Analyst

Accepted Manuscript

This is an *Accepted Manuscript*, which has been through the Royal Society of Chemistry peer review process and has been accepted for publication.

Accepted Manuscripts are published online shortly after acceptance, before technical editing, formatting and proof reading. Using this free service, authors can make their results available to the community, in citable form, before we publish the edited article. We will replace this *Accepted Manuscript* with the edited and formatted *Advance Article* as soon as it is available.

You can find more information about *Accepted Manuscripts* in the Information for Authors.

Please note that technical editing may introduce minor changes to the text and/or graphics, which may alter content. The journal's standard Terms & Conditions and the Ethical guidelines still apply. In no event shall the Royal Society of Chemistry be held responsible for any errors or omissions in this *Accepted Manuscript* or any consequences arising from the use of any information it contains.

ROYAL SOCIETY OF CHEMISTRY

www.rsc.org/analyst

Multiple metabolomics of uropathogenic *E. coli* reveal different information content in terms of metabolic potential compared to virulence factors.

Haitham AlRabiah[1], Yun Xu[1], Nicholas J. Rattray[1], Andrew A. Vaughan[1], Tarek Gibreel[2,3], Ali Sayqal[1], Mathew Upton[2,3], J. William Allwood[1,4] and Royston Goodacre[1]

[1]School of Chemistry and Manchester Institute of Biotechnology, University of Manchester, 131 Princess Street, Manchester, M1 7DN, UK.

[2]School of Medicine, University of Manchester, Stopford Building, Oxford Road, Manchester, M13 9PL, UK.

[3]Current address – Plymouth University, Peninsula Schools of Medicine and Dentistry, Plymouth University, Drake Circus, Plymouth, PL4 8AA, UK.

[4]Currrent address: Clinical and Environmental Metabolomics, School of Bioscience, University of Birmingham, Edgbaston, Birmingham, B15 2TT, UK

Abstract

No single analytical method can cover the whole metabolome and the choice of which platform to use may inadvertently introduce chemical selectivity. In order to investigate this we analysed a collection of uropathogenic *Escherichia coli*. The selected strains had previously undergone extensive characterisation using classical microbiological methods for a variety of metabolic tests and virulence factors. These bacteria were analysed using Fourier transform infrared (FT-IR) spectroscopy; gas chromatography mass spectrometry (GC-MS) after derivatisation of polar non-volatile analytes; as well as reversed-phase liquid chromatography mass spectrometry in both positive (LC-MS(+)) and negative (LC-MS(-)) electrospray ionisation modes. A comparison of the discriminatory ability of these four methods with the metabolic test and virulence factors was made using Procrustes transformations to ascertain which methods produce congruent results. We found that FT-IR and LC-MS(-), but not LC-MS(+), were comparable with each other and gave highly similar clustering compared with the virulence factors tests. By contrast, FT-IR and LC-MS(-) were not comparable to the metabolic tests, and we found that the GC-MS profiles were significantly more congruent with the metabolic tests than the virulence determinants. We conclude that metabolomics investigations may be biased to the analytical platform that is used and reflects the chemistry employed by the methods. We therefore consider that multiple platforms should be employed where possible and that the analyst should consider that there is a danger of false correlations between the analytical data and the biological characteristics of interest if the full metabolome has not been measured.

Analyst Accepted Manuscript

1. Introduction

Metabolomics aims to categorise the small molecular weight complement of cells, tissue and biofluids [1-3], and although arguably an 'ancient' science [4] a plethora of analytical platforms, mainly based on mass spectrometry (MS) and various molecular separation techniques including gas chromatography (GC) and liquid chromatography (LC), have made it possible to detect small molecules in biological matrices.[5]

In practice, the detection of the full metabolome is still unachievable by a single analytical tool due to the chemical complexity of metabolites, great variations in their concentration levels and various other reasons such as analyte lability.[6] Therefore, in addition to MS, other detection techniques such as NMR spectroscopy and vibrational spectroscopies (*viz*. FT-IR and Raman) are used as complementary analytical approaches. In particular, FT-IR spectroscopy is considered to be a low cost, high-throughput technique making it a first option for preliminary experiments to give a preview of the experiment direction before more advanced tools are employed.[7]

The question arises as to exactly how complementary these methods are. For example, in FT-IR spectroscopy sample extraction is usually not performed and the method provides chemical information at the level of molecular vibrations, not isolated metabolites *per se*. By contrast, MS-based studies are performed usually after extraction and usually after GC or LC. All of these processes introduce selectivity into the analysis and hence potential analytical bias. If we consider GC-MS using methanol extraction followed by a two-stage methoxamination and silyation [1, 8], one is generally selecting metabolites from central metabolism such as sugars, sugar phosphates, amino acids and small fatty acids etc. For LC-MS, using reversed-phase chromatography one targets more lipophilic species and another choice made is the polarity of the ion source in terms of positive or negative electrospray ionisation. It is currently unlikely that people have the resources to include all possible analytical approaches and therefore choices are made on which are the most appropriate or accessible to select.

Therefore in this study, we used a range of metabolomics platforms on a microbiologically characterised set of uropathogenic *Escherichia coli* (UPEC) isolates that all belong to the same sequence type (ST131), an important and globally disseminated clone.[9] Due to the platforms available in our laboratory, we selected FT-IR spectroscopy, GC-MS of polar non-

Analyst Accepted Manuscript

volatile analytes, and reversed-phase LC-MS in both positive and negative ESI modes. Both GC-MS and LC-MS analysis were performed on 80% methanol (80:20 methanol-water (vol/vol)) extracts. Once the data were collected, we used a series of chemometric methods to compare the differentiation ability of all four methods. Moreover, these were compared with genotypic and phenotypic characteristics that are measured during investigation of the pathogenic potential of UPEC and included data for a panel of metabolic tests and virulence factor carriage.

2. Experimental

2.1 General Chemicals

Unless otherwise stated, all chemicals were supplied by Fisher Scientific (Fisher Scientific Ltd., Loughborough, UK), and all solvents and acids were obtained from Sigma Aldrich (Sigma Aldrich, Dorset, UK).

2.2 Microorganisms

The 11 uropathogenic *Escherichia coli* (UPEC) isolates examined were obtained from bacteriuria urine samples submitted to the bacteriology laboratory at the Central Manchester Foundation Trust. The isolates were all from the ST131 lineage and resistant to quinolones due to different genetic mechanisms (Table S1). Identification of virulence capacity, metabolic profile and antibiotic susceptibility have been previously described [10, 11] and these are provided in Tables S2 and S3.

2.3 Preparation of *Escherichia coli* inoculates for metabolic fingerprinting and metabolic profiling

Samples were prepared according to the protocols described in [12] with the only exception being that samples were incubated for 21 h rather than 18 h (see Figure S1 for details). After cultivation of the bacteria (see Supplementary Information) each of the 4 biological replicates were split for FT-IR, GC-MS and LC-MS to ensure that results were obtained from the same biological cultures.

For GC-MS and LC-MS, 15 mL from each replicate was collected, quenched and extracted according to the procedures developed by [8]. The only difference in this study is that for

Analyst Accepted Manuscript

metabolite extraction 80% methanol (80:20 methanol-water (vol/vol)) was used rather than 100% methanol to enhance the recovery of polar small molecules. Samples for GC-MS and LC-MS, including quality control samples (QCs), were normalised to optical density (OD) and made up with 80% methanol (80:20 methanol-water (vol/vol)). Further sample processing steps were applied to the GC-MS samples (adding internal standards, a two-step chemical derivatisation and adding retention index marker solutions). LC-MS samples were reconstituted in 100 µL HPLC grade water, vortex mixed and centrifuged before instrumental analysis. Full details of sample preparations are available in the Supplementary Information.

2.4 FT-IR spectroscopy

A Bruker Equinox 55 infrared spectrometer (Bruker Ltd., Coventry, UK) equipped with a HTX™ module was used for FT-IR spectroscopic analysis using the method described in [12], [13]. Spectra were collected in the range of 4000-600 cm^{-1}, with 64 co-adds and at a resolution of 4 cm^{-1}.

2.5 GC-MS

A LECO Pegasus III TOF/MS was used to conduct GC-TOF/MS and its mode of operation is provided in the Supplementary Information following our established GC-MS protocol [14, 15], which follows Metabolomics Standards Initiative (MSI) guidelines.[16] After GC-MS, data were processed via the deconvolution method of [14]. QC samples were used before statistical analysis, as described by [17], to give quality assurance of data by evaluating and removing mass features exhibiting high deviation within the QC samples.

2.6 LC-MS

UHPLC-MS analysis was carried out on an Accela UHPLC autosampler system coupled to an electrospray LTQ-Orbitrap XL hybrid mass spectrometry system (ThermoFisher, Bremen, Germany) as previously described by [15, 17] and highlighted in the Supplementary Information. Note that the same samples were analysed twice: once in positive and again in negative ESI modes. QCs were also used as detailed in [17] to provide quality assurance of the LC-MS data.

Analyst Accepted Manuscript

2.7 Data analysis

The pre-processed FT-IR, GC-MS and LC-MS data (see Supplementary Information for full details) were first analysed using principal component analysis (PCA). The first 1:*n* PCs scores which explained ~75% of the total variance were then subjected to discriminant function analysis (DFA). DFA was calibrated with 11 classes (one for each of the 11 *E. coli* isolates) and the first 3 discriminant functions (ordinates) were retained. In order to make visualisation easier, and more importantly to balance the number of samples for Procrustes analysis (*vide infra*), as each class contained 36 FT-IR spectra (4 biological replicates, 3 spots for each and 3 measurements off each spot) these were mean-averaged to generate 11 DFA coordinates for the 11 isolates. In a similar fashion for GC-MS and LC-MS (in both ion source modes) where each sample was represented by 4 injections (1 for each of the 4 biological replicates), the resulting DFA scores were also averaged.

In addition to the analytical metabolomics data, the *E. coli* strains had also been subjected to classical microbiological testing. Metabolic activity was probed via 47 biochemical tests (Table S3) designed to measure carbon source utilisation and enzymatic activity using the Vitek 2 ID-GNB card and the Vitek 2 compact Automated Expert System (Biomérieux).[11] The virulence capabilities (Table S2) of these strains were investigated through genetic screening for the presence of 29 ExPEC associated Virulence Factors (VF) encompassing five categories (adhesins, toxins, siderophores, capsule and "miscellaneous").[10]

These metabolic tests (MT) and VF tests are characters that are both represented as present/absent data. These are clearly very different to the FT-IR, GC-MS and LC-MS quantitative data which are all continuous data. To make these two different data types comparable with each other, the pattern of the MT and the VF test data sets were also projected into ordination space using the following procedure: first a pair-wise distance matrix was calculated to measure the dissimilarity between every pair of the isolates using the Jaccard distance [18]; next principal coordinate analysis (PCoA) was performed on the square rooted distance matrix and the first 3 PCs were retained.[19]

The result of the above analysis was six different ordination analyses: PC-DFA from the four metabolomics data sets, and PCoA from the metabolic tests and virulence factors. In order to compare the similarity in the discriminatory ability generated by these different analyses Procrustes analysis was performed on all possible data set pairs.[20] In this process, the

Analyst Accepted Manuscript

similarity is measured in terms of the Procrustes error, which varies from 0 to 1; where 0 indicates a perfect match and 1 indicates that the two sets of clusters are completely different. The statistical significance level of the levels of these similarities were assessed using a Procrustean test procedure.[21] For each comparison, 10,000 permutation tests were performed by permuting the order of the samples in the data sets and subsequently performing the Procrustes analysis. The Procrustes errors of these permutations were recorded to form a null distribution. The observed Procrustes error was then compared against the null distribution and an empirical *p*-value was derived by counting the number of cases when the Procrustes error obtained from the permuted data sets was lower than the observed error; this was then divided by 10,000 (the total number of the permutation tests).

If any of the pair-wise comparisons indicated comparable clusters, it would also be interesting to investigate which variables in the metabolomics data sets (i.e., FT-IR, GC-MS and LC-MS in both +ve and -ve ionisation mode data sets) were mainly responsible for the matched patterns revealed after the Procrustes rotation. This was achieved by first projecting the loadings of the PCA into the PC-DFA space using the DFA loadings and then rotating these again using the Procrustes orthogonal rotation matrix. The resultant loadings were denoted as Procrustes rotated loadings. The variables with significantly high loadings were the ones that contributed most to the matched pattern after the Procrustes rotation.

3. Results and discussions

In clinical microbiology, bacterial characterisation is largely dependent on phenotypic methods such as biochemical tests and bacterial morphology. These are time consuming and often provide limited information when compared with modern bioanalytical techniques. The two most common biochemical tests that microbiologists use are (i) those based on metabolic tests which involve growth on selective media to test for specific enzymes and (ii) assays for virulence factors which often reflect how the microorganism interacts with its environment and include its adhesins and capsule as well as any toxins produced. In general terms, metabolic tests reflect the organism's metabolic potential whilst some virulence tests probe the surface of the microorganism, as it is this surface that interacts with the environment.

To assess the level of information that metabolomics data may generate from microbiological samples, we compare four metabolomics approaches with each other and, importantly, with

Analyst Accepted Manuscript

these two classical microbiology tests from a range of UPEC isolated from a local hospital. The results from the metabolomics methods, MT and VF, were analysed using cluster analysis and these generated six different ordination scores plots: four PC-DFA plots from the FT-IR, GC-MS and LC-MS in both +ve and -ve ionisation modes and PCoA from the MT and VF. The resulting cluster plots then need to be compared and this is very difficult by eye. For example, the comparison of two sets of clusters in three dimensions requires one to: (i) first translate the spatial clustering (arrangement of samples) of one sample set onto the other, so that they are now both centred together; (ii) next, the clusters are scaled so that they are of equivalent size; (iii) finally, the clusters are aligned by rotation. Of course for simple shapes, this can be done by eye. The problem is that for the comparison of clusters generated from six different methods (as in this study) the number of unique comparisons that needs to be made is 15, and these need to be ranked and objectively assessed. Therefore in this study, we used a series of Procrustes transformations.

The Procrustes errors with the associated *p*-values of the pattern comparisons were calculated as described above and are presented in Table 1. In this table the comparisons which revealed very similar spatial arrangements of the clusters from the PCoA and PC-DFA are highlighted in yellow. A Venn diagram-like figure reflecting these overall comparisons is shown in Figure 1. This figure was constructed by first performing PCoA on the Procrustes errors table and converting it to a 2-D X-Y scatter scores plot. Next, we calculated the 95% χ^2 confident regions (these are the ellipses shown in the plot) around each class, assuming that each have the same size of covariance matrices; this presumes that following the Procrustes transformation all resulting data transformations would have the same scale. It is clear from this comparison in Figure 1 that there are mainly four congruent pairs of clusters. In Table 1, these can be judged by having a low *p*-value (<0.01; from multiple testing). These are highlighted below:

1. The LC-MS profiles in negative mode and the virulence factor test data had the highest similarity level with a Procrustes error of 0.4533 and the associated *p*-value was 0.0002 (i.e. only 2 out of 10,000 permutations had obtained a higher Procrustes error).
2. The FT-IR spectra also obtained a statistically significant similarity to the VF test data with a *p*-value of 0.0072. By contrast, GC-MS and LC-MS in positive mode did not have a significant similarity to the VF test data ($p > 0.01$).

Analyst Accepted Manuscript

3. The GC-MS metabolite data obtained a very significant similarity to the classical metabolic tests (MT; p=0.0006), while the other 3 data sets had no significant similarity to this type of data (p> 0.01). We note that in Figure 1 there is some congruence between GC-MS with the VF but this is not as strong as the MT.
4. For the comparisons between the four metabolomics data types, the FT-IR data and LC-MS profiles in the negative mode had similar shapes, and this was to be expected as both were very similar to the VF test.
5. Finally there was low similarity between the VF test and the metabolic test as p>0.01.

3.1 Interpretation of FT-IR spectra

FT-IR spectroscopy is not a particularly popular metabolic fingerprinting method but it has been extensively used for so called 'whole-organism fingerprinting' [22] for bacterial characterisations due to its high-throughput nature with minimal sample preparation.[23-26] In this study, FT-IR was applied to discriminate between isolates with the same sequence type and the FT-IR clusters had similar scores to those from virulence factor tests (Figure 2 and Table 1). Figure 2 shows the results from both the FT-IR (in red) and VF (in blue) where it can be seen that, in FT-IR, isolate 48 forms a cluster that is distinct from the other isolates, but is collocated with results from its VF test. Inspection of Table S2, which shows the scores of the different virulence tests, reveals isolate 48 is the only isolate with a negative score for *PAI*. *PAI* is an acronym for pathogenicity islands, which are mobile genetic elements that carry the genes responsible for the production of many virulence factors, including protein secretion systems, toxins, adhesins and many others.[27] FT-IR spectra from intact bacteria contain information on fatty acids, amides, polysaccharides, proteins and amino acids. As these virulence factors may be located in the membrane (outer surface of the organism), it is likely that FT-IR spectroscopy is detecting the loss of these as the whole organism is analysed and hence that is why it is located away from the other 10 isolates.

Isolates 52 and 75, 160 and 164 share the same VF profile, with the exception of strains 160 and 164 being negative for *traT* (Table S2), a cell surface molecule involved in resistance to the activity of complement (serum). All four isolates cluster together in the FT-IR data and are located reasonably close to their respective clusters from VF; they are located in the positive side of PC1 (Figure 2) and this may reflect that these isolates are all positive for the *afa/draBC* surface adhesins. Isolate 2 is also coincident in terms of FT-IR spectra with these

Analyst Accepted Manuscript

four isolates but is very different for VF and this disparity was also observed for the LC-MS in negative ionisation mode comparison with VF (*vide infra*).

Capsular association factors (*kpsMT K5* and *kpsMT II*) are extracellular and this may be reflected in the FT-IR spectra. Isolates 2, 25, 48, 183, 184 and 230 are positive for both these factors and, with the exception of isolate 2, are located on the negative part of PC1. Isolate 124 is also associated with these isolates and this may be a consequence of it being negative for *afa/draBC* as discussed above.

Finally, no relationship between FT-IR spectra and *traT* was evident from this analysis and this was also observed for the LC-MS conducted in the negative ionisation mode.

3.2 Interpretation of LC-MS profiles

The same 11 *E. coli* isolates from uropathogenic infections were also analysed by reversed-phased LC-MS. As discussed above, 80% methanol (80:20 methanol-water (vol/vol)) extracts were prepared from these bacterial cultures and MS was performed in both positive ($LC\text{-}MS^{+ve}$) and negative ($LC\text{-}MS^{-ve}$) ionisation modes. Comparisons were made with VF and MT and it was found that LC-MS in the negative ionisation mode shows a higher level of similarity with VF tests than FT-IR spectroscopy did (Table 1 and Figure 3). Moreover, because of these congruencies between [$LC\text{-}MS^{-ve}$ and VF] and [FT-IR and VF] it was not surprising that the [$LC\text{-}MS^{-ve}$ and FT-IR] comparison was also very similar (Table 1).

There were, however, two minor differences between the $LC\text{-}MS^{-ve}$ comparison with VF (Figure 3) compared with the FT-IR spectroscopic comparison (Figure 2) and these are briefly highlighted below:

- The first significant disparity is the observation that isolates 2, 25 and 184 were collocated in $LC\text{-}MS^{-ve}$ mode whereas they were significantly spread in PC1 in FT-IR. We note that they possess identical VF tests (Table S2) and a possible explanation for this is that $LC\text{-}MS^{-ve}$ is detecting these preferentially compared with FT-IR (Table 1).
- The second difference is that in FT-IR, isolates 2, 52, 75, 160 and 164 were very closely clustered together. By contrast, in $LC\text{-}MS^{-ve}$ isolates 160 and 164 ‘moved’ to the positive parts of PC1 and PC2 and cluster very closely with their respective VF

Analyst Accepted Manuscript

tests, whilst isolates 2, 52 and 75 are now collocated near the origin with isolates 124 and 183 (Figure 3).

It is possible that some of these small differences are because in LC-MS a methanolic extract is used compared to FT-IR where whole-organism fingerprinting is used. The similarity between the differentiation ability of FT-IR and LC-MS^{-ve} with VF is interesting and this may reflect that both metabolomics methods are preferentially detecting cell wall components. As discussed above, FT-IR analyses the intact bacteria and certainly contains information on proteins and lipids, amongst other cellular components. In LC-MS, as reversed-phase LC is used with the negative ionisation mode more lipophilic species are analysed that may be associated with the cell wall and this has been reported before for direct infusion MS.[28, 29]

In the positive mode of LC-MS, very little similarity with VF was observed (Table 1). By contrast, although comparison of LC-MS^{+ve} with MT (Figure S2) showed some congruence; this was not statistically significant and so will not be discussed here.

3.3 Interpretation of GC-MS profiles

The GC-MS approach used here [30] generates information-rich metabolite profiles of polar analytes and so mainly covers metabolites involved in the central metabolism. As can be seen from Figure 4, there is high similarity between GC-MS profiles for the 11 bacteria (highlighted in red) with the metabolic tests (in blue) and the similarity match is 0.5681 and is highly significant with p=0.0006 (Table 1).

Isolates 160 and 164 share exactly the same results from MT and they are located closer to each other in the positive side of PC1 with isolate 183. Following Procrustes transformation of the PC-DFA from the GC-MS data, isolates 160 and 164 are very similar and are recovered far from all other isolates, which are congruent with their MT except for 230 which is positive in PC2. Inspection of the MT (Table S3) reveals that 160 and 164 are unique from all other isolates in that they scored positive in the GlyA test, which detects the glycine arylamidase enzyme. Arylamidase enzymes mainly hydrolyse peptides containing L-amino acids with an unsubstituted α-amino group in the N-terminal residue [31] and one of the main amino acids released by this enzyme is leucine.[32] Therefore, the PC-DFA loading plots from GC-MS were produced (Figure S3) and it was found that two variables were highly

Analyst Accepted Manuscript

discriminatory (variables 17 and 49). Variable 17 is identified by in house database matching to leucine and shows a much higher level in these two isolates than in the other *E. coli* (Figure 5a). Moreover, arylamidase enzymes are involved in 8 of the metabolic tests in this experiment (Table S3) and isolates 160 and 164 have the highest scores in these tests compared with others.

The other variable that was identified as significant (Figure S3) was variable 49, which unfortunately we are unable to identify. When this feature is plotted for the 11 isolates (Figure 5b) it is also elevated in isolates 160 and 164 confirming its importance as a discriminatory metabolite feature. We note also that isolate 183 also has increased levels compared with all the other isolates, although its level is not as high as the levels generated by 160 and 164.

In terms of metabolic tests, isolate 183 is closer to isolates 160 and 164 as can be seen from its blue coding in Figure 4, and in GC-MS it is recovered to the right of the other 8 isolates and in the positive part of PC1. It shares the same metabolic results with these two isolates in all tests except GlyA (glycine arylamidase) and PHOS (phosphatase) tests. It is expected to observe a notable signal by phosphatase as the production of alkaline phosphatase is induced by alkaline environment generated by peptide metabolism.[33] Although phosphate is produce in many metabolic reactions this elevation is generally reflected for most of the strains that express phosphatase (Table S3) and this is generally reflected in the phosphate levels measured by GC-MS (Figure 5c).

4. Concluding remarks

In metabolomics, investigation choices of the most appropriate analytical method have to be made. To date, most of these are based on early decisions to do with analytical procurement due to the expense of metabolomics instrumentation. The question arises as to whether equivalent results are generated by all platforms. In this investigation, we attempted to address this by analysing a set of well characterised uropathogenic *E. coli* that had been analysed by a battery of metabolic tests (n = 47) and virulence factor determinations (n = 30). These tests probe different parts of the bacterial cell. Obviously, metabolic tests probe the enzyme component of the bacteria and are usually focused on central metabolism and carbon

Analyst Accepted Manuscript

utilisation. By contrast, virulence factors tend to be cell wall associated and include adhesins, capsules and toxins.

Four different approaches for metabolomics were investigated. FT-IR spectroscopy was employed directly on intact bacteria for metabolic fingerprinting, or what is often described as whole-organism fingerprinting. Following quenching and extraction using methanol, GC-MS was performed following a two-stage derivatisation, and LC-MS was performed in reversed-phased LC mode directly on the methanolic extracts in both positive and negative ionisation source modes.

In order to compare the clustering patterns from the six different analyses with one another, Procrustes transformations were performed and this allowed objective assessment of the similarity of the cluster patterns in terms of the spatial arrangement of the 11 *E. coli* isolates in either PCoA or PC-DFA scores space. We found that FT-IR and LC-MS in negative ionisation mode were comparable with each other and also with the virulence factors tests but not comparable to the metabolic tests. By contrast, GC-MS compared well with metabolic tests but not the virulence determinants. Although LC-MS in the positive ionisation mode was not statistically correlated with either, visual inspection of clusters with the metabolic tests suggested there may be some loose congruence between the two methods.

In conclusion, we believe that whenever possible more than one metabolomics modality should be used, and the analyst should consider carefully the analytical technique employed and these will certainly reflect the chemical bias of the methods used. We know for example that LC-MS mainly targets lipophilic species when reversed phase is used; by contrast, GC-MS mainly focuses on polar small molecules. It is possible that there is a danger of false correlations between the analytical data and the biological characteristics of interest if the full metabolome has not been measured. This is clearly demonstrated in this study where the GC-MS data predominantly correlates with the metabolic tests, whilst LC-MS in negative ionisation mode and FT-IR spectroscopy correlate with the virulence determinants. Of course if we did not know about these two different types of inherent characteristics we may have jumped to false conclusions, and the same rules are likely to be manifest when metabolomics is employed to study higher organisms like mammalian systems and plants as well as complex body fluids.

Analyst Accepted Manuscript

5. Acknowledgments and Notes

HR thanks The Saudi Ministry of higher education and King Saud University for funding. YX, AAV, JWA and RG are also indebted to the Cancer Research UK for funding. NR and RG are also grateful to the UK MRC for funding.

† Electronic Supplementary Information (ESI) available: See DOI: XXXXX

Analyst Accepted Manuscript

Table and Figures

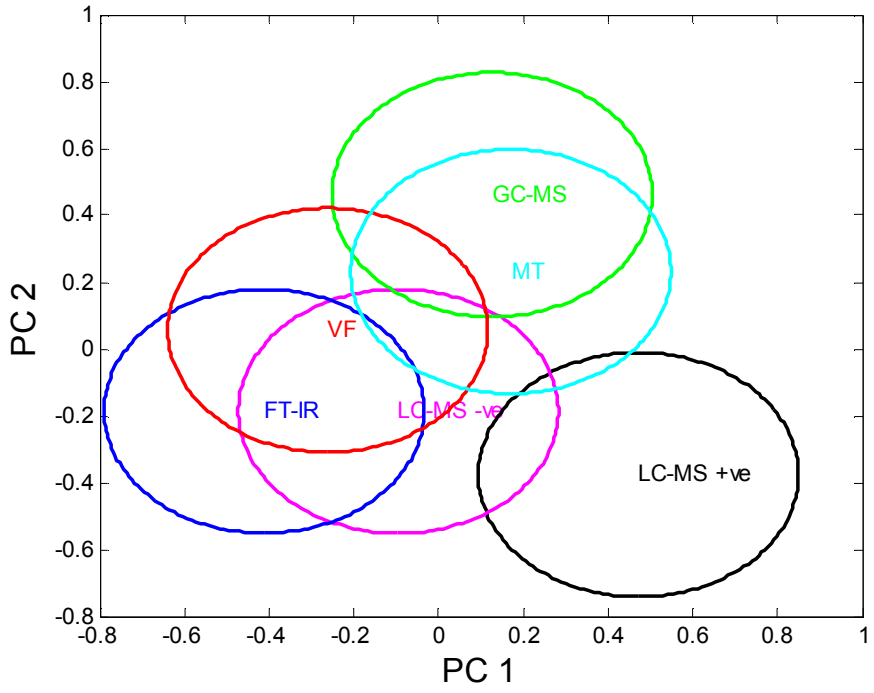


Figure 1 Venn diagram-like plotted showing the overall clustering congruence between the four analytical approaches and the two microbiological tests. See text for explanation of its construction.

Analyst Accepted Manuscript

Table 1 The Procrustes errors with the associated *p*-values of the pair-wise comparisons

	LC-MS (pos)	LC-MS (neg)	GC-MS	FT-IR	VF	Metabolic test
LC-MS (pos)	-					
LC-MS (neg)	0.6699 (*p*=.0543)	-				
GC-MS	0.9239 (*p*=0.7521)	0.7423 (*p*=0.0903)	-			
FT-IR	0.9344 (*p*=0.8118)	0.5333 (*p*=0.0059)	0.8973 (*p*=0.3701)	-		
VF	0.8855 (*p*=0.5633)	0.4533 (*p*=0.0002)	0.6603 (*p*=0.0107)	0.5429 (*p*=0.0072)	-	
Metabolic test	0.7782 (*p*=0.2021)	0.6737 (*p*=0.072)	0.5681 (*p*=0.0006)	0.7843 (*p*=0.2195)	0.6653 (*p*=0.091)	-

Values highlighted in yellow are considered significant ($p < 0.01$) and indicate pairs of methods that provide equivalent clusters/shapes.

Analyst Accepted Manuscript

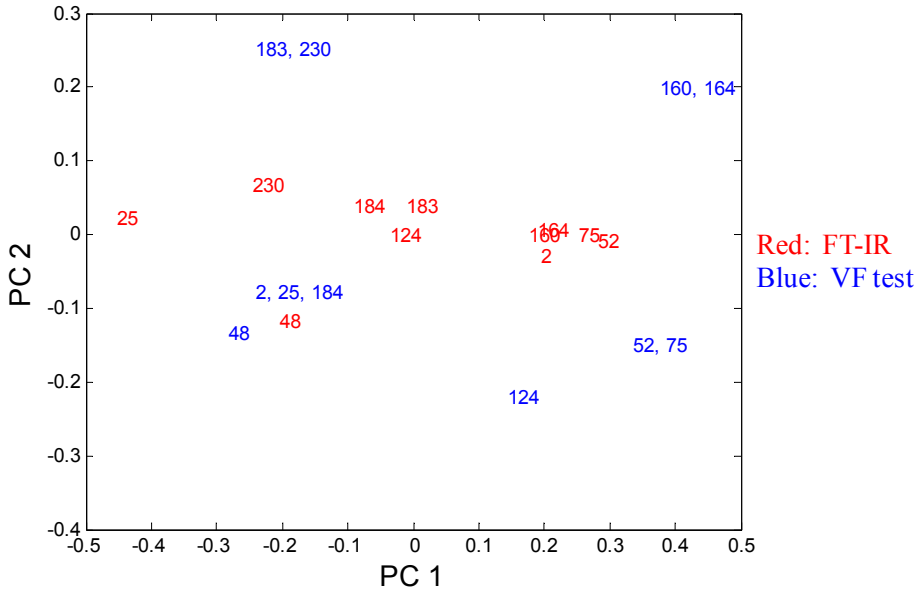


Figure 2 Super-imposed scatter plots of PCoA scores of the first two components of the VF tests and Procrustean-transformed FT-IR spectra.

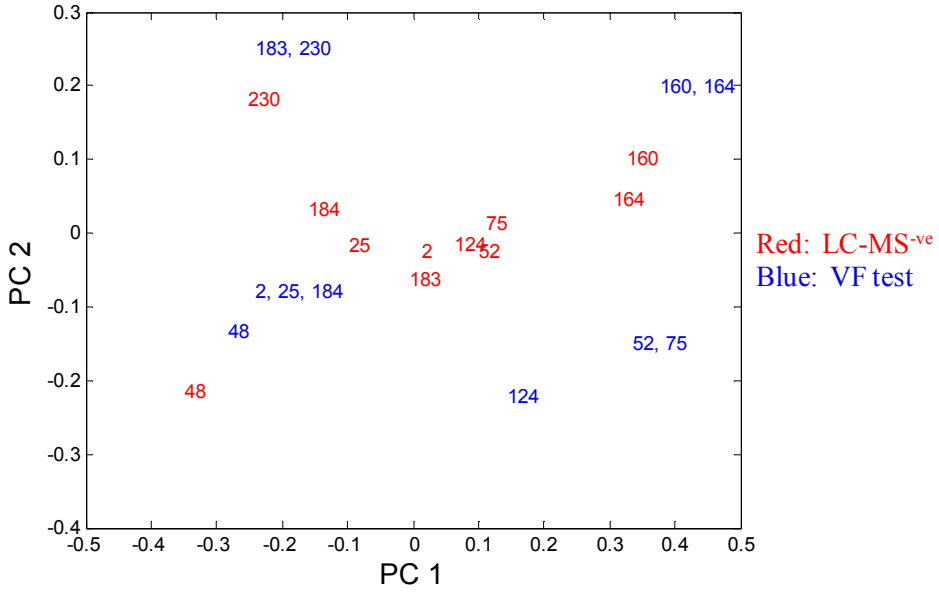


Figure 3 Superimposed scatter plots of PCoA scores of the first two components of the VF tests and Procrustean-transformed LC-MS negative mode data.

Analyst Accepted Manuscript

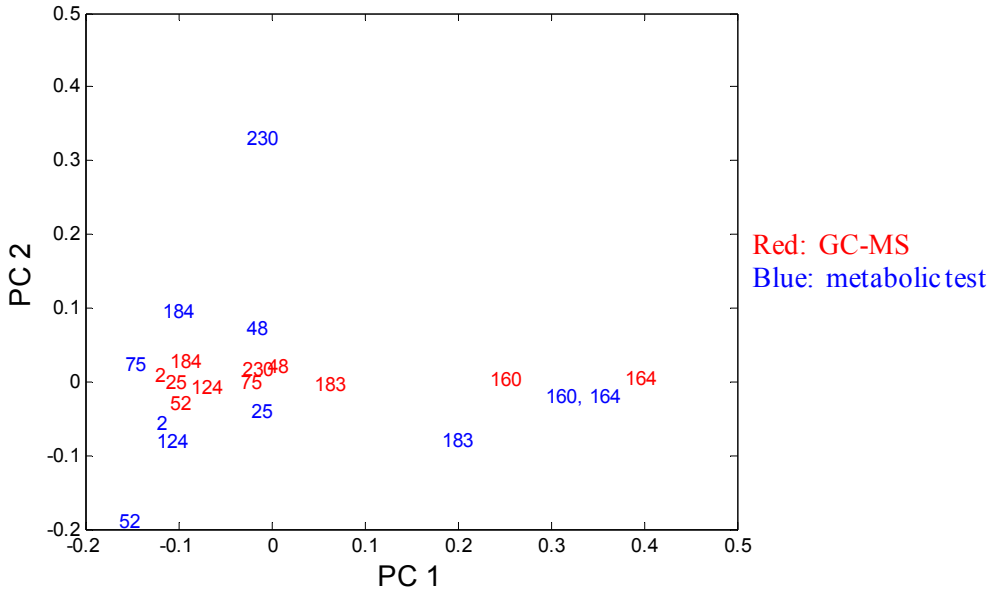


Figure 4 Superimposed scatter plots of PCoA scores of the first two components of the metabolic tests and Procrustean-transformed GC-MS data.

Analyst Accepted Manuscript

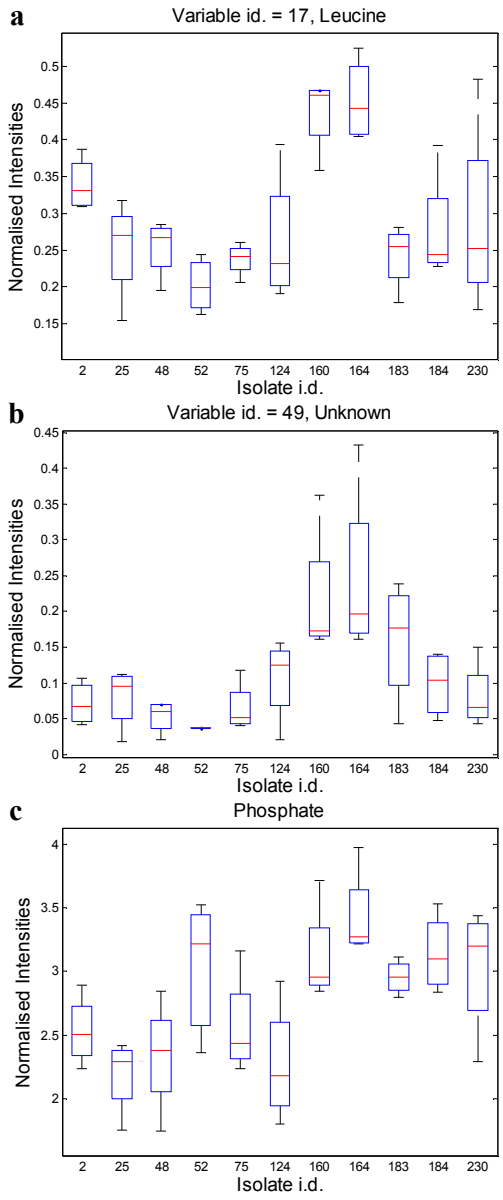


Figure 5 Box-whisker plots for each isolate demonstrating the concentration level of candidate intracellular metabolites from (a) variable 17 (leucine), (b) variable 49 (unknown), and (c) phosphate

Analyst Accepted Manuscript

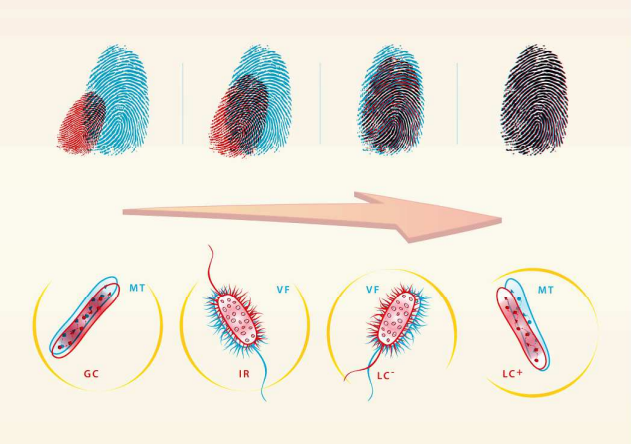
References

1. O. Fiehn, *Plant Mol.Biol.*, 2002, **48**, 155-171.
2. S. G. Oliver, M. K. Winson, D. B. Kell and F. Baganz, *Trends Biotechnol.*, 1998, **16**, 373-378.
3. W. B. Dunn, D. I. Broadhurst, H. J. Atherton, R. Goodacre and J. L. Griffin, *Chem. Soc. Rev.*, 2011, **40**, 387-426.
4. M. Oresic, *Nutrition Metabolism and Cardiovascular Diseases*, 2009, **19**, 816-824.
5. W. Dunn, A. Erban, R. M. Weber, D. Creek, M. Brown, R. Breitling, T. Hankemeier, R. Goodacre, S. Neumann, J. Kopka and M. Viant, *Metabolomics*, 2013, **9**, 44-66.
6. R. Goodacre, S. Vaidyanathan, W. B. Dunn, G. G. Harrigan and D. B. Kell, *Trends Biotechnol.*, 2004, **22**, 245-252.
7. W. B. Dunn, N. J. C. Bailey and H. E. Johnson, *Analyst*, 2005, **130**, 606-625.
8. C. L. Winder, W. B. Dunn, S. Schuler, D. Broadhurst, R. Jarvis, G. M. Stephens and R. Goodacre, *Anal. Chem.*, 2008, **80**, 2939-2948.
9. S. H. Lau, M. E. Kaufmann, D. M. Livermore, N. Woodford, G. A. Willshaw, T. Cheasty, K. Stamper, S. Reddy, J. Cheesbrough, F. J. Bolton, A. J. Fox and M. Upton, *Journal of Antimicrobial Chemotherapy*, 2008, **62**, 1241-1244.
10. T. M. Gibreel, A. R. Dodgson, J. Cheesbrough, F. J. Bolton, A. J. Fox and M. Upton, *Journal of Clinical Microbiology*, 2012, **50**, 3202-3207.
11. T. M. Gibreel, A. R. Dodgson, J. Cheesbrough, A. J. Fox, F. J. Bolton and M. Upton, *Journal of Antimicrobial Chemotherapy*, 2012, **67**, 346-356.
12. H. AlRabiah, E. Correa, M. Upton and R. Goodacre, *Analyst*, 2013, **138**, 1363-1369.
13. C. L. Winder, S. V. Gordon, J. Dale, R. G. Hewinson and R. Goodacre, *Microbiology*, 2006, **152**, 2757-2765.
14. P. Begley, S. Francis-McIntyre, W. B. Dunn, D. I. Broadhurst, A. Halsall, A. Tseng, J. Knowles, R. Goodacre and D. B. Kell, *Anal. Chem.*, 2009, **81**, 7038-7046.
15. W. B. Dunn, D. Broadhurst, P. Begley, E. Zelena, S. Francis-McIntyre, N. Anderson, M. Brown, J. D. Knowles, A. Halsall and J. N. Haselden, *Nat. Protoco.*, 2011, **6**, 1060-1083.
16. L. W. Sumner, A. Amberg, D. Barrett, M. H. Beale, R. Beger, C. A. Daykin, T. W. M. Fan, O. Fiehn, R. Goodacre, J. L. Griffin, T. Hankemeier, N. Hardy, J. Harnly, R. Higashi, J. Kopka, A. N. Lane, J. C. Lindon, P. Marriott, A. W. Nicholls, M. D. Reily, J. J. Thaden and M. R. Viant, *Metabolomics*, 2007, **3**, 211-221.
17. D. C. Wedge, J. W. Allwood, W. Dunn, A. A. Vaughan, K. Simpson, M. Brown, L. Priest, F. H. Blackhall, A. D. Whetton, C. Dive and R. Goodacre, *Anal. Chem.*, 2011, **83**, 6689-6697.
18. P. Jaccard, *New Phytologist*, 1912, **11**, 37-50.
19. I. Borg and P. J. F. Groenen, *Modern Multidimensional Scaling: Theory and Applications*, Springer, London, 2005.
20. J. M. Andrade, M. P. Gomez-Carracedo, W. Krzanowski and M. Kubista, *Chemom. Intell. Lab. Syst.*, 2004, **72**, 123-132.
21. D. A. Jackson, *Ecoscience*, 1995, **2**, 297-303.
22. R. Goodacre, E. M. Timmins, R. Burton, N. Kaderbhai, A. M. Woodward, D. B. Kell and P. Rooney, *Microbiology* 1998, **144**, 1157-1170.
23. S. Garip, F. Bozoglu and F. Severcan, *Appl. Spectrosc.*, 2007, **61**, 186-192.
24. L. Mariey, J. P. Signolle, C. Amiel and J. Travert, *Vib. Spectrosc.*, 2001, **26**, 151-159.
25. D. Naumann, D. Helm and H. Labischinski, *Nature*, 1991, **351**, 81-82.
26. R. Goodacre, E. M. Timmins, P. J. Rooney, J. J. Rowland and D. B. Kell, *FEMS Microbiol. Lett.*, 1996, **140**, 233-239.
27. J. Hacker and J. B. Kaper, *Annu. Rev. of Microbiol.*, 2000, **54**, 641-679.
28. S. Vaidyanathan, J. J. Rowland, D. B. Kell and R. Goodacre, *Anal. Chem.*, 2001, **73**, 4134-4144.

Analyst Accepted Manuscript

29. S. Vaidyanathan, D. B. Kell and R. Goodacre, *Journal of the American Society for Mass Spectrometry*, 2002, **13**, 118-128.
30. O. Fiehn, J. Kopka, P. Dormann, T. Altmann, R. N. Trethewey and L. Willmitzer, *Nat. Biotechnol.*, 2000, **18**, 1157-1161.
31. F. J. Behal and S. T. Cox, *J. of Bacteriol.*, 1968, **96**, 1240-1248.
32. P. S. Riley and F. J. Behal, *J. of Bacteriol.*, 1971, **108**, 809-816.
33. S. J. Van Dien and J. D. Keasling, *Journal of Theoretical Biology*, 1998, **190**, 37-49.

For TOC only:



Analyst Accepted Manuscript