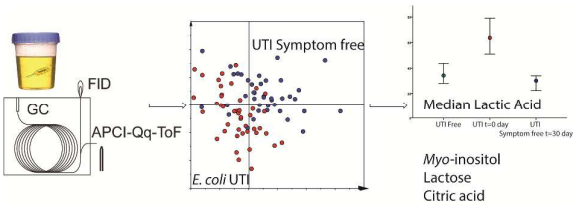
Analyst

Accepted Manuscript

ROYAL SOCIETY OF CHEMISTRY

www.rsc.org/analyst

Table of contents



This study shows the first application of GC-APCI-MS in a clinical setting specifically in the context of urinary tract infection.

Exploratory analysis of urinary tract infection using a GC-APCI-MS platform

Tiziana Pacchiarotta[1]*, Rico J. Derks[1], Ekaterina Nevedomskaya[1], Willize van der Starre[2], Jaap van Dissel[2], André Deelder[1], Oleg A. Mayboroda[1]

[1] Center for Proteomics and Metabolomics, Leiden University Medical Center, Leiden, The Netherlands

[2] Department of Infectious Diseases, Leiden University Medical Center, Leiden, The Netherlands

* Author to whom correspondence should be addressed:

Dr. Tiziana Pacchiarotta

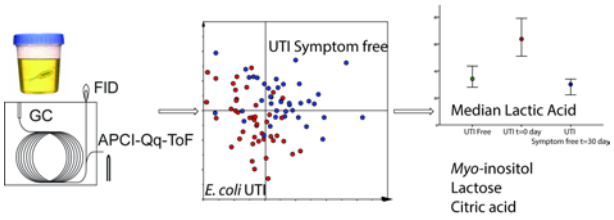
Center for Proteomics and Metabolomics, Leiden University Medical Center, Leiden, The Netherlands

PO Box 9600, 2300RC Leiden

Phone: +31-71-526-8705

Email: T.Pacchiarotta@lumc.nl

Table of contents



This study shows the first application of GC-APCI-MS in a clinical setting specifically in the context of urinary tract infection.

Abstract

Urinary tract infection (UTI) is among the most common bacterial infections worldwide. The understanding of the physiological mechanisms affected by UTI may need modern integrative '–omics' technologies, and metabolomics in particular. Here we present the first GC-APCI-MS-based explorative metabolomics study of UTI, using MS and FID detectors simultaneously. This provides high quality mass spectral data as well as semi-quantitative information demonstrating the feasibility of the GC-APCI-MS platform for non-targeted approach. The work is part of a bigger project aiming at providing a comprehensive overview of UTI-induced changes in urine.

Taking advantage of a fully clinically characterized cohort that offers the possibility of both case-control and longitudinal modelling, we can define UTI-induced change as a list of urinary metabolites which distinguish *E. coli* UTI patients from the subjects with no signs of an active infection. The list of molecular descriptors includes compounds related to bacterial activity such as lactic acid and lactose while other molecules show an association with the physiological status (inositol, citric acid).

Introduction

Urinary Tract Infection (UTI) is among the most common bacterial infections worldwide. Clinical presentation of UTI is broad: from a mild cystitis to severe pyelonephritis, which can quickly evolve into life-threatening conditions such as urosepsis and multiple organ failure[1, 2]. In a way, a positive bacterial urine culture is the only common denominator which unifies a variety of clinical syndromes under the term 'urinary tract infection' even though almost 20% of women presenting UTI symptoms will have negative urine culture [3, 4]. The understanding of the physiological mechanisms affected by UTI may need modern integrative '–omics' technologies, and metabolomics in particular. Indeed, a few proof of principle studies have demonstrated the feasibility of the metabolomics approaches for the analysis of urines from UTI patients[5, 6] and even demonstrated metabolomics-based routines for uropathogen identification[7-9]. Each of these studies advocated a specific technological solution but it is well known that the use of different analytical platforms often results in complementary data. Our aim is to create a comprehensive overview of UTI-induced changes in the metabolic patterns of patients' urine by applying several explorative methodologies on a carefully matched selection of patients from a multicentre prospective observational study of UTI[10]. To this end, in our previously published papers, we used well-established techniques such as liquid chromatography-mass spectrometry (LC-MS)[11] and nuclear magnetic resonance (NMR)[12] in order to investigate the UTI-induced changes in the metabolite profile.

GC-MS equipped with atmospheric pressure chemical ionization (GC-APCI-MS) has only recently been included into the toolbox for metabolite analysis. The detection at atmospheric pressure rather than with a vacuum stage (e.g. chemical ionization and electronic impact) has demonstrated its potential in a wide variety of fields ranging from pharmaceutical applications[13] to studies of bacterial metabolism[14], food[15-19] and environmental analysis[20]. In this manuscript, we present the results obtained with a recently described GC platform equipped with two detectors running in parallel – namely an atmospheric pressure chemical

ionization source – mass spectrometer (GC-APCI-MS) and a flame ionization detector (FID). We applied this platform to the investigation of changes in metabolic profiles of urinary tract infection (UTI) patients which may assist in monitoring the infectious status of the patients.

Results and discussion

GC-APCI-MS is a relatively recent addition to the metabolomics' toolbox. A number of publications showing the feasibility of this new technique for non-targeted approach is growing consistently. Yet, to our best knowledge, there is no report demonstrating how it performs in exploratory studies in clinical application. Here we present the first GC-APCI-MS-based explorative investigation of urines from UTI patients. This work is part of a larger study in which we aim to create a comprehensive overview of UTI-induced changes in the metabolic patterns within the urine of these patients. A part of the sample set used for the current study has previously been characterized with two other analytical platforms, namely NMR[12] and LC-MS[11]. Those two studies provided us with complementary pictures of the metabolic changes related to the UTI and the recovery process. NMR has revealed a set of metabolites indicative for the degree of the infection, while the LC-MS study resulted in the identification of *O*-glycosylated fragments of fibrinogen alpha chain as possible morbidity markers.

Thus, as first step of the data analysis, we performed Principal Component Analysis (PCA) on the entire data set which included subjects' samples and pooled quality control samples (193 samples in total). The PCA score scatter plot in the Supplementary Figure 1 shows the spread of the samples along the first two principal components (out of a total of 24 components) which together cover the 16% of the total variability. We have tested the possible correlation with known source of the analytical variability such as the acquisition block, the order of injection or the influence of the column and we could conclude that these factors have eventually a negligible effect on the entire dataset.

Supplementary Figure 2 (A and B) shows the PCA score plots for a model built with the 155 samples, including the controls with no clinical signs of the active infection, the baseline samples of UTI patients (t=0 day) and the follow-up samples – 30 days after the day of enrollment (t=30 day)– is shown. No trend related to the infectious status is present in the first two principal components. However, a trend separating the UTI-subjects from UTI-free and recovered patients(t=30 day) is present along the second and the third components.

In the next step, we built two separate PCA models: one for the baseline (UTI patients t=0 days and controls) and one for UTI patients at baseline and at 30 days after the treatment (Supplementary Figure 3 A and B). Despite a clear visual trend indicating the expected clustering in both PCA models, a fraction of the variance covered by the first two principal components is just 16% that suggests that there is no main trend driving the dataset but a combination of confounders which contribute to the model. Consequently, the statistical quality of the regression models (goodness of fit and prediction power) built on the same selection of samples was poor (data not shown). This observation appears to be contradictory to our previously published reports. However, in comparison to the previous studies, where we restricted our selection to the patients with culture-confirmed febrile *E.coli* UTI, the current selection includes the subjects with different uropathogens (Table 1) in order to mimic the real microbiological incidence that show as main pathogenic source extra intestinal *Escherichia coli*; this organism is isolated in 80% to 85% of episodes [21-23]. Thus, with the aim to explore whether the quality of the models is affected by the heterogeneity of the pathogens, we restricted the selection to UTI subjects with *E. coli* as pathogen. Figure 1A and 1B display the cross-validate score plots of the PLS-DA models respectively for the baseline (UTI patients t=0 days and controls) and for UTI patients at baseline and at 30 days after the treatment. Judging on the basis of the model parameters (R2Y=0.922 and Q2=0.519), the longitudinal (t=0, t=30 day) model (Fig. 1B) is the strongest of the two.

To get an insight into the metabolic differences between the day of enrolment (t=0 day) and after the recovery (t=30 days, we focused on *m/z* signals with Variable Importance in Projection (VIP) values above 1.5 for the given PLS-DA model. We selected 25 *m/z* values

which could be grouped into 11 chromatographic peaks, when sorted by retention time. For the identification of the underlying compounds, we used two strategies. First, we checked whether the metabolites were present in our publicly available in-house developed database (http://metams.lumc.nl), using the spectra matching (MS and MS^2 spectra were compared by visual inspection) [18, 24]. If the compounds were not present in the library, *de novo* identification was performed according to the protocol provided elsewhere[24]. Briefly, the protocol includes the calculation of the molecular ion according to the number of silylated group and fragments (those information can be extrapolated from both MS and MS^2 spectra), the use of accurate mass (mass error < 5ppm) and isotopic distribution. In both cases, library-matched or *de novo* identification, the possible candidates were compared against urine metabolome database (http://www.urinemetabolome.ca/) and literature data in order to ensure the sustainability of the findings. Table 2 summarizes the results of these analyses. The spectra of the identified and unknown compounds are shown in Supplementary Figure 4.

Even a brief overview of the literature shows that all compounds listed in the Table 2 (with exception of the unknowns) are well known constituents of urine[17]. For instance, hyrdroxyhippuric acid is a product of the gut microbiota while 3-(3-hydroxyphenyl)-3-hydroxypropionic acid (HPHPA) was reported as an excretion product of anaerobic bacteria of the *Clostridium* genus[25].

Another interesting compound is inositol. There is no single source for urinary inositol and its isomers. A fraction of *myo*-inositol has a dietary origin but substantial part of the pool is synthetized *de novo*. Under normal conditions, up to 98% of *myo*-inositol is reabsorbed in the renal tubules. Consequently, increased concentrations of *myo*-inositol could be considered as an indication of ongoing pathological changes in the renal system: increased concentrations of urinary *myo*-inositol were reported for patients with diabetes and renal failure[26-28]. For our case *myo*-inositol is interesting because its alteration during recovery from UTI would eventually be related not to the bacterial metabolism but to the renal function. Lactose and lactate are very well known markers of the bacterial metabolism: the ability of

E.coli to metabolize lactose to lactate was an essential part in discovery of the operons[29]. Recently Gupta *et al.* developed a NMR-based method for the selective identification and quantification of the *E. coli* colonies in urine of UTI patients using the conversion of lactose to lactate. In their procedure, the authors incubated the infected urine samples with the substrate for 6 hours and after that they could measure a decreased intensity of lactose and an increased intensity of lactic acid[8]. Figure 2 shows the FID integrated signals of both lactic acid (A) and lactose (B). The levels of lactic acid are significantly higher in the group of the patients with an active infection and, even though, these data are semi-quantitative, they resemble the quantitative data acquired with NMR[12]. In addition, the calculation of statistical significance of the lactose responses between the classes provided p-values which indicate high degree of significance; specifically $p = 3.76\ 10^{-6}$ for UTI free against UTI and $p = 1.66\ 10^{-5}$ for UTI t=0 against Recovered t=30. On the contrary, the levels of lactose appear to be more or less constant. The last observation is not really contradictory to Gupta's report. In this last case, 6 hours long incubation of the samples leads to the depletion of lactose pool within a sample and results in a perfect inverse relation between concentrations of the lactate and lactose. Our samples are just the spot urines, where lactose pool is dependent not only of the rate of its consumption by the pathogen, but also by its renal excretion.

Since some of the classifiers seem to be indicative for the presence of pathogens, we tried to build a model for prediction of the infectious status. To this end, we selected culture-confirmed febrile *E.coli* UTI subjects who presented samples at both time point (a total of 50 samples) and we built a PLS-DA model which showed comparable statistical values presented for the model in Figure 1B (Supplementary Figure 5). We used this last model as basis to predict the outcome of the remained UTI t=0 samples at baseline which included the rest of UTI t= 0 day with positive *E.coli* urine culture (that did not have the sample at t=30 day) as well as the rest of the other pathogen. The results of the prediction, reported in Table 3, confirm that the so built model has a good sensitivity for *E.coli* urine samples (75%) while the prediction fails for the majority of the urine contaminated with other pathogen regardless

their classification (Gram positive or negative) even though a more accurate test of the predictive power for other pathogens would require a greater number of samples.

Another important parameter that characterizes UTI is the bacterial load in urine. Thus we built a PLS-regression model using urinary bacterial load (CFU/ml) as a response variable. The plots shows the classification of the samples (Figure 3A) is most of the time in agreement with the bacterial load (Figure 3B). The *m/z* and retention time responsible for the correlation of the GC-APCI-MS data and the bacterial count were chosen on the basis of the corresponding VIP values. The list of these variable is reported in Table 4. This table show a number of unknown compounds and some sugar which were not rising with the previous models. In addition, with this model we can detect the effect of the exogenous compounds such as paracetamol metabolites (paracetamol glucoronide). However on this data matrix, the influence of such compounds was much less pronounced than in both previous reports. The spectra of the identified and unknown compounds are shown in Supplementary Figure 6.

Citrate is another 'simple' metabolite with a complex functional network. Citrate is often solely considered as one of the intermediates of the TCA cycle. However, as a constituent of body fluids, it plays a different but equally important role in the regulation of homeostasis. The three carboxylic groups of a citrate molecule may form complexes with metal ions and/or other divalent cations; as such, citrate has been implicated in the control of renal stone formation by forming soluble complexes with calcium[30]. For many years, citrate excretion in urine is used for the assessment of the renal function: for example, reduction of urinary citrate, which we observed for UTI patients at t=0 days, is an alerting signal for pathological changes in the renal metabolism. There are multiple reports showing the decrease of the urinary citrate in nephrolithiasis[31], glomerulonephritis[32] and more recently in infectious diseases[12, 33].

Thus, the results obtained with our GC-APCI-MS platform are in line with previous findings and biological meaningful. However there are a few points to improve such as the limited identification. Table 2 and 4 show a considerable number of unknown compounds.

In some cases we failed in identifying compounds *de novo*; even though we had MS and MS^2 spectra data, this did not lead to an unequivocal identification. In particular, carbohydrates need special attention. The identification of this class is quite challenging because of the high number of isomers present in body fluids. In the case of lactose we could unequivocally discriminate the compound since the limited number of isomers present in urine which are all included in the library by matching the MS spectra with the chromatographic data and select only the compounds with similar chromatographic behavior (same Kovats retention index). On the contrary this was not the case for Xylose and Fucose. Their spectra investigation revealed a good matching with the compounds in the database. However not all the possible isomers eventually present in urine are reported. Thus, despite the good matching of the experimental spectra with the reference spectra, we could not unequivocally annotate the mentioned compounds. Obviously this indicated a limitation that might be overcome with the enlargement of the in-house library.

Experimental

Chemicals. Methanol, pyridine, iso-propanol, toluene (all at analytical grade) and urease were purchased from Sigma (Zwijndrecht, the Netherlands). N-methyl-N-trimethylsilyltrifluoroacetamide with 1% trimethylchlorosilane (MSTFA + 1% TMCS) was used as a derivatization reagent (Pierce; Rockford. IL. USA) and taken from freshly opened 1 ml bottles. Methoxyamine hydrochloride was purchased from Supelco (Zwijndrecht, the Netherlands). APCI tune mix (containing purine, hexamethoxyphosphazine, hexakis (2,2-difluoroethoxy) phosphazine and hexakis (1H,1H,3H-tetrafluoropropoxy)phosphazine) and pyridine (99%. ultra-pure GC grade) were acquired from Fluka.

Clinical samples. Urine samples were collected during a prospective observational multicenter cohort study in which eight emergency departments (ED) of seven hospitals and 35 affiliating primary health care centers participated. From January 2004 till November 2008, patients who presented with a diagnosis of febrile UTI, were considered for enrolment in the study. The study was approved by the local ethics committees and all patients gave written

informed consent. Inclusion criteria and exclusion criteria have been described in detail elsewhere[34, 35]. In short, patients of 18 years or older, with fever and presenting at least one symptom of UTI (dysuria, increased frequency of urination, perineal pain, flank pain or costovertebral tenderness) and a positive nitrite dipstick test or a positive leukocyte esterase dipstick test were enrolled. Exclusion criteria were current treatment for urolithiasis or hydronephrosis, pregnancy, hemo- or peritoneal dialysis, a history of kidney transplantation or known presence of polycystic kidney disease. Details with respect to empiric therapy and follow-up are provided elsewhere [10]. Urine samples were collected at enrolment as baseline samples (t=0). After four and thirty days (t=30) urine samples of the same patients were collected for follow-up. At all the time points, clean midstream-catch urine samples were cultured and analyzed using standard microbiological methods. A positive urine culture was defined as bacterial growth of over 10^3 colony-forming units per ml. Within 2 hours after collection all samples were stored at -80° C until further analysis.

For the current study, a group of 67 subjects with culture-confirmed bacteria in the urine was selected from a database of approximately 700 subjects. Samples from volunteers without UTI or symptoms of another infection were taken as controls. Baseline characteristics of the samples are presented in the Table 1. The study design includes the control samples and samples of UTI patients collected at baseline (t=0 days) as well as patient follow up samples collected after the antibiotic treatment (t=30 days). In this last point of the sampling, the subjects did not show any sign of active ongoing infection.

Sample preparation. Urine samples were prepared according to a reference method with minor modifications[36]. Briefly, 100 µl of urine were incubated with 20 Units of urease in 100 µl of water for 45 minutes at 37 ºC to deplete the excess of urea. Protein precipitation was carried out by adding 300 µl of cold methanol and by incubation for 10 min on ice. After centrifugation at 19.600g for 10 min, 100 µl of the supernatant was transferred to a GC/MS vial and evaporated to dryness. All the samples were stored at -80 ºC until the moment of the GC analysis.

The samples were derivatized online with a double step reaction. Firstly, we added 50 µl of methoxyamine (20 mg/ml in pyridine) and incubated for 1 hour at 60 °C. Secondly, we added 50 µl of MSTFA and incubated at 60 °C for 40 minutes. Subsequently, we added 100 µl of toluene.

Hundread µl of each urine sample were pooled, aliquoted and then used as quality control which were prepared with the same protocol.

GC-APCI-MS-FID analysis. The samples (1 µl) were applied by splitless injection with a programmable Multi Purpose Sampler MPS2 (Gerstel, Mülheim an der Ruhr, Germany) into an Agilent 7890A GC (Agilent, Palo Alto USA) equipped with one single column (HP-5-MS column 30 m. 0.25 mm ID. 0.25 µm film thickness) and two detectors running in parallel; a column flow splitter with deactivated capillaries (0.6 m. 0.25 mm ID. 0 µm film thickness) allowed the simultaneous acquisition with both the MS and FID detector. For sample injection, a septumless PTV injector was used. For each analysis the purge time of the PTV injector was set to 60s at a purge flow rate of 20 ml/min and an equilibration time of 1 min. Helium was used as carrier gas at a constant flow rate of 1 ml/min through the column. Urine samples were injected with the following temperature program: 70 °C held for 1 minute, the temperature increased till 280 °C at 5.5 °C/min and held for 10 minutes.

The transfer line to the mass spectrometer was kept at 300 °C. The APCI source was operated in positive mode. The temperature and the flow rate of the dry gas (nitrogen) were 300°C and 2.00 l/min, respectively. The APCI vaporizer temperature was 300 °C; the pressure of the nebulizer gas (nitrogen) was set to 3.5 bar and the voltage of the corona discharge needle was +2000 nA. Capillary voltage was set to -1000V and the end-plate offset to -500 V. The source and detector parameters have been set in order to achieve a minimal resolution of 40000.

A maXis 4G (Bruker Daltonics, Bremen, Germany), an orthogonal-accelerated Qq-ToF mass spectrometer, was used as mass analyzer. The spectra were acquired in a mass range from *m/z* 100-1000 at an acquisition rate of 1Hz. Before each chromatographic run, a tune mix solution was used as external calibrant and injected in Flow Injection mode.

For compound identifications Auto MS^2 experiments were performed. Collision energies were the following: *m/z* 100 at 25 V, *m/z* 1000 at 50 V. All the *m/z* values within this interval were fragmented with interpolated values of collision energy.

The FID flows and temperature were set as follows: front detector 300 °C, hydrogen flow: 40 ml/min and air flow: 370 ml/min.

Statistical analysis. The GC-MS data files were exported as mzXML files. The peak picking was performed with XCMS package (The Scripps Research Institute, La Jolla, USA) using the following commands: method="centWave", scanrange=c(480,2900), ppm=10, snthresh=10, peakwidth=c(5,15), prefilter=c(3,3000), noise=3000. The grouping was performed using band width=10, minfrac=1, minsamp=37. The final data matrix was normalized with Probabilistic Quotient Normalization algorithm[37]. The generated data matrix was imported to the SIMCA-P 13.0 software package (Umetrics, Umeå, Sweden). The data were mean centred, unit variance-scaled and Power transformed prior to the statistical analysis. The validity and the degree of overfitting of the PLS-DA models were checked using a 20 permutations test.

To identify metabolites of interest, rational chemical formulas were generated based on internally calibrated monoisotopic masses within 2 mDa mass error, using the SmartFormula tool within the DataAnalysis software package (Bruker Daltonics).

Conclusion

Here, we have presented the first application of an explorative metabolomics study on a GC-APCI-MS platform in a clinical setting. This work has demonstrated the feasibility of the GC-APCI-MS platform for non-targeted approach tracing results already present in literature and confirmed with our previously published paper on the same dataset in the context of UTI. Here we have presented how this platform can simultaneously provide high quality mass spectral data and semi-quantitative information taking advantage of the parallel acquisition of high-resolution mass spectrometer acquisition with FID which benefits of a wide linear dynamic range. Moreover, as part of a bigger project aimed at the comprehensive

overview of UTI-induced changes in urine, this study points some physiological marker of the febrile UTI status. In accordance with the NMR study, this approach has revealed a number of compounds that might be associated with bacterial activity. On the other side, this also shows some unique feature that needs to be investigated further for eventual association to the physiological status. Although obviously no firm biological conclusions can yet be drawn, this study nevertheless has provided findings which may further on contribute to the understanding of the physiological mechanisms underlying the progression of UTI.

References

1. A. G. Pinson, J. T. Philbrick, G. H. Lindbeck and J. B. Schorling, *American Journal of Emergency Medicine*, 1997, **15**, 148-151.
2. C. M. Kunin, *Archives of Internal Medicine*, 2003, **163**, 2393-2393.
3. M. L. Wilson and L. Gaido, *Clin Infect Dis*, 2004, **38**, 1150-1158.
4. B. Foxman, *Nat Rev Urol*, 2010, **7**, 653-660.
5. H. Lv, C. S. Hung, K. S. Chaturvedi, T. M. Hooton and J. P. Henderson, *Analyst*, 2011, **136**, 4752-4763.
6. R. Ramautar, O. A. Mayboroda, R. J. E. Derks, C. van Nieuwkoop, J. T. van Dissel, G. W. Sornsen, A. M. Deelder and G. J. de Jong, *Electrophoresis*, 2008, **29**, 2714-2722.
7. A. Gupta, M. Dwivedi, G. A. N. Gowda, A. A. Mahdi, A. Jain, A. Ayyagari, R. Roy, M. Bhandari and C. L. Khetrapal, *Nmr in Biomedicine*, 2006, **19**, 1055-1061.
8. A. Gupta, M. Dwivedi, A. A. Mahdi, G. A. N. Gowda, C. L. Khetrapal and M. Bhandari, *Bju International*, 2009, **104**, 236-244.
9. A. Gupta, M. Dwivedi, A. A. Mahdi, C. L. Khetrapal and M. Bhandari, *Journal of Proteome Research*, 2012, **11**, 1844-1854.
10. C. van Nieuwkoop, T. N. Bonten, J. W. van't Wout, E. J. Kuijper, G. H. Groeneveld, M. J. Becker, T. Koster, G. H. Wattel-Louis, N. M. Delfos, H. C. Ablij, E. M. S. Leyten and J. T. van Dissel, *Critical Care*, 2010, **14**.
11. T. Pacchiarotta, P. J. Hensbergen, M. Wuhrer, C. van Nieuwkoop, E. Nevedomskaya, R. J. E. Derks, B. Schoenmaker, C. A. M. Koeleman, J. van Dissel, A. M. Deelder and O. A. Mayboroda, *Journal of Proteomics*, 2012, **75**, 1067-1073.
12. E. Nevedomskaya, T. Pacchiarotta, A. Artemov, A. Meissner, C. van Nieuwkoop, J. T. van Dissel, O. A. Mayboroda and A. M. Deelder, *Metabolomics*, 2012, **8**, 1227-1235.
13. T. Bristow, M. Harrison and M. Sims, *Rapid Communications in Mass Spectrometry*, 2010, **24**, 1673-1681.
14. C. J. Wachsrnuth, M. F. Almstetter, M. C. Waldhier, M. A. Gruber, N. Nurnberger, P. J. Oefner and K. Dettmer, *Analytical Chemistry*, 2011, **83**, 7514-7522.
15. T. Portoles, L. Cherta, J. Beltran and F. Hernandez, *Journal of Chromatography A*, 2012, **1260**, 183-192.
16. T. Portoles, J. G. J. Mol, J. V. Sancho and F. Hernandez, *Analytical Chemistry*, 2012, **84**, 9802-9810.
17. R. Garcia-Villalba, T. Pacchiarotta, A. Carrasco-Pancorbo, A. Segura-Carretero, A. Fernandez-Gutierrez, A. M. Deelder and O. A. Mayboroda, *Journal of Chromatography A*, 2011, **1218**, 959-971.
18. E. Hurtado-Fernandez, T. Pacchiarotta, E. Longueira-Suarez, O. A. Mayboroda, A. Fernandez-Gutierrez and A. Carrasco-Pancorbo, *Journal of Chromatography A*, 2013, **1313**, 228-244.

19. E. Hurtado-Fernandez, T. Pacchiarotta, O. A. Mayboroda, A. Fernandez-Gutierrez and A. Carrasco-Pancorbo, *Food Research International*, 2014, **62**, 801-811.
20. F. David, P. Sandra and P. Hancock, *Lc Gc Europe*, 2011, **24**, 16-+.
21. S. D. Fihn, *New England Journal of Medicine*, 2003, **349**, 259-266.
22. T. M. Hooton, *Infectious Disease Clinics of North America*, 2003, **17**, 303-+.
23. G. Kahlmeter, *Journal of Antimicrobial Chemotherapy*, 2003, **51**, 69-76.
24. T. Pacchiarotta, R. J. E. Derks, E. Hurtado-Fernandez, P. van Bezooijen, A. Henneman, R. Schiewek, A. Fernández-Gutiérrez, A. Carrasco-Pancorbo, A. M. Deelder and O. A. Mayboroda, *Bioanalysis*, 2013, **5**, 1515-1525.
25. W. Shaw, *Nutritional Neuroscience*, 2010, **13**, 135-143.
26. W. H. Daughaday, J. Larner and E. Houghton, *Journal of Clinical Investigation*, 1954, **33**, 1075-1080.
27. G. Sarashina, M. Yamakoshi, M. Noritake, M. Takahashi, M. Kure, Y. Katsura, H. Shiomi, I. Tsuboi, S. Kawazu, F. Yamagata, M. Tominaga and T. Matsuoka, *Clinica Chimica Acta*, 2004, **344**, 181-188.
28. J. H. Hong, H. W. Jang, Y. E. Kang, J. H. Lee, K. S. Kim, H. J. Kim, K. R. Park and B. J. Ku, *Disease Markers*, 2012, **33**, 193-199.
29. F. Jacob, D. Perrin, C. Sanchez and J. Monod, *Comptes Rendus Hebdomadaires Des Seances De L Academie Des Sciences*, 1960, **250**, 1727-1729.
30. D. Rudman, J. L. Dedonis, M. T. Fountain, J. B. Chandler, G. G. Gerron, G. A. Fleming and M. H. Kutner, *New England Journal of Medicine*, 1980, **303**, 657-661.
31. M. Angel Arrabal-Polo, M. Arrabal-Martin, S. Arias-Santiago, J. Garrido-Gomez, A. Poyatos-Andujar and A. Zuluaga-Gomez, *Bju International*, 2013, **111**, 622-627.
32. N. G. Psihogios, R. G. Kalaitzidis, S. Dimou, K. I. Seferiadis, K. C. Siamopoulos and E. T. Bairaktari, *Journal of Proteome Research*, 2007, **6**, 3760-3770.
33. C. I. A. Balog, A. Meissner, S. Goraler, M. R. Bladergroen, B. J. Vennervald, O. A. Mayboroda and A. M. Deelder, *Molecular Biosystems*, 2011, **7**, 1473-1480.
34. C. van Nieuwkoop, T. N. Bonten, J. W. van't Wout, M. J. Becker, G. H. Groeneveld, C. L. Jansen, E. R. van der Vorm, E. P. Ijzerman, P. H. Rothbarth, E. M. TerMeer-Veringa, E. J. Kuijper and J. T. Van Dissel, *Clinical Infectious Diseases*, 2010, **50**, E69-E72.
35. W. E. van der Starre, J. T. van Dissel and C. van Nieuwkoop, *Current infectious disease reports*, 2011, **13**, 571-578.
36. A. Jiye, Q. Huang, G. J. Wang, W. B. Zha, B. Yan, H. C. Ren, S. H. Gu, Y. Zhang, Q. Zhang, F. Shao, L. S. Sheng and J. G. Sun, *Anal. Biochem.*, 2008, **379**, 20-26.
37. C. A. Smith, E. J. Want, G. O'Maille, R. Abagyan and G. Siuzdak, *Analytical Chemistry*, 2006, **78**, 779-787.

Table 1. Baseline characteristics. Data are presented as n (%) unless otherwise stated. IQR: interquartile range.

	Controls	Cases
Characteristics	n=53	n=67
Age, years, median	59[47–70]	68[54–78]
Male	25 (47)	32 (47)
Smoking	8 (1 missing value)	11
Uropathogen (1 missing value)		
E.coli	2	48
Haemolytic Streptococcus Group B		2
Enterococcus faecalis		3
Klebisella pneumoniae		4
Pseudomonas aeruginosa		2
Staphylococcus aureus		1
Staphylococcus saprophyticus		2
Klebsiella oxytoca		1
Aerococcus urinae		1
Acinetobacter species		1
Candida spcies	1	
Co-morbidity		
Urinary tract disorder	13	16
Malignancy	4	10
Renal insufficiency	3	4
Heart failure	8	7
Diabetes mellitus	0	0
Urine dipstick results		
Nitrites	4 (5 missing values)	38 (8 missing values)
Leucocyte esterase	13 (3 missing values)	62 (3 missing values)

Table 2. Identification of the features with VIP values > 1.5 describing the PLS-DA model built on UTI t=0 (*E.coli*) + UTI t=30 days.

Compound	Experimental ion parent	Mass Error (ppm)	Identification type
Lactic acid (3 TMS)	307.1575	0.2	In-house library
Phosphoric Acid (3TMS)	315.1025	0.9	In-house library
Malic Acid (3 TMS)	351.1473	0.3	In-house library
Glutamic Acid (3 TMS)	364.1788	0.5	In-house library
Unknown 1	364.1789		
Unknown 2	300.1669.		
Unknown 3	281.1245		
HPHPA or isomer (	398.1759	1.7	*De novo*
Myo-inositol	613.3083	-0.8	In-house library
3-hydroxyhippuric acid (2TMS)	340.1397	-0.7	*De novo*
Lactose (1 MeOX + 8 TMS)	948.4669	-0.7	In-house library

Table 3. Prediction of the outcome of the baseline samples on the basis of the PLS-DA model built using the UTI subjects (*E.coli*) who presented samples at both time points. The model is shown in Supplementary figure 5.

	Predicted as UTI	Predicted as Recovered	Total predicted samples	Gram positive/negative
E.coli	18	6	24	Negative
Klebisiella	4	1	5	Negative
Proteus mirabilis	0	1	1	Negative
Enterococcus Faecalis	2	2	4	Positive
Pseudomas aeruginsa	1	2	3	Negative
Staphilococcus	2	1	3	Positive
Aerococcus urinae	1	0	1	Negative
Acinobaster species	0	1	1	Negative
Candida species	1	0	1	Fungus

Table 4. Identification of the features with VIP values > 1.5 describing the PLS regression model built using the bacterial load as response variable.

Compound	Experimental ion parent	Mass Error (ppm)	Identification type
Lactic acid (3 TMS)	307.1575	0.2	In-house library
Unknown 1	221.1022		
Phosphoric Acid (3TMS)	315.1031	-1.1	In-house library
C4H8O4 (3TMS)	337.1669	0.5	*De novo*
Unknown 2	226.0891		
Unknown 3	289.1283		
Xylitol or Isomer (5 TMS)	513.2732	0.4	*De-novo*
Fucose or Isomer (MeOX + 4 TMS)	482.2601	0.6	*De novo*
$C_9H_{16}O_4$ (4TMS)	333.1910	0.6	*De novo*
Citric Acid (4TMS)	481.1921	0.6	In-house library
Unknown 4	309.1332		
Myo-inositol	613.3083	-0.8	In-house library
Paracetamol glucuronide (5TMS)	687.3020	-0.5	*De novo*

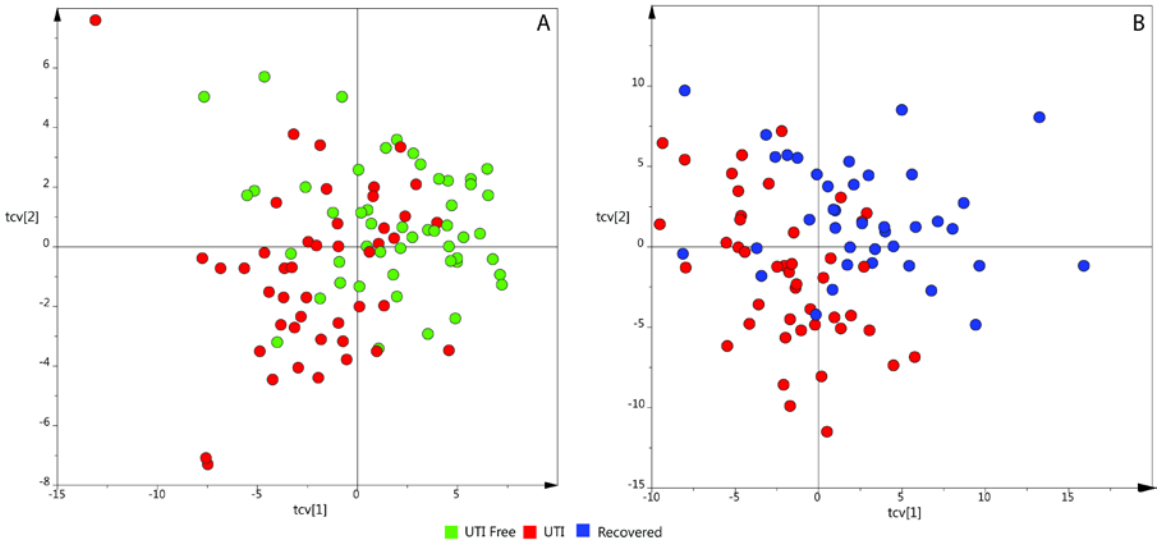


Figure 1. Cross-validated score plots of selected samples from subjects who presented *E. coli* as uropathogen. A) UTI Free + UTI (*E.coli*); R2X=0.0929, R2Y=0.77, Q2=0.27 B) UTI t=0 (*E. coli*) + UTI t=30; R2X=0.241, R2Y=0.922, Q2=0.519.

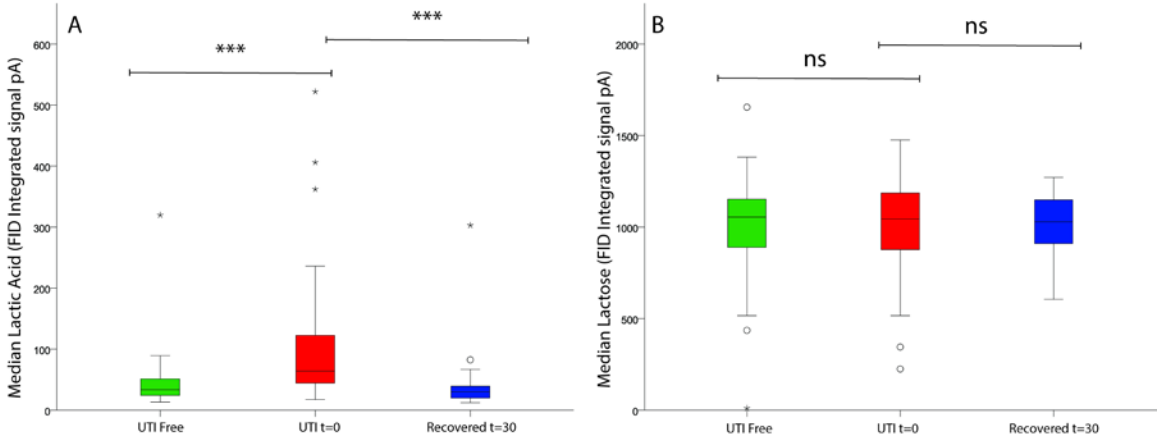


Figure 2. Distribution of Lactic Acid (A) and Lactose (B) along the three classes UTI Free, UTI t= 0 (*E. coli*) and UTI t=30 day. The bar plots refer to the median obtained from the FID integrated signal. p-vales are calculated with Mann-Whitney test for UTI Free against UTI t=0 and with signed test for UTI t=0 against Recovered t=30. The p-value have been corrected for multiple testing using Bonferroni correction. *** $p<0.001$, ◦ outlier, ⋆ far outlier.

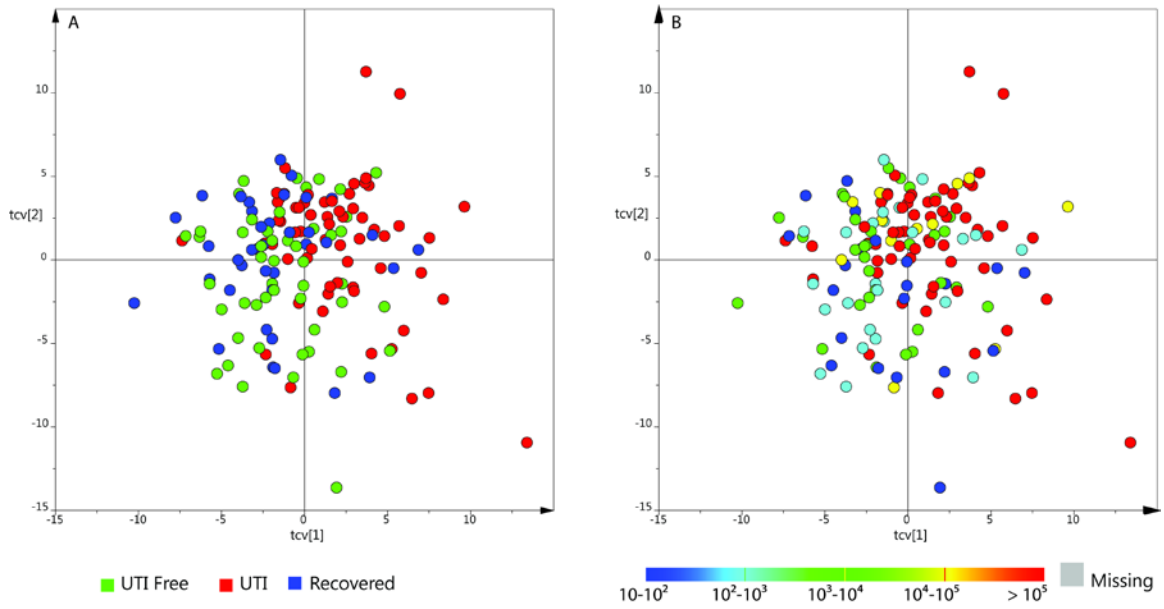
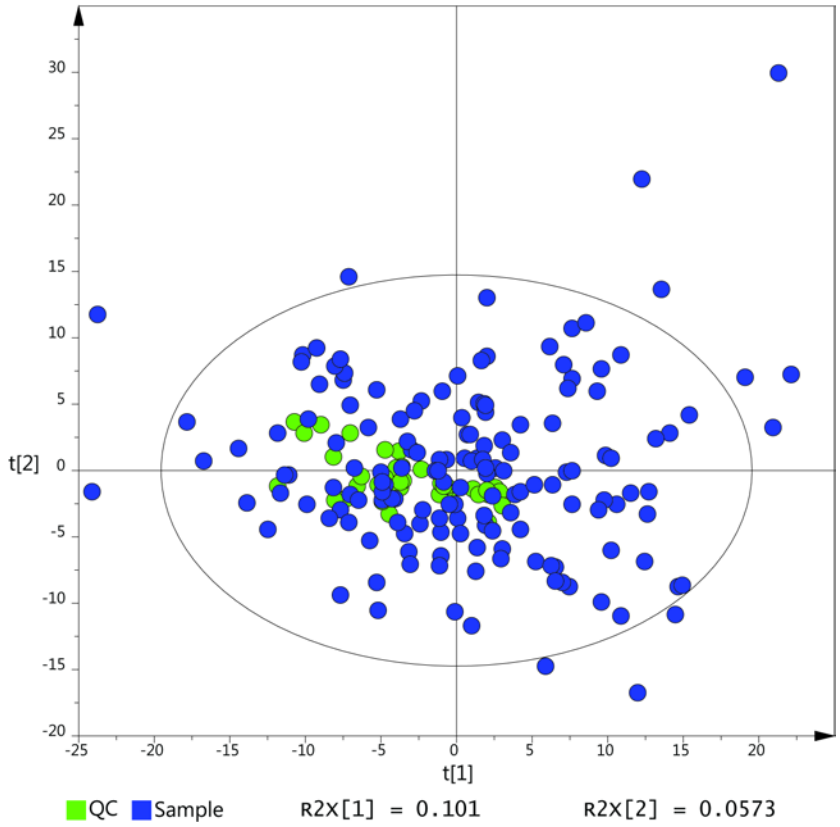
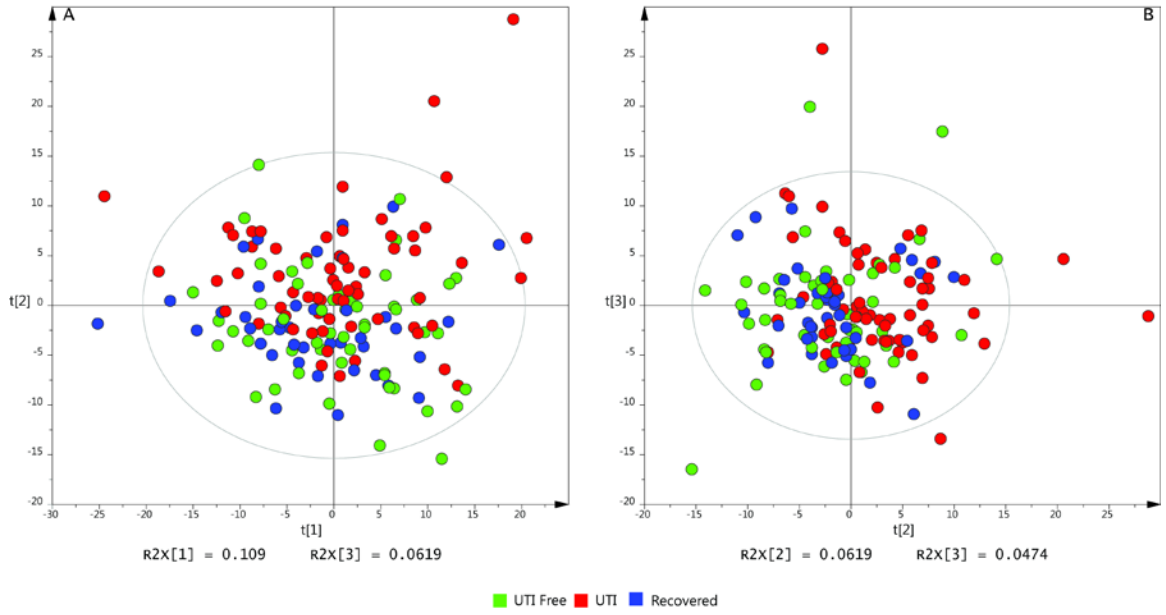


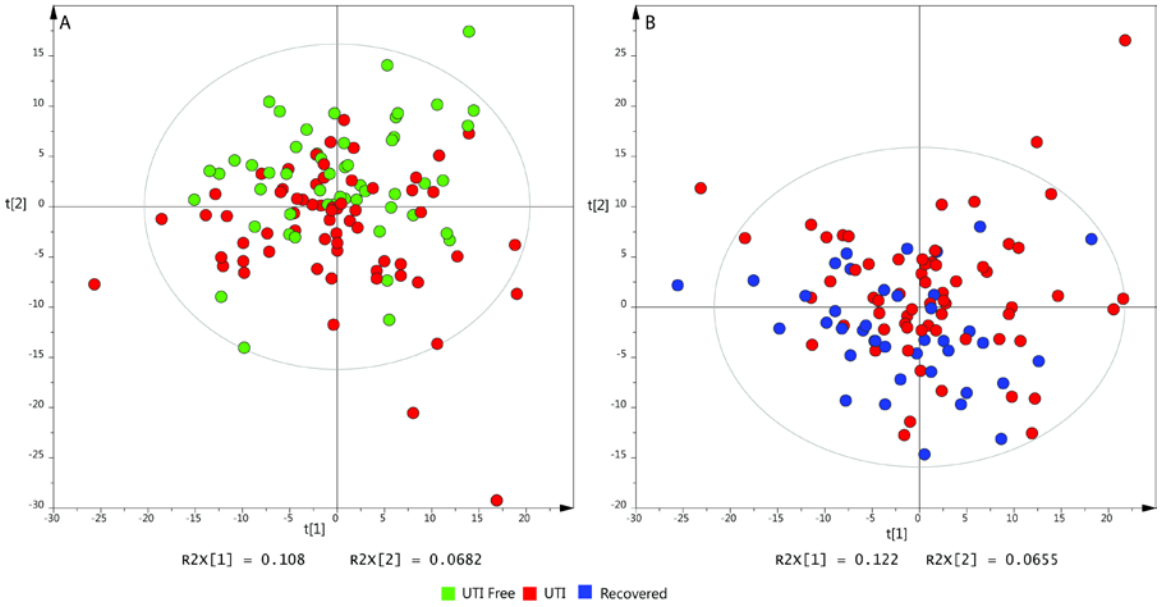
Figure 3. Score plot of the PLS regression model built on spectral data from urine and urinary bacteria load (CFU/ml). R2Y=0.811 and Q2=0.198. Colored by infectious status (A) or number of bacteria (B). Fisher and p-values are 7.049 and 1.82 e-6 respectively.



Supplementary Figure 1. The score plot of a PCA model built on the entire data set, 24 components cover 56% of variance, dots are colored according to their ID: QC and samples.

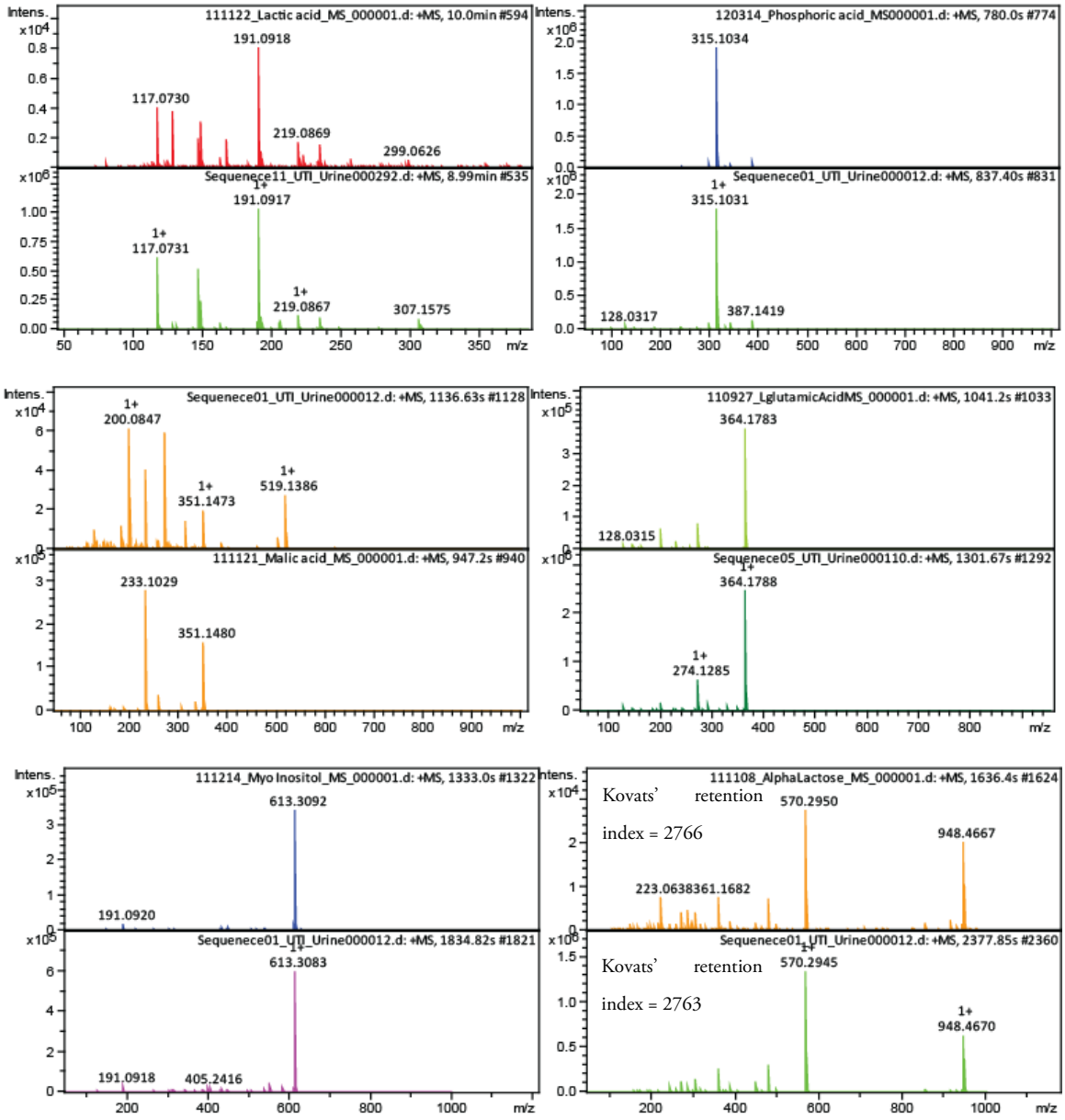


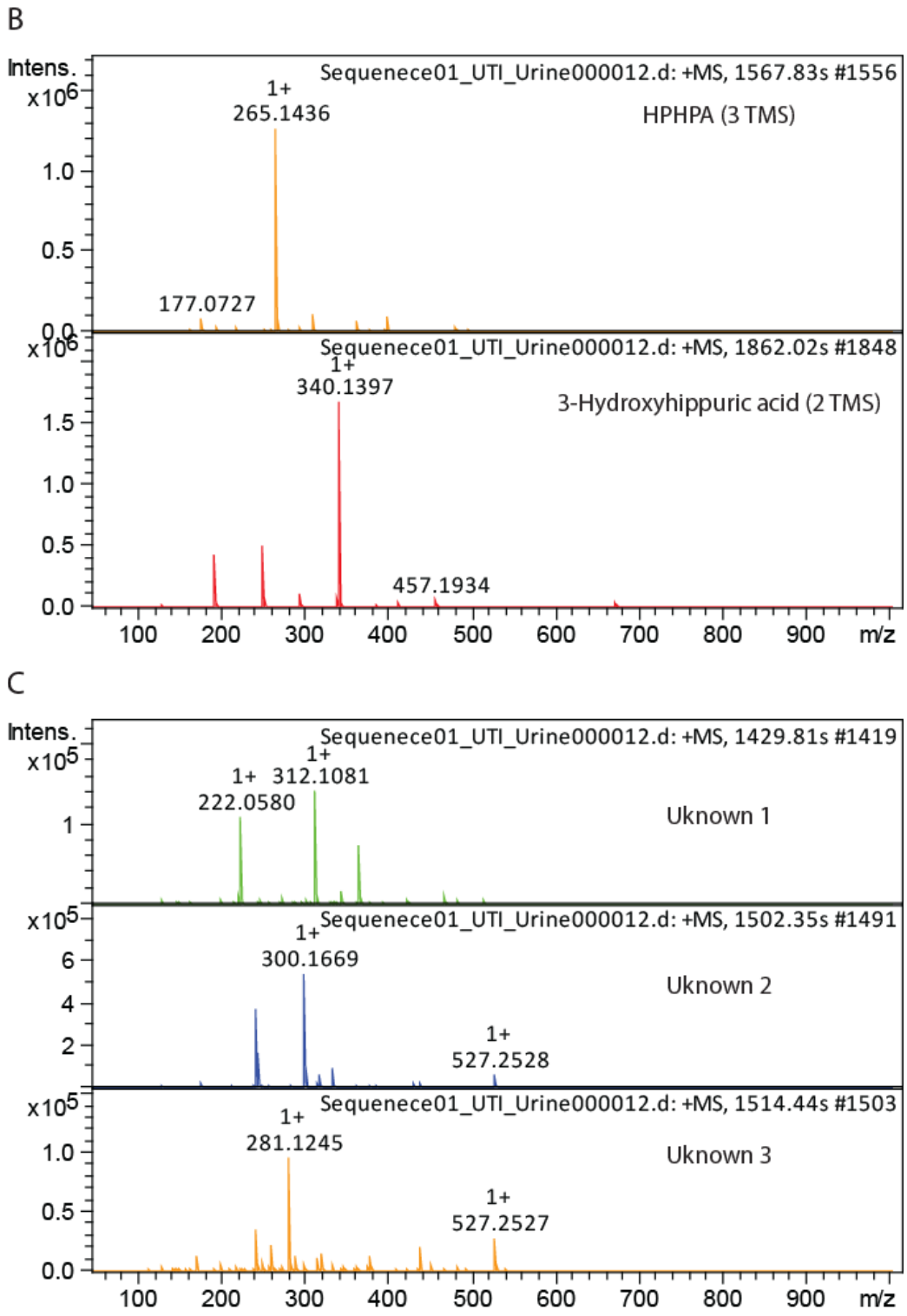
Supplementary Figure 2. A score plots showing the first and second (A) and the second and third (B) components of the PCA model (24 components cover 59% of the variance) built of 155 clinical samples. The samples are coloured according to the infectious status. The first and the second components display a major trend which cover more than 16% of the total variability (A). However, the second and the third component show a tendency in the cluster of the controls and UTI patients t=30 days against UTI patients t=0 day (B).



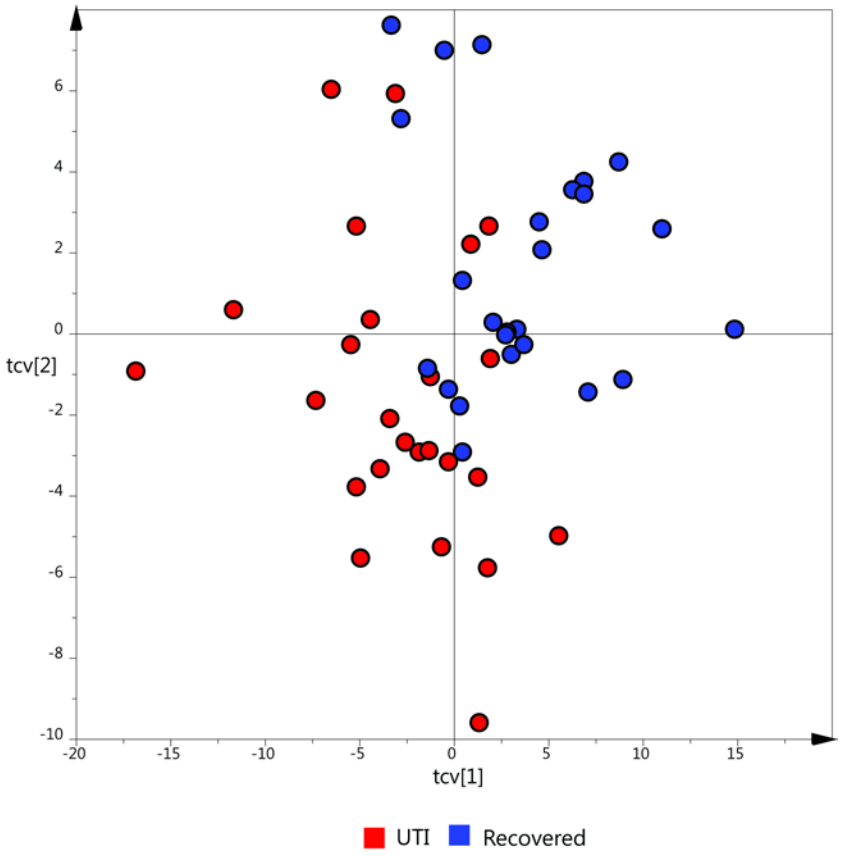
Supplementary Figure 3. PCA score plots for the models built for cross-sectional (A) and longitudinal data (B) parts separately. A) UTI Free + UTI t=0; 9 components cover 57% of the variance. B) UTI t=0 + UTI t=30; 16 components cover 55% of the variance .

A



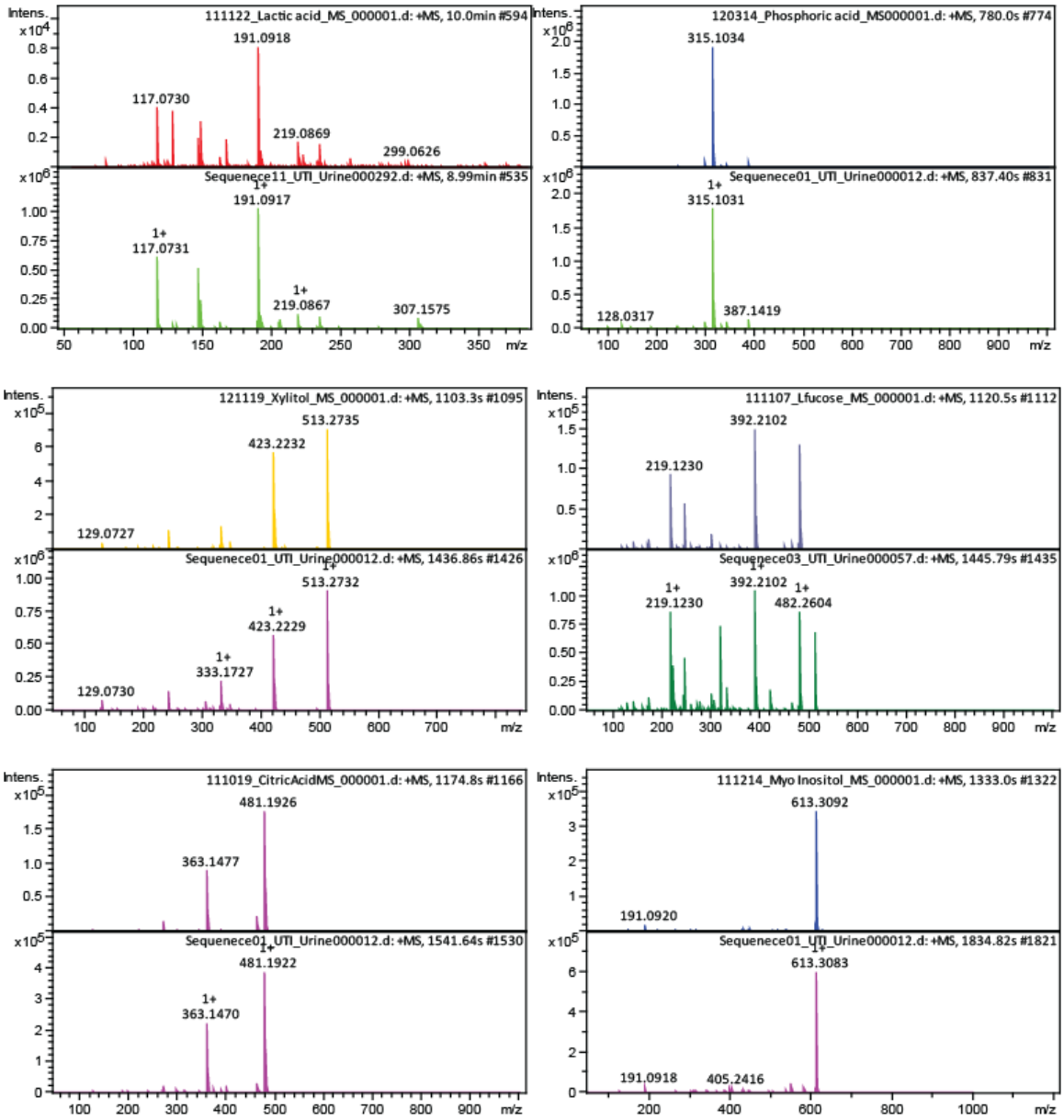


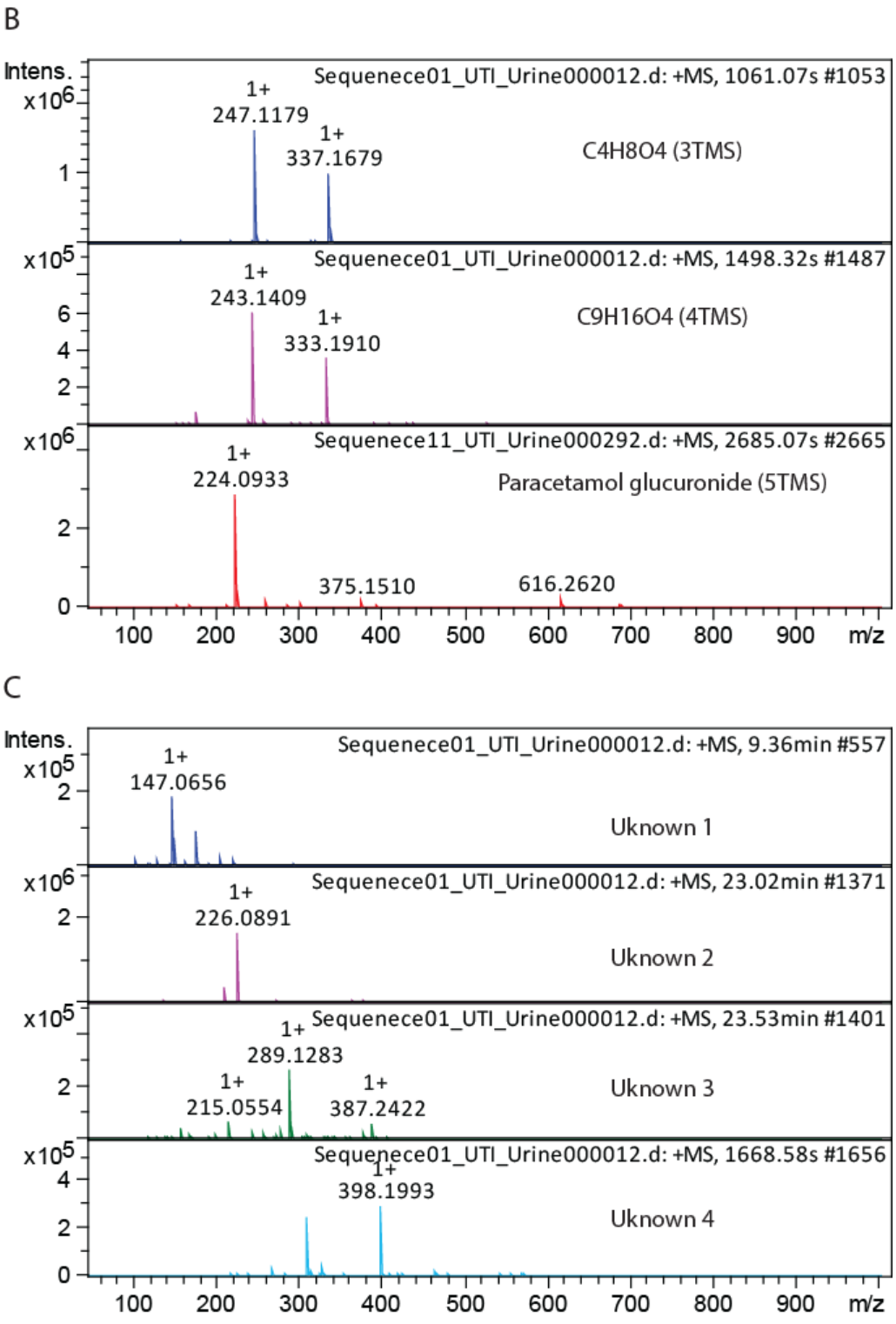
Supplementary Figure 4. MS spectra of classifiers reported in Table 2. A) Compounds identified with library matching B) compounds identified with *de novo* identification protocol and C) Unknown compounds.



Supplementary Figure 5. Cross-validates score plot of a PLS-DA built on UTI subjects (*E.coli*) who presented samples at both time point t=0 and t=30 days (50 samples in total). R2X=0.186, R2Y=0.816 Q2=0.46.

A





Supplementary Figure 6. MS spectra of classifiers reported in Table 4. A) compounds identified with library matching B) Compounds identified with *de novo* identification protocol and C) Unknown compounds.