Discrimination of Escherichia coli and Shigella spp. by Nuclear Magnetic Resonance Based Metabolomic Characterization of Culture Media

Gilles Rautureau, Tony L Palama, Isabelle Canard, Caroline Mirande, Sonia Chatellier, Alex van Belkum, Bénédicte Elena-Herrmann

Discrimination of *Escherichia coli* and *Shigella* spp. by Nuclear Magnetic Resonance-based Metabolomic Characterization of Culture Media

Gilles J. P. Rautureau[†‡], Tony L. Palama[†ø‡], Isabelle Canard[§], Caroline Mirande[§], Sonia Chatellier[§¶], Alex van Belkum[§] and Bénédicte Elena-Herrmann[†⊥*]

† Univ Lyon, CNRS, Université Claude Bernard Lyon 1, ENS de Lyon, Institut des Sciences Analytiques, UMR 5280, 5 rue de la Doua, F-69100 Villeurbanne, France

§ bioMérieux, Innovation Unit – Microbiology Research, 38390 La Balme-les-Grottes, France

⊥ Univ Grenoble Alpes, CNRS, INSERM, IAB, Allée des Alpes, 38000 Grenoble, France

* to whom correspondence should be addressed: Bénédicte Elena-Herrmann (benedicte.elena@univ-grenoble-alpes.fr)

KEYWORDS

Bacterial identification, *Shigella*, *E. coli*, NMR, metabolomics, metabolic footprint, exo-metabolome

Dysentery is a major health threat that dramatically impacts childhood morbidity and mortality in developing countries. Various pathogenic agents cause dysentery such as Shigella spp. and Escherichia coli, which are very closely related if not identical species. Sensitive and precise detection and identification of the infectious agent is important to target the best therapeutic strategy but the differential diagnosis of these two groups remains a challenge using conventional methods. Here we present a nuclear magnetic resonance (NMR) based multivariate classification model employing bacterial metabolic footprints in post-culture growth media with remarkable segregation

capability, including the discrimination of lactose negative E. coli and Shigella species. Our results confirm the potential of metabolomic markers in the field of bacterial identification for the distinction of even very closely related species.

Shigella spp. and *Escherichia coli* belong to the family of the Enterobacteriaceae. They represent species with a very high genetic relatedness [1-2]. In fact, they could even be classified as one distinctive species in the genus *Escherichia* [3-5]. Many *Shigella* isolates are responsible for shigellosis (bacillary dysentery) that can lead to life-threatening dysentery, with a global impact on childhood morbidity and mortality [6]. Clinical isolates of *E. coli* species can be commensal or pathogenic, with some isolates such as entero-invasive *E. coli* (EIEC) causing illnesses similar to shigellosis. Hence, most *E. coli* are commensals normally found in the human gut flora, whereas *Shigella* spp. are generally considered real pathogens. *Shigella* spp. are considered distinct from *E. coli* mainly from a clinical perspective.

The *Shigella* and *E. coli* initial differentiation based on phenotypic and biochemical tests [7] has rapidly shown limitations and remains a challenge for clinical laboratories [4, 8]. The triple sugar iron (TSI) test separates *E. coli* and *Shigella spp* on the basis of acidification of a pH indicator-containing growth medium, but with poor sensitivity and specificity regarding discrimination of *E. coli* and *Shigella* from other species [9]. 16S rRNA gene sequencing cannot differentiate these species and identification relies on a limited number of phenotypic and biochemical characteristics which may still not correctly identify all isolates [10]. In particular, a sub-group of *E. coli* isolates that does not ferment lactose, usually termed lactose negative, is biochemically very similar to *Shigella* spp. and consequently difficult to separate from *Shigella* [8]. Matrix-assisted laser desorption/ionization-time of flight mass spectrometry (MALDI-TOF MS) has become the microbial identification method of choice over the past few years [11]. However, *E. coli* and Shigella usually do not separate with the conventional MS standard operating procedures and additional classification methods and algorithms have recently emerged to tackle the distinction between these closely related species [12-13]. Molecular tests do exist to identify these species and pathovars but can be expensive, preventing their use in routine

settings [3, 14]. Nowadays serological identification remains a key test for the diagnostic of infections by *Shigella* but certain species remain non-serotypeable [15]. Recent advances in next-generation sequencing open a more comprehensive and high-resolution analysis of their genome differences and evolution [16-18]. However, they still need to be translated into easy-to-use diagnostic tests available in routine settings.

NMR is an intrinsically quantitative, non-destructive and information-rich technique for the determination of molecular structures that is uniquely suited for the analysis of complex mixtures such as bio-fluids. Significant advances in NMR technology and automation have positioned this technique as a method of choice for metabolomic investigations [19-20] that aim at comprehensive and quantitative analysis of metabolites present in biological samples [21]. Various metabolomic approaches have been adapted to study bacteria [22] and have been successfully used to investigate intra- and extracellular bacterial imprints [23-27], metabolic pathways [28-30], or antibiotic modes of action [31-32]. NMR-based metabolomic approaches have also been proposed for targeted bacterial identification and discrimination [33-37]. We notably established a method to rapidly discriminate and identify bacterial species that relies on untargeted metabolic profiling of bacterial culture supernatants [38]. In the present study, we evaluate the potential of this approach to distinguish the closely related, and hard to discriminate *Shigella* and *E. coli* species. The various groups of bacteria studied are well discriminated from supervised multivariate data analyses of their metabolic footprints. We present a robust classification signature based on a limited set of metabolite concentrations, which may be measurable by other analytical techniques, thus allowing application of our approach to microbial clinical diagnosis.

MATERIALS AND METHODS

Bacterial samples. We studied a total of 144 samples of bacterial growth media collected after 1.5 hour (T_e) of culture initiation. These samples correspond to 48 cultivated strains from the bioMérieux collection, with 3 independent cultures (biological replicates) each, which

classify into 3 main groups for the present study: *Shigella* species, *E. coli* lactose (+) and lactose (-) strains (*i.e.* lactose fermenting and non-fermenting isolates, respectively), with 16 strains per group. The group of *Shigella* species includes sets of *S. boydii*, *S. flexneri*, and *S. sonnei* strains. The detailed list of the 48 bacterial strains is provided in the supplementary material (Table S1, ESI). T_e corresponds to the middle of the exponential growth phase as was determined previously when applying the bacterial growth conditions used in this study. A blind investigation was conducted on an additional cohort of 60 *E. coli* lactose (-) and *Shigella* growth media samples, as independent validation of this work.

Bacterial culture and sample preparation. Clinical isolates of bacterial species, stored at -80°C, were thawed, and pre-cultured on Columbia agar + Sheep blood 5% (2 consecutive sub-cultures). Cultures were then realized at 37°C in Mueller-Hinton (MH) (5 mL) liquid medium until T_e, from a suitable inoculum calibrated to obtain the same numbers of bacteria for all strains at T_e. This corresponds to a concentration at T_0 between 2 and 2.5 MacFarland units (McF). One mL of the culture medium was collected at time T_0 and T_e in Eppendorf tubes and centrifuged for 10 min at 4000g. Supernatants were collected in Eppendorf tubes and stored at -80°C until NMR analysis.

NMR analysis. The samples for NMR were prepared as follows: 60 µL of phosphate buffer (1.25M KH_2PO_4, 2mM NaN_3 and 0.1% trimethylsilyl propionate-2,2,3,4-d_4 (TMSP) in D_2O, pH=7.4) were added to 540 µL of bacterial culture supernatants and mixed thoroughly. Finally, 550 µL were then transferred to 5 mm NMR tubes. Samples were kept at 4 °C until analysis. All NMR spectra were acquired at 27°C on a 600 MHz Bruker Avance III NMR spectrometer equipped with a 5 mm TCI cryoprobe. A SampleJet auto-sampler enabled high throughput data acquisition. A standard ^{1}H 1D NMR NOESY (nuclear Overhauser effect spectroscopy) experiment with z-gradient and water pre-saturation (Bruker pulse program *noesygppr1d*) was carried out on each sample, with 128 transient free induction decays (FID) co-added, an acquisition time of 2 s and a spectral width of 20 ppm. The relaxation delay was set to 4 s, the NOESY mixing time was 10 ms and the 90° pulse length was automatically determined for

each sample (around 14.5 µs). The total acquisition time for each spectrum was 13 min 48 sec. We note that ^{1}H CPMG NMR spectra recorded for these types of samples (data not shown) are fairly identical to the ^{1}H NOESY fingerprints exploited in this study.

Data processing. Prior Fourier transform, all NMR FIDs were multiplied by an exponential function corresponding to a 0.3 Hz line-broadening factor. ^{1}H-NMR spectra were phased and referenced to the TMSP signal (-0.016 ppm) using Topspin 3.2 (Bruker GmbH, Rheinstetten, Germany). Extraction of a data matrix for multivariate statistical analysis from the ^{1}H NMR profiles was done using the software AMIX (Bruker GmbH). Spectra were integrated and bucketed from 0.3 to 10 ppm in steps of 0.01 ppm, excluding the region of residual water signal (5.10 to 4.50 ppm), and normalized to total intensity. The resulting data matrix contained 970 NMR variables.

Metabolite identification and quantification. NMR peak assignments were obtained from comparisons with metabolites databases such as HMDB, Chenomx NMR Suite (Chenomx Inc., Edmonton, Canada) and BBIOREFCODE (Bruker GmbH) and verified with homonuclear and heteronuclear 2D NMR experiments (^{1}H-^{13}C HSQC, ^{1}H-^{13}C HMBC, ^{1}H-^{1}H TOCSY and J-resolved experiments). Individual metabolite concentrations were determined by manual fitting of the proton resonance lines for the compounds available in the Chenomx database. The line-width used for deconvolution with the reference database was adjusted to the width of one component of the alanine doublet.

Multivariate data analysis. Principal component analysis (PCA), O-PLS discriminant analysis (O-PLS-DA) and Receiver of Operator (ROC) were performed using SIMCA-P 15 (Umetrics, Umea, Sweden) with centered variables (no scaling). O-PLS-DA analyses [39-40] were used to build predictive sample classification models based on either 0.01 ppm bucketed NOESY NMR spectra (950 variables) or metabolite concentrations (26 metabolites). Results were visualized on score plots, corresponding to sample projections onto the predictive axis and the first orthogonal component of the model, and the associated loadings plot. The optimal number of

orthogonal components was selected using a 7-fold cross validation procedure. The R^2 and Q^2 parameters were computed as a measure of the goodness of fit and prediction, i.e. the explained and predicted variances, respectively. The O-PLS-DA models were validated using permutations under the null hypothesis (1000 times); for each permutated classification labels, R^2 and Q^2 were recalculated and compared to the original ones, their decrease indicating the good quality of the model [41]. For each O-PLS-DA model, variable importance values in the projection (VIP) were computed [42]. Receiver Operating Characteristic (ROC) curves [43] and corresponding area under the curve (AUC) were calculated for individual metabolites based on univariate testing, or for multivariate analysis based on O-PLS-DA cross-validation. Subsets of discriminant metabolites were obtained using the feature selection algorithm in Metaboanalyst 4.0 [44], a multivariate exploratory ROC analysis based on random sub-sampling and cross-validation performance of PLS-DA.

RESULTS AND DISCUSSION

E. coli and Shigella species NMR metabolic footprints in culture media.

Bacteria were cultured in a classical, rich Mueller Hinton medium during 1.5 hour, a delay sufficient to reach exponential growth for both *Shigella* spp. and *E. coli* species. As metabolites were consumed or produced, a complex metabolic footprint was detected by untargeted NMR analysis. Well-resolved ^{1}H NMR metabolic profiles were obtained for each sample of bacterial culture medium (Fig. 1a). The spectra displayed typical sharp lines corresponding to small metabolites, overlaid with broad signals at baseline from lipids or larger proteins, which appeared as minor contributors in the case of our culture supernatants. Detailed analysis of the ^{1}H 1D as well as additional two-dimensional ^{1}H-^{1}H and ^{1}H-^{13}C 2D NMR correlation spectra recorded for a subset of representative samples delivered the identification of 43 metabolites (Table S2, ESI) that belong to a variety of biochemical classes (amino-acids, sugars, nucleotides and metabolic intermediates). The library of 138 ^{1}H 1D NMR footprints (6 samples

were excluded due to poor quality of spectra) was then exploited by multivariate statistical analysis.

PCA of the bucketed NMR profiles was first used to evaluate the dataset homogeneity and the potential of unsupervised sample class discrimination concerning *Shigella* spp. and *E. coli* footprints, following the approach proposed in our previous study [38]. No strong outliers were identified on the PCA score plot (Fig. 1b), confirming the reproducibility of our experimental approach. A straightforward species-based discrimination could be observed on this unsupervised model between *E. coli lactose (+)* and *Shigella* spp. but not between *E. coli lactose (-)* and *Shigella*. An independent PCA model of *E. coli lactose (+)* and *Shigella* samples is presented in Fig. 1c to highlight the corresponding discrimination. The first two principal components of this model explain 57.1% of the variance within the dataset, and this model soundly represents the data structure as attested by high values of goodness-of-fit parameters R^2 = 0.985 and Q^2= 0.915, related respectively to the variance explained and predicted by the model. In contrast, *E. coli lactose (-)* and *Shigella* could not be discriminated from unsupervised analysis. Altogether these results confirm our previous observations that NMR analysis of culture supernatants is adapted to distinguish bacterial species [38], but also illustrate, at the exo-metabolome level, the detailed phenotype resemblances between *E. coli lactose (-)* and *Shigella* species that lead to identification issues using regular analytical techniques.

Discrimination of E. coli lactose (-) and Shigella spp. using O-PLS-DA.

To address the discrimination challenge between *E. coli lactose (-)* and *Shigella* species, a supervised analysis of the NMR metabolic profiles was conducted by O-PLS-DA [39-40]. While the principal objective of this study was the global binary discrimination of *E. coli lactose (-)* and *Shigella* species, supervised O-PLS-DA multivariate analysis of the ^{1}H NMR bacterial culture footprints was able to robustly distinguish the four types of lactose negative species investigated, i.e. *E. coli lactose (-)*, *S. boydii*, *S. flexneri*, and *S. sonnei*, as shown in Figure 2. Corresponding model validation from permutations of the Y values is provided in the supplementary material (Fig. S1, ESI). Subsequently, focusing on the 2-class discrimination of

E. coli lactose (-) vs. *Shigella* species, we obtained a robust predictive O-PLS-DA model ($R^2(X)$= 0.861, $R^2(Y)$=0.962 and Q^2 = 0.893) based on the full NMR fingerprint (Fig. S2, ESI).

This analysis, carried out using the 950 NMR variables to exploit the full dynamic range of spectral information, was then repeated on a set of 26 quantified metabolites to broaden the applicability of our study. Indeed, metabolite concentrations could potentially be determined without NMR using a wide range of biochemical methods. Here, metabolites concentrations were obtained from NMR spectrum deconvolution of each sample of culture supernatant for *E. coli lactose (-)* and *Shigella* spp. (91 samples). A significant discrimination was obtained from the O-PLS-DA score plot between *E. coli lactose (-)* and *Shigella* samples (Fig. 3a), associated with high values of goodness-of-fit model parameters ($R^2(X)$= 0.99, $R^2(Y)$= 0.75 and Q^2 = 0.639). Robustness of the model was validated by permutation testing (1000 permutations) under the null hypothesis, showing a clear decrease of R^2 and Q^2 with the correlation between original and permuted class information in the Y matrix (Fig. 3b). The reliability of our multivariate model was also assessed by a *p*-value of 7.53 $x10^{-11}$ from analysis of variance (CV-ANOVA). Out of the 26 quantified metabolites, 7 metabolites (succinate, acetate, aspartate, formate, lysine, propionate and threonine) appeared to have a significant contribution to the statistical model as shown by Variable Importance in Projection for independent variables (VIP) values superior to one. Most of these metabolites were differently secreted in the medium by the two species (succinate, acetate, formate, and propionate), lysine and threonine were only secreted by *Shigella* spp. and one (aspartate) was consumed by both at different levels.

Combinations of small sets of metabolites provide significant discrimination between E. coli lactose (-) and Shigella species

Classification models that rely on only a few metabolites have the potential to be easily implemented using analytical data from various chemical and biochemical platforms, and therefore to be globally more accessible to the scientific and diagnostic communities. We thus evaluated the classification potential of individual metabolites, as well as combinations of small

sets of metabolites. Individual metabolites were chosen according to highest VIP values described above, for which consumption or secretion patterns differ between *E. coli* lactose (-) and *Shigella* spp. (succinate, acetate, aspartate, formate, lysine, propionate and threonine). Mean concentrations and standard deviations for individual metabolites are reported in the supplementary material (Table S3 and Fig. S3, ESI). Combination of metabolites were obtained using an objective feature selection method based on multivariate exploratory ROC analysis [44]. Multivariate models classification performance were evaluated based on AUC values of the ROC curves [43] constructed from O-PLS-DA cross-validation, as well as O-PLS-DA Q^2 values. Results are summarized in Table 1. For the entire 26 metabolites' dataset, we obtained a remarkable AUC of 0.9995, confirming the powerful classification ability of the model. As expected, individual metabolites displayed modest classification capacities with AUCs under 0.77. We obtained higher AUCs of 0.86 and 0.97 when using limited sets of 5 or 10 metabolites, respectively. These results show that simple combinations of discriminating metabolites can be designed to derive effective classification models between *E. coli* lactose (-) and *Shigella* spp. While individual metabolites fail to provide any satisfying discrimination between these closely related types of bacteria, multivariate models based on five, or ten metabolites could be exploited, depending on the analytical capacity and classification tolerance, retaining a diagnostic potential close to models constructed with the full ensemble of 26 quantified metabolites.

Independent validation of E. coli lactose (-) and Shigella multivariate discrimination

A second, independent series of 60 bacterial cultures was investigated to validate the classification model obtained for *E. coli lactose (-)* and *Shigella* species discrimination. NMR analysis, subsequent metabolite quantification and multivariate statistics were conducted without knowledge of the samples' class membership, which was revealed to the operator at readout stage only. One sample was discarded as an outlier from PCA analysis. Predicted scores (Fig. S4a, ESI), *i.e.* projection of this independent dataset onto the model of Fig. 3, show a clear separation of the two classes of samples along the predictive latent variable

(horizontal axis). Yet, individual class membership prediction is not accurate, with a clear shift of the whole validation cohort towards the left side of the diagram. This systematic error conveys a clear batch effect associated with the use of different batches of bacterial growth media for the two independent studies, whereas the good separation of the two classes along the predictive component stresses the robustness of this latent variable towards *Shigella* vs. *E. coli* lac (-) discrimination. When considering the above selected subset of 5 metabolites only, OPLS-DA blind prediction of individual class membership is noticeably enhanced (Fig. S4b-c, ESI), and predicted samples project onto their counterparts from the model cohort. Despite the presence of batch differences between the two cohorts, an OPLS-DA analysis carried out on the merged datasets of 26 quantified metabolites (Fig. S5, ESI), provides a robust discrimination of *Shigella* vs. *E. coli* lac (-) growth media, with homogenous groups of samples for each class, and the same set of main VIPs contributing to the model as the one determined from Fig. 3c in the initial cohort. In this work, the diversity of strains studied (16 strains per group) largely accounted for intra-species variability . Classification models based on larger cohorts will be certainly required to cover further variability among bacterial growth media sources, and generalize the proposed approach to robust classification of *E. coli* lac (-) and *Shigella* culture media of unknow origin.

CONCLUSION

This study demonstrates that untargeted proton NMR metabolomics of bacterial culture supernatants can robustly classify *E. coli* lactose (+) and *Shigella* spp. by unsupervised data analysis, and *E. coli* lactose (-) and *Shigella* spp. using supervised approaches. Lactose negative species have traditionally shown to be difficult to discriminate using classical (biochemical) techniques but also including the more modern ones such as mass spectrometry and nucleic acid amplification and sequencing. Statistical evaluation of ^{1}H NMR profiles of the bacterial exo-metabolome, as detected from culture media footprints, provides a robust and

predictive species classification between *E. coli* lactose (-) and *Shigella* species. Simple classification models, based on limited sets (5 or 10 metabolites) of easily measurable metabolite concentrations from supernatants of bacterial cultures in Mueller Hinton medium, delivered significant discrimination for *E. coli* lactose (-) vs. *Shigella* spp. Our approach demonstrates the relevance of NMR footprinting in the field of bacterial identification and discrimination, even in the case of traditionally challenging bacterial species discrimination. We stress however that the proposed classification based on small sets of metabolites does not require NMR as an exclusive analytical workhorse, and can potentially rely on other technologies or biochemical methods for metabolite quantification. The proposed approach opens up new prospects for accurate and cost-effective microbial clinical diagnosis.

AUTHOR INFORMATION

Present address

ø T.L.P.: Université Paris 13, Sorbonne Paris Cité, Laboratoire CSPBAT, Team NBD, CNRS (UMR 7244), UFR-SMBH, 74 rue Marcel Cachin, 93017 Bobigny, France

¶ S.C.: MacoPharma, 200 Chaussée Fernand Forest, 59200 Tourcoing, France

Author Contributions

‡ G.J.P.R. and T.L.P. contributed equally to this work.

The manuscript was written through contributions of all authors. All authors have given approval to the final version of the manuscript.

Notes

I.C., C.M. and A.vB. are bioMerieux employees. T.L.P. was employed by the CNRS with funding from BioMerieux. BioMerieux develops, markets and sells diagnostics tests in the infectious disease domain. The company had no direct influence on test design and data interpretation. S.C. is currently an employee of MacoPharma.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website. List of the 48 bacterial strains cultured and analyzed by NMR spectroscopy; list of 43 identified metabolites from NMR analysis of bacterial culture media; O-PLS-DA permutations testing for *E. coli* and individual *Shigella* species (*S. sonnei*, *S. flexneri* and *S. boydii*) discrimination; 2-classes O-PLS-DA model discriminating *E. coli* lac (-) and *Shigella* species based on 950 NMR variables; Individual metabolite concentrations in *E. coli* lac (-) and *Shigella* strains; Models validation using an independent cohort; O-PLS-DA model discriminating *E. coli* lac (-) and *Shigella* species for the merged test and validation cohorts.

REFERENCES

1. Chaudhuri, R. R.; Henderson, I. R., The evolution of the Escherichia coli phylogeny. *Infect Genet Evol* **2012,** *12* (2), 214-26. DOI: 10.1016/j.meegid.2012.01.005.

2. Pettengill, E. A.; Pettengill, J. B.; Binet, R., Phylogenetic Analyses of Shigella and Enteroinvasive Escherichia coli for the Identification of Molecular Epidemiological Markers: Whole-Genome Comparative Analysis Does Not Support Distinct Genera Designation. *Front Microbiol* **2016,** *6*, 1573-1573. DOI: 10.3389/fmicb.2015.01573.

3. van den Beld, M. J. C.; Friedrich, A. W.; van Zanten, E.; Reubsaet, F. A. G.; Kooistra-Smid, M. A. M. D.; Rossen, J. W. A., Multicenter evaluation of molecular and culture-dependent diagnostics for Shigella species and Entero-invasive Escherichia coli in the Netherlands. *J Microbiol Meth* **2016,** *131*, 10-15. DOI: 10.1016/j.mimet.2016.09.023.

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

5. Zuo, G.; Xu, Z.; Hao, B., Shigella strains are not clones of Escherichia coli but sister species in the genus Escherichia. *Genomics Proteomics Bioinformatics* **2013,** *11* (1), 61-65. DOI: 10.1016/j.gpb.2012.11.002.

6. Anderson, M.; Sansonetti, P. J.; Marteyn, B. S., Shigella Diversity and Changing Landscape: Insights for the Twenty-First Century. *Front Cell Infect Microbiol* **2016,** *6*, 45. DOI: 10.3389/fcimb.2016.00045.

7. Wallace, H. A.; Jacobson, A., Shigella. In *Bacteriological Analytical Manual* [Online] 2001. https://www.fda.gov/food/laboratory-methods-food/bam-shigella.

8. Devanga Ragupathi, N. K.; Muthuirulandi Sethuvel, D. P.; Inbanathan, F. Y.; Veeraraghavan, B., Accurate differentiation of Escherichia coli and Shigella serogroups: challenges and strategies. *New Microbes New Infect* **2018,** *21*, 58-62. DOI: 10.1016/j.nmni.2017.09.003.

9. Stager, C. E.; Erikson, E.; Davis, J. R., Rapid method for detection, identification, and susceptibility testing of enteric pathogens. *J Clin Microbiol* **1983,** *17* (1), 79-84.

10. Khot, P. D.; Couturier, M. R.; Wilson, A.; Croft, A.; Fisher, M. A., Optimization of matrix-assisted laser desorption ionization-time of flight mass spectrometry analysis for bacterial identification. *J Clin Microbiol* **2012,** *50* (12), 3845-3852. DOI: 10.1128/JCM.00626-12.

11. Lavigne, J. P.; Espinal, P.; Dunyach-Remy, C.; Messad, N.; Pantel, A.; Sotto, A., Mass spectrometry: a revolution in clinical microbiology? *Clin Chem Lab Med* **2013,** *51* (2), 257-70. DOI: 10.1515/cclm-2012-0291.

12. Paauw, A.; Jonker, D.; Roeselers, G.; Heng, J. M. E.; Mars-Groenendijk, R. H.; Trip, H.; Molhoek, E. M.; Jansen, H. J.; van der Plas, J.; de Jong, A. L.; Majchrzykiewicz-Koehorsta, J. A.; Speksnijder, A. G. C. L., Rapid and reliable discrimination between Shigella species and

Escherichia coli using MALDI-TOF mass spectrometry. *Int J Med Microbiol* **2015,** *305* (4-5), 446-452. DOI: 10.1016/j.ijmm.2015.04.001.

13. Khot, P. D.; Fisher, M. A., Novel approach for differentiating Shigella species and Escherichia coli by matrix-assisted laser desorption ionization-time of flight mass spectrometry. *J Clin Microbiol* **2013,** *51* (11), 3711-6. DOI: 10.1128/jcm.01526-13.

14. Lobersli, I.; Wester, A. L.; Kristiansen, A.; Brandal, L. T., Molecular Differentiation of Shigella Spp. from Enteroinvasive E. Coli. *Eur J Microbiol Immunol (Bp)* **2016,** *6* (3), 197-205. DOI: 10.1556/1886.2016.00004.

15. Muthuirulandi Sethuvel, D. P.; Devanga Ragupathi, N. K.; Anandan, S.; Walia, K.; Veeraraghavan, B., Molecular diagnosis of non-serotypeable Shigella spp.: problems and prospects. *J Med Microbiol* **2017,** *66* (2), 255-257. DOI: 10.1099/jmm.0.000438.

16. The, H. C.; Thanh, D. P.; Holt, K. E.; Thomson, N. R.; Baker, S., The genomic signatures of Shigella evolution, adaptation and geographical spread. *Nat Rev Microbiol* **2016,** *14* (4), 235-250. DOI: 10.1038/nrmicro.2016.10.

17. Chattaway, M. A.; Schaefer, U.; Tewolde, R.; Dallman, T. J.; Jenkins, C., Identification of Escherichia coli and Shigella Species from Whole-Genome Sequences. *J Clin Microbiol* **2017,** *55* (2), 616-623. DOI: 10.1128/JCM.01790-16.

18. Camelena, F.; Birgy, A.; Smail, Y.; Courroux, C.; Mariani-Kurkdjian, P.; Le Hello, S.; Bonacorsi, S.; Bidet, P., Rapid and Simple Universal Escherichia coli Genotyping Method Based on Multiple-Locus Variable-Number Tandem-Repeat Analysis Using Single-Tube Multiplex PCR and Standard Gel Electrophoresis. *Appl Environ Microb* **2019,** *85* (6). DOI: 10.1128/aem.02812-18.

19. Markley, J. L.; Bruschweiler, R.; Edison, A. S.; Eghbalnia, H. R.; Powers, R.; Raftery, D.; Wishart, D. S., The future of NMR-based metabolomics. *Curr Opin Biotechnol* **2017,** *43*, 34-40. DOI: 10.1016/j.copbio.2016.08.001.

20. Nagana Gowda, G. A.; Raftery, D., Recent Advances in NMR-Based Metabolomics. *Anal Chem* **2017,** *89* (1), 490-510. DOI: 10.1021/acs.analchem.6b04420.

21. Lindon, J. C.; Holmes, E.; Nicholson, J. K., So whats the deal with metabonomics? Metabonomics measures the fingerprint of biochemical perturbations caused by disease, drugs, and toxins. *Anal Chem* **2003,** *75* (17), 384A-391A. DOI: Doi 10.1021/Ac031386+.

22. Grivet, J. P.; Delort, A. M., NMR for microbiology: In vivo and in situ applications. *Prog Nucl Mag Res Sp* **2009,** *54* (1), 1-53. DOI: 10.1016/j.pnmrs.2008.02.001.

23. Kusch, H.; Engelmann, S., Secrets of the secretome in Staphylococcus aureus. *Int J Med Microbiol* **2014,** *304* (2), 133-141. DOI: 10.1016/j.ijmm.2013.11.005.

24. Meyer, H.; Liebeke, M.; Lalk, M., A protocol for the investigation of the intracellular Staphylococcus aureus metabolome. *Anal Biochem* **2010,** *401* (2), 250-259. DOI: 10.1016/j.ab.2010.03.003.

25. Resmer, K. L.; White, R. L., Metabolic footprinting of the anaerobic bacterium Fusobacterium varium using H-1 NMR spectroscopy. *Mol Biosyst* **2011,** *7* (7), 2220-2227. DOI: 10.1039/c1mb05105a.

26. Ye, Y. F.; Zhang, L. M.; An, Y. P.; Hao, F. H.; Tang, H. R., Nuclear Magnetic Resonance for Analysis of Metabolite Composition of Escherichia coli. *Chinese J Anal Chem* **2011,** *39* (8), 1186-1194. DOI: 10.3724/SP.J.1096.2011.01186.

27. Zandomeneghi, G.; Ilg, K.; Aebi, M.; Meier, B. H., On-Cell MAS NMR: Physiological Clues from Living Cells. *J Am Chem Soc* **2012,** *134* (42), 17513-17519. DOI: 10.1021/ja307467p.

28. Bartholomeusz, T. A.; Molinie, R.; Mesnard, F.; Robins, R. J.; Roscher, A., Optimisation of 1D and 2D in vivo H-1 NMR to study tropane alkaloid metabolism in Pseudomonas. *Cr Chim* **2008,** *11* (4-5), 457-464. DOI: 10.1016/j.crci.2007.09.009.

29. Meier, S.; Jensen, P. R.; Duus, J. O., Real-time detection of central carbon metabolism in living Escherichia coli and its response to perturbations. *Febs Lett* **2011,** *585* (19), 3133-3138. DOI: 10.1016/j.febslet.2011.08.049.

30. Sadykov, M. R.; Zhang, B.; Halouska, S.; Nelson, J. L.; Kreimer, L. W.; Zhu, Y. F.; Powers, R.; Somerville, G. A., Using NMR Metabolomics to Investigate Tricarboxylic Acid Cycle-dependent Signal Transduction in Staphylococcus epidermidis. *J Biol Chem* **2010,** *285* (47), 36616-36624. DOI: 10.1074/jbc.M110.152843.

31. Hoerr, V.; Duggan, G. E.; Zbytnuik, L.; Poon, K. K. H.; Große, C.; Neugebauer, U.; Methling, K.; Löffler, B.; Vogel, H. J., Characterization and prediction of the mechanism of action of antibiotics through NMR metabolomics. *BMC Microbiol* **2016,** *16*, 82-82. DOI: 10.1186/s12866-016-0696-5.

32. Vincent, I. M.; Ehmann, D. E.; Mills, S. D.; Perros, M.; Barrett, M. P., Untargeted Metabolomics To Ascertain Antibiotic Modes of Action. *Antimicrob Agents Chemother* **2016,** *60* (4), 2281-2291. DOI: 10.1128/AAC.02109-15.

33. Bourne, R.; Himmelreich, U.; Sharma, A.; Mountford, C.; Sorrell, T., Identification of Enterococcus, Streptococcus, and Staphylococcus by multivariate analysis of proton magnetic resonance spectroscopic data from plate cultures. *J Clin Microbiol* **2001,** *39* (8), 2916-2923. DOI: Doi 10.1128/Jcm.39.8.2916-2923.2001.

34. Bundy, J. G.; Willey, T. L.; Castell, R. S.; Ellar, D. J.; Brindle, K. M., Discrimination of pathogenic clinical isolates and laboratory strains of Bacillus cereus by NMR-based metabolomic profiling. *Fems Microbiol Lett* **2005,** *242* (1), 127-136. DOI: 10.1016/j.femsle.2004.10.048.

35. Gupta, A.; Dwivedi, M.; Mahdi, A. A.; Khetrapal, C. L.; Bhandari, M., Broad Identification of Bacterial Type in Urinary Tract Infection Using H-1 NMR Spectroscopy. *J Proteome Res* **2012,** *11* (3), 1844-1854. DOI: 10.1021/pr2010692.

36. Gupta, A.; Dwivedi, M.; Mahdi, A. A.; Roy, R.; Elhandari, M.; Khetrapal, C. L., Metabonic approach to the diagnosis of proteus mirabilis: H-1 NMR spectroscopic study. *J Urology* **2008,** *179* (4), 82-82. DOI: Doi 10.1016/S0022-5347(08)60239-6.

37. Lussu, M.; Camboni, T.; Piras, C.; Serra, C.; Del Carratore, F.; Griffin, J.; Atzori, L.; Manzin, A., 1H NMR spectroscopy-based metabolomics analysis for the diagnosis of symptomatic E. coli-associated urinary tract infection (UTI). *BMC Microbiol* **2017,** *17* (1), 201. DOI: 10.1186/s12866-017-1108-1.

38. Palama, T. L.; Canard, I.; Rautureau, G. J. P.; Mirande, C.; Chatellier, S.; Elena-Herrmann, B., Identification of bacterial species by untargeted NMR spectroscopy of the exo-metabolome. *Analyst* **2016,** *141* (15), 4558-4561. DOI: 10.1039/c6an00393a.

39. Fonville, J. M.; Richards, S. E.; Barton, R. H.; Boulange, C. L.; Ebbels, T. M. D.; Nicholson, J. K.; Holmes, E.; Dumas, M. E., The evolution of partial least squares models and related chemometric approaches in metabonomics and metabolic phenotyping. *J Chemometr* **2010,** *24* (11-12), 636-649. DOI: 10.1002/cem.1359.

40. Trygg, J.; Wold, S., Orthogonal projections to latent structures (O-PLS). *J Chemometr* **2002,** *16* (3), 119-128. DOI: 10.1002/cem.695.

41. Westerhuis, J. A.; Hoefsloot, H. C. J.; Smit, S.; Vis, D. J.; Smilde, A. K.; van Velzen, E. J. J.; van Duijnhoven, J. P. M.; van Dorsten, F. A., Assessment of PLSDA cross validation. *Metabolomics* **2008,** *4* (1), 81-89. DOI: 10.1007/s11306-007-0099-6.

42. Chong, I.-G.; Jun, C.-H., Performance of some variable selection methods when multicollinearity is present. *Chemometr Intell Lab Syst* **2005,** *78* (1–2), 103-112. DOI: 10.1016/j.chemolab.2004.12.011.

43. Hanley, J. A.; McNeil, B. J., The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* **1982,** *143* (1), 29-36. DOI: 10.1148/radiology.143.1.7063747.

44. Chong, J.; Soufan, O.; Li, C.; Caraus, I.; Li, S.; Bourque, G.; Wishart, D. S.; Xia, J., MetaboAnalyst 4.0: towards more transparent and integrative metabolomics analysis. *Nucleic Acids Research* **2018,** *46* (W1), W486-W494. DOI: 10.1093/nar/gky310.

FIGURE LEGENDS

Fig. 1: NMR analyses of the exo-metabolome. (a) Typical ^{1}H NMR spectrum of a *Shigella boydii* sample (culture supernatant) at exponential growth, i.e. after 1.5 hours of culture in a Mueller Hinton medium. Unsupervised multivariate data analysis based on the 950 variables derived from the NMR 1D spectra of culture media samples (centered variables, with no scaling). (b) Score plot of the PCA model (PC1 and PC2) including both *E. coli* lactose (+) and (-) and *Shigella* (N=138, R^2= 0.986 and Q^2 = 0.92 on 32 principal components). *E. coli* lactose (-) and *Shigella* cannot be discriminated. (c) Score plot (PC1 and PC2) of the PCA model of *E. coli lactose* (+) and *Shigella* samples (N=91, R^2= 0.95 and Q^2 = 0.854 on 15 principal components). The discrimination between those samples is straightforward.

Fig. 2. Supervised multivariate discriminant analysis (O-PLS-DA) of *S. boydii*, *S. flexneri*, *S. sonnei* and *E. coli* lactose (-) culture supernatants. The 4-classes model is built from the full NMR data matrix (950 variables, no scaling) with 3 predictive and 12 orthogonal components; $R^2(X)$ = 0.994, $R^2(Y)$ = 0.846 and Q^2 = 0.666. Score plots represent data projections on planes defined by (a) the first 2 predictive components, and (b) the 1st and 3rd predictive components that displays optimum discrimination of species, respectively.

Fig. 3. O-PLS-DA model based on 26 exo-metabolite concentrations derived from ^{1}H NMR profiles of *Shigella* and *E. coli* lactose (-) culture supernatants. (a) Score plot of the (1+7) O-PLS-DA model discriminating 46 *Shigella* samples (in purple) and 45 *E. coli* lactose (-) (in blue); $R^2(X)$ = 0.99, $R^2(Y)$ = 0.75 and Q^2 = 0.639, CV-ANOVA *p*-value= 7.53 x10^{-11}. (b) The O-PLS model was validated by re-sampling under the null hypothesis. (c) VIP value of each metabolite.

Table 1: AUC of the ROC curves based on the cross-validated score values from O-PLS-DA models for the discrimination of *E. coli* lac(-) and *Shigella* species, for individual metabolites presented by predictive VIP order, and combination of metabolites obtained by feature selection from multivariate ROC exploratory analysis [44].

Individual metabolites	**Individual AUC**
Succinate (Suc)	0.674
Acetate (Ace)	0.680
Aspartate (Asp)	0.589
Formate (Form)	0.588
Lysine (Lys)	0.762
Proprionate (PrP)	0.717
Threonine (Thr)[a]	0.636

Full Model	**AUC**	**R2X**	**R2Y**	**Q2**	**Number of components**
All 26 quantified metabolites	**0.9995**	0.99	0.75	**0.639**	1+7
Combination by ROC analysis	**AUC**	**R2X**	**R2Y**	**Q2**	**Number of components**
3 metabolites: Ace-Suc-Form	0.779	1	0.249	0.203	1+2
5 metabolites: Ace-Suc-Asp-Form-PrP	0.859	1	0.429	0.383	1+4
10 metabolites: Ace-Suc-Asp-Form-PrP-Thr-Lys-Glu-Leu-Pro	0.971	0.982	0.612	0.546	1+3
20 metabolites	0.998	0.991	0.744	0.64	1+7

[a]Other abbreviations : Glu: Glutamate, Leu: Leucine, Pro : Proline

Figure 1.

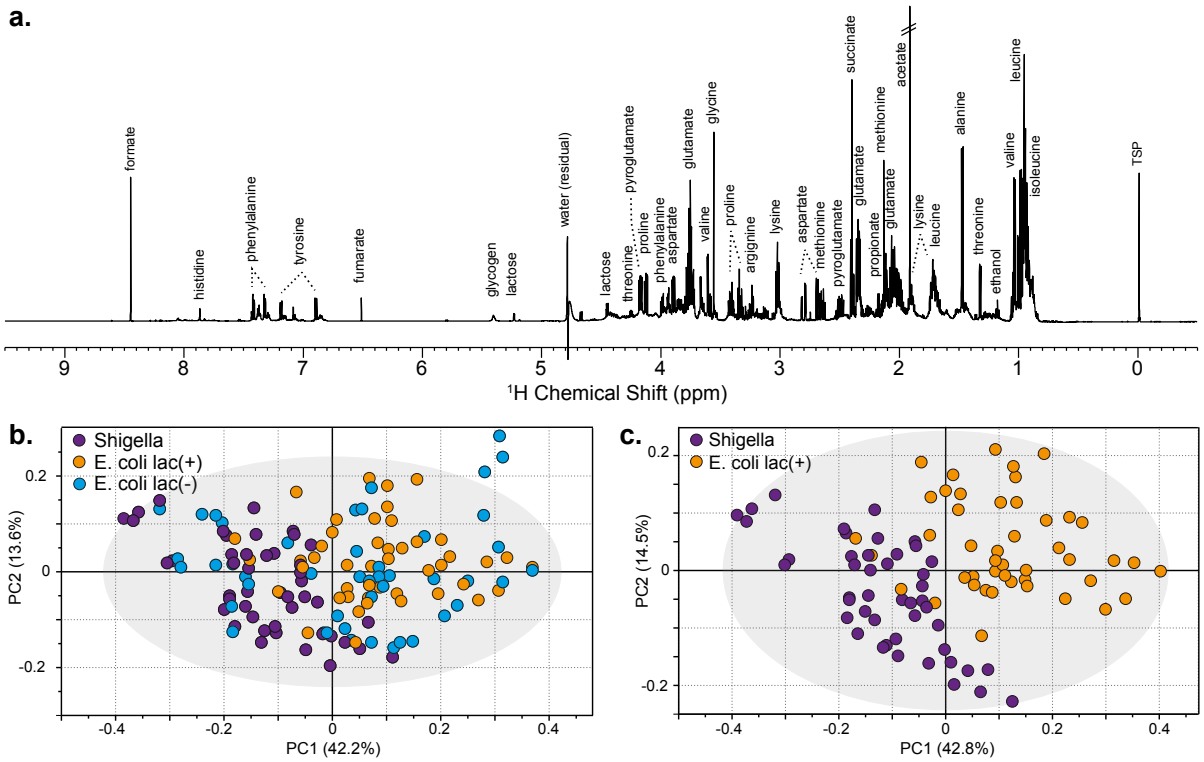

a.
formate
histidine
phenylalanine
tyrosine
fumarate
glycogen
lactose
water (residual)
lactose
threonine
pyroglutamate
proline
phenylalanine
aspartate
glutamate
valine
glycine
proline
arigignine
lysine
aspartate
methionine
pyroglutamate
succinate
glutamate
propionate
methionine
glutamate
acetate
lysine
leucine
alanine
threonine
ethanol
valine
leucine
isoleucine
TSP
9
8
7
6
5
4
3
2
1
0
¹H Chemical Shift (ppm)
b.
Shigella
E. coli lac(+)
E. coli lac(-)
PC2 (13.6%)
PC1 (42.2%)
c.
Shigella
E. coli lac(+)
PC2 (14.5%)
PC1 (42.8%)

Figure 2.

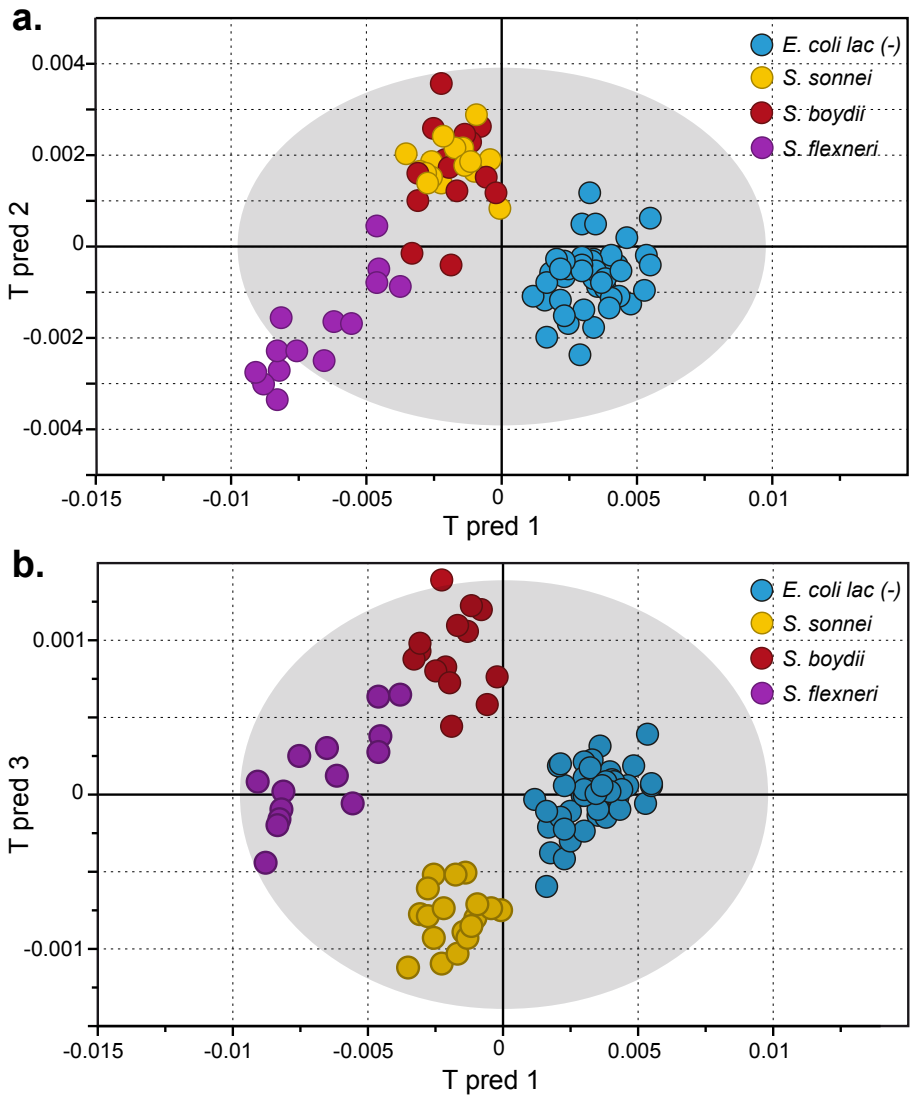
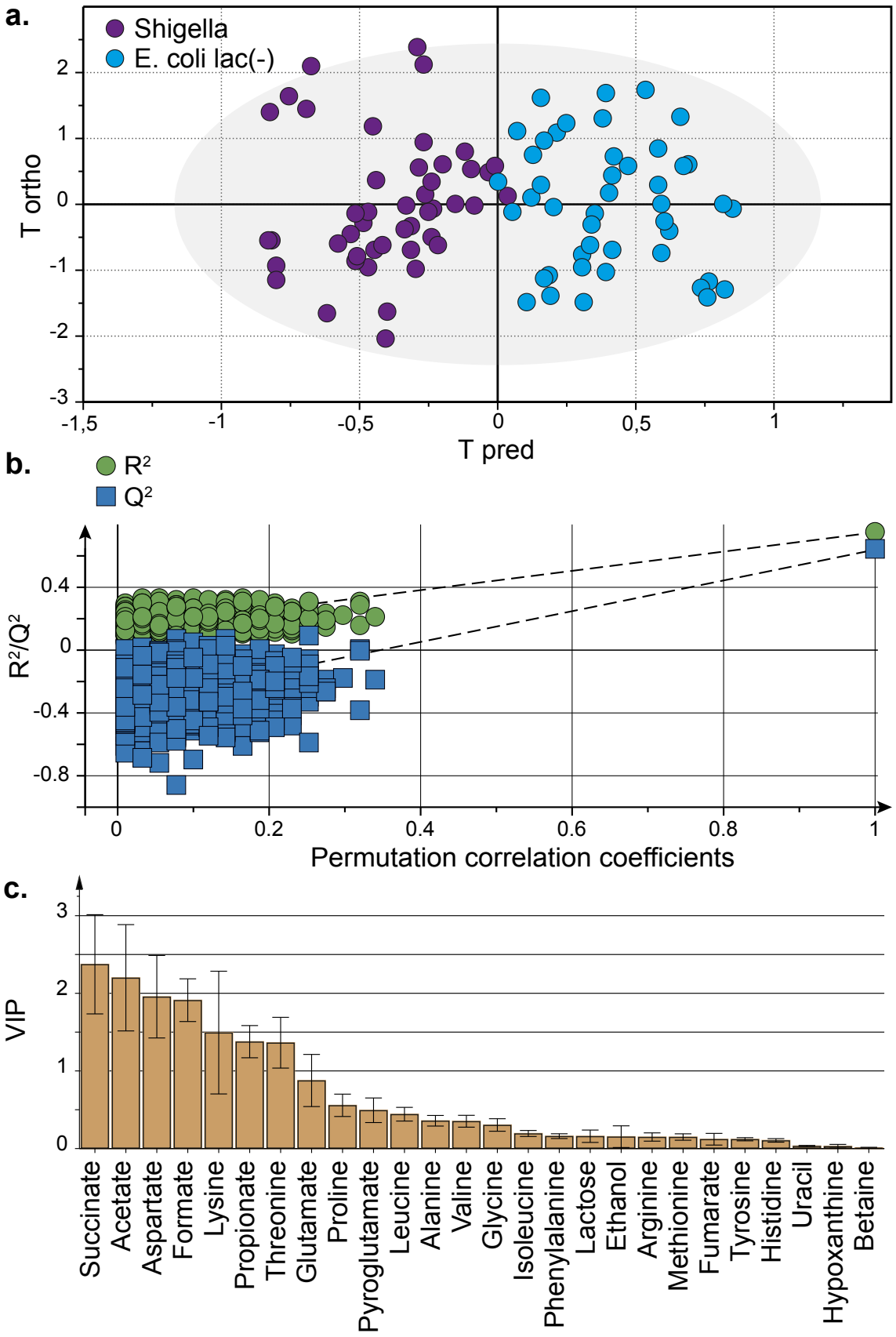

a.
E. coli lac (-)
S. sonnei
S. boydii
S. flexneri
T pred 2
T pred 1
0.004
0.002
0
-0.002
-0.004
-0.015
-0.01
-0.005
0
0.005
0.01
b.
E. coli lac (-)
S. sonnei
S. boydii
S. flexneri
T pred 3
T pred 1
0.001
0
-0.001
-0.015
-0.01
-0.005
0
0.005
0.01

Figure 3.


a.
Shigella
E. coli lac(-)
T ortho
T pred
b.
R²
Q²
R²/Q²
Permutation correlation coefficients
c.
VIP
Succinate
Acetate
Aspartate
Formate
Lysine
Propionate
Threonine
Glutamate
Proline
Pyroglutamate
Leucine
Alanine
Valine
Glycine
Isoleucine
Phenylalanine
Lactose
Ethanol
Arginine
Methionine
Fumarate
Tyrosine
Histidine
Uracil
Hypoxanthine
Betaine

For Table of Contents Use Only