Statistical HOmogeneous Cluster SpectroscopY (SHOCSY): an optimized statistical approach for clustering of ^{1}H NMR spectral data to reduce interference and enhance robust biomarkers selection

(Supplementary Information)

*Xin Zou[1], Elaine Holmes[2, 3], Jeremy K Nicholson[2, 3], Ruey Leng Loo[1, 2]**

[1] Medway School of Pharmacy, Universities of Kent and Greenwich, Chatham Maritime, Kent, UK

[2] Section of Biomolecular Medicine, Department of Surgery and Cancer, Imperial College London, UK

[3] MRC-HPA Centre for Environment and Health, Imperial College London, UK

*Email: r.loo@kent.ac.uk

Table S-1. The means and variances in the signal intensities of metabolites used to generate the simulated spectral dataset to represent the Paraquat poisoning and control classes based on N=30 in each biological class and 50% idiosyncratic spectra in the intervention class.

Metabolites	Control class mean (variance)		Paraquat toxicity class mean (variance)			
	Homogeneous control		Homogeneous responders to toxicity		Idiosyncratic responders to toxicity	
	N=15	N=15	N=8	N=7	N=7	N=8
Lactate	441 (127)	441 (127)	6279.8 (1808.5)	6279.8 (1808.5)	441 (127)	441 (127)
Creatinine	13200 (4100)	13200 (4100)	2376 (287)	2376 (287)	13200 (4100)	13200 (4100)
Citrate	2022 (1081)	2022 (1081)	30.3 (16.2)	30.3 (16.2)	2022 (1081)	2022 (1081)
L-alanine	290 (73)	290 (73)	1531.2 (385.4)	1531.2 (385.4)	290 (73)	290 (73)
Hippurate	5280 (130)	1320 (32.5)	5280 (130)	1320 (32.5)	5280 (130)	1320 (32.5)
Glycine	2058 (440)	514 (110)	2058 (440)	514 (110)	2058 (440)	514 (110)
Trimethylamine *N*-oxide	1964 (101.1)	491 (25.2)	1964 (101.1)	491 (25.2)	1964 (101.1)	491 (25.2)
L-Histidine	1896 (371)	474 (92.7)	1896 (371)	474 (92.7)	1896 (371)	474 (92.7)
Phenylacetylglutamine	1874 (79.2)	468 (19.8)	1874 (79.2)	468 (19.8)	1874 (79.2)	468 (19.8)

Table S-2. The number of useable spectra in the high and low dose hydrazine study.

Time points	90mg/kg		30mg/kg	
	Control	Hydrazine	Control	Hydrazine
t_1	35	29	35	33
t_2	38	39	38	37
t_3	39	45	39	47
t_4	47	40	47	47
t_5	24	16	24	23
t_6	18	17	18	18
t_7	21	24	21	23
t_8	23	18	23	22
t_9	21	17	21	22

Assessment of SHOCSY algorithm

The key goal of SHOCSY algorithm is to remove idiosyncratic spectra from the dataset and improve the extraction of potential discriminatory biomarkers. We evaluated the performance of our algorithm using simulated datasets of different sizes (N = 30, 100 and 500 per class) and with different proportion of idiosyncratic responders (5%, 10%, 33% and 50%) by their sensitivity, specificity and overall accuracy (Supplementary information, Table S-3).

The sensitivity, specificity, and overall accuracy are calculated as below:

$$sensitivity = \frac{\#TP}{\#P}$$

$$specificity = 1 - \frac{\#FP}{\#N}$$

$$accuracy = \frac{\#TP + \#TN}{\#N + \#P}$$

where *#P* and *#N* are the total number of homogeneous and idiosyncratic spectra, respectively. *#TP*, *#FP* and *#TN* are the correctly identified homogenous spectra, incorrectly identified homogenous spectra and correctly identified idiosyncratic spectra, respectively.

Table S-3. The comparison of sensitivity, specificity and accuracy by standard OPSL-DA and SHOCSY approaches. In the standard OPLSDA approach, the spectra in the same biological class are assumed to be homogeneous.

Proportion of non-responders in toxicity class	Number of samples in each class	Standard OPLSDA approach (based on whole dataset)			SHOCSY approach (based on homogeneous subsets)		
		Specificity	Sensitivity	Accuracy	Specificity	Sensitivity	Accuracy
5%	30*	0	1	0.97	1	1	1
	100	0	1	0.97	0.80	0.95	0.95
	500	0	1	0.97	0.96	1	0.96
10%	30	0	1	0.95	1	1	1
	100	0	1	0.95	0.8	0.95	0.95
	500	0	1	0.95	0.98	0.96	0.96
33%	30	0	1	0.83	1	1	1
	100	0	1	0.83	0.9	0.95	0.94
	500	0	1	0.83	0.94	0.95	0.95
50%	30	0	1	0.75	1	1	1
	100	0	1	0.75	0.92	0.94	0.94
	500	0	1	0.75	0.93	0.94	0.94

Key * = 2 samples were used; this represents 6.7% of the dataset instead of 5%.

We also assessed whether seven-fold cross-validation is a reliable measure for the performance of the models by comparing to the double cross-validation. The double cross-validation method involves optimizing the number of orthogonal components of OPLS-DA model using an inner loop and evaluating the model Q^2 using an outer loop. The results showed that seven-fold cross-validation is as reliable as the double cross-validation, Table S-4.

Table S-4. The seven-fold and double cross-validation comparison of Q^2 values for standard OPLS-DA and SHOCSY approaches. For the double cross-validation of Q^2, the inner and outer loops contain 100 random-split iterations and in each random-split, a seventh of the spectra were used as test sets and the remaining spectra were used as training sets.

Proportions of idiosyncratic responders in toxicology class	Number of samples in each class	Q^2 of a standard OPLSDA approach		Q^2 of SHOCSY approach	
		7-fold CV (time in s)	double CV (time in s)	7-fold CV (time in s)	double CV (time in s)
5%	30*	0.71 (3)	0.73 (5300)	0.81 (2.8)	0.8 (5300)
	100	0.82 (9.8)	0.81 (17800)	0.89 (9)	0.89 (17200)
	500	0.77 (47)	0.81 (>15h)	0.91 (44)	0.93 (>15h)
10%	30	0.67 (2.9)	0.71 (5000)	0.81 (2.6)	0.79 (6200)
	100	0.76 (10)	0.76 (19000)	0.88 (9)	0.87 (17000)
	500	0.63 (56)	0.65 (>15h)	0.9 (52)	0.92 (>15h)
33%	30	0.48 (2.8)	0.52 (4400)	0.81 (2.3)	0.8 (4800)
	100	0.49 (9.9)	0.48 (19400)	0.86 (8)	0.86 (14000)
	500	0.19 (56)	0.2 (>15h)	0.87 (44)	0.89 (>15h)
50%	30	0.34 (2.8)	0.35 (4600)	0.78 (2.1)	0.79 (4500)
	100	0.3 (10.3)	0.32 (17800)	0.84 (7)	0.84 (14000)
	500	-0.06 (52)	-0.01 (>15h)	0.82 (32)	0.84 (>15h)

Key * = 2 samples were used; this represents 6.7% of the dataset instead of 5%.

CV stands for Cross-validation

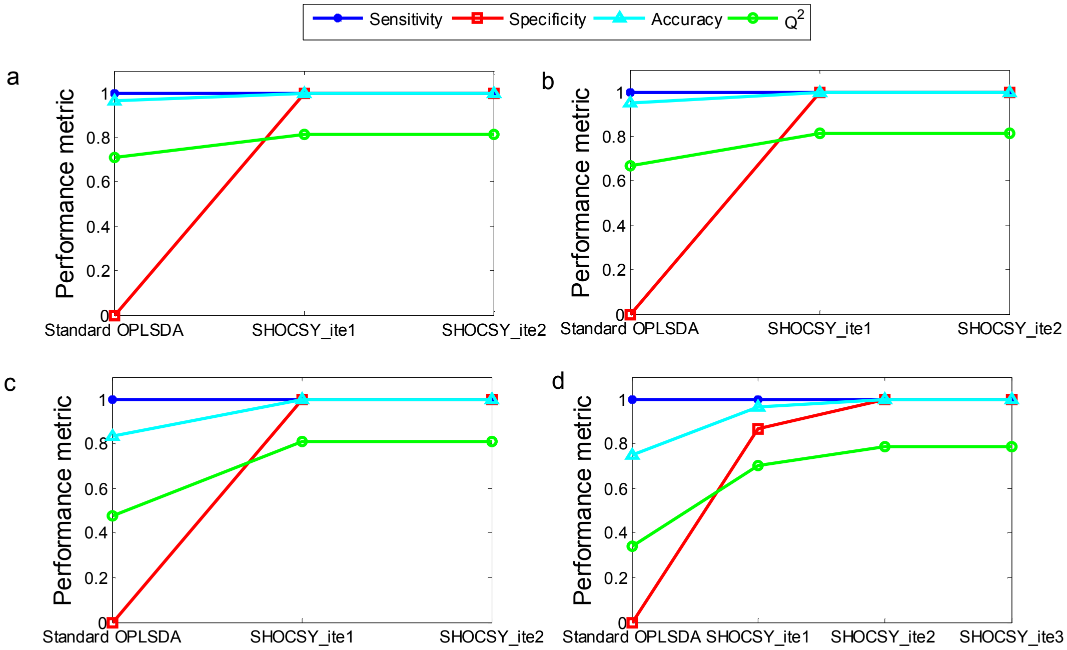

Figure S-1. The comparison for sensitivity, specificity, accuracy and Q^2 for standard OPLSDA approach to SHOCSY algorithm based on a simulated dataset of 30 spectra in each biological class with a) 6.7%; b) 10%; c) 33.3%; and d) 50% of idiosyncratic responders in the Paraquat toxicity class. The x-axes indicate the results of standard OPLS-DA and SHOCSY after each iteration. The SHOCSY approach stops when the Q^2 reaches a maximum value, indicating that all homogenous and idiosyncratic responders have been identified. Results for N = 100 and 500 were not shown here as their performance were similar to the simulated datasets of N = 30.

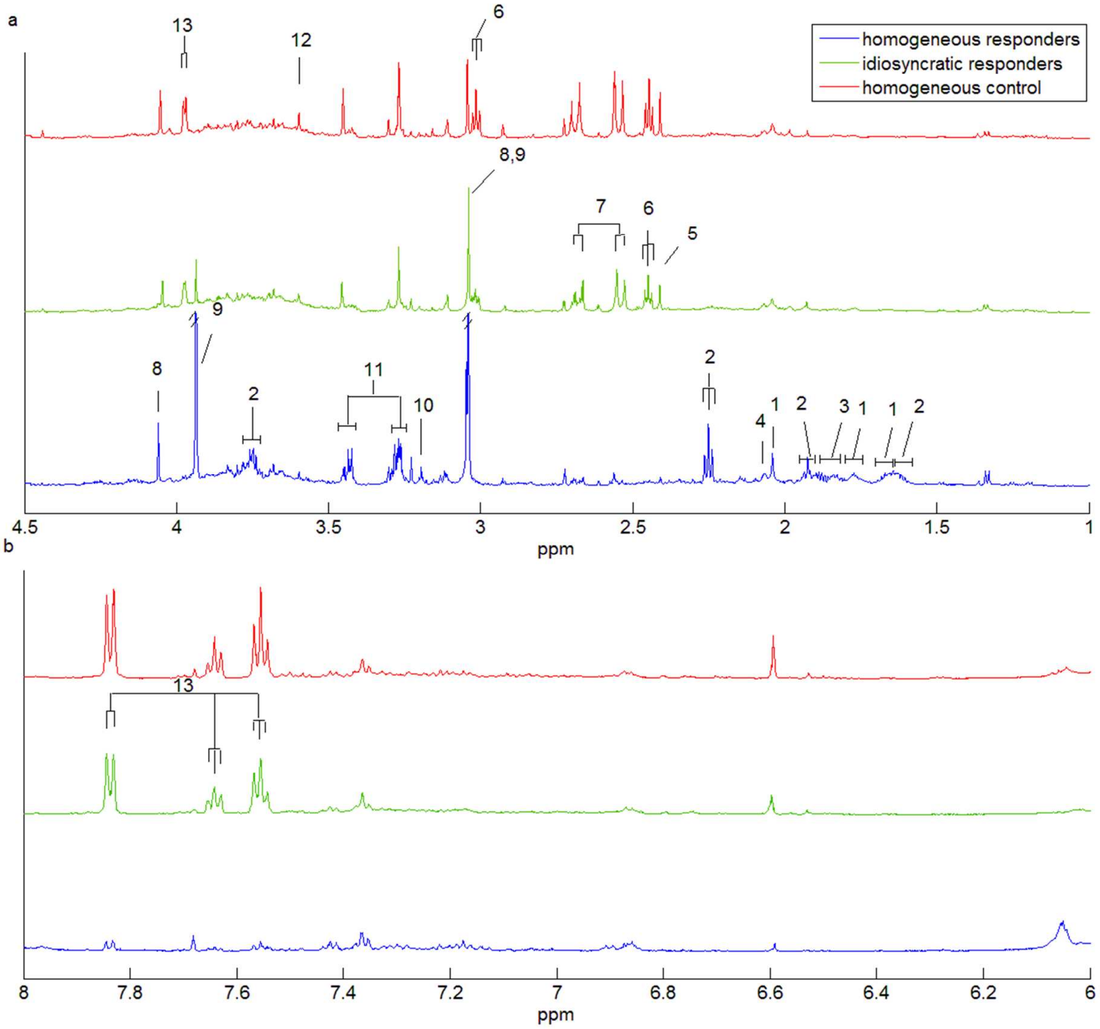

Figure S-2. The median ^{1}H NMR spectra of the homogeneous subsets and idiosyncratic responders for data representing 120-144h post dose for a) the aliphatic region from 1 to 4.5 ppm; and b) the aromatic region from 6 – 8 ppm. Key: 1, Nα-acetyl-citrulline; 2, 2-aminoadipic acid; 3, Citruline; 4, Diacetyl-hydrazine; 5, Succinate; 6, 2-oxoglutarate; 7, Citrate; 8, Creatinine; 9, Creatine; 10, Beta-alanine; 11, Taurine; 12, Glycine; 13, Hippurate.