**Supporting Information:**

# Liquid Chromatography-Mass Spectrometry Calibration Transfer and Metabolomics Data Fusion

*Andrew A. Vaughan[1], Warwick B. Dunn[2,3], J. William Allwood[1], David C. Wedge[1,4], Fiona H. Blackhall[5], Anthony D. Whetton[6], Caroline Dive[5] and Royston Goodacre[1,3]*

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

[2] Centre for Advanced Discovery & Experimental Therapeutics (CADET), Central Manchester NHS Foundation Trust and School of Biomedicine, University of Manchester, Manchester Academic Health Science Centre, York Place, Oxford Road, Manchester, M13 9WL, U.K.

[3] Manchester Centre for Integrative Systems Biology, Manchester Institute of Biotechnology, University of Manchester, 131 Princess Street, Manchester, M1 7DN, U.K.

[4] Cancer Genome Project, Wellcome Trust Sanger Institute, Hinxton, Cambridgeshire, CB10 1SA, U.K.

[5] 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, U.K.

[6] School of Cancer and Enabling Sciences, Manchester Academic Health Science Centre, University of Manchester, Manchester, M20 3LJ, U.K.

## Contents

## FIGURES

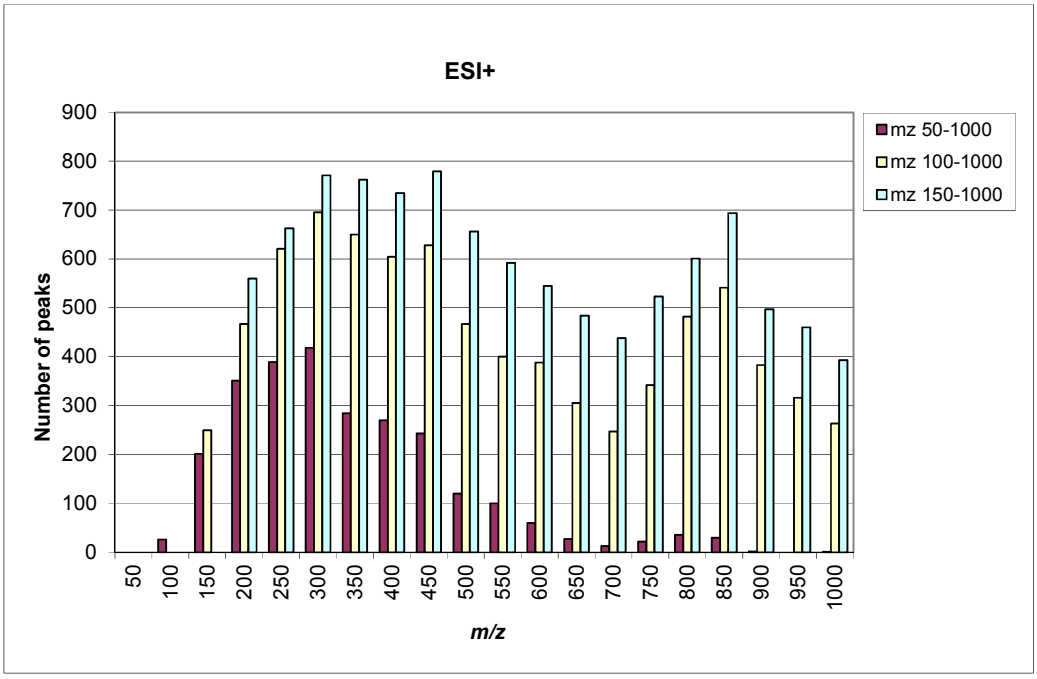


**Figure S-1A.** Effect of *m/z* scan range on the number of features detected by instrument B in ESI+ ionisation mode (using serum QCs).

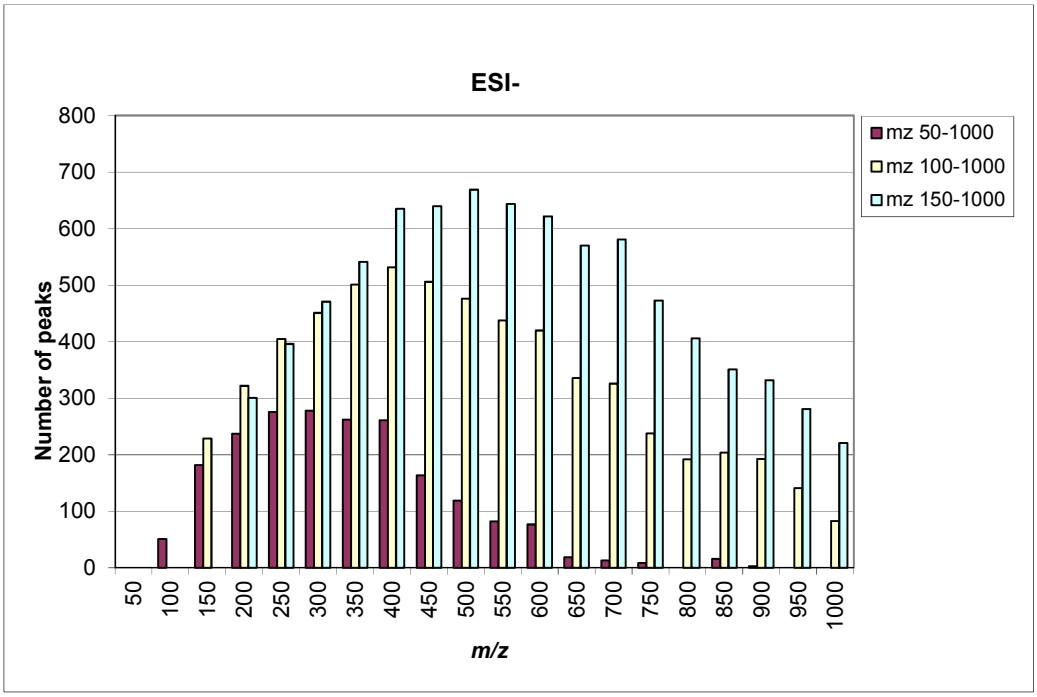


**Figure S-1B.** Effect of *m/z* scan range on the number of features detected by instrument B in ESI- ionisation mode (using serum QCs).

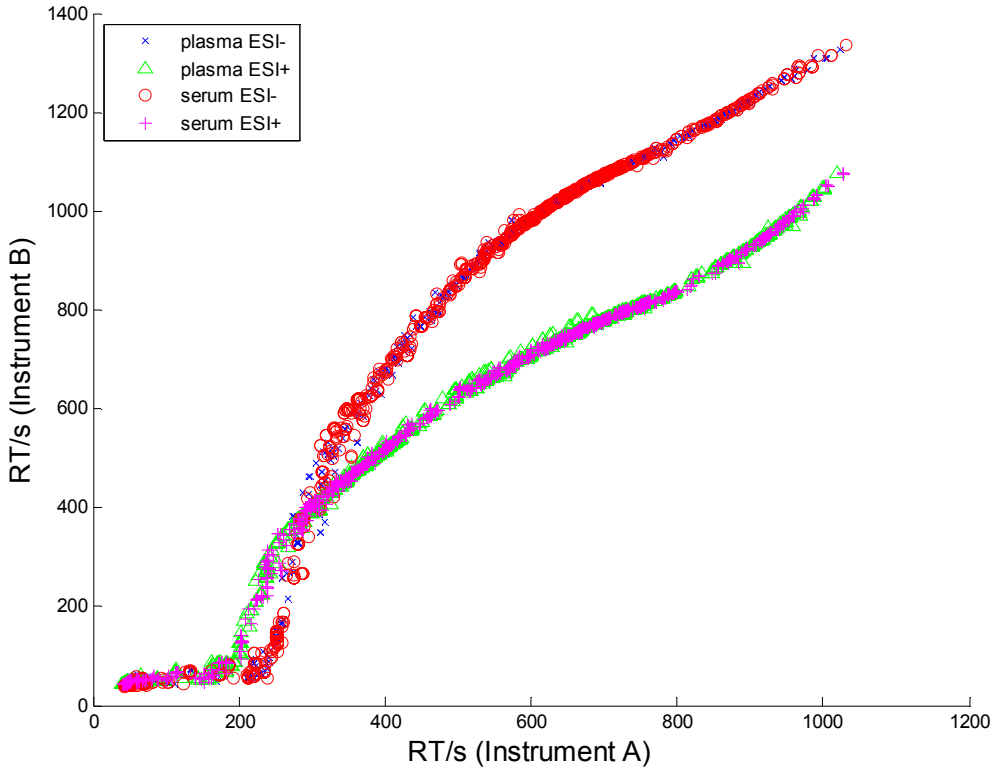


**Figure S-2.** Retention times for features from plasma or serum matched between instruments A and B using both ESI source polarities

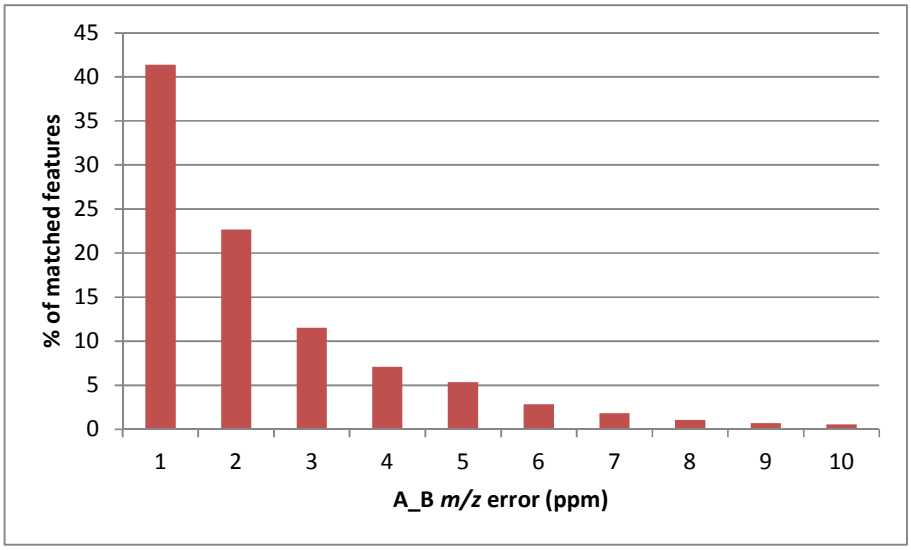


**Figure S-3.** Mass difference (*m/z* ppm) of features matched between instruments A and B (includes both serum and plasma using both ESI source polarities)

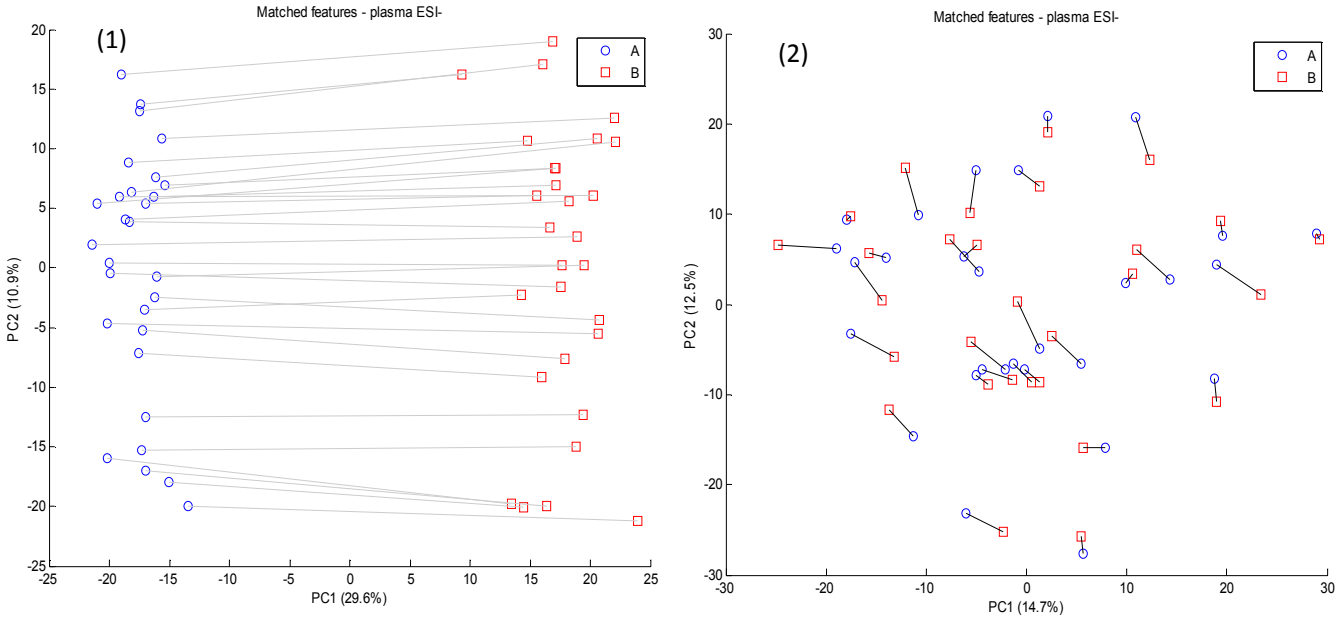


**Figure S-4A.** PCA scores plots for all samples using features matched between instruments for plasma ESI- (1) combining A and B samples BEFORE autoscaling together, (2) combining A and B samples AFTER separate autoscaling. Identical samples pairs (analyzed on different instruments) are joined by grey/black lines.

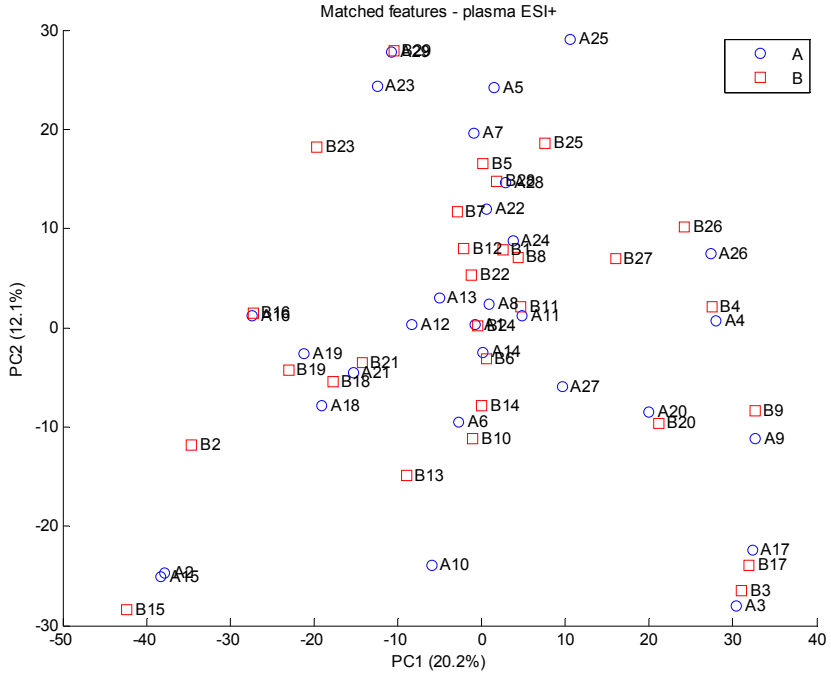


**Figure S-4B.** PCA scores plots for all samples using features matched between instruments for plasma ESI+

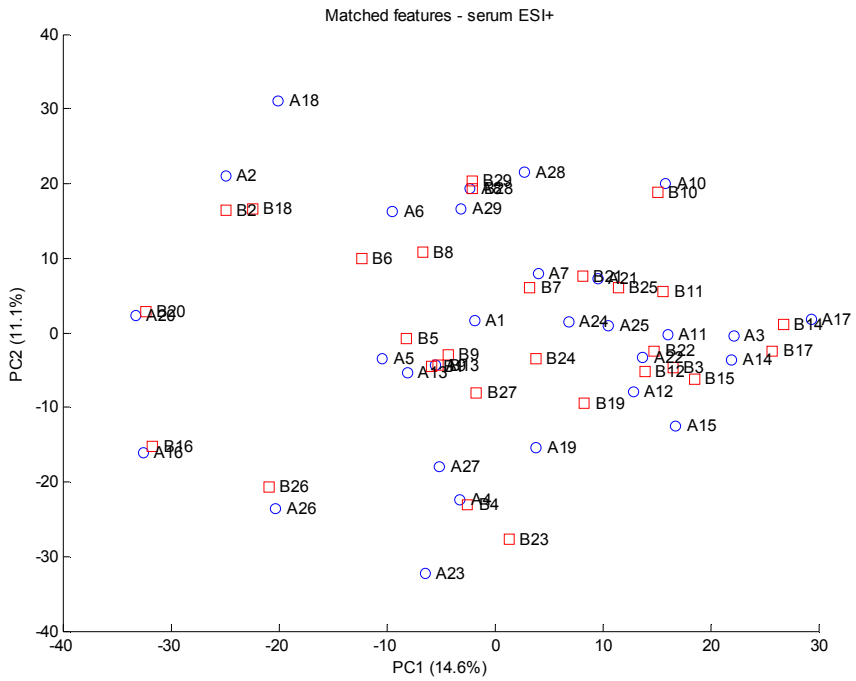


**Figure S-4C.** PCA scores plots for all samples using features matched between instruments for serum ESI+

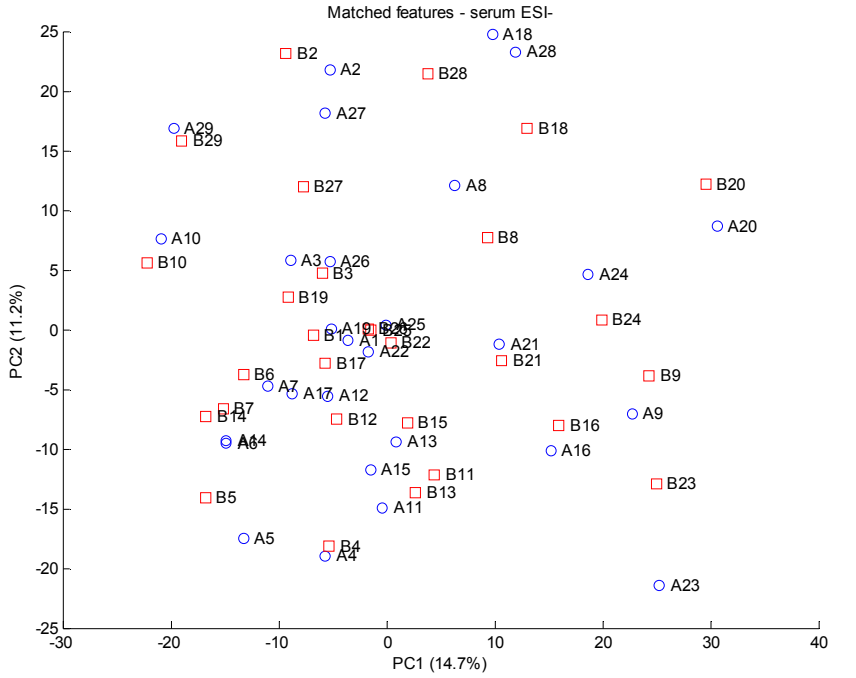


**Figure S-4D.** PCA scores plots for all samples using features matched between instruments for serum ESI-

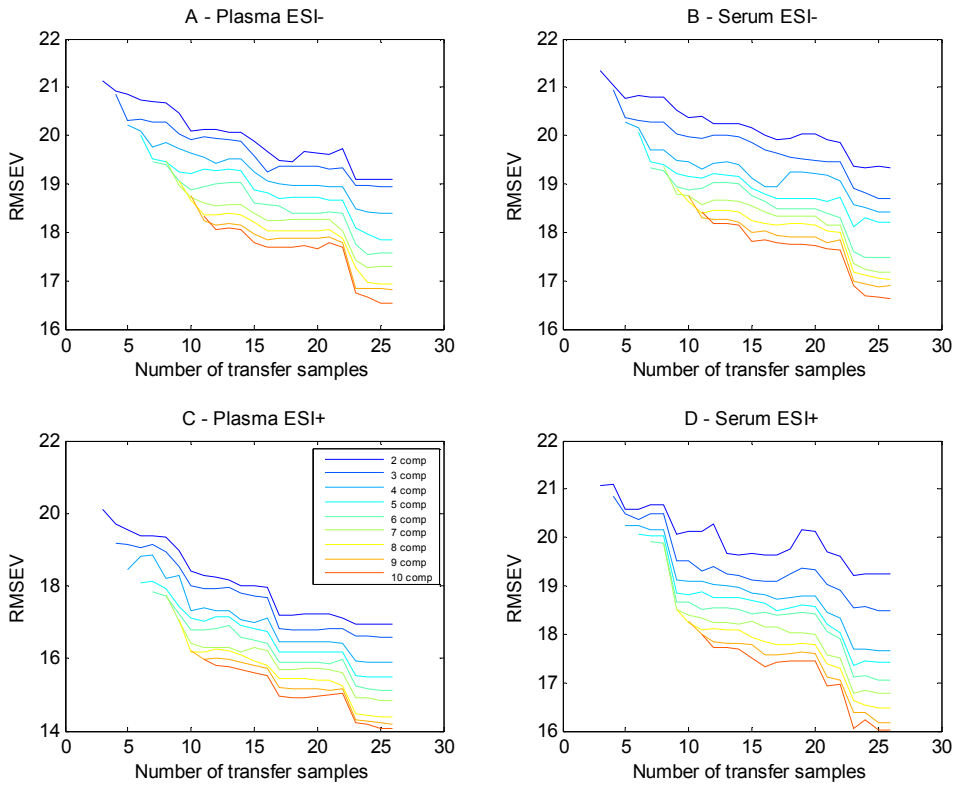


**Figure S-5.** RMSEV plotted against number of samples used for calibration transfer for: (A) plasma ESI-, (B) serum ESI-, (C) plasma ESI+, and (D) serum ESI+

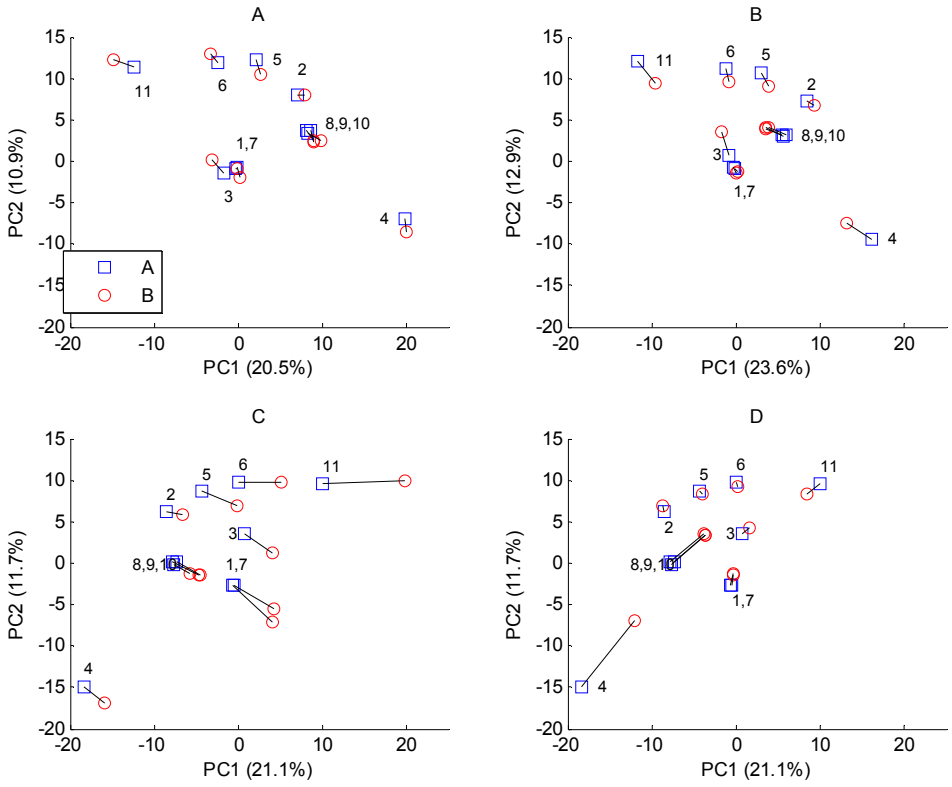


**Figure S-6.** Scores plots for the test samples (plasma ESI-) from PCA of (A) all original A samples and all original B samples, (B) all original A samples and all PLS-fitted B samples, (C) $A_{test}$, $A_{cal}$, $B_{cal}$ and $B_{val}$, with the original $B_{test}$ samples projected, and (D) $A_{test}$, $A_{cal}$, $B_{cal}$ and $B_{val}$, with the PLS-fitted $B_{test}$ samples projected (500 features)

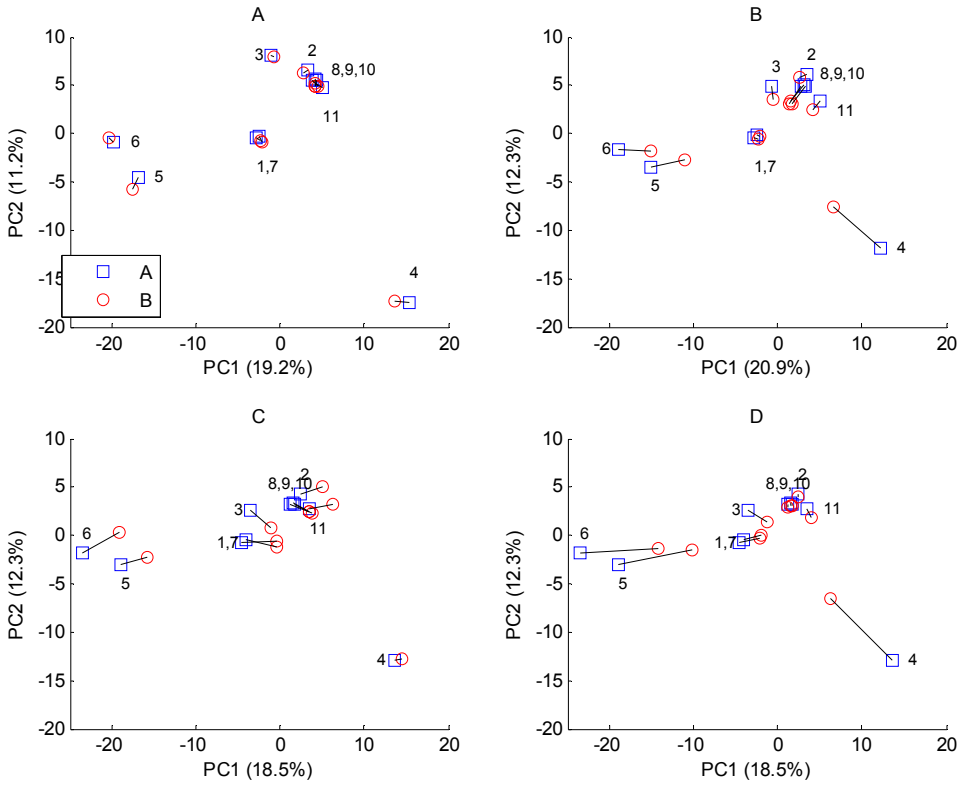


**Figure S-7.** Scores plots for the test samples (serum ESI+) from PCA of (A) all original A samples and all original B samples, (B) all original A samples and all PLS-fitted B samples, (C) $A_{test}$, $A_{cal}$, $B_{cal}$ and $B_{val}$, with the original $B_{test}$ samples projected, and (D) $A_{test}$, $A_{cal}$, $B_{cal}$ and $B_{val}$, with the PLS-fitted $B_{test}$ samples projected (500 features)

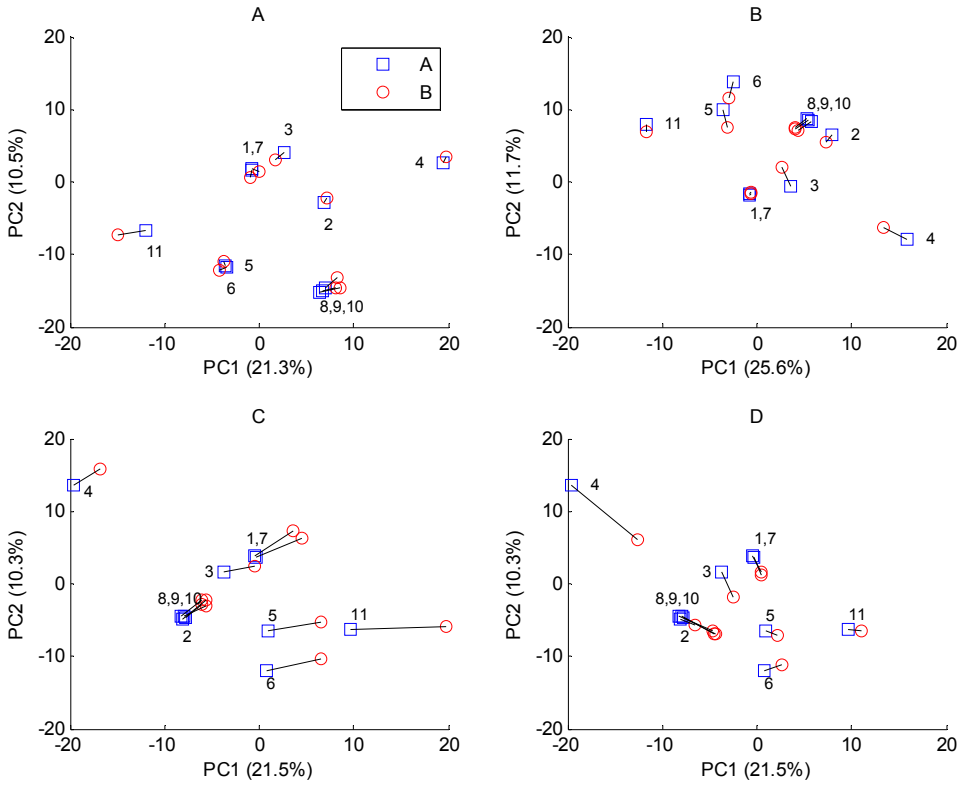


**Figure S-8.** Scores plots for the test samples (serum ESI-) from PCA of (A) all original A samples and all original B samples, (B) all original A samples and all PLS-fitted B samples, (C) $A_{test}$, $A_{cal}$, $B_{cal}$ and $B_{val}$, with the original $B_{test}$ samples projected, and (D) $A_{test}$, $A_{cal}$, $B_{cal}$ and $B_{val}$, with the PLS-fitted $B_{test}$ samples projected (500 features)

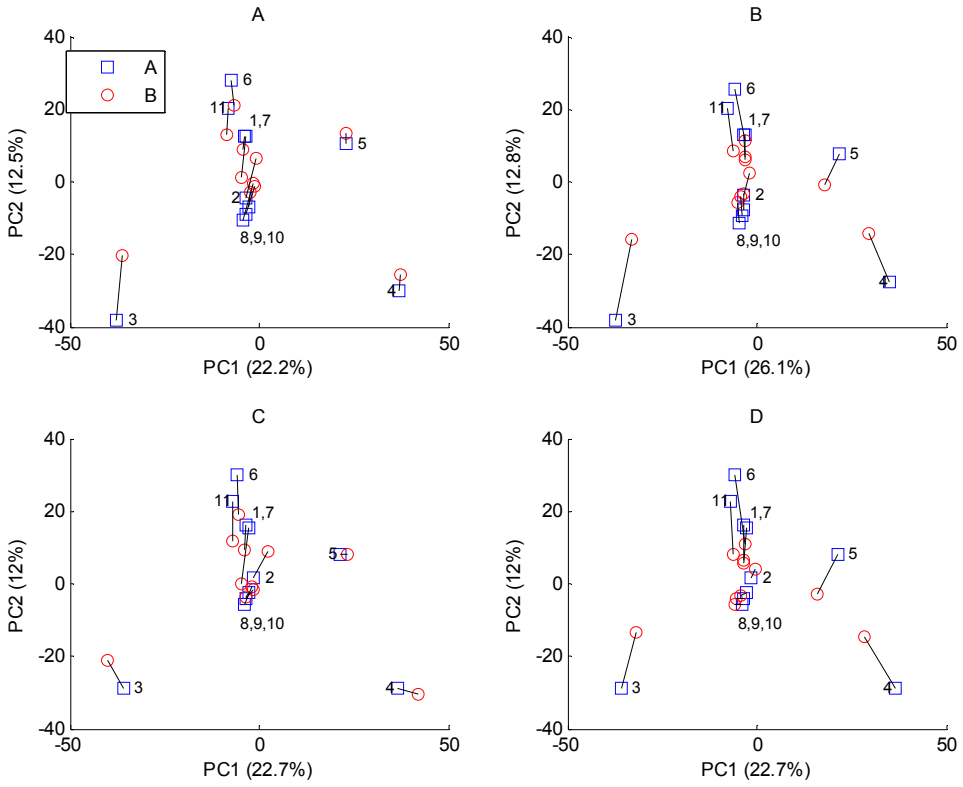


**Figure S-9.** Scores plots for the test samples (plasma ESI+) from PCA of (A) all original A samples and all original B samples, (B) all original A samples and all PLS-fitted B samples, (C) $A_{test}$, $A_{cal}$, $B_{cal}$ and $B_{val}$, with the original $B_{test}$ samples projected, and (D) $A_{test}$, $A_{cal}$, $B_{cal}$ and $B_{val}$, with the PLS-fitted $B_{test}$ samples projected (all 1851 features)

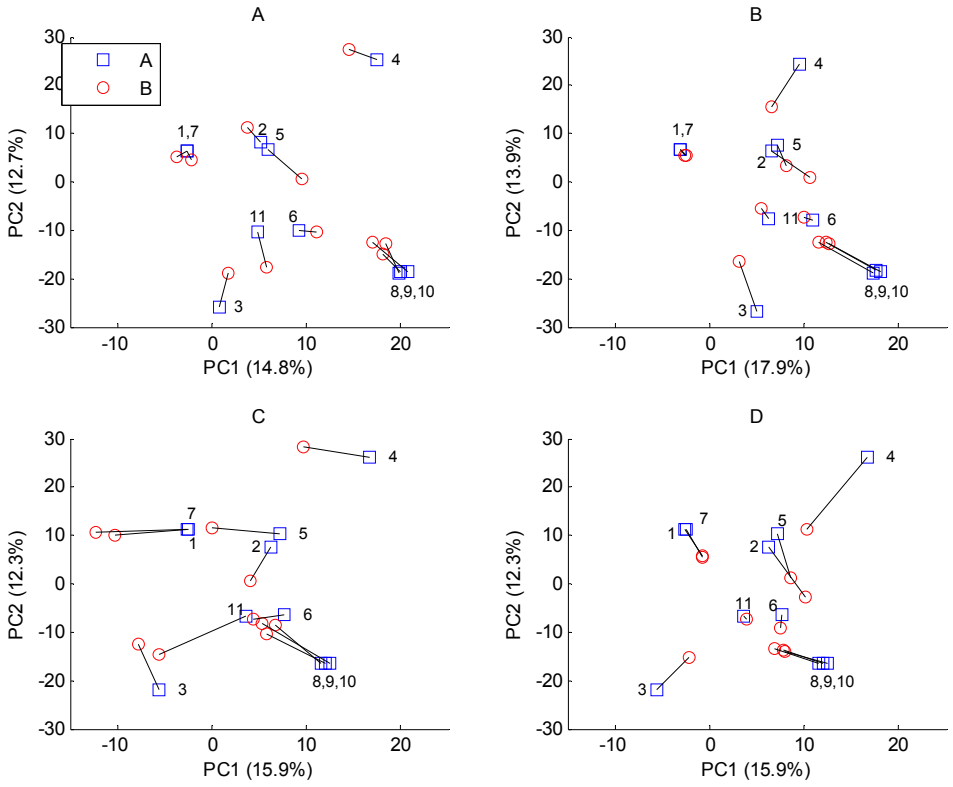


**Figure S-10.** Scores plots for the test samples (plasma ESI-) from PCA of (A) all original A samples and all original B samples, (B) all original A samples and all PLS-fitted B samples, (C) $A_{test}$, $A_{cal}$, $B_{cal}$ and $B_{val}$, with the original $B_{test}$ samples projected, and (D) $A_{test}$, $A_{cal}$, $B_{cal}$ and $B_{val}$, with the PLS-fitted $B_{test}$ samples projected (all 1100 features)

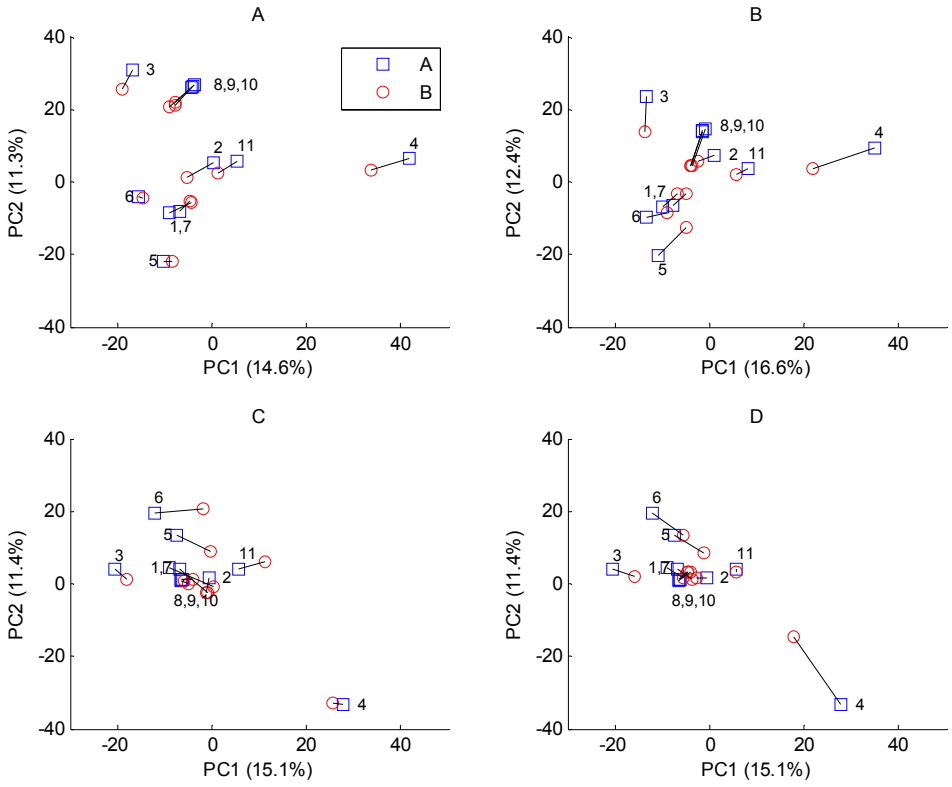


**Figure S-11.** Scores plots for the test samples (serum ESI+) from PCA of (A) all original A samples and all original B samples, (B) all original A samples and all PLS-fitted B samples, (C) $A_{test}$, $A_{cal}$, $B_{cal}$ and $B_{val}$, with the original $B_{test}$ samples projected, and (D) $A_{test}$, $A_{cal}$, $B_{cal}$ and $B_{val}$, with the PLS-fitted $B_{test}$ samples projected (all 1755 features)

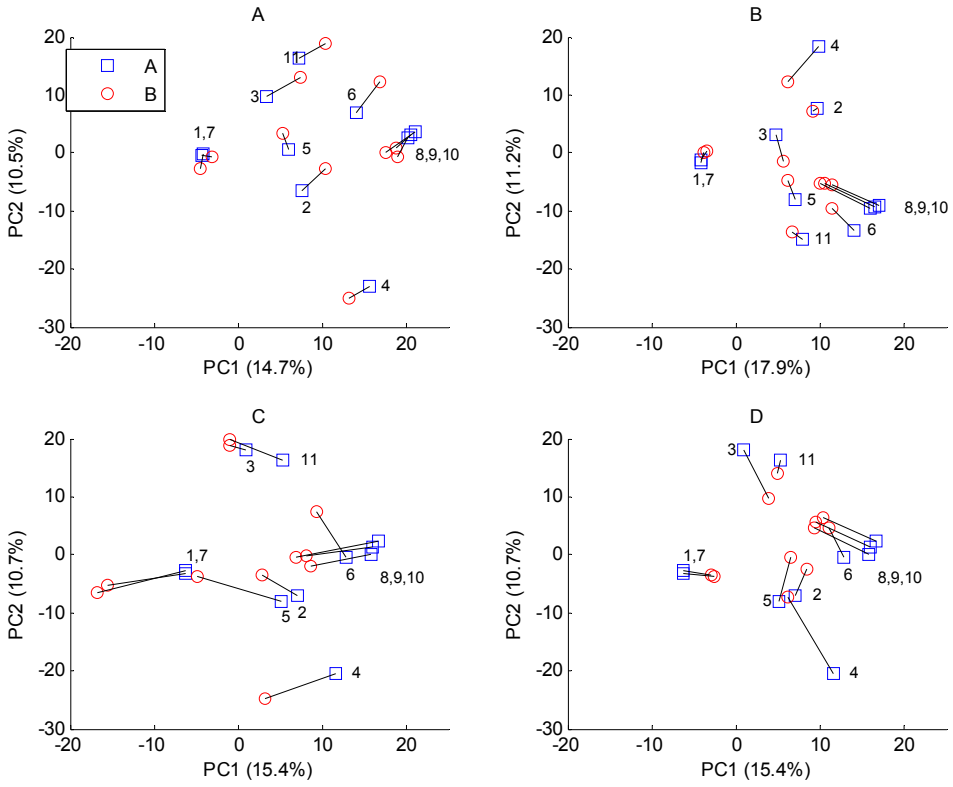


**Figure S-12.** Scores plots for the test samples (serum ESI-) from PCA of (A) all original A samples and all original B samples, (B) all original A samples and all PLS-fitted B samples, (C) $A_{test}$, $A_{cal}$, $B_{cal}$ and $B_{val}$, with the original $B_{test}$ samples projected, and (D) $A_{test}$, $A_{cal}$, $B_{cal}$ and $B_{val}$, with the PLS-fitted $B_{test}$ samples projected (all 1261 features)

# TABLES

**Table S-1.** Tuning parameters for instruments A and B

| Parameter | A (ESI+) | A (ESI-) | B (ESI+) | B (ESI-) |
|---|---|---|---|---|
| Capillary Temp (°C) | 275 | 275 | 275 | 275 |
| APCI Vaporizer Temp (°C) | 0 | 0 | 0 | 0 |
| Sheath Gas Flow (AU) | 30 | 30 | 30 | 30 |
| Aux Gas Flow (AU) | 8 | 8 | 8 | 8 |
| Sweep Gas Flow (AU) | 0 | 0 | 0 | 0 |
| Source Voltage (kV) | 4 | 5 | 4 | 4 |
| Source Current (μA) | 100 | 100 | 100 | 100 |
| Capillary Voltage (V) | 17 | -35 | 47 | -25 |
| Tube Lens (V) | 75 | -110 | 180 | -180 |
| Skimmer Offset (V) | 0 | 0 | 0 | 0 |
| Multipole RF Amplifier (Vp-p) | 400 | 400 | 400 | 400 |
| Multipole 00 Offset (V) | -5.25 | 4 | -2.25 | 4.5 |
| Lens 0 Voltage (V) | -4 | 4.2 | -1 | 3.5 |
| Multipole 0 Offset (V) | -4.5 | 4.5 | -4.5 | 5.25 |
| Lens 1 Voltage (V) | -34 | 15 | -29 | 11 |
| Gate Lens Offset (V) | -22 | 35 | -30 | 44 |
| Multipole 1 Offset (V) | -9.5 | 8 | -9 | 6 |
| Front Lens (V) | -6 | 5.25 | -5 | 5.5 |
| FTMS Full AGC Target | 500000 | 500000 | 1000000 | 1000000 |

**Table S-2.** Analysis order for ESI+ and ESI-. The same orders were used for both serum and plasma. In columns 2 and 3, the number refers to the subject ID (technical replicates are suffixed .1, .2 and .3). QC = pooled quality control sample.

| Analysis order | ESI+ | ESI- |
|---|---|---|
| 1 | QC | QC |
| 2 | 145.1 | 88.1 |
| 3 | 196 | 196 |
| 4 | 203.1 | 130.1 |
| 5 | 271 | 271 |
| 6 | 136 | 97 |
| 7 | QC | QC |
| 8 | 131.1 | 8.1 |
| 9 | 131.2 | 8.2 |
| 10 | 131.3 | 8.3 |
| 11 | 141 | 141 |
| 12 | 124 | 124 |
| 13 | QC | QC |
| 14 | 214.1 | 89.1 |
| 15 | 41 | 41 |
| 16 | 140 | 140 |
| 17 | 154 | 154 |
| 18 | 214.2 | 89.2 |
| 19 | QC | QC |
| 20 | 88 | 145 |
| 21 | 139 | 139 |
| 22 | 145.2 | 88.2 |
| 23 | 200 | 200 |
| 24 | 89 | 214 |
| 25 | QC | QC |
| 26 | 216 | 270 |
| 27 | 108 | 108 |
| 28 | 97.1 | 136.1 |
| 29 | 97.2 | 136.2 |
| 30 | 97.3 | 136.3 |
| 31 | QC | QC |
| 32 | 265 | 265 |
| 33 | 185 | 185 |
| 34 | 203.2 | 130.2 |
| 35 | 190 | 190 |
| 36 | 186 | 186 |
| 37 | QC | QC |
| 38 | 130 | 203 |
| 39 | 132 | 132 |
| 40 | 270.1 | 216.1 |
| 41 | 270.2 | 216.2 |
| 42 | 270.3 | 216.3 |
| 43 | QC | QC |
| 44 | 8 | 131 |
| 45 | 203.3 | 130.3 |
| 46 | 214.3 | 89.3 |
| 47 | 30 | 30 |
| 48 | 146 | 146 |
| 49 | QC | QC |
| 50 | 145.3 | 88.3 |
| 51 | QC | QC |
| 52 | QC | QC |

**Table S-3.** Mantel correlations and Procrustes *d* values for the test A/B sample pairs from PCA models: (1) all original A samples and all original B samples, (2) all original A samples and all PLS-fitted B samples, (3) $A_{test}$, $A_{cal}$, $B_{cal}$ and $B_{val}$, with the original $B_{test}$ samples projected, and (4) $A_{test}$, $A_{cal}$, $B_{cal}$ and $B_{val}$, with the PLS-fitted $B_{test}$ samples

| | Mantel correlation (model 1 &2) | Procrustes d (model 1) | Procrustes d (model 2) | Mantel correlation (model 3 & 4) | Procrustes d (model 3) | Procrustes d (model 4) |
|---|---|---|---|---|---|---|
| **Plasma ESI+ (Top 500 features)** | 0.96 | 0.005 | 0.018 | 0.95 | 0.008 | 0.041 |
| **Plasma ESI- (Top 500 features)** | 0.93 | 0.013 | 0.04 | 0.87 | 0.02 | 0.056 |
| **Serum ESI+ (Top 500 features)** | 0.95 | 0.005 | 0.018 | 0.94 | 0.011 | 0.027 |
| **Serum ESI- (Top 500 features)** | 0.88 | 0.014 | 0.015 | 0.90 | 0.024 | 0.025 |
| **Plasma ESI+ (1851 features)** | 0.97 | 0.091 | 0.105 | 0.95 | 0.102 | 0.112 |
| **Plasma ESI- (1100 features)** | 0.92 | 0.077 | 0.043 | 0.86 | 0.133 | 0.059 |
| **Serum ESI+ (1755 features)** | 0.88 | 0.024 | 0.059 | 0.94 | 0.068 | 0.063 |
| **Serum ESI- (1261 features)** | 0.78 | 0.074 | 0.067 | 0.75 | 0.092 | 0.141 |

**Table S-4.** Pearson and Spearman correlation coefficients between the original B dataset and (1) PLS-fitted $B_{test}$ samples merged with original $B_{cal}$ and $B_{val}$ samples, and (2) $A_{test}$ samples merged with original $B_{cal}$ and $B_{val}$ samples (after correlation with survival time)

| | **Full, original B dataset *vs*. PLS-fitted $B_{test}$ samples merged with original $B_{cal}$ and $B_{val}$ samples** | | | **Full, original B dataset *vs*. $A_{test}$ samples merged with original $B_{cal}$ and $B_{val}$ samples** | | |
|---|---|---|---|---|---|---|
| | **Pearson correlation** | **Spearman correlation** | **Number of top 10 correlated features matched** | **Pearson correlation** | **Spearman correlation** | **Number of top 10 correlated features matched** |
| **Plasma ESI+ (Top 500 features)** | 0.94 | 0.93 | 7 | 0.99 | 0.99 | 9 |
| **Plasma ESI- (Top 500 features)** | 0.89 | 0.87 | 3 | 0.97 | 0.96 | 8 |
| **Serum ESI+ (Top 500 features)** | 0.95 | 0.94 | 7 | 0.99 | 0.99 | 9 |
| **Serum ESI- (Top 500 features)** | 0.91 | 0.90 | 8 | 0.98 | 0.98 | 9 |

**Table S-5.** Pearson and Spearman correlation coefficients between the original datasets and a merged dataset comprising autoscaled samples A (1:26) and autoscaled B (27:52) (after correlation with survival time)

| | **Full, original B dataset *vs.* A 1:26 samples merged with B 27:52 samples** | | | **Full, original A dataset *vs.* A 1:26 samples merged with B 27:52 samples** | | |
|---|---|---|---|---|---|---|
| | **Pearson correlation** | **Spearman correlation** | **Number of top 10 correlated features matched** | **Pearson correlation** | **Spearman correlation** | **Number of top 10 correlated features matched** |
| **Plasma ESI+ (Top 500 features)** | 0.96 | 0.95 | 9 | 0.96 | 0.94 | 8 |
| **Plasma ESI- (Top 500 features)** | 0.94 | 0.92 | 4 | 0.96 | 0.95 | 5 |
| **Serum ESI+ (Top 500 features)** | 0.98 | 0.97 | 7 | 0.97 | 0.97 | 7 |
| **Serum ESI- (Top 500 features)** | 0.96 | 0.95 | 7 | 0.95 | 0.95 | 6 |

## SCRIPTS

**Script S-1.** Matlab script for retention time mapping: performs linear interpolation on the retention times of *matched* features, and evaluates the resulting piecewise polynomial on the full set of retention times from instrument B (uses supplementary file "RT_data for ac_2012_02227.txt")

```
% Retention time mapping
%
% Linear interpolation algorithm to create piecewise polynomial
% Input: RT_A_match, RT_B_match
% (lists of retention times for features matched by m/z & correlation)
% Output: pp (piecewise polynomial form of curve plotted between
%             RT_A_match & RT_B_match)
pp = interp1(RT_B_match,RT_A_match,'linear','pp');
% Evaluate pp for full set of instrument B retention times to map B to A
% Input: pp, RT_B_full
% Output: RT_B_mapped
RT_B_mapped = ppval(pp,RT_B_full);
```

**Script S-2.** Matlab script for optimizing the number of transfer samples and PLS components for use in the PLS-R calibration transfer model (uses supplementary file "spos_data for ac_2012_02227.txt")

```
% PLS-R optimization script
% Input: spos_data
%        500 rows (metabolite features)
%        110 columns (1:3 - peak ID, m/z, RT for instrument A)
%                    (4:55 - peak areas for instruments A)
%                    (56:58 - peak ID, m/z, RT for instrument B)
%                    (59:110 - peak areas for instruments B)
%       * Columns 1:55 correspond to the same samples in columns 56:110
% Output: rmsev (table of root mean squared errors of validation)
data_A=spos_data(:,4:55);
data_B=spos_data(:,59:110);
% replace NaNs with low values
dims_A=size(data_A);
dims_B=size(data_B);
lowvals_A=(min(data_A,[],2)./4)*ones(1,dims_A(2));
lowvals_B=(min(data_B,[],2)./4)*ones(1,dims_B(2));
A_noNaNs=data_A;
A_noNaNs(isnan(A_noNaNs))=lowvals_A(isnan(A_noNaNs));
B_noNaNs=data_B;
B_noNaNs(isnan(B_noNaNs))=lowvals_B(isnan(B_noNaNs));
% normalise to mean peak area
A_mean=ones(dims_A(1),1)*nanmean(A_noNaNs);
norm_A=A_noNaNs./A_mean;
B_mean=ones(dims_B(1),1)*nanmean(B_noNaNs);
norm_B=B_noNaNs./B_mean;
% transpose
norm_A=norm_A';
norm_B=norm_B';
% validation datasets
A_val=zscore(norm_A(27:52,:));
B_val=zscore(norm_B(27:52,:));
```

```
% preallocate output matrix for RMSEV values
rmsev=zeros(24,9);
% initialise variables
ncomp=2; % PLS components
last=24; % transfer samples
% PLS regression:
% vary PLS components (outer loop) and transfer samples (inner loop)
while ncomp < 11
for i=1:last
A_train=zscore(norm_A(i:26,:));
B_train=zscore(norm_B(i:26,:));
[A_L,B_L,A_S,B_S,beta,PCTVAR,MSE,stats] =
plsregress(A_train,B_train,ncomp);
B_fit_val = [ones(size(A_val,1),1) A_val]*beta;
B_residuals = B_val-B_fit_val;
mse_calc=sum(mean(B_residuals.^2));
rmsev(i,ncomp-1)=sqrt(mse_calc);
end
last=last-1;
ncomp=ncomp+1;
end;
```