expression in *Clostridium difficile* 630ΔErm. *PloS One* 7, e43247. doi:10.1371/journal.pone.0043247.

Barbut, F., Decre, D., Lalande, V., Burghoffer, B., Noussair, L., Gigandon, A., et al. (2005). Clinical features of *Clostridium difficile*-associated diarrhoea due to binary toxin (actin-specific ADP-ribosyltransferase)-producing strains. *J. Med. Microbiol.* 54, 181–185.

Berger, F. K., Rasheed, S. S., Araj, G. F., Mahfouz, R., Rimmani, H. H., Karaoui, W. R., et al. (2018). Molecular characterization, toxin detection and resistance testing of human clinical *Clostridium difficile* isolates from Lebanon. *Int. J. Med. Microbiol.* 308, 358–363.

Brazier, J. S., Raybould, R., Patel, B., Duckworth, G., Pearson, A., Charlett, A., et al. (2008). Distribution and antimicrobial susceptibility patterns of

Clostridium difficile PCR ribotypes in English hospitals, 2007-08. *Eurosurveillance* 13, 19000.

Burns, D. A., Heap, J. T., and Minton, N. P. (2010). The diverse sporulation characteristics of *Clostridium difficile* clinical isolates are not associated with type. *Anaerobe* 16, 618–622.

Byl, B., Jacobs, F., Struelens, M. J., and Thys, J.-P. (1996). Extraintestinal *Clostridium difficile* infections. *Clin. Infect. Dis.* 22, 712–712.

Carter, G. P., Chakravorty, A., Nguyen, T. A. P., Mileto, S., Schreiber, F., Li, L., et al. (2015). Defining the roles of TcdA and TcdB in localized gastrointestinal disease, systemic organ damage, and the host response during *Clostridium difficile* infections. *MBio* 6, e00551–15.

Cartman, S. T., Kelly, M. L., Heeg, D., Heap, J. T., and Minton, N. P. (2012). Precise manipulation of the *Clostridium difficile* chromosome reveals a lack of association between the *tcdC* genotype and toxin production. *Appl. Environ. Microbiol.* 78, 4683–4690. doi:10.1128/AEM.00249-12.

Cassini, A., Plachouras, D., Eckmanns, T., Sin, M. A., Blank, H.-P., Ducomble, T., et al. (2016). Burden of six healthcare-associated infections on European population health: estimating incidence-based disability-adjusted life years through a population prevalence-based modelling study. *PLoS Med.* 13, e1002150.

Cheng, J.-W., Liu, C., Kudinha, T., Xiao, M., Yu, S.-Y., Yang, C.-X., et al. (2018). Use of matrix-assisted laser desorption ionization-time of flight mass spectrometry to identify MLST clade 4 *Clostridium difficile* isolates. *Diagn. Microbiol. Infect. Dis.* 92, 19–24. doi:10.1016/j.diagmicrobio.2018.04.011.

Collins, J., Robinson, C., Danhof, H., Knetsch, C. W., van Leeuwen, H. C., Lawley, T. D., et al. (2018). Dietary trehalose enhances virulence of epidemic *Clostridium difficile*. *Nature*. doi:10.1038/nature25178.

Corver, J., Sen, J., Hornung, B. V. H., Mertens, B. J., Berssenbrugge, E. K. L., Harmanus, C., et al. (2018). Identification and validation of two peptide markers for the recognition of *Clostridioides difficile* MLST-1 and MLST-11 by MALDI-MS. *Clin. Microbiol. Infect.* doi:10.1016/j.cmi.2018.10.008.

Curry, S. R., Marsh, J. W., Muto, C. A., O'Leary, M. M., Pasculle, A. W., and Harrison, L. H. (2007). *tcdC* genotypes associated with severe TcdC truncation in an epidemic clone and other strains of *Clostridium difficile*. *J. Clin. Microbiol.* 45, 215–221. doi:10.1128/JCM.01599-06.

Dannheim, H., Riedel, T., Neumann-Schaal, M., Bunk, B., Schober, I., Spröer, C., et al. (2017). Manual curation and reannotation of the genomes of *Clostridium difficile* 630Δ*erm* and *C. difficile* 630. *J. Med. Microbiol.* 66, 286–293. doi:10.1099/jmm.0.000427.

Dawson, L. F., Stabler, R. A., and Wren, B. W. (2008). Assessing the role of *p*-cresol tolerance in *Clostridium difficile*. *J. Med. Microbiol.* 57, 745–749. doi:10.1099/jmm.0.47744-0.

Dingle, K. E., Elliott, B., Robinson, E., Griffiths, D., Eyre, D. W., Stoesser, N., et al. (2014). Evolutionary history of the *Clostridium difficile* pathogenicity locus. *Genome Biol. Evol.* 6, 36–52. doi:10.1093/gbe/evt204.

Donskey, C. J., Kundrapu, S., and Deshpande, A. (2015). Colonization versus carriage of *Clostridium difficile*. *Infect. Dis. Clin.* 29, 13–28.

Drudy, D., Kyne, L., O'Mahony, R., and Fanning, S. (2007). *gyrA* mutations in fluoroquinolone-resistant *Clostridium difficile* PCR-027. *Emerg. Infect. Dis.* 13, 504.

Durighello, E., Bellanger, L., Ezan, E., and Armengaud, J. (2014). Proteogenomic biomarkers for identification of *Francisella* species and subspecies by matrix-assisted laser desorption ionization-time-of-flight mass spectrometry. *Anal. Chem.* 86, 9394–9398.

Elliott, B., Androga, G. O., Knight, D. R., and Riley, T. V. (2017). *Clostridium difficile* infection: Evolution, phylogeny and molecular epidemiology. *Infect. Genet. Evol.* 49, 1–11.

Fagerquist, C. K., Bates, A. H., Heath, S., King, B. C., Garbus, B. R., Harden, L. A., et al. (2006). Sub-speciating *Campylobacter jejuni* by proteomic analysis of its protein biomarkers and their post-translational modifications. *J Proteome Res* 5, 2527–38. doi:10.1021/pr050485w.

Fawley, W. N., Davies, K. A., Morris, T., Parnell, P., Howe, R., and Wilcox, M. H. (2016). Enhanced surveillance of *Clostridium difficile* infection occurring outside hospital, England, 2011 to 2013. *Eurosurveillance* 21, 30295.

Fawley, W. N., Knetsch, C. W., MacCannell, D. R., Harmanus, C., Du, T., Mulvey, M. R., et al. (2015). Development and validation of an internationally-standardized, high-resolution capillary gel-based electrophoresis PCR-ribotyping protocol for *Clostridium difficile*. *PloS One* 10, e0118150. doi:10.1371/journal.pone.0118150.

Frottin, F., Martinez, A., Peynot, P., Mitra, S., Holz, R. C., Giglione, C., et al. (2006). The Proteomics of N-terminal Methionine Cleavage. *Mol. Cell. Proteomics* 5, 2336–2349. doi:10.1074/mcp.M600225-MCP200.

Gerding, D. N., Johnson, S., Rupnik, M., and Aktories, K. (2014). *Clostridium difficile* binary toxin CDT. *Gut Microbes* 5, 15–27. doi:10.4161/gmic.26854.

Gonzales, T., and Robert-Baudouy, J. (1996). Bacterial aminopeptidases: properties and functions. *FEMS Microbiol Rev* 18, 319–44.

Govind, R., and Dupuy, B. (2012). Secretion of *Clostridium difficile* toxins A and B requires the holin-like protein TcdE. *PLoS Pathog.* 8, e1002727. doi:10.1371/journal.ppat.1002727.

Govind, R., Fitzwater, L., and Nichols, R. (2015). Observations on the Role of TcdE Isoforms in *Clostridium difficile* Toxin Secretion. *J. Bacteriol.* 197, 2600–2609. doi:10.1128/JB.00224-15.

Gu, S.-L., Chen, Y.-B., Lv, T., Zhang, X., Wei, Z.-Q., Shen, P., et al. (2015). Risk factors, outcomes and epidemiology associated with *Clostridium difficile* infection in patients with haematological malignancies in a tertiary care hospital in China. *J. Med. Microbiol.* 64, 209–216.

Gürtler, V. (1993). Typing of *Clostridium difficile* strains by PCR-amplification of variable length 16S-23S rDNA spacer regions. *Microbiology* 139, 3089–3097.

He, M., Miyajima, F., Roberts, P., Ellison, L., Pickard, D. J., Martin, M. J., et al. (2013). Emergence and global spread of epidemic healthcare-associated *Clostridium difficile*. *Nat. Genet.* 45, 109.

Hubert, B., Loo, V. G., Bourgault, A.-M., Poirier, L., Dascal, A., Fortin, É., et al. (2007). A portrait of the geographic dissemination of the *Clostridium difficile* North American pulsed-field type 1 strain and the epidemiology of *C. difficile*-associated disease in Quebec. *Clin. Infect. Dis.* 44, 238–244.

Hunt, J. J., and Ballard, J. D. (2013). Variations in virulence and molecular biology among emerging strains of *Clostridium difficile*. *Microbiol. Mol. Biol. Rev.* 77, 567–581.

Indra, A., Huhulescu, S., Schneeweis, M., Hasenberger, P., Kernbichler, S., Fiedler, A., et al. (2008). Characterization of *Clostridium difficile* isolates using capillary gel electrophoresis-based PCR ribotyping. *J. Med. Microbiol.* 57, 1377–1382.

Jacobs, A., Barnard, K., Fishel, R., and Gradon, J. D. (2001). Extracolonic manifestations of *Clostridium difficile* infections: presentation of 2 cases and review of the literature. *Medicine (Baltimore)* 80, 88–101.

Janezic, S. (2016). Direct PCR-Ribotyping of *Clostridium difficile*. *Methods Mol. Biol. Clifton NJ* 1476, 15–21. doi:10.1007/978-1-4939-6361-4_2.

Janezic, S., and Rupnik, M. (2010). Molecular typing methods for *Clostridium difficile*: pulsed-field gel electrophoresis and PCR ribotyping. *Methods Mol. Biol. Clifton NJ* 646, 55–65. doi:10.1007/978-1-60327-365-7_4.

Jin, D., Luo, Y., Huang, C., Cai, J., Ye, J., Zheng, Y., et al. (2016). Molecular epidemiology of *Clostridium difficile* infection in hospitalized patients in eastern China. *J. Clin. Microbiol.*, JCM–01898.

Karlsson, R., Gonzales-Siles, L., Boulund, F., Svensson-Stadler, L., Skovbjerg, S., Karlsson, A., et al. (2015). Proteotyping: Proteomic characterization, classification and identification of microorganisms--A prospectus. *Syst. Appl. Microbiol.* 38, 246–257. doi:10.1016/j.syapm.2015.03.006.

Keel, M. K., and Songer, J. G. (2006). The comparative pathology of *Clostridium difficile*-associated disease. *Vet. Pathol.* 43, 225–240.

Kentache, T., Jouenne, T., Dé, E., and Hardouin, J. (2016). Proteomic characterization of *Nα*- and *Nε*-acetylation in *Acinetobacter baumannii*. *J. Proteomics* 144, 148–158. doi:10.1016/j.jprot.2016.05.021.

Knetsch, C. W., Kumar, N., Forster, S. C., Connor, T. R., Browne, H. P., Harmanus, C., et al. (2018). Zoonotic Transfer of *Clostridium difficile* Harboring Antimicrobial Resistance between Farm Animals and Humans. *J. Clin. Microbiol.* 56. doi:10.1128/JCM.01384-17.

Knetsch, C. W., Lawley, T. D., Hensgens, M. P., Corver, J., Wilcox, M. W., and Kuijper, E. J. (2013). Current application and future perspectives of molecular typing methods to study *Clostridium difficile* infections. *Euro Surveill. Bull. Eur. Sur Mal. Transm. Eur. Commun. Dis. Bull.* 18, 20381.

Knetsch, C. W., Terveer, E. M., Lauber, C., Gorbalenya, A. E., Harmanus, C., Kuijper, E. J., et al. (2012). Comparative analysis of an expanded *Clostridium difficile* reference strain collection reveals genetic diversity and evolution through six lineages. *Infect. Genet. Evol. J. Mol. Epidemiol. Evol. Genet. Infect. Dis.* 12, 1577–1585. doi:10.1016/j.meegid.2012.06.003.

Knight, D. R., Elliott, B., Chang, B. J., Perkins, T. T., and Riley, T. V. (2015). Diversity and evolution in the genome of *Clostridium difficile*. *Clin. Microbiol. Rev.* 28, 721–741.

Kuhns, M., Zautner, A. E., Rabsch, W., Zimmermann, O., Weig, M., Bader, O., et al. (2012). Rapid discrimination of *Salmonella enterica* serovar Typhi from other serovars by MALDI-TOF mass spectrometry. *PLoS One* 7, e40004. doi:10.1371/journal.pone.0040004.

Kumar, S., Stecher, G., Li, M., Knyaz, C., and Tamura, K. (2018). MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. *Mol. Biol. Evol.* 35, 1547–1549. doi:10.1093/molbev/msy096.

Kuo, S.-F., Wu, T.-L., You, H.-L., Chien, C.-C., Chia, J.-H., and Lee, C.-H. (2015). Accurate detection of binary toxin producer from *Clostridium difficile* by matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. *Diagn. Microbiol. Infect. Dis.* 83, 229–231. doi:10.1016/j.diagmicrobio.2015.07.013.

Lanis, J. M., Barua, S., and Ballard, J. D. (2010). Variations in TcdB activity and the hypervirulence of emerging strains of *Clostridium difficile*. *PLoS Pathog.* 6, e1001061.

Lartigue, M. F. (2013). Matrix-assisted laser desorption ionization time-of-flight mass spectrometry for bacterial strain characterization. *Infect Genet Evol* 13, 230–5. doi:10.1016/j.meegid.2012.10.012.

Lavigne, J.-P., Espinal, P., Dunyach-Remy, C., Messad, N., Pantel, A., and Sotto, A. (2013). Mass spectrometry: a revolution in clinical microbiology? *Clin. Chem. Lab. Med.* 51, 257–270.

Lawson, P. A., Citron, D. M., Tyrrell, K. L., and Finegold, S. M. (2016). Reclassification of *Clostridium difficile* as *Clostridioides difficile* (Hall and O'Toole 1935) Prévot 1938. *Anaerobe* 40, 95–99. doi:10.1016/j.anaerobe.2016.06.008.

Leffler, D. A., and Lamont, J. T. (2015). *Clostridium difficile* infection. *N. Engl. J. Med.* 372, 1539–1548. doi:10.1056/NEJMra1403772.

Li, R., Xiao, D., Yang, J., Sun, S., Kaplan, S., Li, Z., et al. (2018). Identification and Characterization of *Clostridium difficile* Sequence Type 37 Genotype by Matrix-Assisted Laser Desorption Ionization-Time of Flight Mass Spectrometry. *J. Clin. Microbiol.* 56. doi:10.1128/JCM.01990-17.

Lo Vecchio, A., and Zacur, G. M. (2012). *Clostridium difficile* infection: an update on epidemiology, risk factors, and therapeutic options. *Curr. Opin. Gastroenterol.* 28, 1–9.

Loo, V. G., Poirier, L., Miller, M. A., Oughton, M., Libman, M. D., Michaud, S., et al. (2005). A predominantly clonal multi-institutional outbreak of *Clostridium difficile*–associated diarrhea with high morbidity and mortality. *N. Engl. J. Med.* 353, 2442–2449.

Maiden, M. C., Bygraves, J. A., Feil, E., Morelli, G., Russell, J. E., Urwin, R., et al. (1998). Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Proc Natl Acad Sci U A* 95, 3140–5.

Mani, N., and Dupuy, B. (2001). Regulation of toxin synthesis in *Clostridium difficile* by an alternative RNA polymerase sigma factor. *Proc. Natl. Acad. Sci.* 98, 5844–5849.

Martin, J. S., Monaghan, T. M., and Wilcox, M. H. (2016). *Clostridium difficile* infection: advances in epidemiology, diagnosis and transmission. *Nat. Rev. Gastroenterol. Hepatol.* 13, 206–216.

Matamouros, S., England, P., and Dupuy, B. (2007). *Clostridium difficile* toxin expression is inhibited by the novel regulator TcdC. *Mol. Microbiol.* 64, 1274–1288.

Matsumura, Y., Yamamoto, M., Nagao, M., Tanaka, M., Machida, K., Ito, Y., et al. (2014). Detection of extended-spectrum-beta-lactamase-producing *Escherichia coli* ST131 and ST405 clonal groups by matrix-assisted laser desorption ionization-time of flight mass spectrometry. *J Clin Microbiol* 52, 1034–40. doi:10.1128/JCM.03196-13.

McDonald, L. C., Killgore, G. E., Thompson, A., Owens Jr, R. C., Kazakova, S. V., Sambol, S. P., et al. (2005). An epidemic, toxin gene–variant strain of *Clostridium difficile*. *N. Engl. J. Med.* 353, 2433–2441.

McEllistrem, M. C., Carman, R. J., Gerding, D. N., Genheimer, C. W., and Zheng, L. (2005). A hospital outbreak of *Clostridium difficile* disease associated with isolates carrying binary toxin genes. *Clin. Infect. Dis.* 40, 265–272.

Mehner-Breitfeld, D., Rathmann, C., Riedel, T., Just, I., Gerhard, R., Overmann, J., et al. (2018). Evidence for an Adaptation of a Phage-Derived

Holin/Endolysin System to Toxin Transport in *Clostridioides difficile*. *Front. Microbiol.* 9, 2446. doi:10.3389/fmicb.2018.02446.

Monot, M., Eckert, C., Lemire, A., Hamiot, A., Dubois, T., Tessier, C., et al. (2015). *Clostridium difficile*: new insights into the evolution of the pathogenicity locus. *Sci. Rep.* 5, 15023.

Mooney, H. (2007). Annual incidence of MRSA falls in England, but *C. difficile* continues to rise. *BMJ* 335, 958.

Nanwa, N., Kendzerska, T., Krahn, M., Kwong, J. C., Daneman, N., Witteman, W., et al. (2015). The economic impact of *Clostridium difficile* infection: a systematic review. *Am. J. Gastroenterol.* 110.

Novais, A., Sousa, C., de Dios Caballero, J., Fernandez-Olmos, A., Lopes, J., Ramos, H., et al. (2014). MALDI-TOF mass spectrometry as a tool for the discrimination of high-risk *Escherichia coli* clones from phylogenetic groups B2 (ST131) and D (ST69, ST405, ST393). *Eur J Clin Microbiol Infect Dis*. doi:10.1007/s10096-014-2071-5.

Olling, A., Seehase, S., Minton, N. P., Tatge, H., Schröter, S., Kohlscheen, S., et al. (2012). Release of TcdA and TcdB from *Clostridium difficile* cdi 630 is not affected by functional inactivation of the *tcdE* gene. *Microb. Pathog.* 52, 92–100. doi:10.1016/j.micpath.2011.10.009.

Ortega, L., Ryberg, A., and Johansson, Å. (2018). HMW-profiling using MALDI-TOF MS: A screening method for outbreaks of *Clostridioides difficile*. *Anaerobe* 54, 254–259. doi:10.1016/j.anaerobe.2018.04.013.

Ouidir, T., Jarnier, F., Cosette, P., Jouenne, T., and Hardouin, J. (2015). Characterization of *N*-terminal protein modifications in *Pseudomonas aeruginosa* PA14. *J. Proteomics* 114, 214–225. doi:10.1016/j.jprot.2014.11.006.

Papatheodorou, P., Carette, J. E., Bell, G. W., Schwan, C., Guttenberg, G., Brummelkamp, T. R., et al. (2011). Lipolysis-stimulated lipoprotein receptor (LSR) is the host receptor for the binary toxin *Clostridium difficile* transferase (CDT). *Proc. Natl. Acad. Sci.* 108, 16422–16427. doi:10.1073/pnas.1109772108.

Patel, R. (2015). MALDI-TOF MS for the diagnosis of infectious diseases. *Clin. Chem.* 61, 100–111.

Pavón, A. B. I., and Maiden, M. C. J. (2009). Multilocus sequence typing. *Methods Mol. Biol. Clifton NJ* 551, 129–140. doi:10.1007/978-1-60327-999-4_11.

Pépin, J., Valiquette, L., Alary, M.-E., Villemure, P., Pelletier, A., Forget, K., et al. (2004). *Clostridium difficile*-associated diarrhea in a region of Quebec from 1991 to 2003: a changing pattern of disease severity. *Can. Med. Assoc. J.* 171, 466–472.

Redelings, M. D., Sorvillo, F., and Mascola, L. (2007). Increase in *Clostridium difficile*–related mortality rates, United States, 1999–2004. *Emerg. Infect. Dis.* 13, 1417.

Reil, M., Erhard, M., Kuijper, E. J., Kist, M., Zaiss, H., Witte, W., et al. (2011). Recognition of *Clostridium difficile* PCR-ribotypes 001, 027 and 126/078 using an extended MALDI-TOF MS system. *Eur J Clin Microbiol Infect Dis* 30, 1431–6. doi:10.1007/s10096-011-1238-6.

Riedel, T., Wetzel, D., Hofmann, J. D., Plorin, S. P. E. O., Dannheim, H., Berges, M., et al. (2017). High metabolic versatility of different toxigenic and non-toxigenic *Clostridioides difficile* isolates. *Int. J. Med. Microbiol. IJMM* 307, 311–320. doi:10.1016/j.ijmm.2017.05.007.

Rizzardi, K., and Åkerlund, T. (2015). High molecular weight typing with MALDI-TOF MS-a novel method for rapid typing of *Clostridium difficile*. *Plos ONE* 10, e0122457.

Robinson, C. D., Auchtung, J. M., Collins, J., and Britton, R. A. (2014). Epidemic *Clostridium difficile* strains demonstrate increased competitive fitness compared to nonepidemic isolates. *Infect. Immun.* 82, 2815–2825. doi:10.1128/IAI.01524-14.

Schwan, C., Kruppke, A. S., Nölke, T., Schumacher, L., Koch-Nolte, F., Kudryashev, M., et al. (2014). *Clostridium difficile* toxin CDT hijacks microtubule organization and reroutes vesicle traffic to increase pathogen adherence.

Proc. Natl. Acad. Sci. U. S. A. 111, 2313–2318. doi:10.1073/pnas.1311589111.

Sebaihia, M., Wren, B. W., Mullany, P., Fairweather, N. F., Minton, N., Stabler, R., et al. (2006). The multidrug-resistant human pathogen *Clostridium difficile* has a highly mobile, mosaic genome. *Nat. Genet.* 38, 779–786. doi:10.1038/ng1830.

Seng, P., Rolain, J. M., Fournier, P. E., La Scola, B., Drancourt, M., and Raoult, D. (2010). MALDI-TOF-mass spectrometry applications in clinical microbiology. *Future Microbiol* 5, 1733–54. doi:10.2217/fmb.10.127.

Seugendo, M., Janssen, I., Lang, V., Hasibuan, I., Bohne, W., Cooper, P., et al. (2018). Prevalence and Strain Characterization of *Clostridioides* (*Clostridium*) *difficile* in Representative Regions of Germany, Ghana, Tanzania and Indonesia - A Comparative Multi-Center Cross-Sectional Study. *Front. Microbiol.* 9, 1843. doi:10.3389/fmicb.2018.01843.

Smits, W. K. (2013). Hype or hypervirulence: a reflection on problematic *C. difficile* strains. *Virulence* 4, 592–596.

Smits, W. K., Lyras, D., Lacy, D. B., Wilcox, M. H., and Kuijper, E. J. (2016). *Clostridium difficile* infection. *Nat. Rev. Dis. Primer* 2, 16020. doi:10.1038/nrdp.2016.20.

Spigaglia, P., Barbanti, F., Dionisi, A. M., and Mastrantonio, P. (2010). *Clostridium difficile* isolates resistant to fluoroquinolones in Italy: emergence of PCR ribotype 018. *J. Clin. Microbiol.* 48, 2892–2896. doi:10.1128/JCM.02482-09.

Spigaglia, P., Barbanti, F., Mastrantonio, P., Brazier, J. S., Barbut, F., Delmée, M., et al. (2008). Fluoroquinolone resistance in *Clostridium difficile* isolates from a prospective study of *C. difficile* infections in Europe. *J. Med. Microbiol.* 57, 784–789. doi:10.1099/jmm.0.47738-0.

Spigaglia, P., and Mastrantonio, P. (2002). Molecular analysis of the pathogenicity locus and polymorphism in the putative negative regulator of toxin production (TcdC) among *Clostridium difficile* clinical isolates. *J. Clin. Microbiol.* 40, 3470–3475.

Stabler, R. A., He, M., Dawson, L., Martin, M., Valiente, E., Corton, C., et al. (2009). Comparative genome and phenotypic analysis of *Clostridium difficile* 027 strains provides insight into the evolution of a hypervirulent bacterium. *Genome Biol.* 10, R102.

Stubbs, S., Rupnik, M., Gibert, M., Brazier, J., Duerden, B., and Popoff, M. (2000). Production of actin-specific ADP-ribosyltransferase (binary toxin) by strains of *Clostridium difficile*. *FEMS Microbiol. Lett.* 186, 307–312.

Suarez, S., Ferroni, A., Lotz, A., Jolley, K. A., Guerin, P., Leto, J., et al. (2013). Ribosomal proteins as biomarkers for bacterial identification by mass spectrometry in the clinical microbiology laboratory. *J Microbiol Methods* 94, 390–6. doi:10.1016/j.mimet.2013.07.021.

Sundriyal, A., Roberts, A. K., Ling, R., McGlashan, J., Shone, C. C., and Acharya, K. R. (2010). Expression, purification and cell cytotoxicity of actin-modifying binary toxin from *Clostridium difficile*. *Protein Expr. Purif.* 74, 42–48.

Valiente, E., Cairns, M. D., and Wren, B. W. (2014). The *Clostridium difficile* PCR ribotype 027 lineage: a pathogen on the move. *Clin. Microbiol. Infect.* 20, 396–404. doi:10.1111/1469-0691.12619.

van Dorp, S. M., Kinross, P., Gastmeier, P., Behnke, M., Kola, A., Delmée, M., et al. (2016). Standardised surveillance of *Clostridium difficile* infection in European acute care hospitals: a pilot study, 2013. *Euro Surveill. Bull. Eur. Sur Mal. Transm. Eur. Commun. Dis. Bull.* 21. doi:10.2807/1560-7917.ES.2016.21.29.30293.

van Leeuwen, H. C., Bakker, D., Steindel, P., Kuijper, E. J., and Corver, J. (2013). *Clostridium difficile* TcdC protein binds four-stranded G-quadruplex structures. *Nucleic Acids Res.* 41, 2382–2393. doi:10.1093/nar/gks1448.

Varland, S., Osberg, C., and Arnesen, T. (2015). *N*-terminal modifications of cellular proteins: The enzymes involved, their substrate specificities and biological effects. *Proteomics* 15, 2385–2401.

Warny, M., Pepin, J., Fang, A., Killgore, G., Thompson, A., Brazier, J., et al. (2005). Toxin production by an emerging strain of *Clostridium difficile* associated with outbreaks of severe disease in North America and Europe. *The Lancet* 366, 1079–1084.

Waslawski, S., Lo, E. S., Ewing, S. A., Young, V. B., Aronoff, D. M., Sharp, S. E., et al. (2013). *Clostridium difficile* ribotype diversity at six health care institutions in the United States. *J. Clin. Microbiol.* 51, 1938–1941. doi:10.1128/JCM.00056-13.

Zautner, A. E., Lugert, R., Masanta, W. O., Weig, M., Groß, U., and Bader, O. (2016). Subtyping of *Campylobacter jejuni* ssp. *doylei* Isolates Using Mass Spectrometry-based PhyloProteomics (MSPP). *JoVE J. Vis. Exp.*, e54165–e54165. doi:10.3791/54165.

Zautner, A. E., Masanta, W. O., Tareen, A. M., Weig, M., Lugert, R., Gross, U., et al. (2013). Discrimination of multilocus sequence typing-based *Campylobacter jejuni* subgroups by MALDI-TOF mass spectrometry. *BMC Microbiol* 13, 247. doi:10.1186/1471-2180-13-247.

Zautner, A. E., Masanta, W. O., Weig, M., Groß, U., and Bader, O. (2015). Mass Spectrometry-based PhyloProteomics (MSPP): A novel microbial typing Method. *Sci. Rep.* 5, 13431. doi:10.1038/srep13431.

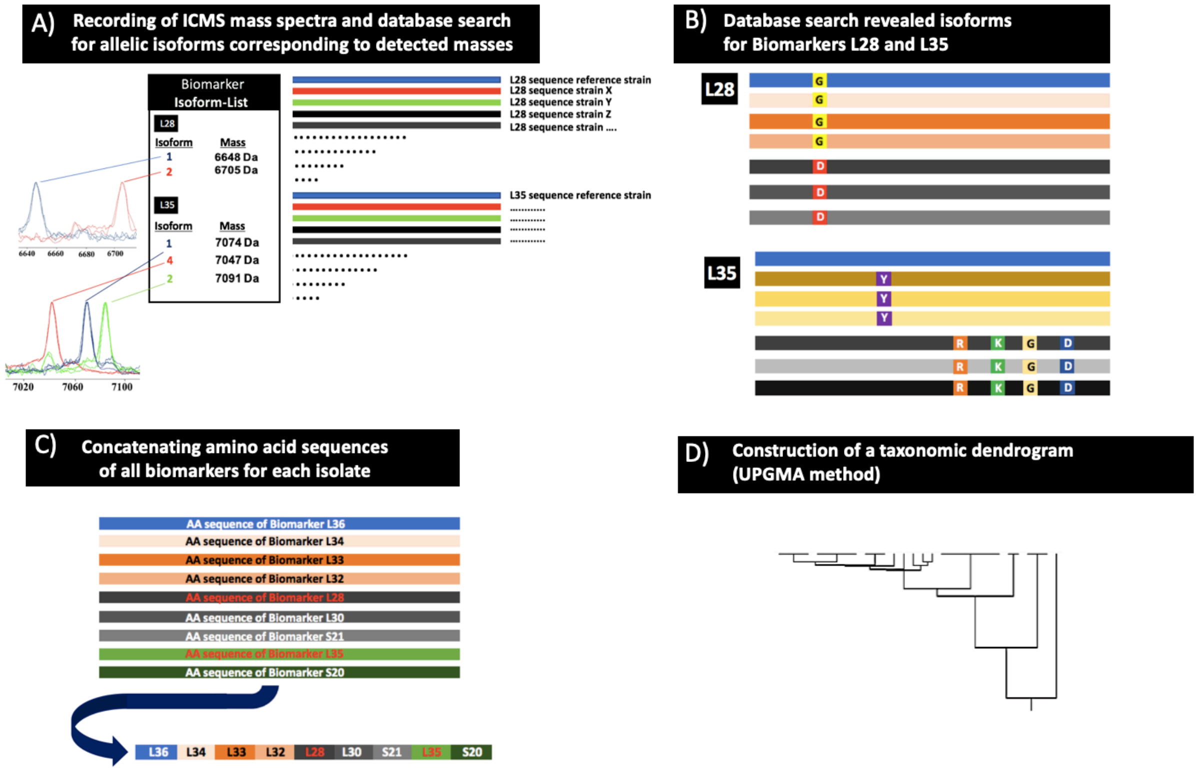

A) Recording of ICMS mass spectra and database search for allelic isoforms corresponding to detected masses
Biomarker Isoform-List
L28
Isoform
Mass
1
2
6648 Da
6705 Da
L35
Isoform
Mass
1
4
2
7074 Da
7047 Da
7091 Da
L28 sequence reference strain
L28 sequence strain X
L28 sequence strain Y
L28 sequence strain Z
L28 sequence strain ....
L35 sequence reference strain
6640
6660
6680
6700
7020
7060
7100
B) Database search revealed isoforms for Biomarkers L28 and L35
L28
G
D
L35
Y
R
K
G
D
C) Concatenating amino acid sequences of all biomarkers for each isolate
AA sequence of Biomarker L36
AA sequence of Biomarker L34
AA sequence of Biomarker L33
AA sequence of Biomarker L32
AA sequence of Biomarker L28
AA sequence of Biomarker L30
AA sequence of Biomarker S21
AA sequence of Biomarker L35
AA sequence of Biomarker S20
L36
L34
L33
L32
L28
L30
S21
L35
S20
D) Construction of a taxonomic dendrogram (UPGMA method)

Fig. 1. Scheme of the proteotyping workflow. A) Recording of MALDI-TOF mass spectra of *C. difficile* isolates (extracts as well as smear preparation) B) Identification of allelic isoforms by comparison with the allelic isoform database that contains the sequence data of the *C. difficile* genomes deposited in public databases. C) Assembly of the concatenated amino acid sequences of the respective isoforms to one continuous sequence D) Calculation of a taxonomic proteotyping-derived UPGMA dendrogram.

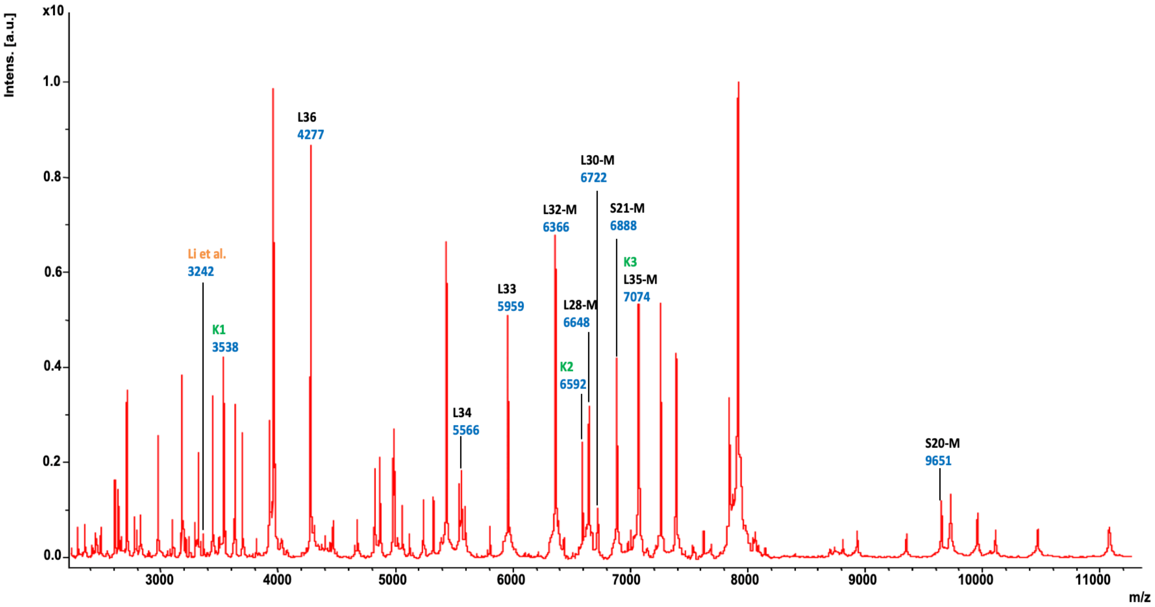


Fig. 2. MALDI-TOF mass spectrum of *C. difficile* reference strain 630 (DSM 27543). In this mass spectrum, all singularly charged biomarkers that were included in the *C. difficile* proteotyping scheme are marked in black and a red arrow; multiply charged ions are not labeled separately. Additionally, the biomarkers used for *C. difficile* subtyping by Li and coworkers (Li et al., 2018) as well as Kuo and coworkers (Kuo et al., 2015) are indicated in orange ("Li") dark and green ("K1", "K2", "K3"), respectively.

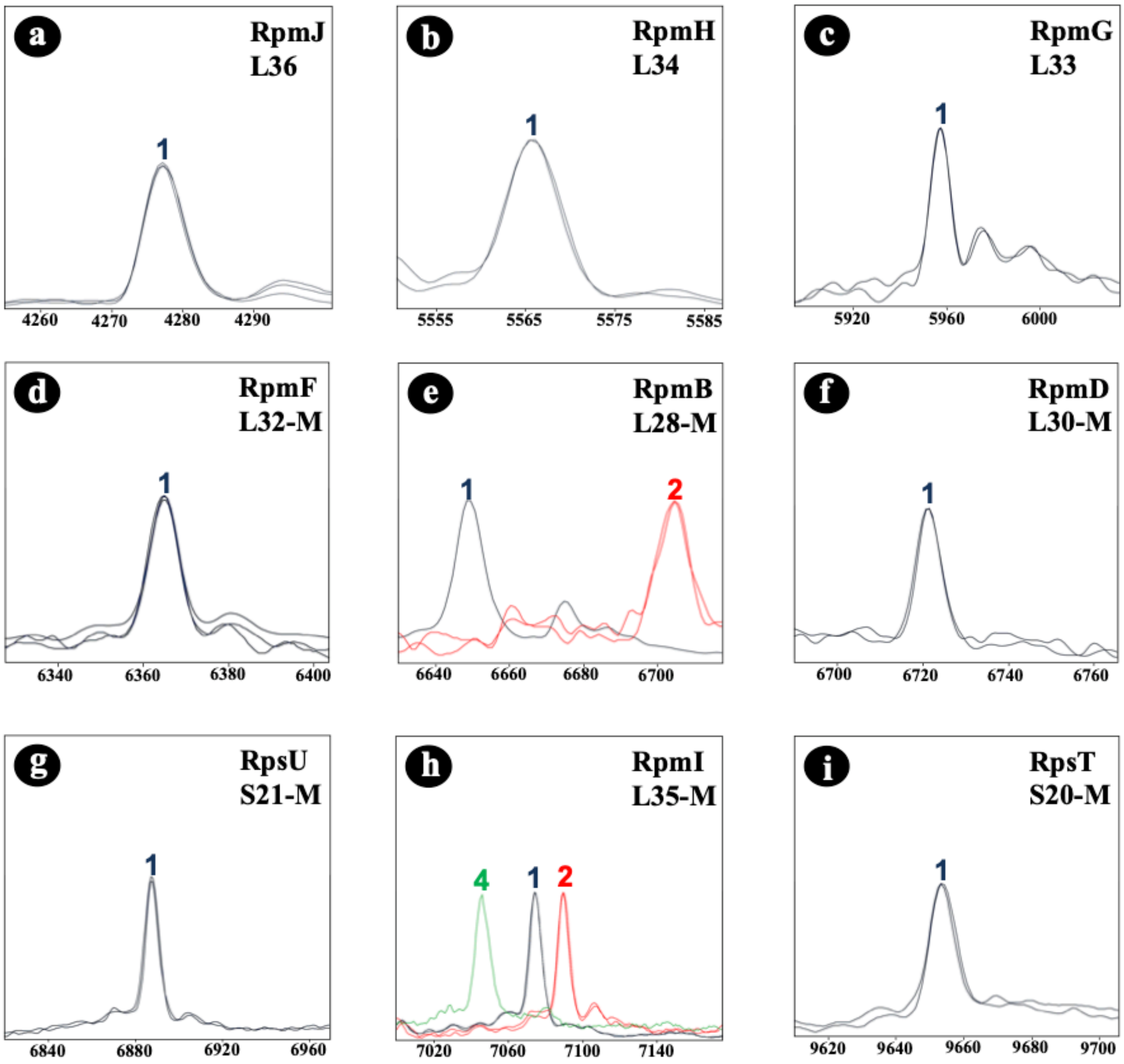


Fig. 3. *C. difficile*-specific proteotyping-derived biomarkers (a-i). In order to demonstrate mass differences between allelic isoforms, spectra of representative *C. difficile* isolates of each of the five detected proteotyping-derived types were overlaid. X-Axis: mass [Da] charge-1 ratio, scale 200 Da. Y-Axis: intensity [10x arbitrary units], spectra were individually adjusted to similar noise in order to improve visualization of peaks with low-intensity. Color codes: the isoform of *C. difficile* reference strain 630 (DSM 27543) is depicted in blue; red and light green indicate isoforms that differ in their mass from the reference strain 630 (DSM 27543). Isoforms lacking *N*-terminal methionine are appended with “-M”.

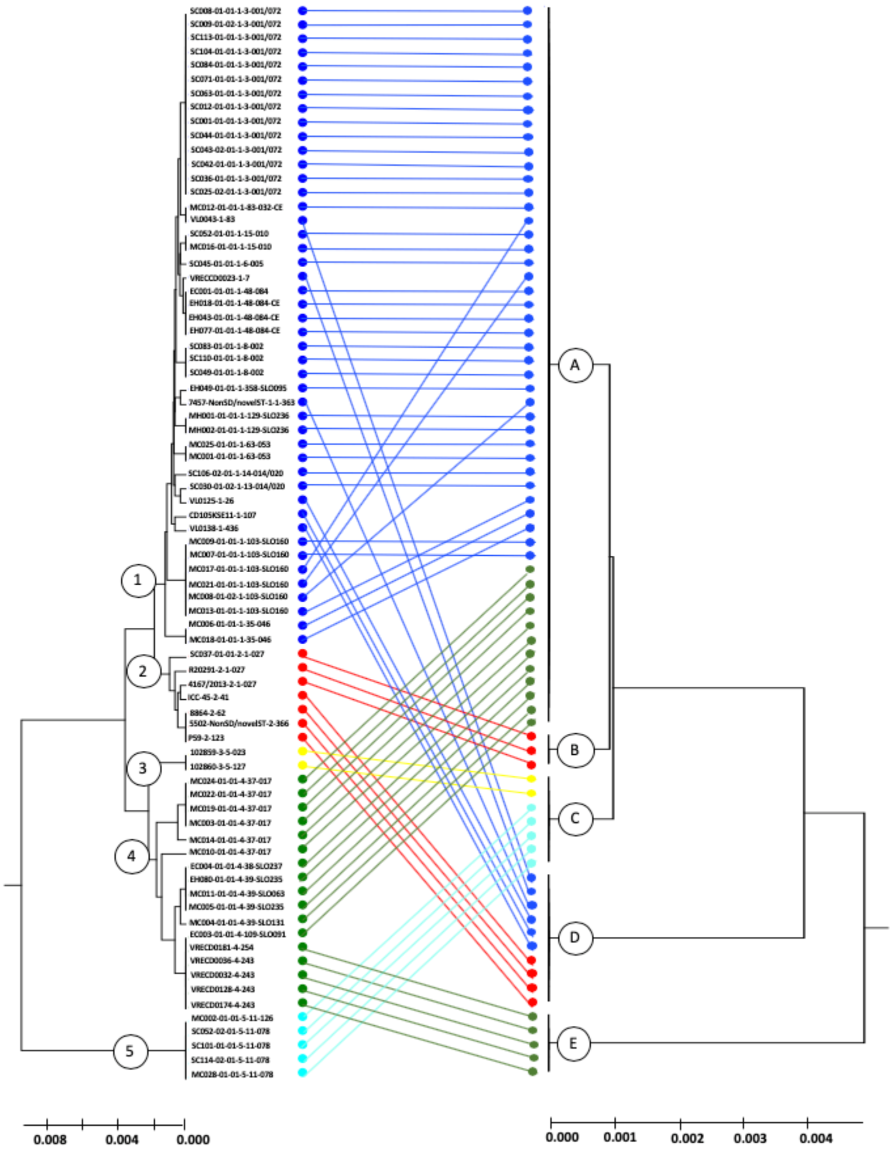


Fig. 4. Comparison of MLST- and proteotyping-derived phylogenies. Left tree: MLST-based evolutionary tree using the UPGMA method (maximum composite likelihood method). The isolates of the clades 1-5, indicated by different colors, form coherent clades. Here, the isolates of the clades 1 and 2 as well as the clades 3 and 4 form superclades while the clade 5 remains for itself.

Right dendrogram: Proteotyping-derived UPGMA-tree. Here, too, the isolates are arranged in five clades, which, however, do not correspond to the MLST clades. Especially noteworthy is the proteotyping-derived Clade B, which consists of exclusively hypervirulent RT027 isolates, also forming a separate MLST clade (clade 2).

3 Diskussion

Diese Arbeit beschäftigt sich mit der Entwicklung von Typisierungs- und Subtypisierungsschemata für klinisch relevante Bakterien auf Grundlage der Proteotypisierung. Genauer wurden Proteotypisierungsschemata für die Enterobakterien *C. jejuni* (siehe 2.1), *C. fetus* (siehe 2.2) und *C. difficile* (siehe 2.3) entwickelt, die das große Potenzial der Methode verdeutlichen.

Da der Einsatz der Methodik für die Subtypisierung von Bakterien noch in den Anfängen steckt, werden Vor- und Nachteile sowie das Abschneiden der Methode im Vergleich zu gängigen Methoden im Folgenden diskutiert.

Beweggrund für die Entwicklung neuer Methoden für die Typisierung unterhalb der Spezies- und Subspeziesebene in klinisch-mikrobiologischen Laboren ist der teilweise sehr hohe Kosten- beziehungsweise Zeitaufwand der gegenwärtig gängigen Prozeduren. Neben dem momentanen Goldstandard MLST ist die Subtypisierung mittels MALDI-TOF MS in den letzten Jahren zu einem Standardverfahren in der klinischen Diagnostik geworden. Die Applikation im Rahmen der Speziesidentifizierung ist bereits etabliert, der Einsatz im Kontext der Subtypisierung unterhalb der Spezies- und Subspeziesebene steckt jedoch noch in den Kinderschuhen.

Verschiedene Studien haben bereits die Existenz subgruppenspezifischer Biomarkermassen gezeigt, jedoch existierte lange Zeit kein standardisiertes Schema für die Evaluation massenspektrometrischer Daten im Rahmen der mikrobiellen Subtypisierung (Spinali *et al.*, 2015). Eine wesentliche Neuerung stellt das von unserer Arbeitsgruppe im Jahr 2015 entwickelte Proteotypisierungsschema dar, welches diese Lücke potenziell füllen kann (Zautner *et al.*, 2015). Im Wesentlichen werden für die erfolgreiche Etablierung der Methode eine Möglichkeit zur Anzucht der Mikroorganismen (Inkubator) sowie ein Massenspektrometer benötigt, also Geräte, die in der Regel in modernen klinisch-mikrobiologischen Routinelaboren vorhanden sind.

Ein wesentlicher Faktor für eine erfolgreiche Anwendung der Methode sind qualitativ hochwertige Spektren, wobei sämtliche Biomarkerionen eines Proteotypisierungsschemas auch im Spektrum des untersuchten Isolates eindeutig erkennbar sein müssen.

Präzision und Verlässlichkeit der Proteotypisierung beruhen im Wesentlichen auf der kombinatorischen Auswertung verschiedener Biomarkermassen. Um das bestmögliche Ergebnis zu erzielen, sollte das jeweilige Proteotypisierungsschema daher eine ausreichende Anzahl detektierbarer Marker beinhalten. Da in dieser Arbeit ausschließlich bakterielle Spezies untersucht wurden, stellte dieser Punkt kein Problem dar.

Bei der Analyse von Eukaryoten zeigte sich jedoch, dass die stärkere Zellwand eine insuffiziente Lyse der Zellen zur Folge hat. Dadurch werden weniger intrazelluläre Proteine freigesetzt, wodurch die im Massenspektrum detektierbaren Biomarkermassen für eine aussagekräftige Typisierung nicht ausreichend sind. Für Hefen wurde bereits ein vielversprechendes Lyse-Verfahren (YOTL - Yeast on target lysis) entwickelt, welches die Anzahl der detektierbaren Biomarkerionen erhöht (Bernhard *et al*., 2014). Dieser Ansatz zeigt darüber hinaus, dass sich die Anwendbarkeit der Methodik keineswegs auf Prokaryoten beschränkt.

Vergleicht man die Ergebnisse der Proteotypisierung mit denen von Clusteralgorithmen wie der Hauptkomponentenanalyse (PCA) oder der *single-linkage*-Clusteralgorithmus-Analyse, zeigt sich ein Vorteil der Proteotypisierung: Die Verwandtschaftsbeziehungen sind in hohem Maße reproduzierbar. In einer Pionierstudie unserer Arbeitsgruppe wurden für das ribosomale Protein L32-M drei verschiedene allele Isoformen nachgewiesen, welche die Differenzierung zweier Subspezies von *C. jejuni* ermöglichten. Resultate, die Wegbereiter für die Entwicklung des in dieser Arbeit verwendeten Proteotypisierungsansatzes waren (Zautner *et al*., 2013). In dieser Studie wurde die Phylogenie noch mittels PCA-basiertem Clusteralgorithmus ermittelt. In weiteren Untersuchungen unserer Arbeitsgruppe zeigte sich die Reproduzierbarkeit der Ergebnisse mit dieser Methode bei Verwendung unterschiedlicher Kulturbedingungen jedoch als unzureichend (Zautner *et al*., 2015). Dieses Problem kann mit der Proteotypisierung umgangen werden, indem die Massenverschiebungen der Biomarkerionen in die Aminosäuresequenz der jeweiligen Biomarkerisoform übertragen werden. Wie bereits angesprochen, gilt im Rahmen der Evaluation der Spektren das klar definierte Kriterium, dass sämtliche im Proteotypisierungsschema enthaltenen Biomarker deutlich sichtbar sein müssen. Ist dies nicht gegeben, erfolgt die Einstufung des Spektrums als qualitativ zu niedrig und wird folglich verworfen.

Es ist problemlos möglich, derartige Spektrendatensätze mit umfangreichen DNA-Sequenzdatensätzen (wgMLST- oder rMLST-Datenbanken) zu kombinieren. Die daraus resultierenden phylogenetischen beziehungsweise phyloproteomischen Relationen sind wesentlich verlässlicher als die aus PCA-Analysen hervorgehenden, da die Intensitäten lokaler Maxima nicht berücksichtigt werden. Auch die Reproduzierbarkeit ist der der PCA-Analysen überlegen.

Ein weiterer Vorteil der Methode ist die Robustheit gegenüber variierenden Kulturbedingungen. Es ist nicht erforderlich, dass sämtliche zu analysierenden Lysate parallel in Kultur genommen werden. Vielmehr ist es möglich, einzelne zu verschiedenen Zeitpunkten und an verschiedenen Orten erfasste Datensätze im Rahmen einer Analyse zu evaluieren, sofern die technische Ausstattung die gleiche ist. Auf diese Weise können mehrere hundert Isolate zeitgleich prozessiert werden.

Betrachtet man DNA-sequenzbasierte Methoden, bei welchen nach der Kultivierung der Mikroorganismen noch DNA-Isolierung und Sequenzierung erfolgen, wird der Vorteil der Proteotypisierung schnell deutlich: Sowohl die Schnelligkeit bei der Datenakquise, als auch bei der Speziesbestimmung im Rahmen der Routinediagnostik ist bei der Proteotypisierung gegeben. Auch von einem wirtschaftlichen Standpunkt aus gesehen, ist die Proteotypisierung ausgesprochen attraktiv, da die Kosten pro Analyse nur wenige Cent betragen.

Im Vergleich zu den zu Beginn von unserer Arbeitsgruppe durchgeführten Analysen, bei denen nur einzelne Biomarker betrachtet wurden, haben die in dieser Arbeit durchgeführten proteotypisierungsbasierten Analysen ein wesentlich größeres Diskriminierungsvermögen, da hier zwischen neun und 16 Biomarkern für die Evaluation zurate gezogen wurden. In der Studie von Zautner *et al*. (2015) wurden gar 19 verschiedene Biomarkerionen in das *C. jejuni*-Proteotypisierungsschema involviert.

Betrachtet man die Biomarker und die posttranslationalen Modifikationen genauer, stellt man fest, dass sich kein genaues Muster erkennen lässt, beziehungsweise dass Spezies derselben Gattung nicht immer die gleichen Biomarker aufweisen.

Die Biomarker L36, L34, L32-M und S20-M sind bei allen untersuchten Bakterienspezies (*C. jejuni jejuni*, *C. jejuni doylei, C. coli*, *C. fetus*, *C. difficile*) vorhanden und unterscheiden sich auch nicht hinsichtlich der PTM. Die Biomarker L29 und

L24-M sind bei allen hier untersuchten Spezies der Gattung *Campylobacter* vorhanden und weisen ebenfalls die gleichen PTM auf.

Die Biomarker L31, L28-M, L27-M, S16, S15-M und Hyp Prot kommen sowohl bei *C. jejuni*, als auch bei *C. coli* vor, fehlen jedoch bei *C. fetus*. Diese Ergebnisse bestätigen die Resultate von Fagerquist *et al*. (2006), wonach die Muster posttranslationaler Modifikationen spezies- und subspeziesspezifisch jedoch nicht isolatspezifisch sind.

Um die Eignung der Methode für die klinische Diagnostik beurteilen zu können, ist das Abschneiden im Vergleich zu aktuell gängigen Methoden wichtig. Die Proteotypisierungsstudie zu *C. coli* (siehe 2.1) zeigte, dass via Proteotypisierung eine Diskriminierung der drei bekannten MLST-Kladen möglich ist. Auch Isolate, die der Spezies ähnlich sind, konnten erfolgreich abgegrenzt werden. Ebenso war eine deutliche Unterscheidung von *C. jejuni*-Stämmen möglich. Da von den drei bekannten *C. coli*-Kladen ausschließlich Klade 1-Isolate klinisch relevant sind, handelt es sich beim *C. coli*-Proteotypisierungsschema um ein äußerst interessantes Tool für die Routinediagnostik in klinisch-mikrobiologischen Laboren. Die Fähigkeit zur Unterscheidung entspricht der der MLST, jedoch ist die Proteotypisierung schneller und günstiger.

Beim *C. fetus*-MLST war das Ziel, die drei bekannten *C. fetus*-Subspezies via Proteotypisierung zu unterscheiden (siehe 2.3): *Cff*, *Cfv* und *Cft.* Das *C. fetus*-Proteotypisierungsschema unterscheidet zwischen vier verschiedenen Typen. Drei davon beinhalten ausschließlich *Cff* und *Cfv* Stämme während der vierte proteotypisierungsbasierte Sequenztyp ausschließlich *Cft* Stämme beinhaltet. Das bedeutet, dass die Unterscheidung von *Cft* Stämmen von anderen *C. fetus*-Subspezies genauso gut funktioniert wie mit der Referenzmethode MLST. Insgesamt wurden für die Untersuchung 41 Stämme verwendet, die alle *C. fetus*-Subspezies sowie 14 MLST-Sequenztypen abdeckten. Da nur vier proteotypisierungsbasierte Sequenztypen ermittelt wurden, ist die diskriminatorische Fähigkeit der Proteotypisierung in diesem Fall der MLST unterlegen.

Betrachtet man die Ergebnisse der dritten Studie, der *C. difficile*-Proteotypisierung (siehe 2.3), zeigt sich, dass die identifizierten proteotypisierungsbasierten Sequenztypen zwar nicht genau mit den MLST-Typen übereinstimmten, jedoch die Subtypisierungstiefe der beiden Methoden äquivalent ist. Besonders interessant war in dieser Studie, dass einer der proteotypisierungsbasierten

Sequenztypen ausschließlich aus RT027 Stämmen bestand, die bekanntermaßen hypervirulent sind. So bietet das *C. difficile*-Proteotypisierungsschema eine ausgezeichnete Möglichkeit für die klinische Diagnostik, diese hochpathogenen Stämme im Fall eines Ausbruchsgeschehens äußerst schnell zu identifizieren. Zukünftige Studien können sich damit befassen, weitere *C. difficile*-Ribotypen zu untersuchen, die mit RT027 Stämmen eng verwandt sind. Hierzu zählen beispielsweise RT016, RT036, RT075, RT122, RT153, RT176. Ein Screening von Stämmen mit diesem PCR-Ribotypen würde die Aussagekraft der Methode noch steigern. Analysen der mit RT027 eng verwandten Stämme konnten im Rahmen dieser Arbeit nicht durchgeführt werden, da entsprechende Stämme nicht zur Verfügung standen.

Grundsätzlich kann gesagt werden, dass die Gleichwertigkeit der Proteotypisierung gegenüber der MLST von der spezifischen epidemiologischen Fragestellung für bestimmte Isolategruppen abhängt. Ist eine eindeutige Unterscheidung auf Grundlage der vorliegenden Massenspektren nicht möglich, ist die Bevorzugung DNA-sequenzbasierter Analysemethoden, aufgrund des besseren Diskriminierungsvermögens, anzuraten. Um eine Verbesserung der diskriminatorischen Fähigkeiten der Proteotypisierung zu erreichen und das Proteotypisierungsschema zu vervollständigen, sollten zukünftige Studien darauf abzielen, zusätzliche bis dato nicht identifizierte Biomarker den jeweiligen Genloci zuzuordnen. Außerdem ist es sinnvoll, das detektierbare Massenspektrum zu erweitern. Ein limitierender Faktor bei Anwendung der Proteotypisierung sind die in der Isoformendatenbank zur Verfügung stehenden Daten. Es sollten sowohl die Proteinsequenzen, als auch die assoziierten Nukleotidsequenzen vorliegen. Diese Isoformentabelle bildet die Grundlage der Auswertung und ist umso aussagekräftiger, je mehr Nukleotid- beziehungsweise Proteinsequenzdaten vorliegen. Um ein erregerspezifisches Proteotypisierungsschema etablieren zu können, sollte deshalb ein umfangreicher Genomsequenzdatensatz verfügbar sein, wie etwa in öffentlich zugänglichen Portalen wie NCBI. Für die Erstellung der Isoformendatenbank sind jedoch *whole genome shotgun*-Sequenzen ausreichend. Liegt kein ausreichend großer Datensatz vor, bedeutet dies einen erheblichen Mehraufwand während des Etablierungsprozesses: Jedes einzelne Biomarkerisoform muss mittels PCR und Sangersequenzierung geprüft werden. Die Anwendung der Methodik auf weitere prokaryotische und eukaryotische Spezies ist gegenwärtig durch einen Mangel an zur Verfügung stehender Sequenzdaten limitiert, weshalb die Etablierung einer ribosomalen MLST beziehungsweise whole genome MLST assoziierten Spektrendatenbank wünschenswert ist.

Prinzipiell kann die Proteotypisierung zur Lösung vieler klinisch relevanter Probleme beitragen. Die gute Adaptierbarkeit der Methode ermöglicht beispielsweise die Differenzierung wichtiger klonaler MRSA-Komplexe, die Unterscheidung von VISA und VSSA, die Differenzierung von MSSA und MRSA sowie die Unterscheidung der SSC*mec*-Typen, insbesondere, wenn auf eine ausreichende Menge Genomsequenzen in rMLST- beziehungsweise wgMLST-Datenbanken zurückgegriffen werden kann.

Die Untersuchungen unserer Arbeitsgruppe haben bereits die Anwendung auf eine Vielzahl von Organismen erfolgreich demonstriert. So wurden neben den in dieser Arbeit untersuchten Organismen *C. coli*, *C. fetus* und *C. difficile* auch *C. jejuni* subsp. *jejuni*, *C. jejuni* subsp. *doylei*, *E. coli*, *S. enterica* spp. *enterica* und *S. aureus* charakterisiert.

Bei der Proteotypisierung wird die Phylogenie einer Spezies anhand von Unterschieden in bestimmten Biomarkern (ribosomalen Proteinen) abgeleitet. Da die Phylogenie mikrobieller Spezies mit Präsenz beziehungsweise Absenz von Resistenz- und Virulenzfaktoren korreliert, stellt die Proteotypisierung also ein sehr vielversprechendes Werkzeug für die frühzeitige Detektion von Resistenzprofilen sowie hochvirulenten Phänotypen dar.

Zukünftige Arbeiten sollten sich mit der Adaption der Methodik auf weitere Erreger beziehungsweise klinische Fragestellungen befassen. Da die Methode mittelfristig in der Routinediagnostik in klinisch-mikrobiologischen Laboren eingesetzt werden soll, ist die Entwicklung einer endnutzerfreundlichen Lösung auf bioinformatischer Ebene notwendig. Die entsprechende Software sollte in der Lage sein, die durch massenspektrometrische Analysen gewonnenen Rohdaten mit epidemiologischen Daten zu verknüpfen und dadurch die unmittelbare Berechnung entsprechender Wahrscheinlichkeiten für nosokomiale, multiresistente oder hochvirulente Keime erlauben.

4 Quellenverzeichnis

Alfredson, David A. und Victoria Korolik (2007) Antibiotic resistance and resistance mechanisms in Campylobacter jejuni and Campylobacter coli. *FEMS microbiology letters* 277(2):123–132.

Amador, Paula P. u. a. (2015) Antibiotic resistance in wastewater: Occurrence and fate of Enterobacteriaceae producers of Class A and Class C β-lactamases. *Journal of Environmental Science and Health, Part A* 50(1):26–39.

Arias, Cesar A. und Barbara E. Murray (2012) The rise of the Enterococcus: beyond vancomycin resistance. *Nature Reviews Microbiology* 10(4):266.

Barbut, F. u. a. (2014) Does a rapid diagnosis of Clostridium difficile infection impact on quality of patient management? *Clinical Microbiology and Infection* 20(2):136–144.

Berendonk, Thomas U. u. a. (2015) Tackling antibiotic resistance: the environmental framework. *Nature Reviews Microbiology* 13(5):310.

Bernhard, Mareike u. a. (2014) YOTL–a procedure for making auxiliary mass spectrum datasets for clinical routine identification of yeasts using the on-target-lysis method. *Journal of clinical microbiology*: JCM–02128.

Boers, Stefan A., Wil A. Van der Reijden und Ruud Jansen (2012) High-throughput multilocus sequence typing: bringing molecular typing to the next level. PloS one 7(7):e39630.

Bolton, Declan J. (2015) Campylobacter virulence and survival factors. *Food microbiology* 48:99–108.

Bowman, Rodney A., Gael L. O'Neill und Thomas V. Riley (1991) Non-radioactive restriction fragment length polymorphism (RFLP) typing of Clostridium difficile. *FEMS microbiology letters* 79(2–3):269–272.

Carrillo, Catherine Dianna u. a. (2012) A framework for assessing the concordance of molecular typing methods and the true strain phylogeny of Campylobacter jejuni and C. coli using draft genome sequence data. *Frontiers in cellular and infection microbiology* 2:57.

Cheng, Jing-Wei u. a. (2018) Use of matrix-assisted laser desorption ionization–time of flight mass spectrometry to identify MLST clade 4 Clostridium difficile isolates. *Diagnostic microbiology and infectious disease*.

Christner, Martin u. a. (2014) Rapid MALDI-TOF mass spectrometry strain typing during a large outbreak of Shiga-toxigenic Escherichia coli. *PLoS One* 9(7): e101924.

Cody, Alison J. u. a. (2013) Real-time genomic epidemiology of human Campylobacter isolates using whole genome multilocus sequence typing. *Journal of clinical microbiology*: JCM–00066.

Conway, Gregory C. u. a. (2001) Phyloproteomics: species identification of Enterobacteriaceae using matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. *Journal of molecular microbiology and biotechnology* 3(1):103–112.

Corver, Jeroen u. a. (2018) Identification and validation of two peptide markers for the recognition of Clostridioides difficile MLST-1 and MLST-11 by MALDI-MS. *Clinical Microbiology and Infection*.

Croxatto, Antony, Guy Prod’hom und Gilbert Greub (2012) Applications of MALDI-TOF mass spectrometry in clinical diagnostic microbiology. *FEMS microbiology reviews* 36(2):380–407.

Dingle, K. E. u. a. (2001) Multilocus Sequence Typing System for Campylobacter jejuni. *Journal of clinical microbiology* 39(1):14–23.

Dingle, Kate E. u. a. (2005) Sequence typing and comparison of population biology of Campylobacter coli and Campylobacter jejuni. *Journal of Clinical Microbiology* 43(1):340–347.

Dodd, Michael C. (2012) Potential impacts of disinfection processes on elimination and deactivation of antibiotic resistance genes during water and wastewater treatment. *Journal of Environmental Monitoring* 14(7):1754–1771.

Durighello, Emie u. a. (2014) Proteogenomic biomarkers for identification of Francisella species and subspecies by matrix-assisted laser desorption ionization-time-of-flight mass spectrometry. *Analytical chemistry* 86(19):9394–9398.

Emele, Matthias Frederik u. a. (2019) Proteotyping as alternate typing method to differentiate *Campylobacter coli* clades. Scientific Reports 9.

Fagerquist, Clifton K. u. a. (2006) Sub-speciating Campylobacter jejuni by proteomic analysis of its protein biomarkers and their post-translational modifications. Journal of proteome research 5(10):2527–2538.

Gibreel, Amera und Diane E. Taylor (2006) Macrolide resistance in Campylobacter jejuni and Campylobacter coli. Journal of Antimicrobial Chemotherapy 58(2):243–255.

Griekspoor, Petra u. a. (2015) Genetic diversity and host associations in Campylobacter jejuni from human cases and broilers in 2000 and 2008. *Veterinary microbiology* 178(1–2):94–98.

Griffiths, David u. a. (2010) Multilocus sequence typing of Clostridium difficile. *Journal of clinical microbiology* 48(3):770–778.

Hugo, Alys u. a. (2012) Proteotyping of Microbial Communities by Optimization of Tandem Mass Spectrometry Data Interpretation. In: Biocomputing 2012. S. 225–234. World Scientific.

Jolley, Keith A. u. a. (2012) Ribosomal multilocus sequence typing: universal characterization of bacteria from domain to strain. *Microbiology* 158(4):1005–1015.

Kachrimanidou, Melina und Nikolaos Malisiovas (2011) Clostridium difficile infection: a comprehensive review. *Critical reviews in microbiology* 37(3):178–187.

Karlsson, Roger u. a. (2015) Proteotyping: proteomic characterization, classification and identification of microorganisms–a prospectus. *Systematic and applied microbiology* 38(4):246–257.

Khan, Hassan Ahmed, Aftab Ahmad und Riffat Mehboob (2015) Nosocomial infections and their control strategies. *Asian pacific journal of tropical biomedicine* 5(7):509–514.

Kinnebrew, Melissa A. u. a. (2010) Bacterial flagellin stimulates toll-like receptor 5—dependent defense against vancomycin-resistant Enterococcus infection. *The Journal of infectious diseases* 201(4):534–543.

Knapp, Charles W. u. a. (2010) Differential fate of erythromycin and beta-lactam resistance genes from swine lagoon waste under different aquatic conditions. *Environmental Pollution* 158(5):1506–1512.

Kuhns, Martin u. a. (2012) Rapid discrimination of Salmonella enterica serovar Typhi from other serovars by MALDI-TOF mass spectrometry. PLoS One 7(6):e40004.

Lartigue, Marie-Frédérique (2013) Matrix-assisted laser desorption ionization time-of-flight mass spectrometry for bacterial strain characterization. *Infection, Genetics and Evolution* 13:230–235.

Leekitcharoenphon, Pimlapas u. a. (2012) Genomic variation in Salmonella enterica core genes for epidemiological typing. *BMC genomics* 13(1):88.

Lemee, Ludovic u. a. (2004) Multilocus sequence typing analysis of human and animal Clostridium difficile isolates of various toxigenic types. *Journal of clinical microbiology* 42(6):2609–2617.

Lindstedt, Bjørn-Arne u. a. (2000) Comparative Fingerprinting Analysis of Campylobacter jejuni subsp. jejuni Strains by Amplified-Fragment Length Polymorphism Genotyping. *Journal of clinical microbiology* 38(9):3379–3387.

Maiden, Martin CJ u. a. (1998) Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. *Proceedings of the National Academy of Sciences* 95(6):3140–3145.

Maiden, Martin CJ (2006) Multilocus sequence typing of bacteria. *Annu. Rev. Microbiol.* 60:561–588.

Maslow, J. N. (1993) Application of pulsed-field gel electrophoresis to molecular epidemiology. *Diagnostic molecular microbiology* :563–572.

Ojima-Kato, Teruyo u. a. (2016) Matrix-assisted Laser Desorption Ionization-Time of Flight Mass Spectrometry (MALDI-TOF MS) can precisely discriminate the lineages of Listeria monocytogenes and species of Listeria. *PloS one* 11(7):e0159730.

Opota, Onya, Guy Prod’hom und Gilbert Greub (2017) Applications of MALDI-TOF Mass Spectrometry in Clinical Diagnostic Microbiology. *MALDI-TOF and Tandem MS for Clinical Microbiology* :55–92.

Ortega, Lucía, Anna Ryberg und \AAsa Johansson (2018) HMW-profiling using MALDI-TOF MS: A screening method for outbreaks of Clostridioides difficile. *Anaerobe*.

Pérez-Losada, Marcos u. a. (2013) Pathogen typing in the genomics era: MLST and the future of molecular epidemiology. *Infection, Genetics and Evolution* 16:38–53.

Pfaller, Michael A. und Mariana Castanheira (2015) Nosocomial candidiasis: antifungal stewardship and the importance of rapid diagnosis. *Medical mycology* 54(1):1–22.

Ragimbeau, Catherine u. a. (2014) Investigating the host specificity of Campylobacter jejuni and Campylobacter coli by sequencing gyrase subunit A. BMC microbiology 14(1):205.

Rizzardi, Kristina und Thomas \AAkerlund (2015) High molecular weight typing with MALDI-TOF MS-a novel method for rapid typing of clostridium difficile. *Plos one* 10(4): e0122457.

Rodriguez, Carlos u. a. (2006) Proteotyping of human haptoglobin by MALDI-TOF profiling: Phenotype distribution in a population of toxic oil syndrome patients. *Proteomics* 6(S1): S272–S281.

Rolfe, RIAL D. (1984) Role of volatile fatty acids in colonization resistance to Clostridium difficile. *Infection and immunity* 45(1):185–191.

Rosef, O. u. a. (1983) Isolation and characterization of Campylobacter jejuni and Campylobacter coli from domestic and wild mammals in Norway. *Applied and Environmental Microbiology* 46(4):855–859.

Sandora, Thomas J., Peter Gerner-Smidt und Alexander J. McAdam (2014) What's your subtype? The epidemiologic utility of bacterial whole-genome sequencing. *Clinical chemistry* 60(4):586–588.

Schwahn, Alexander B., Jason WH Wong und Kevin M. Downard (2010) Rapid differentiation of seasonal and pandemic H1N1 influenza through proteotyping of viral neuraminidase with mass spectrometry. *Analytical chemistry* 82(11):4584–4590.

Schwartz, David C. und Charles R. Cantor (1984) Separation of yeast chromosome-sized DNAs by pulsed field gradient gel electrophoresis. cell 37(1):67–75.

Seng, Piseth u. a. (2010) MALDI-TOF-mass spectrometry applications in clinical microbiology. *Future microbiology* 5(11):1733–1754.

Sheppard, Samuel K., John F. Dallas, Marion MacRae, u. a. (2009a) Campylobacter genotypes from food animals, environmental sources and clinical disease in Scotland 2005/6. *International journal of food microbiology* 134(1–2):96–103.

Sheppard, Samuel K., John F. Dallas, Norval JC Strachan, u. a. (2009b) Campylobacter genotyping to determine the source of human infection. *Clinical Infectious Diseases* 48(8):1072–1078.

Shillingford, Jonathan M. u. a. (2003) Proteotyping of mammary tissue from transgenic and gene knockout mice with immunohistochemical markers: a tool to define developmental lesions. *Journal of Histochemistry & Cytochemistry* 51(5):555–565.

Spinali, Sébastien u. a. (2015) Microbial typing by matrix-assisted laser desorption ionization–time of flight mass spectrometry: do we need guidance for data interpretation? *Journal of clinical microbiology* 53(3):760–765.

Suarez, Stéphanie u. a. (2013) Ribosomal proteins as biomarkers for bacterial identification by mass spectrometry in the clinical microbiology laboratory. *Journal of microbiological methods* 94(3):390–396.

Tang, Yizhi u. a. (2017) Emergence of a plasmid-borne multidrug resistance gene cfr(C) in foodborne pathogen Campylobacter. Journal of Antimicrobial Chemotherapy 72(6):1581–1588.

Van Bergen, Marcel AP u. a. (2005) Clonal nature of Campylobacter fetus as defined by multilocus sequence typing. *Journal of clinical microbiology* 43(12):5888–5898.

Velappan, Nileena u. a. (2001) Rapid identification of pathogenic bacteria by single-enzyme amplified fragment length polymorphism analysis. *Diagnostic microbiology and infectious disease* 39(2):77–83.

Vos, Pieter u. a. (1995) AFLP: a new technique for DNA fingerprinting. *Nucleic acids research* 23(21):4407–4414.

Waldenström, Jonas u. a. (2002) Prevalence of Campylobacter jejuni, Campylobacter lari, and Campylobacter coli in different ecological guilds and taxa of migrating birds. *Applied and Environmental Microbiology* 68(12):5911–5917.

Wolters, Manuel u. a. (2011) MALDI-TOF MS fingerprinting allows for discrimination of major methicillin-resistant Staphylococcus aureus lineages. *International Journal of Medical Microbiology* 301(1):64–68.

Yan, William, Nicholas Chang und Diane E. Taylor (1991) Pulsed-field gel electrophoresis of Campylobacter jejuni and Campylobacter coli genomic DNA and its epidemiologic application. *Journal of Infectious Diseases* 163(5):1068–1072.

Zautner, Andreas Erich u. a. (2013) Discrimination of multilocus sequence typing-based Campylobacter jejuni subgroups by MALDI-TOF mass spectrometry. *BMC microbiology* 13(1):247.

Zautner, Andreas Erich u. a. (2015) Mass Spectrometry-based PhyloProteomics (MSPP): A novel microbial typing Method. *Scientific reports* 5:13431.

Zautner, Andreas E. u. a. (2016) Subtyping of Campylobacter jejuni ssp. doylei Isolates Using Mass Spectrometry-based PhyloProteomics (MSPP). *Journal of visualized experiments: JoVE* (116).

Publikationen

Emele, M. F. u. a. (2019) Proteotyping as alternate typing method to differentiate Campylobacter coli clades. Scientific Reports 9.

Zautner, A. E., Groß, U., **Emele, M. F.** u. a. (2017) More Pathogenicity or Just More Pathogens? —On the Interpretation Problem of Multiple Pathogen Detections with Diagnostic Multiplex Assays. *Frontiers in microbiology* 8:1210.

Ergebnisse dieser Dissertation wurden zwischen Februar 2017 und Mai 2019 in folgenden Kongressbeiträgen und Publikationen präsentiert:

Vorträge

02/2018 **M.F. Emele,** Andreas E. Zautner
"Mass Spectrometry-based PhyloProteomics (MSPP) of *Campylobacter coli*"
70. Jahrestagung der Deutschen Gesellschaft für Hygiene und Mikrobiologie (Bochum)

02/2019 **M.F. Emele,** Andreas E. Zautner
"Mass spectrometry-based phyloproteomics of *Clostridioides difficile* as alternate typing method to ribotyping is able to differentiate the Ribotype 027"
71. Jahrestagung der Deutschen Gesellschaft für Hygiene und Mikrobiologie (Göttingen)

Posterpräsentationen

02/2018 **M.F. Emele,** Andreas E. Zautner
"Mass Spectrometry-based PhyloProteomics (MSPP) of *Campylobacter coli*"
National Symposium on Zoonoses Research 2018 (Berlin)

Erklärung

Hiermit versichere ich, dass ich die vorliegende Dissertation mit dem Titel

„Charakterisierung klinisch-relevanter Bakterien mittels Proteotypisierung“

selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet habe.

Matthias Frederik Emele

Göttingen, den 25.03.2019