

1 **Fecal host biomarkers predicting severity of *Clostridioides difficile* infection**

2 Makan Golizeh¹, Kaitlin Winter^{1,2}, Marija Landekic^{1,2}, Lucie Roussel¹, Mélanie Langelier¹,
3 Vivian G. Loo^{1,2,3}, Momar Ndao^{1,2,3}, and Donald C. Vinh^{1,2,3},

4 ¹ *Infectious Diseases and Immunity in Global Health (IDIGH) Program, Research Institute of the McGill*
5 *University Health Centre (RI-MUHC), Montreal, Quebec, Canada*

6 ² *Department of Microbiology and Immunology, McGill University, Montreal, Quebec, Canada*

7 ³ *Department of Medicine, McGill University, Montréal, Québec, Canada*

8
9 Corresponding authors: Momar Ndao, momar.ndao@mcgill.ca; Donald C. Vinh, donald.vinh@mcgill.ca.

10
11 **Supplemental Information**

12
13 **Index**

	Section	Page
S1	Proteomics analysis	2
S2	<i>Clostridioides difficile</i> infection diagnosis	4
S3	Mouse experiment	4
S4	Immunoblotting procedure	5
	References for Supplemental information.....	6
S5	Supplemental Figures	7
S5	Supplemental Tables	11

1 **S1. Proteomics analysis**

2 Fecal material (~ 1.0 g) was transferred into a sterile low protein binding 1.5-mL plastic tube and
3 suspended in 0.5 mL 8 M urea, 0.1 M NaCl, 0.025 M Tris pH 8.2 solution containing Pierce protease
4 inhibitor (Thermo Fisher Scientific, Waltham, MA). Samples were mixed vigorously (4 x 30 s) and
5 centrifuged (2500 g, 5 min). The supernatant was pipetted into a clean tube and re-centrifuged (20,000 g,
6 30 min). The supernatant was then filtered on a Pierce ~30 µm polyethylene centrifugal filter (8000 g, 10
7 min) and re-filtered using a Corning Spin-X 0.45 µm cellulose acetate centrifugal filter (Thermo Fisher
8 Scientific, 12,000 g, 15 min). The filtrate (0.15 mL, 0.4 mg total protein, quantified by Pierce BCA protein
9 assay kit) was loaded on to a re-hydrated reverse-phase C₄ solid-phase extraction (SPE) spin column
10 (Harvard Apparatus, Holliston, MA). The unbound fraction was collected (2000 g, 4 min) and discarded.
11 C₄ material-bound species were eluted from the column (2000 g, 4 min) with acetonitrile (ACN, 1%
12 formic acid) and evaporated to dryness under vacuum (2 h, 37°C). Samples were kept at ≤ 4°C during
13 processing.

14 Liquid chromatography mass spectrometry (LC-MS) analysis was performed on a Dionex Ultimate 3000
15 UHPLC system (Thermo Fisher Scientific) coupled to a Maxis II high-resolution quadrupole-time of flight
16 (Q-TOF) tandem mass spectrometer (Bruker, Billerica, MA) equipped with an Apollo electrospray
17 ionization (ESI) source (Bruker). Briefly, the cleaned-up fecal protein extract (**Section S1.2**) was
18 reconstituted in 1% formic acid (60 µL) and injected (20 µL) on to a Jupiter C₄ 1 x 150 mm 5 µm (300 Å)
19 HPLC column (Phenomenex, Torrance, CA) and separated using a gradient elution with ACN (0.8%
20 formic acid) in ultra-pure water (1% formic acid) at 125 µL/min (50°C). % ACN was at 5% within the
21 first 4 min, during which sample was being loaded on to the column, then increased to 20% at 10 min, to
22 50% at 30 min, to 70% at 40 min, to 90% at 45 min, held for 5 min, and then changed back to 5% to re-
23 equilibrate the column for an additional 10 min. MS analysis was conducted in positive ionization mode
24 with a source voltage of 4.5 kV for the mass-to-charge ratio (*m/z*) of 400–3000 with an acquisition rate of
25 1.00 Hz. A continuous flow of 99.5% pure N₂ was provided (Parker, Mayfield Heights, OH, USA) as dry
26 gas (200°C) at a flow rate of 6 L/min. A low-concentration ESI tuning mix (Agilent Technologies, Santa
27 Clara, CA) was used for post-column infusion external calibration within the first 3 min of every LC-MS
28 analysis. Samples were analyzed in random order in the sequence to avoid biases due to the injection order
29 and one blank (5 µL 1% formic acid in water) was inserted between each two analyses to minimize
30 carryover.

31 Proteoform detection was conducted by Compass DataAnalysis software (Bruker) using Sophisticated
32 Numerical Annotation Procedure (SNAP) algorithm [1] for LC-MS features within a MW range of 1.0–
33 100.0 kDa and a maximum charge state of 40+. The signal-to-noise (S/N) and quality factor thresholds
34 were set to 5.0 and 0.65, respectively. Data processing was performed on a Hewlett-Packard PC

1 workstation (Palo Alto, CA) equipped with an Intel (Santa Clara, CA) Xenon 8-core 3.70 GHz CPU and
2 16 GB RAM. The list of proteoforms was imported into the Compass ProfileAnalysis software (Bruker)
3 for statistical analysis. The software generated a list of biomarker candidates with m/z 1500–30,000 and
4 retention time 10–55 min present in at least 60% of the samples within each cohort using the time
5 alignment option to account for LC retention shifts. Candidates elucidating $\geq \pm 50\%$ change ($p < 0.05$,
6 Student's t -test) in signal intensity between the cohorts were selected for protein identification by liquid
7 chromatography tandem mass spectrometry (LC-MS/MS). To this end, samples with relatively high
8 abundance of each candidate were re-analyzed (20 μ L) using the LC parameters employed in the
9 discovery phase coupled with a targeted MS/MS method to selectively fragment the biomarker candidate
10 at the given m/z value and retention time. The MS/MS analysis was performed at an acquisition rate of
11 1.08 Hz using collision-induced dissociation. LC-MS/MS data were processed in DataAnalysis software
12 and the MS/MS spectra associated with selected biomarker candidates were sent to BioTools software
13 (Bruker) for protein identification using Mascot search engine (Matrix Science, London, UK) against the
14 UniProt/Swiss-Prot protein database (downloaded on 28 July 2017; 555,100 sequences). Top-down
15 sequencing was conducted by adding methionine oxidation and asparagine/glutamine deamidation as
16 variable modifications with a mass tolerance of 50 ppm and 50 mDa at the MS and MS/MS levels,
17 respectively. Other statistical testing including principal component analysis (PCA), partial least squares
18 (PLS) and receiver operating characteristic (ROC) analysis were performed by the ProfileAnalysis
19 software (Bruker).

20 Gene ontology information was obtained from the UniProt knowledgebase. Pathway over-representation
21 analysis (hypergeometric algorithm & Benjamini-Hochberg correction) and protein-protein interaction
22 analysis were performed by InnateDB integrated analysis platform [2] with UniProt identifiers as the
23 cross-reference database. The results were subsequently visualized with the Cerebral plugin [3] of
24 Cytoscape software to organize the interaction network in the context of subcellular localization for all of
25 the proteins displayed and visualized by Cerebral.

26 The LC-MS method was optimized for proteins within 1.5–80.0 kDa using an intact protein mix standard
27 solution (Sigma-Aldrich). The limit of detection (LOD) was determined using 0.1–1.5 mg/mL solutions
28 from the intact protein mix. The LOD for proteins within the range of 5.0–30.0 kDa was found to be
29 approximately 0.2–0.5 μ g for the LC column used in this study. This LOD may be insufficient for the
30 detection of some immunologically important but usually low-abundance fecal proteins such as
31 calprotectin (50 μ g/g feces) [4]. Low-end micro-flow and nano-flow LC-MS is known to ameliorate
32 signal-to-noise ratio and thus method sensitivity via decreased analyte diffusion due to smaller column
33 diameter and lower flow rate [5]. Initial tests on 0.3 mm and 0.075 mm i.d. HPLC columns at reduced

1 flow rates (6.0 and 0.3 $\mu\text{L}/\text{min}$, respectively) yielded promising results however the method failed to
2 deliver reproducible results.

3 Initial tests also demonstrated that the peak finding script that employed SNAP algorithm occasionally
4 failed to detect complex MS spectra associated with excessively charged high-MW proteins such as serum
5 albumin (69.3 kDa) and transferrin (77.1 kDa). An alternate technique, maximum entropy (MaxEnt), has
6 been successfully applied to deconvolute electrospray mass spectra of multiply charged proteins [6]. In
7 contrast to SNAP, MaxEnt is known for its effectiveness in the detection of super-complex MS spectra of
8 larger proteins albeit at the expense of longer data processing time [6]. While a regular peak detection
9 process using SNAP lasted 30–45 min, MaxEnt processing took 6–8 h to complete on the same computer.
10 A hybrid SNAP/MaxEnt method did not considerably improve peak detection efficiency. However,
11 excluding known primary contaminants, such as common detergents and polymeric species, from the MS
12 acquisition method led to faster generation of a cleaner LC-MS peak list.

13 **S2. *Clostridioides difficile* infection diagnosis**

14 Stool samples were submitted by the treating physician on their real-world, clinical suspicion for CDI.
15 Blood-based laboratory parameters were measured by the treating physicians as deemed necessary by
16 them and retrospectively analyzed to grade severity as per SHEA/IDSA criteria. Diagnostic testing for *C.*
17 *difficile* was performed on stool using the standard operating procedure of the clinical diagnostic
18 microbiology laboratory of the McGill University Health Centre. In particular, the acceptance and
19 rejection criteria for stool form is based on the Bristol stool chart [7], such that stool samples acceptable
20 for testing are unformed or liquid (i.e. Bristol types 5 to 7, inclusive), while other stool forms are not
21 tested. Testing is performed on a commercial, FDA-cleared, real-time PCR diagnostic platform targeting
22 the *C. difficile* tcdB gene, with an established negative predictive value of $\geq 99\%$. The medical charts of
23 all subjects were reviewed independently by clinical health care workers to ensure that the criteria of CDI
24 were met (diarrhea and either a stool test positive for *C. difficile* toxins, or colonoscopic or histopathologic
25 findings revealing pseudomembranous colitis [8]), in the absence of any other cause for diarrhea (e.g.
26 iatrogenic, including laxative or feeding solutions), and that positive *C. difficile* results were treated. Of 54
27 patients with CDI, 50 had blood-based parameters measured within 24 hours of their diagnostic stool
28 sample, allowing for assessment of their CDI severity.

29 **S3. Mouse experiment**

30 *C. difficile* Strain VPI10463 (ATCC 43255) was obtained from Cedarlane Labs (Burlington, ON).
31 Bacteria were maintained in meat broth (Sigma-Aldrich, Saint Louis, MO) containing 0.1% (w/v) L-
32 cysteine in an anaerobic jar. For colony counts, *C. difficile*-containing media was serially diluted and
33 streaked onto pre-reduced brain heart infusion plates (BD Biosciences, Mississauga, ON) containing 0.1%

1 L-cysteine. Plates were incubated at 37°C in an anaerobic jar overnight. Eight to eleven-week-old male
2 C57BL/6J mice were given 2.15 µL/mL acetic acid in drinking water one week prior to antibiotic
3 treatment. Six days prior to infection, mice were given an antibiotic cocktail in drinking water for 3 days.
4 The cocktail included metronidazole (0.215 mg/mL), gentamicin (0.035 mg/mL), vancomycin (0.045
5 mg/mL), kanamycin (0.400 mg/mL) and colistin (0.042 mg/mL). Regular water was returned and, 24
6 hours prior to infection, mice were given clindamycin at 32 mg/kg based on the average mouse weight
7 intraperitoneally, in 0.2 mL phosphate buffer saline. A 0.5-cc syringe with 28 G ½ needle was used for
8 injection. Mice were given 0 (control; N=4), 2.32×10^5 (low dose; N=3) or 1.62×10^6 (high dose; N=5)
9 cfu/mouse of *C. difficile* from fresh culture by gavage. Doses were estimated the day of infection using
10 OD600 values and confirmed by colony counts. In previous studies, the high dose generates a lethal
11 infection in mice and the low dose generates a sublethal infection [9]. During pre-treatment and infection,
12 the groups were housed separately to avoid contamination between groups due to coprophagy. Twenty-
13 four hours after infection, mice were euthanized with isoflurane anesthesia followed by CO₂ and
14 subsequent cardiac puncture. The intestine was cut just above the cecum and at the anus. It was placed in
15 clean Petri dish and the fecal matter was pushed out of the cecum and colon and stored in the protein
16 solubilization cocktail described in **Section S1**. Samples were briefly spun (4000 g) and stored at -80°C
17 until ready for use. LC-MS analyses were performed on each mouse in duplicate.

18 **S4. Immunoblotting procedure**

19 Fecal protein (25 µg) was denatured using NuPAGE reducing agent (Thermo Fisher Scientific) and
20 separated on a NuPAGE 4–12% polyacrylamide Bis-Tris gel (200 V, 35 min, MES buffer). Protein bands
21 were transferred to a nitrocellulose membrane using an iBLOT dry transfer system (Thermo Fisher
22 Scientific) for immunoblot detection. Successful transfer was confirmed by Ponceau Red staining.
23 Membranes were blocked (1 h) with 5% bovine serum albumin (BSA) in Tris-buffered saline, 0.05%
24 Tween-20 (TBST), and incubated (4°C, overnight) with primary antibody (in 1% BSA/TBST v/v)
25 followed by a second incubation (1 h) with horseradish peroxidase-conjugated anti-mouse or anti-rabbit
26 IgG (1:15,000 v/v; GE Healthcare, Mississauga, ON). Membranes were developed by Pierce Super Signal
27 West Pico chemiluminescence detection solution (Thermo Fisher Scientific) and exposed to radiography
28 films for visualization. Primary antibody working solutions were prepared based on manufacturer's
29 recommendations.

30

1 **References**

- 2 1. Rockwood AL, Orden SL Van. Ultrahigh-Speed Calculation of Isotope Distributions. *Anal Chem.*
3 1996; 68(13):2027–2030.
- 4 2. Breuer K, Foroushani AK, Laird MR, et al. InnateDB: systems biology of innate immunity and
5 beyond—recent updates and continuing curation. *Nucleic Acids Res.* 2013; 41(D1):D1228–D1233.
- 6 3. Barsky A, Gardy JL, Hancock REW, Munzner T. Cerebral: a Cytoscape plugin for layout of and
7 interaction with biological networks using subcellular localization annotation. *Bioinformatics.*
8 2007; 23(8):1040–1042.
- 9 4. Joshi S, Lewis SJ, Creanor S, Ayling RM. Age-related faecal calprotectin, lactoferrin and tumour
10 M2-PK concentrations in healthy volunteers. *Ann Clin Biochem.* 2010; 47(3):259–263.
- 11 5. Nazario CED, Silva MR, Franco MS, Lanças FM. Evolution in miniaturized column liquid
12 chromatography instrumentation and applications: An overview. *J Chromatogr A.* 2015; 1421:18–
13 37.
- 14 6. Ferrige AG, Seddon MJ, Jarvis S, Skilling J, Aplin R. Maximum entropy deconvolution in
15 electrospray mass spectrometry. *Rapid Commun Mass Spectrom.* 1991; 5(8):374–377.
- 16 7. O’Donnell LJ, Virjee J, Heaton KW. Detection of pseudodiarrhoea by simple clinical assessment
17 of intestinal transit rate. *BMJ.* 1990; 300(6722):439–440.
- 18 8. Cohen SH, Gerding DN, Johnson S, et al. Clinical Practice Guidelines for *Clostridium difficile*
19 Infection in Adults: 2010 Update by the Society for Healthcare Epidemiology of America (SHEA)
20 and the Infectious Diseases Society of America (IDSA). *Infect Control Hosp Epidemiol.* 2010;
21 31(05):431–455.
- 22 9. Winter K, Xing L, Kassardjian A, Ward BJ. Vaccination against *Clostridium difficile* by Use of an
23 Attenuated *Salmonella enterica* Serovar *Typhimurium* Vector (YS1646) Protects Mice from Lethal
24 Challenge. *Infect Immun.* 2019; 87(8): :e00089-19.
- 25 10. Warren CA, Opstal E van, Ballard TE, et al. Amixicile, a Novel Inhibitor of Pyruvate:Ferredoxin
26 Oxidoreductase, Shows Efficacy against *Clostridium difficile* in a Mouse Infection Model.
27 *Antimicrob Agents Chemother.* 2012; 56(8):4103–4111.

1 **S5. Supplemental Figures**

2

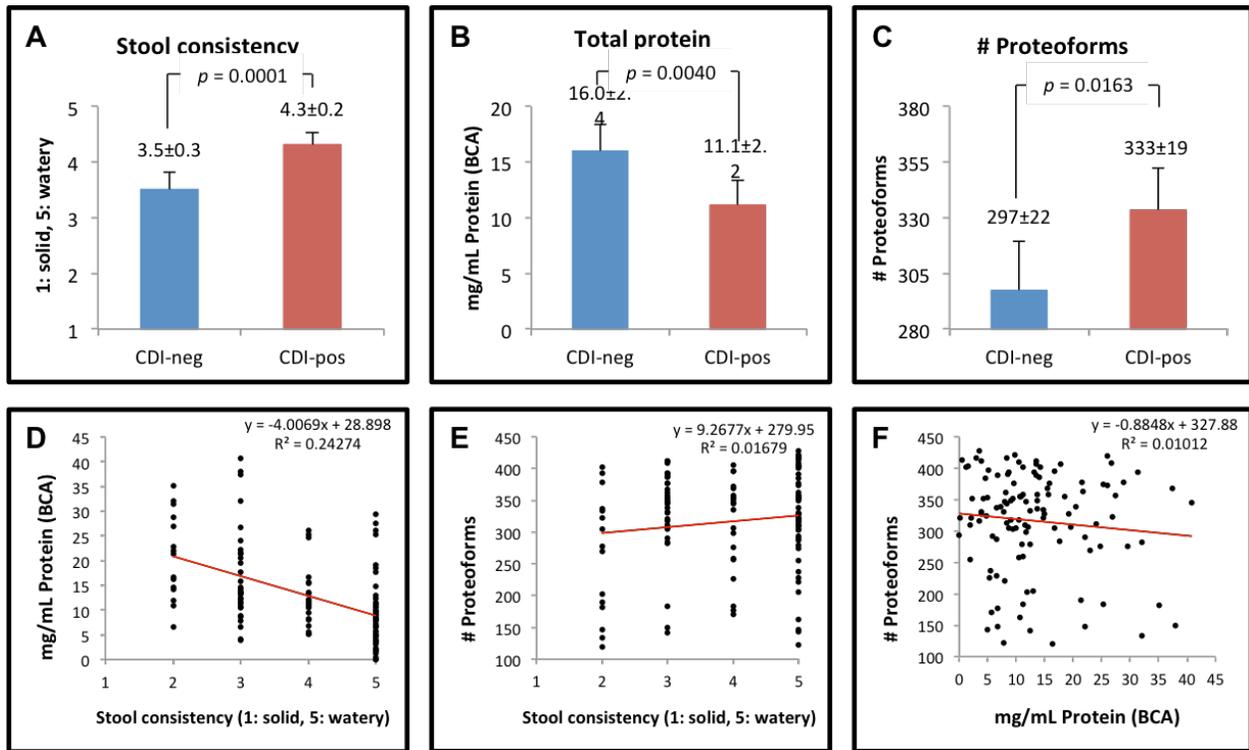


Figure S1. Meta-analysis of stool samples used for biomarker discovery of CDI in humans. **A–C:** Bar charts illustrating stool consistency, total protein concentration (mg/mL) and number of identified proteoforms in CDI-negative and CDI-positive samples. **D–F:** Linear regression plots demonstrating potential correlation between stool consistency, total protein concentration (mg/mL) and number of identified proteoforms in analyzed samples. Error bars represent 95% confidence interval. Significance levels were assessed by Student’s *t*-test. Protein concentration was measured by the BCA protein assay.

3

4

5

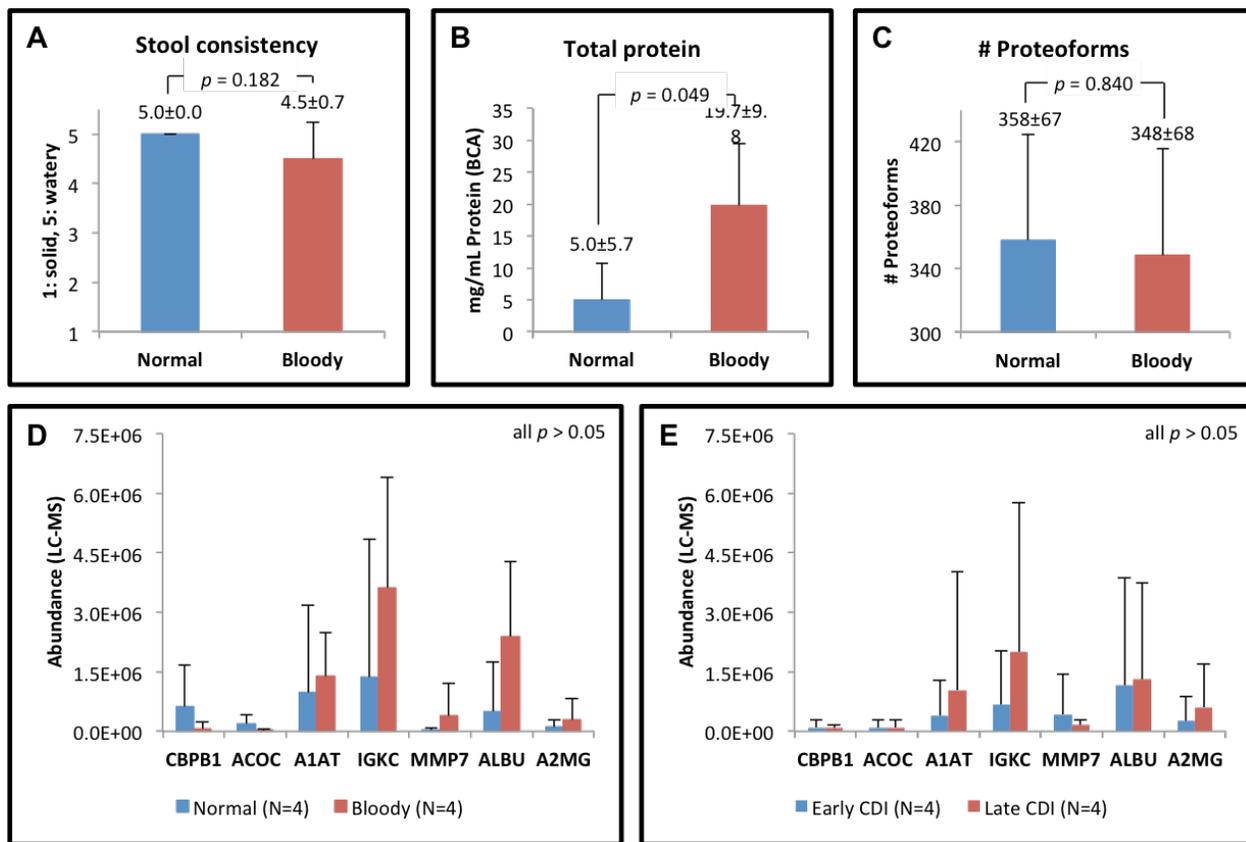
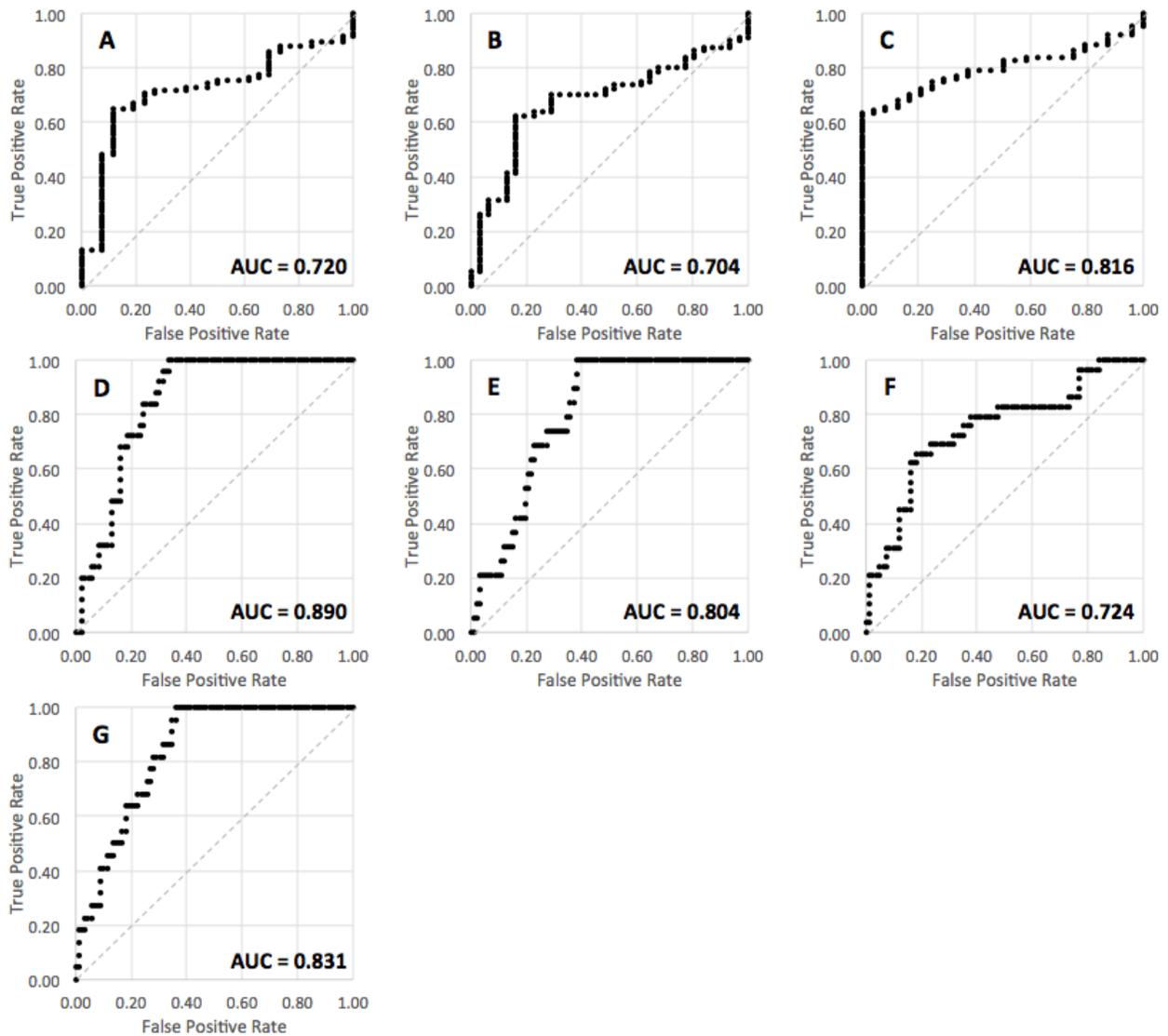


Figure S2. Potential effect of bleeding and disease progression on CDI stool properties. **A–D:** Bar charts illustrating stool consistency, total protein concentration (mg/mL), number of identified proteoforms and biomarker abundances (determined by LC-MS signal intensity) in patients with (N = 4) and without (N = 4) gastrointestinal bleeding. **E:** CDI biomarker abundances in four CDI patients at early-stage vs. late-stage CDI. Error bars represent 95% confidence interval. Significance levels were assessed by Student’s *t*-test. Protein concentration was measured by the BCA protein assay.

1

2

3



1

Figure S3. Receiver operating characteristic (ROC) plots for human fecal CDI biomarkers suggests that the identified biomarkers distinguish CDI-positive from CDI-negative samples with fair to good (AUC = 0.7–0.9) sensitivity and specificity. **A:** CBPB1; **B:** ACOC; **C:** A1AT; **D:** IGKC; **E:** MMP7; **F:** ALBU; **G:** A2MG.

2

3 (Next page):

Figure S4. Cerebral network in the context of subcellular localization illustrating known interactions between *Clostridioides difficile* infection biomarkers identified in humans (marked in red). ALB, A2M, SERPINA1, ACO1, MMP7, IGKC and CPB1 are each involved in 173, 137, 44, 14, 11, 1 and 1 interactions, respectively. Physical association and physical interaction are the most abundant interactions accounting for 248 and 103 interactions, respectively. A2M: alpha-2-macroglobulin; ACO1: cytoplasmic aconitate hydratase; ALB: serum albumin; CPB1: carboxypeptidase B; IGKC: immunoglobulin kappa constant; MMP7: matrix metalloproteinase 7; SERPINA1: alpha-1-antitrypsin.

1 **S6. Supplemental Tables**

2 **Table S1.** Pearson's coefficients of correlation between age, stool consistency (1 for solid to 5 for watery
3 stool), protein content (BCA assay) and number of identified proteoforms in stool.

	Age	Stool consistency	Stool protein content	# Proteoforms
Age	1.00	-0.11	-0.03	0.11
Stool consistency	-0.11	1.00	-0.51	0.14
Stool protein content	-0.03	-0.51	1.00	-0.09
# Proteoforms	0.11	0.14	-0.09	1.00

4

- 1 **Table S2.** Pearson's coefficients of correlation between biomarker level, patient characteristics and
- 2 selected clinical parameters. LC-MS signal intensity was used as a surrogate for biomarker level in stool.
- 3 Moderate ($|r| > 0.45$) and strong ($|r| > 0.90$) correlations were marked in orange and red, respectively.

	CBPB1	ACOC	A1AT	IGKC	MMP7	ALBU	A2MG
CBPB1	1.00	0.46	-0.10	-0.24	-0.18	-0.24	-0.08
ACOC	0.46	1.00	0.03	-0.06	-0.01	-0.04	-0.03
A1AT	-0.10	0.03	1.00	0.25	0.01	0.19	-0.10
IGKC	-0.24	-0.06	0.25	1.00	0.50	0.91	0.28
MMP7	-0.18	-0.01	0.01	0.50	1.00	0.61	0.11
ALBU	-0.24	-0.04	0.19	0.91	0.61	1.00	0.27
A2MG	-0.08	-0.03	-0.10	0.28	0.11	0.27	1.00
Age	-0.05	0.11	-0.03	0.03	-0.05	0.02	-0.03
Sex	0.01	0.10	-0.06	-0.08	-0.01	-0.05	0.13
Stool consistency	0.00	0.08	0.22	0.24	0.08	0.26	-0.02
Stool protein content	-0.01	-0.09	-0.11	-0.19	-0.04	-0.18	-0.08
WBC count	-0.25	-0.16	0.01	0.06	0.07	0.16	-0.08
Serum creatinine	0.00	0.20	-0.03	-0.05	0.15	0.14	-0.11
Serum albumin	0.17	0.28	-0.08	0.18	0.07	-0.11	0.27
Hypotension	-0.15	-0.22	-0.03	-0.04	-0.08	0.06	0.35
Shock	-0.10	-0.15	0.12	-0.01	0.08	0.20	-0.09
Ileus	-0.07	-0.19	-0.03	-0.15	-0.05	-0.02	-0.06
Megacolon	-0.11	-0.16	0.05	-0.08	0.03	0.10	-0.10
Antibiotic treatment	0.15	-0.18	-0.15	-0.41	-0.30	-0.35	-0.13

4

1 **Table S3.** List of significantly ($p < 0.05$) over-represented pathways in *Clostridioides difficile* infected
 2 patients. p -Values were calculated by InnateDB analysis platform and corrected using the Benjamini-
 3 Hochberg correction method.

Pathway Name	Altered genes	Pathway genes	Pathway p-value (corrected)
Response to elevated platelet cytosolic Ca ²⁺	3	83	1.47E-05
Platelet degranulation	3	78	2.44E-05
HDL-mediated lipid transport	2	13	4.03E-05
Platelet activation, signaling and aggregation	3	219	1.10E-04
Lipoprotein metabolism	2	25	1.16E-04
Lipid digestion, mobilization, and transport	2	43	2.32E-04
Hemostasis	3	508	8.41E-04
Degradation of the extracellular matrix	2	116	1.13E-03
Extracellular matrix organization	2	266	5.28E-03
Metabolism of lipids and lipoproteins	2	554	2.00E-02
Metabolism	2	1535	1.22E-01

4
 5 **Table S4.** List of significantly ($p < 0.05$) over-represented pathways in *Clostridioides difficile* infected
 6 mice. p -Values were calculated by InnateDB analysis platform and corrected using the Benjamini-
 7 Hochberg correction method.

Pathway Name	Altered genes	Pathway genes	Pathway p-value (corrected)
Degradation of the extracellular matrix	2	100	3.74E-03
Extracellular matrix organization	2	216	1.35E-02
Innate Immune System	2	450	4.57E-02
Immune System	2	884	1.32E-01
Metabolism	2	1414	2.50E-01

8
 9 **Table S5.** List of protein-protein, protein-DNA and DNA-DNA interactions among the identified fecal
 10 human biomarkers of CDI (**Microsoft Excel file**).

