Supporting Information for the paper *Is serum or plasma more appropriate for inter-subject comparisons in metabolomic studies? An assessment in patients with small-cell lung cancer*

David C. Wedge[1], J. William Allwood[1], Warwick Dunn [1,2,3], Andrew A. Vaughan[1], Kathryn Simpson[4], Marie Brown[3], Lynsey Priest[4], Fiona H. Blackhall[4], Anthony D. Whetton[5], Caroline Dive[4] and Royston Goodacre[1,2]

[1] School of Chemistry, Manchester Interdisciplinary Biocentre, University of Manchester, 131 Princess Street, Manchester, M1 7DN, UK

[2] Manchester Centre for Integrative Systems Biology, Manchester Interdisciplinary Biocentre, University of Manchester, 131 Princess Street, Manchester, M1 7DN, UK

[3] Centre for Advanced Discovery & Experimental Therapeutics (CADET), Central Manchester NHS Foundation Trust and School of Biomedicine, University of Manchester, Manchester Academic Health Sciences Centre (MAHSC); York Place, Oxford Road, Manchester, M13 9WL, UK

[4] Clinical and Experimental Pharmacology Group, Paterson Institute for Cancer Research and Manchester Cancer Research Centre (MCRC), Manchester Academic Health Science Centre, University of Manchester, Wilmslow Road, Withington, Manchester, M20 4BX, UK

[5] School of Cancer and Enabling Sciences, Manchester Academic Health Science Centre, University of Manchester, Manchester, M20 3LJ, UK

Corresponding Author
Andrew A. Vaughan, email: andrew.vaughan-2@manchester.ac.uk, address: School of Chemistry, Manchester Interdisciplinary Biocentre, University of Manchester, 131 Princess Street, Manchester, M1 7DN, UK, tel.: +44 (0)161 3065145

Abstract

In clinical analyses, the most appropriate biofluid should be analysed for optimal assay performance. For biological fluids, the most readily accessible is blood, and metabolomic analyses can be performed either on plasma or serum. To determine the optimal agent for analysis, metabolic profiles of matched human serum and plasma were assessed by Gas Chromatography-Time of Flight-Mass Spectrometry and Ultra High Performance Liquid Chromatography Mass Spectrometry (in positive and negative electrospray ionisation modes). Comparison of the two metabolomes, in terms of reproducibility, discriminative ability and coverage, indicated that they offered similar analytical opportunities. An analysis of the variation between 29 small-cell lung cancer (SCLC) patients revealed that the differences between individuals are markedly similar for the two biofluids. However, significant differences between the levels of some specific metabolites were identified, as were differences in the inter-subject variability of some metabolite levels. Glycerophosphocholines, erythritol, creatinine, hexadecanoic acid and glutamine in plasma, but not in serum, were shown to correlate with life expectancy for SCLC patients, indicating the utility of metabolomic analyses in clinical prognosis and the particular utility of plasma in relation to the clinical management of SCLC.

Supporting Information

Sample collection, storage and selection

Blood was collected in either Monovette serum gel tubes (for processing to serum) or in Monovette Li-heparin tubes (for processing to plasma). Serum samples were left to clot for up to 120 min at room temperature, centrifuged at 2000×*g* for 10 min and then stored at -80°C. Plasma samples were stored at room temperature and were processed within 120 min of collection by centrifuging at 1000×*g* for 10 min followed by storage at -80°C. Samples were obtained at baseline, prior to any chemotherapy treatment. Aliquots were selected from patients with a wide range of survival times (average = 319 days, standard deviation = 260 days, minimum = 9 days, maximum = 676 days). Within these samples there were patients with differing genders (11 male, 18 female), ages (68.6±7.8 years) and body mass index (BMI) values (27.6±6.0). The clinical information for all patients and sample preparation and analysis orders of the samples are included in tables S-1, S-2 and S-3.

Materials

All chemicals used were of analytical reagent (AR) or a higher purity grade. All materials were purchased from Sigma-Aldrich (Gillingham, U.K.) unless otherwise stated. Pyridine (extra dry), hexane, methoxylamine hydrochloride, and *N*-methyl-*N*-(trimethylsilyl)trifluoroacetamide (MSTFA) were obtained from Acros Organics (Loughborough, U.K.). The internal standard succinic acid-d_4 was purchased from Sigma-Aldrich (Gillingham, U.K.), as were HPLC grade methanol and water. Formic acid (BDH Aristar 1) was purchased from VWR International (East Grinstead, U.K.).

Extraction of metabolites from plasma and serum

Extraction involved thawing samples on ice, transferring 400 µL aliquots to 2 mL microcentrifuge tubes (Eppendorf, Cambridge, U.K.) and adding 1200 µL of methanol (room temperature), followed by vortex mixing for 15 s and centrifugation at 13,363x*g* for 15 min. Four aliquots of the supernatant (each of volume 370 µL) were transferred to separate 2 mL microcentrifuge tubes. To the sample set designated for GC-TOF/MS analysis 100 µL of an internal standard was added (0.167 ± 0.009 mg.mL^{-1} of succinic acid-d_4). All four aliquots were lyophilised by speed vacuum concentration at 45°C for 16 h (HETO VR MAXI vacuum centrifuge attached to a Thermo Svart RVT 4104 refrigerated vapour trap; Thermo Life Sciences, Basingstoke, U.K.) and stored at -80°C prior to analysis. Following published protocols[1], pooled quality control (QC) samples were prepared by the combination of 100µL aliquots from each subject sample, and extraction as described above. Separate QC samples were prepared for serum and plasma.

Derivatisation of samples for GC-TOF/MS

Samples were removed from -80°C storage and place in a speed vacuum concentrator for 30 min to remove any residual condensation of water. The extracts were redissolved in 50 μL of 20 mg.mL^{-1} *O*-methoxylamine hydrochloride in pyridine, vortex mixed, and incubated at 60 °C for 30 min in a dri-block heater. Subsequently, 50 μL of N-methyl-N-(trimethylsilyl)trifluoroacetamide (MSTFA) was added and the extracts incubated at 60 °C for a further 30 min. On completion, 20 μL of retention index solution was added (0.3 mg.mL^{-1} n-docosane, n-nonadecane, n-decane, n-dodecane, and n-pentadecane in pyridine) prior to centrifugation at 13,363x*g* for 15 min. The resulting supernatant (120 μL) was transferred to GC-MS vials for analysis.

Table S-1. Clinical information for the patients from whom samples were analysed

patient ID	gender	BMI	age (years)	treatment	survival time (days)
8	Male		67	Carboplatin/etoposide	378
30	Female		67	Carboplatin/etoposide	9
41	Female		77	Cisplatin/etoposide	36
88	Male		66	Carboplatin/etoposide	128
89	Female	42.7	71	Carboplatin/etoposide	31
97	Female	22.8	74	Carboplatin/etoposide	357
108	Male		71	Carboplatin/Etoposide	48
124	Male		65	Carboplatin/etoposide	288
130	Male	39.1	56	Carboplatin/etoposide	273
131	Male	26.0	54	Carboplatin/etoposide	314
132	Female	28.1	65	Carboplatin/etoposide	378
136	Female		64	Carboplatin/etoposide	372
139	Female	20.0	69	Carboplatin/etoposide	364
140	Male		82	Carboplatin/etoposide	364
141	Female	27.1	72	carboplatin/etoposide	187
145	Male	31.4	71	Carboplatin/etoposide	37
146	Female	26.3	74	Carboplatin/etoposide	357
154	Female		78	Carboplatin/etoposide	18
185	Male	26.0	63	Cisplatin/etoposide	322
186	Female	27.8	68	Carboplatin/etoposide	20
190	Female		74	Carboplatin/etoposide	301
196	Female	31.1	68	Carboplatin/etoposide	294
200	Male	21.2	63	Carboplatin/etoposide	287
203	Male	26.2	64	Carboplatin/etoposide	167
214	Female	29.1	69	Carboplatin/etoposide	266
216	Female	23.1	58	Carboplatin	287
265	Female		54	Carboplatin/etoposide	45
270	Female	21.2	48	Carboplatin/etoposide	21
271	Female		57	Carboplatin/etoposide	14

Table S-2. Extraction orders for serum and plasma. Numbers in columns 2-3 refer to the patient IDs in table S-1.

extraction order	serum	plasma
1	190	88
2	265	145
3	132	154
4	185	270
5	203	41
6	141	8
7	41	203
8	270	265
9	145	89
10	214	139
11	108	131
12	140	214
13	186	124
14	30	146
15	271	130
16	8	132
17	200	141
18	154	185
19	131	216
20	97	97
21	136	140
22	130	190
23	196	186
24	139	200
25	124	271
26	216	196
27	88	30
28	89	108
29	146	136

Table S-3. Analysis orders for GC-TOF/MS and UHPLC-MS. For UHPLC-MS-ESI+ and UHPLC-MS-ESI-, the same analysis order was used for serum and plasma. Numbers in columns 2-5 refer to the patient IDs in table S-1.

run order	serum GC-TOF/MS	UHPLC-MS-ESI+	plasma GC-TOF/MS	UHPLC-MS-ESI-
1	8	145	145	88
2	270	196	139	196
3	270	203	89	130
4	270	271	203	271
5	203	136	196	97
6	214	131	97	8
7	108	131	97	8
8	136	131	97	8
9	88	141	88	141
10	145	124	200	124
11	216	214	214	89
12	41	41	145	41
13	131	140	190	140
14	131	154	154	154
15	131	214	185	89
16	271	88	141	145
17	146	139	265	139
18	203	145	203	88
19	190	200	130	200
20	200	89	214	214
21	265	216	30	270
22	141	108	131	108
23	185	97	131	136
24	214	97	131	136
25	132	97	216	136
26	30	265	41	265
27	145	185	140	185
28	196	203	186	130
29	203	190	145	190
30	130	186	132	186
31	97	130	124	203
32	97	132	136	132
33	97	270	108	216
34	186	270	146	216
35	124	270	203	216
36	89	8	271	131
37	139	203	8	130
38	140	214	270	89
39	145	30	270	30
40	214	146	270	146
41	154	145	214	88

Clinical metadata

Table S-1 lists clinical metadata (gender, BMI, survival time after sample collection in days and age in years) for the 29 patients whose samples were analysed in this study.

Extraction and analysis orders

Tables S-2 and S-3 show the order in which samples were extracted and analysed during GC-TOF-MS and UHPLC-MS analyses. For UHPLC-MS, the same run order was used for the analyses of serum and plasma samples.

GC-TOF/MS analysis

Initially 5 injections of QC sample were performed in order to condition the chromatographic system to the specific sample matrix, after which 5 injections of subject samples were performed followed by a QC injection. This was repeated until all samples were analysed. Finally 2 QC injections were performed at the end of the block run. Six subject samples were analysed in triplicate within each block, 3 injected 3 times consecutively and the remaining 3 injected at points spread across the analytical batch. QC samples and repeat injections of subject samples allowed assessments of system stability across both analytical batches.

Processing of raw GC-TOF/MS data

The processing of raw GC-TOF/MS data followed a previously described method[2] and was based on the "Compare" capability of LECO's ChromaTOF v3.25 software (Leco Corp., St. Joseph, MO). Following deconvolution for each datafile, the response ratio was calculated as

> Response ratio = Peak area (metabolite) / peak area (internal standard, succinic d_4 acid).

All data were integrated in MS Excel into a single data matrix of metabolite (with associated retention index and mass spectrum) *vs*. sample, with the response ratio included for detected metabolites in each sample, as described in previous work .[3]

UHPLC analysis

The samples were reconstituted in 90 µL of water, vortex mixed and centrifuged for 15 min at 13,363x*g*. Each supernatant was transferred to a single analytical vial with 200 µL fixed insert, stored in the autosampler at 5°C and analysed within 48 h of reconstitution. The UHPLC was operated at a flow rate of 360 $\mu L.min^{-1}$ in ESI+ and 400 $\mu L.min^{-1}$ in ESI- mode. The column was maintained at a temperature of 50 ^{o}C. The solvent A (water and 0.1% formic acid) and solvent B (methanol and 0.1% formic acid) gradient programme was as follows: ESI+ 100% A held for 1 min, 0-100% B over 11 min, 100% B held for 8 min, returning to 100% A over 2 min (total run time of 22 min); ESI- 100% A held for 1 min, 0-100% B over 16 min, 100% B held for 5 min, returning to 100% A over 2 min (total run time of 24 min). Prior to analysis of the first analytical block, a new UHPLC column was conditioned for 40 min under the same initial gradient conditions applied to sample analysis. A sample injection volume of 10 µL was employed in wasteless mode. The LTQ-Orbitrap XL MS system was controlled under Xcalibur software (Thermo-Fisher Ltd. Hemel Hempsted, U.K.), precisely following the method described in Dunn *et al*.[4] Prior to the first analytical batch, the LTQ-Orbitrap MS system was tuned to optimise conditions for the detection of ions in the *m/z* range 50-1000 and calibrated according to the manufacturers predefined methods in both ESI polarities with the manufacturer's recommended calibration mixture, consisting of caffeine, sodium dodecyl sulphate, sodium taurocholate, the tetrapeptide MRFA and Ultramark 1621. Data were acquired in the Orbitrap mass analyser operating at a mass resolution of 30,000 (FWHM defined at *m/z* 400) and a scan speed of 0.4 s. For each analytical block, initially 20 injections of QC sample were performed for column conditioning. QC and triplicate injections were performed throughout each block, as

described above for GC-TOF/MS. Between each analytical block the ESI ion tube and spray deflector were cleaned using methanol acidified with 0.1 % formic acid and ultrasonication for 15 min.

Processing of raw UHPLC-MS data

The UHPLC-MS raw data profiles were first converted into a NetCDF format within the Xcalibur software's file converter programme. Peak deconvolution was performed using XCMS software[5] as described previously.[4] Deconvolution results in the production of an MS Excel based XY matrix of metabolic feature (with related accurate *m/z* and retention time) *vs*. sample and with peak area inputted where the metabolic feature was detected in each sample.

Metabolite identification from GC-MS data

An identification was confirmed if the match between sample and library mass spectra was greater than 70%. Definitive identification was performed by comparison of the retention index and mass spectrum acquired for samples to those present in a mass spectral library constructed from authenticated reference standards at The University of Manchester.[6] An identification was only confirmed as definitive when the match between sample and in-house library mass spectra was greater than 70% and the retention indices differed by less than 10 s.

Analysis of Variance

2-way analysis of variance (ANOVA) calculations were carried out on those samples that were analysed in triplicate. The independent variables were fluid type (serum / plasma) and subject ID (1 of 6 different subjects for each analytical technique). The *anovan* function in Matlab was used to calculate sum of squares and *p*-values for the main and interaction effects. ANOVA was performed separately on each of the features common to serum and plasma. The amount of variance explained by each of the main effects and by the interaction effect was calculated for each feature and averaged across all features. Each factor was considered to have a significant effect on a feature if it had a p-value ≤ 0.05 after the Bonferroni correction and the total number of features significantly affected by each factor was obtained by counting these significant effects. The ANOVA results are shown in table S-4.

Data preprocessing

Alignment and quality control were performed using a previously reported method.[1, 3] Peaks were aligned by fitting to a low-order nonlinear locally weighted spline (LOESS). All peaks were then removed for which more than 40% of the QCs (excluding lead-in QCs) were either missing or had an area that differed from the mean area by a specified percentage - the allowed percentages were 20% for UHPLC-MS and 30% for GC-TOF/MS. After alignment, peak areas were normalised by dividing all values by the average peak area per sample to remove dilution effects.[7]

Table S-4 ANOVA applied to samples analysed in triplicate: percentage of variance explained by fluid type and subject (number of significant features, after applying the Bonferroni correction, in brackets).

analytical technique	subject	fluid type	subject × fluid type	error (technical variability)
LC-MS-ESI+	54.3 (1695)	24.3 (1150)	12.0 (809)	9.4
LC-MS-ESI-	58.1 (1239)	21.0 (828)	9.4 (452)	11.5
GC-TOF/MS	48.4 (72)	6.3 (20)	12.2 (13)	33.1

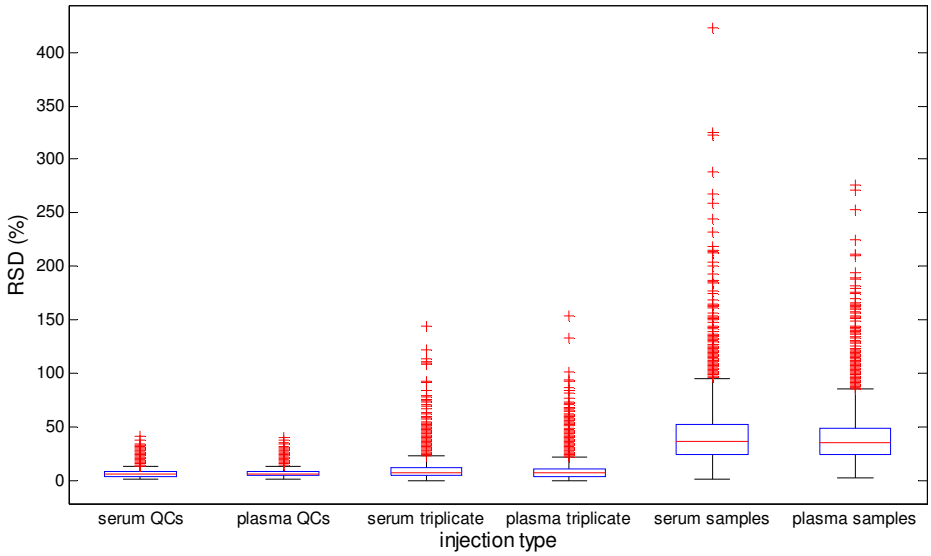
Technical reproducibility of data

For serum, 259, 3152 and 2689 metabolic features were detected on the GC-ToF/MS, UHPLC-MS+ and UHPLC-MS- analytical platforms, respectively. Following signal correction and quality assurance procedures 170, 2877 and 2469 metabolic features remained for data acquired on the respective platforms. For plasma, the corresponding numbers of metabolic features were 259, 3406 and 2433 before signal correction and quality assurance and 178, 3080 and 2258 after these procedures. All results described below apply to data after signal correction and quality assurance procedures.

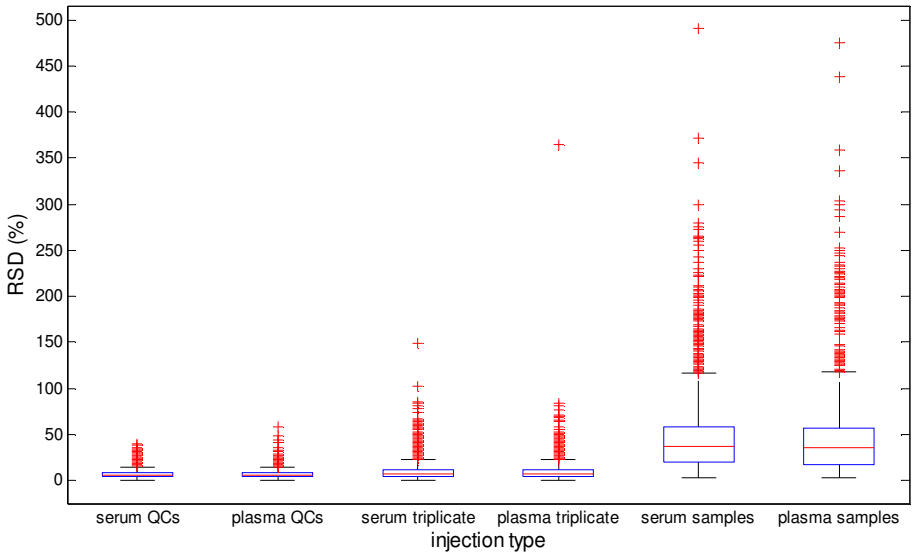
The technical reproducibility was assessed with three methods. The first calculated pairwise Pearson correlation coefficients, r, between the signal corrected peak areas of the features obtained from the injections of the samples analysed in triplicate, after quality assurance and ignoring missing features. (The number of features included in each calculation, *n*, varied between calculations as a result of the exclusion of missing features: $n \approx 160$ for GC-TOF/MS, 2500 for UHPLC-MS+ and 2000 for UHPLC-MS-.) Very high average values were achieved for LC-MS (averaged across positive and negative ion modes) - 0.992(±0.018) for samples analysed non-consecutively and 0.998(±0.002) for samples analysed consecutively - showing that the technical reproducibility was good. The corresponding values for GC-TOF/MS were 0.952(±0.036) and 0.915(±0.111). The lower r values observed for GC-TOF/MS most likely arise from the greater technical variability introduced to GC-TOF/MS analysis by (i) the greater number of sample preparation steps required, particularly the prerequisite of chemical derivatisation to introduce volatility to otherwise non-volatile metabolites and (ii) the accuracy of small volume injections (1µL) compared to 10µL for UHPLC-MS.

The second method determined the relative standard deviation ranges for replicate injections (6 × 3 injections), QC samples (excluding QCs used for column conditioning, resulting in 11 injections) and for sample injections (excluding second and third replicate injections, resulting in 29 injections). Data have been aggregated across all features. Figures S-1(A-C) show box and whisker plots for these classes of injections for data obtained from, respectively, UHPLC-MS+, UHPLC-MS- and GC-TOF/MS. These plots demonstrate that

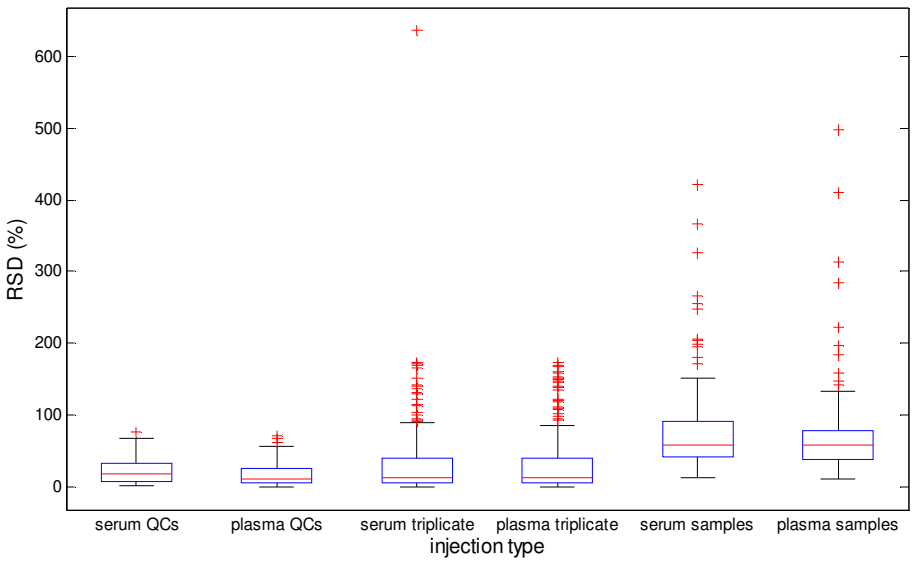
- The data from serum and plasma have similar levels of reproducibility, as evidenced by both the QCs and the replicate injections.
- Repeat injections of the same sample have reproducibility approaching that of the QCs.
- There is considerably more variability between samples than between QCs or replicate injections of the same sample, for both serum and plasma, indicating that UHPLC-MS and GC-TOF/MS have the potential to define biological differences between samples where the biological variance is greater than 20%.
- The RSDs for replicate samples have been included only for non-consecutively injected replicates to allow a fair comparison with respect to QC injections. Consecutive replicate injections have slightly lower RSDs, comparable to or lower than those for the QC samples (see tables S-5, S-6 and S-7).



A



B



C

Figure S-1. Distributions of relative standard deviations for different groups of samples analysed with (A) UHPLC-MS-ESI+ (B) UHPLC-MS-ESI- (C) GC-TOF-MS

Tables S-5, S-6 and S-7 show lower quartile (Q1), median (Q2) and upper quartile (Q3) values for different classes of injection for, respectively, UHPLC-MS+, UHPLC-MS- and GC-TOF/MS platforms. There are no significant differences between the values for serum and for plasma. Non-consecutive replicates have slightly higher RSDs than consecutive replicates. However, the RSD of all replicate samples is much lower than the RSD of non-replicate samples, indicating that most of the variance in the data arises from between-sample differences rather than from technical variability.

Table S-5. Upper quartile, median and lower quartile values of the relative standard deviations for different groups of samples analysed with UHPLC-MS-ESI+

Sample type	Quartile	serum	plasma
QC	Q1	0.044	0.046
	Q2	0.061	0.063
	Q3	0.083	0.085
Consecutively injected replicates	Q1	0.026	0.031
	Q2	0.046	0.054
	Q3	0.077	0.088
Non-consecutively injected replicates	Q1	0.051	0.042
	Q2	0.081	0.072
	Q3	0.124	0.115
Non-replicate samples	Q1	0.242	0.247
	Q2	0.369	0.354
	Q3	0.532	0.491

Table S-6. Upper quartile, median and lower quartile values of the relative standard deviations for different groups of samples analysed with UHPLC-MS-ESI-

Sample type	Quartile	serum	plasma
QC	Q1	0.043	0.041
	Q2	0.060	0.058
	Q3	0.086	0.082
Consecutively injected replicates	Q1	0.026	0.026
	Q2	0.045	0.044
	Q3	0.075	0.070
Non-consecutively injected replicates	Q1	0.045	0.041
	Q2	0.076	0.069
	Q3	0.120	0.113
Non-replicate samples	Q1	0.201	0.177
	Q2	0.367	0.355
	Q3	0.586	0.577

Table S-7. Upper quartile, median and lower quartile values of the relative standard deviations for different groups of samples analysed with GC-TOF/MS

Sample type	Quartile	serum	plasma
QC	Q1	0.072	0.054
	Q2	0.183	0.120
	Q3	0.336	0.267
Consecutively injected replicates	Q1	0.036	0.049
	Q2	0.074	0.102
	Q3	0.264	0.295
Non-consecutively injected replicates	Q1	0.060	0.054
	Q2	0.134	0.130
	Q3	0.401	0.401
Non-replicate samples	Q1	0.420	0.387
	Q2	0.579	0.585
	Q3	0.911	0.794

The third method for assessing technical reproducibility investigated the presence of missing features in UHPLC-MS. The presence of missing features (or zero values where the metabolite is not detected) in UHPLC-MS is very consistent across technical replicates. If a feature is missing for one of the injections of a sample analysed in triplicate, there is a 66% probability that it will also be missing from the other 2 injections of the same sample. This compares to the 2.4% figure expected if the missing features were randomly distributed and suggests that the bulk of the missing features correspond to metabolites that are not present above the detection limit of the mass spectrometer, rather than being the result of technical variability or of errors during spectrum deconvolution. It should be noted that the data from GC-TOF/MS was deconvolved and library-matched before data analysis, leading to very few missing features. Analysis of missing features for this data was therefore considered unnecessary.

Comparison of metabolite groups between serum and plasma

Each metabolite group observed in serum that contained at least 3 features was compared against the features observed in plasma and *vice versa*. For each group, the proportion of features that were also observed in the alternative fluid was calculated. These comparisons were carried out separately on data from positive and negative ion modes. The minimum metabolite group size for these comparisons was chosen to be 3 in order to focus on metabolites for which there is strong supporting evidence. Forty-four percent of metabolite features were assigned to metabolite groups of size ≥3 and the average number of metabolic features in each metabolite group analysed was 4.7.

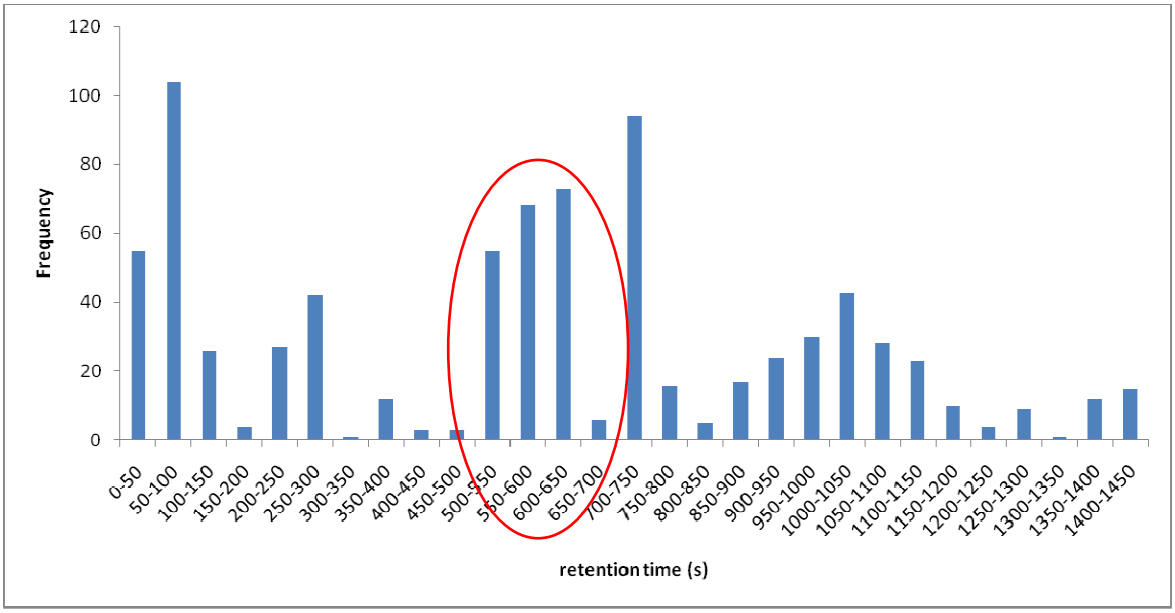
Table S-8 shows the numbers of metabolite groups that are unique to serum or plasma in ESI+ and ESI- mode. Due to the uncertainty caused by partially overlapping metabolite groups, there is a range in the number of metabolites in each category. The maximum and minimum numbers are calculated by assuming that, respectively, all or none of the partially overlapping groups are unique to one fluid.

Distributions across the *m/z* and RT range

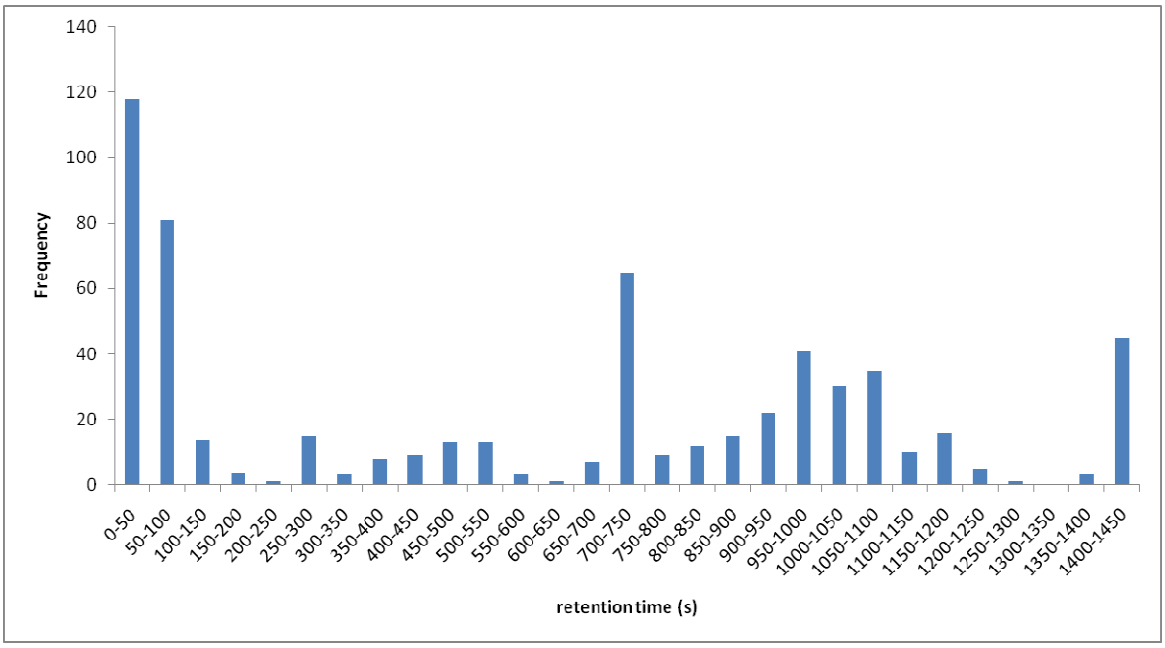
After identifying metabolite features present in serum-only, plasma-only and both fluids, histograms of the RT and *m/z* distributions were constructed for the three sets of features, for each analytical technique. For UHPLC-MS-ESI+ and GC-TOF/MS, there were no significant differences in the distributions for the different subsets. However, for UHPLC-MS-ESI- a significant number of features were observed in the RT range 500-650 s in serum and not in plasma (figures S-2(A-C)). This observation appears to be independent of the m/z distribution, which was similar for the three subsets (Figures S-3(A-C)).

Table S-8 Number of metabolite groups containing at least 3 features found uniquely in one biofluid (serum or plasma). The total numbers of metabolite groups are shown in brackets. The numbers have a range of values due to the presence of partially overlapping metabolite groups.

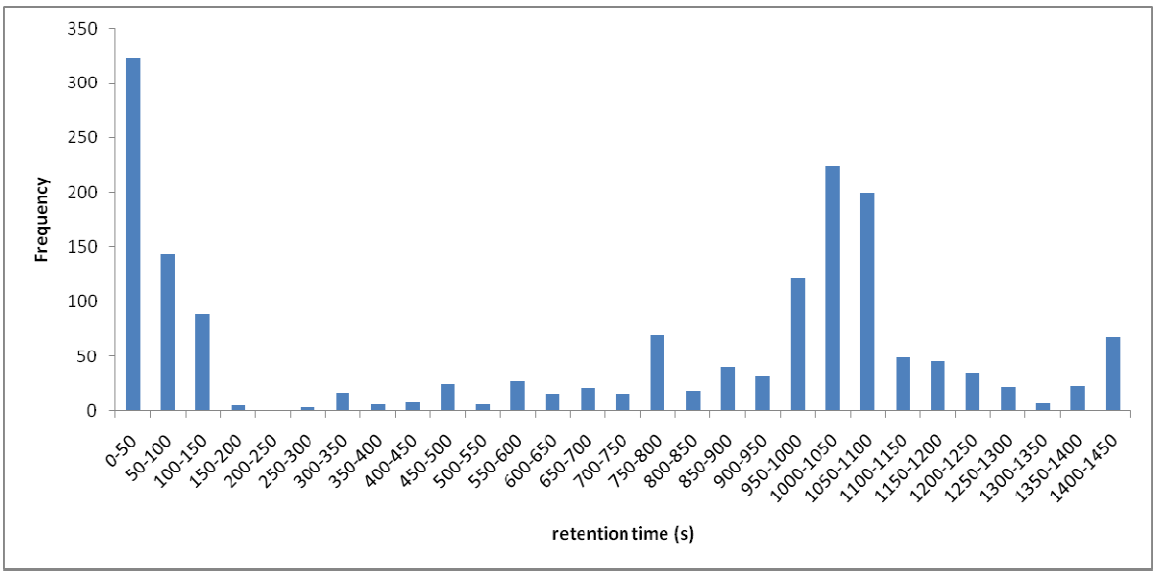
ESI mode	biofluid	unique metabolite groups
ESI+	serum	77-125 (157)
ESI-	serum	88-126 (166)
ESI+	plasma	74-164 (189)
ESI-	plasma	83-120 (140)



A

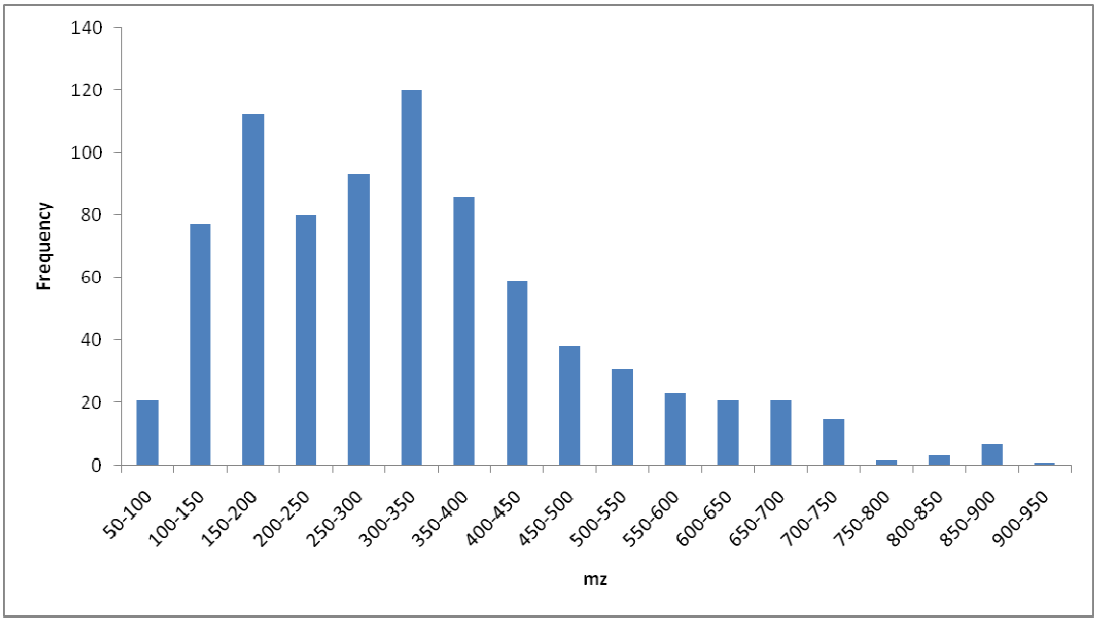


B

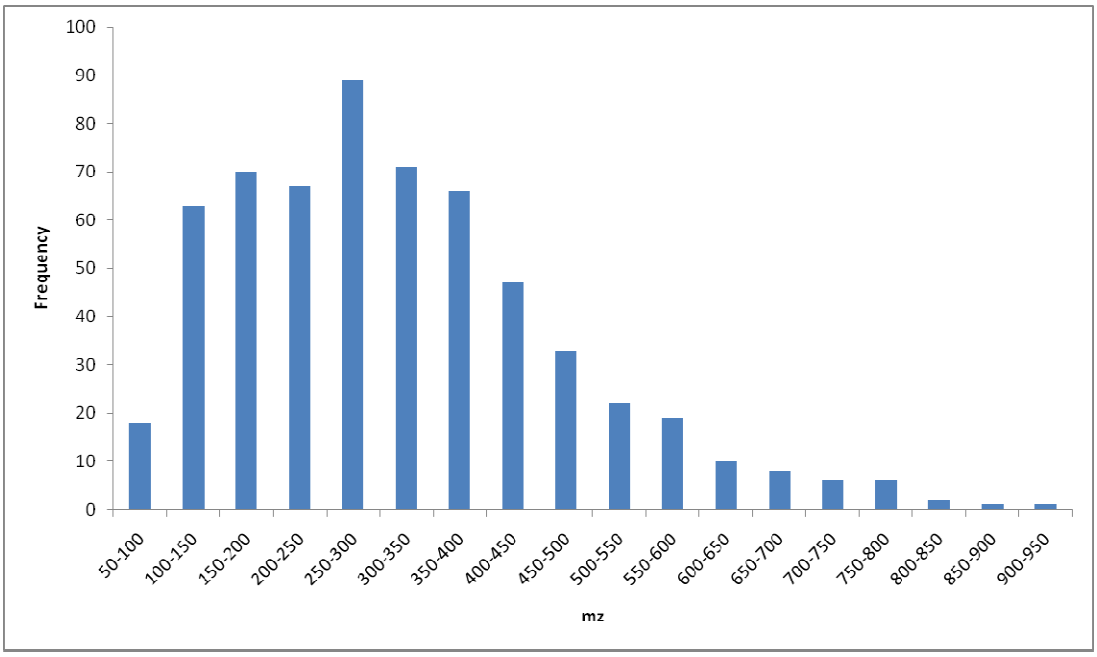


C

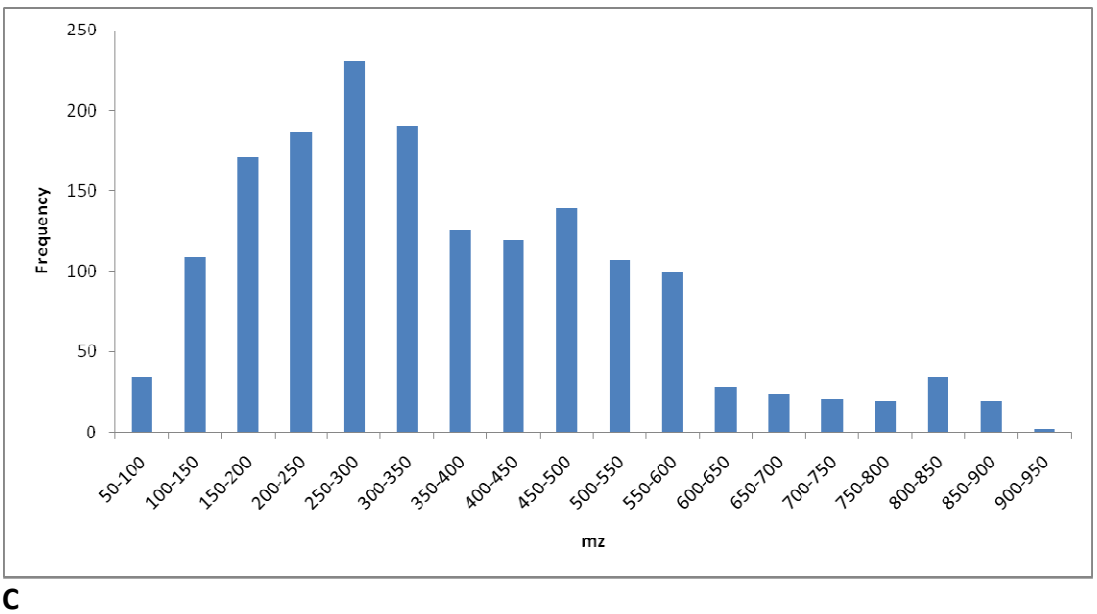
Figure S-2. Distribution of retention times for features identified from UHPLC-MS- in (A) serum only (B) plasma only (C) both serum and plasma A large number of features, marked with a red oval, were seen in serum but not plasma, in the range 500-650 s.



A



B



C

Figure S-3. Distribution of *m/z* values for features identified from UHPLC-MS- in (A) serum only (B) plasma only (C) both serum and plasma.

Principal component analysis

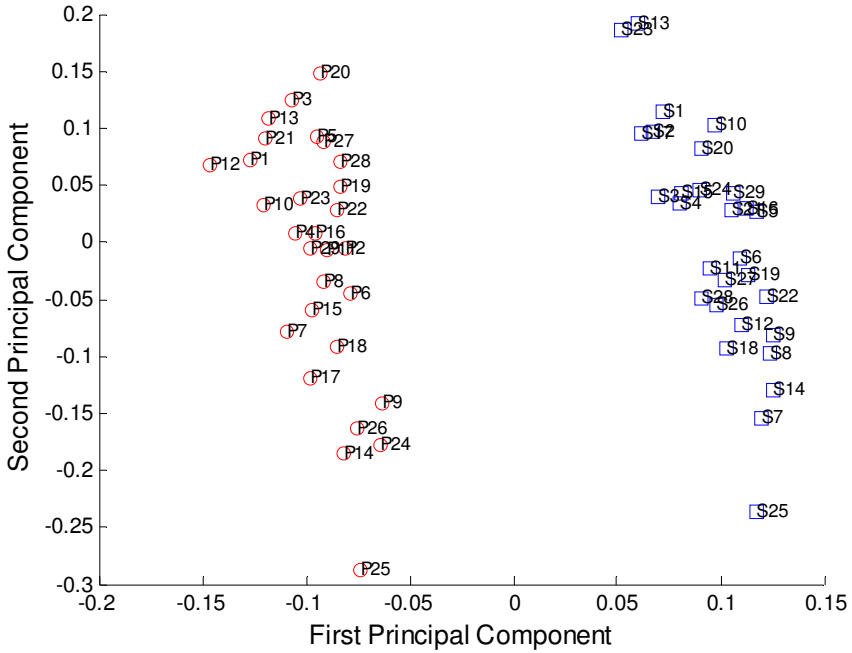
Separately for UHPLC-MS-ESI+, UHPLC-MS-ESI- and GC-TOF/MS data, principal component analysis (PCA) was performed, using the Numerical Algorithms Group (NAG) toolbox (mark 22) for Matlab, on:

- the matched features. Each of the injections (serum and plasma) was included in a single PCA, resulting in a single set of z-scores, covering both serum and plasma data. Auto-scaling (mean centring / standard deviation) was carried out prior to PCA in 2 different ways, as described in the Results and Discussion section.
- all features. The serum and plasma injections were input to 2 different PCAs and the resulting z-scores saved. The Procrustes transformation was applied to conform the plasma z-scores to the serum z-scores and a Mantel correlation[8] was calculated on the most important dimensions of the z-scores. The number of dimensions included in the calculation was sufficient to explain 75% of the variance of the LC-MS data and 50% of the variance of the GC-TOF/MS data. A lower threshold was used for the GC-TOF/MS data in order to avoid fitting to noise, which makes up a larger proportion of the variance in the GC-TOF/MS data. The significance of both the dissimilarity measure, d, obtained from the Procrustes transformation and of the Mantel correlation coefficient were assessed using permutation testing (n=1,000,000).

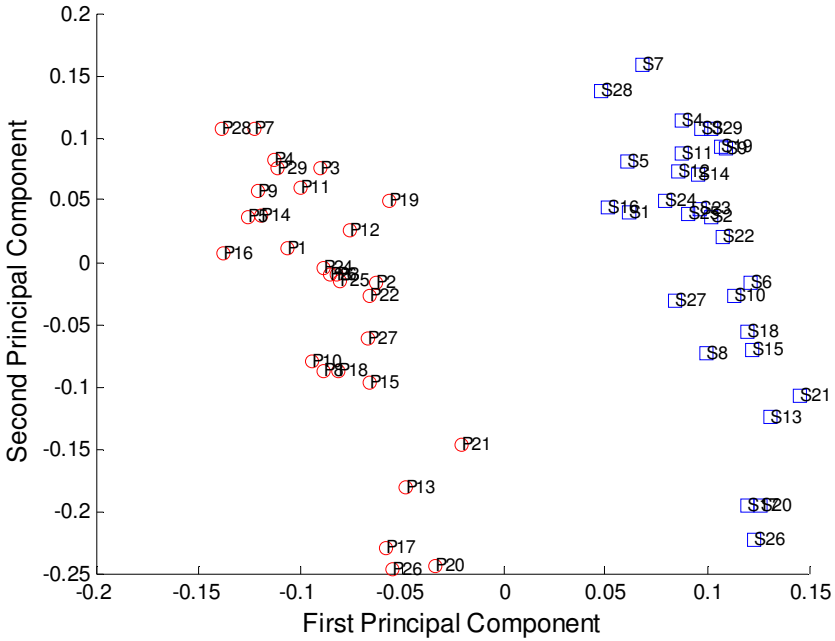
PCA plots with all samples autoscaled together

Figures S-4(A-C) are PCA plots of LC-MS-ESI+, LC-MS-ESI- and GC-MS data, respectively, after autoscaling all samples together. For both LC-MS modes, the serum and plasma samples are clearly separable using just the first principal component (PC1). However, the PC1 loadings indicate that it is dominated by features that show significant differences in peak area between serum and plasma. The 100 features with the highest loadings in PC1 in LC-MS-ESI- all have p-values below 4×10^{-6} from t-tests to assess the difference in mean peak intensities between serum and plasma. In LC-MS-ESI+ the first 400 features with the highest loadings all have p-values below 5×10^{-7}. Once PC1 has been taken into account, the matched serum and plasma samples show considerable similarities. With respect to PC2, the ordering of samples is very similar for the serum and plasma samples. For example, in figure S-4B, which is a PCA plot for LC-MS-ESI-, subjects 14, 7, 11 and 28 all have high PC2 values whereas subjects 13, 20, 21 and 26 all have low PC2 values, for both serum and plasma.

A



B



C

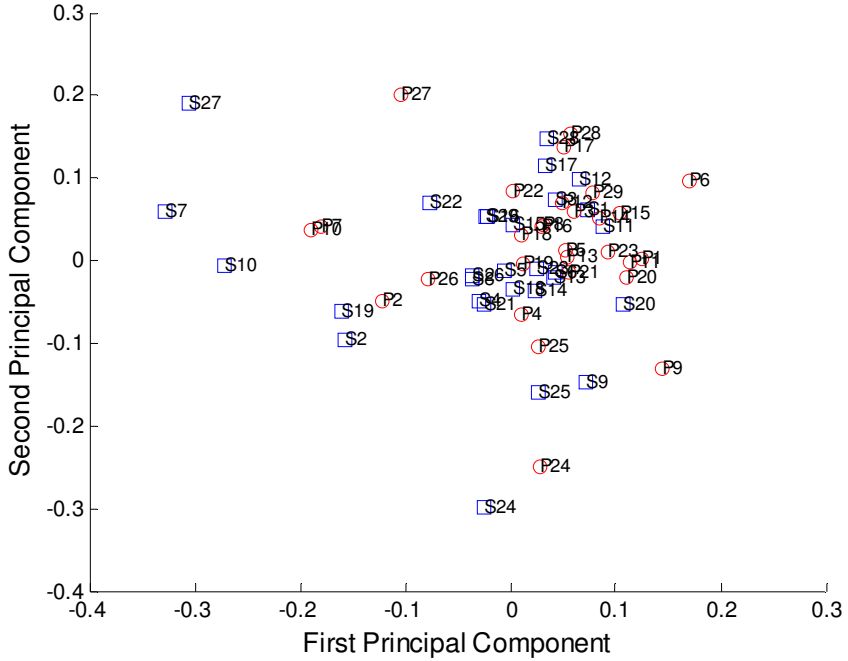


Figure S-4. Plot of the first 2 z-scores resulting from PCA on serum (blue squares) and plasma (red circles) data, with serum and plasma data autoscaled together, obtained from (A) UHPLC-MS-ESI+ (B) UHPLC-MS-ESI- (C) GC-TOF/MS.

References

(1) Dunn, W. B.; Broadhurst, D.; Begley, P.; Zelena, E.; McIntyre, S.; Anderson, N.; Brown, M.; Knowles, J. D.; Haselden, J. N.; Nicholls, A. W.; Wilson, I. D.; The Husermet consortium; Kell, D. B.; Goodacre, R. *Nature Protocols* **2011**, *In Press*.

(2) Begley, P.; Francis-McIntyre, S.; Dunn, W. B.; Broadhurst, D. I.; Halsall, A.; Tseng, A.; Knowles, J.; Goodacre, R.; Kell, D. B. *Anal Chem* **2009**, *81*, 7038-7046.

(3) Zelena, E.; Dunn, W. B.; Broadhurst, D.; Francis-McIntyre, S.; Carroll, K. M.; Begley, P.; O'Hagan, S.; Knowles, J. D.; Halsall, A.; Wilson, I. D.; Kell, D. B. *Anal Chem* **2009**, *81*, 1357-1364.

(4) Dunn, W. B.; Broadhurst, D.; Brown, M.; Baker, P. N.; Redman, C. W. G.; Kenny, L. C.; Kell, D. B. *Journal of Chromatography B-Analytical Technologies in the Biomedical and Life Sciences* **2008**, *871*, 288-298.

(5) Smith, C. A.; Want, E. J.; O'Maille, G.; Abagyan, R.; Siuzdak, G. *Analytical Chemistry* **2006**, *78*, 779-787.

(6) Brown, M.; Dunn, W. B.; Dobson, P.; Patel, Y.; Winder, C. L.; Francis-McIntyre, S.; Begley, P.; Carroll, K.; Broadhurst, D.; Tseng, A.; Swainston, N.; Spasic, I.; Goodacre, R.; Kell, D. B. *Analyst* **2009**, *134*, 1322-1332.

(7) Goodacre, R.; Broadhurst, D.; Smilde, A. K.; Kristal, B. S.; Baker, J. D.; Beger, R.; Bessant, C.; Connor, S.; Calmani, G.; Craig, A.; Ebbels, T.; Kell, D. B.; Manetti, C.; Newton, J.; Paternostro, G.; Somorjai, R.; Sjostrom, M.; Trygg, J.; Wulfert, F. *Metabolomics* **2007**, *3*, 231-241.

(8) Mantel, N. *Cancer Res* **1967**, *27*, 209-220.