SUPPORTING INFORMATION FOR

^{1}H NMR-Linked Metabolomics Analysis of Liver from a Mouse Model of NP-C1 Disease

Victor Ruiz-Rodado[1], Elena-Raluca Nicoli[2], Fay Probert[1], David A. Smith[2], Lauren Morris[2], Christopher A. Wassif[2,3], Frances M. Platt[2] and Martin Grootveld[1*].

[1]Leicester School of Pharmacy, De Montfort University, The Gateway, Leicester LE1 9BH, UK.
[2]Department of Pharmacology, Oxford University, Mansfield Road, Oxford OX1 3QT, UK.
[3]Eunice Kennedy Shriver National Institute of Child Health and Human Development, NIH, Bethesda, USA.

Supplementary Material

Contents:

Table S1: ^{1}H-NMR resonance assignments with chemical shifts, multiplicities, and coupling constant values of signals identified in aqueous liver sample extracts (page S2).

Figure S1. Receiver operating characteristic (ROC) curve exploration and testing probability view for direct comparisons of the NP-C1 disease classification with the (a) WT and (b) HET ones (page S4).

Figure S2: 400 MHz 2D ^{1}H-^{1}H COSY NMR profiles of aqueous liver sample extracts (page S5).

Figure S3: Box plots of TSP-normalised and Pareto-scaled intensities of GSSG and GSH ^{1}H NMR resonances (Table S1), and the GSH:GSSG molar ratio (page S6)

Section S1: Metabolomic advantages offered by the Random Forests (RFs) multivariate analysis strategy employed (page S7).

Figure S4: Mean decrease in accuracy (MDA) values computed for the 5 most effective discriminatory variables throughout 100 iterations (page S8).

Supplementary References (page S8)

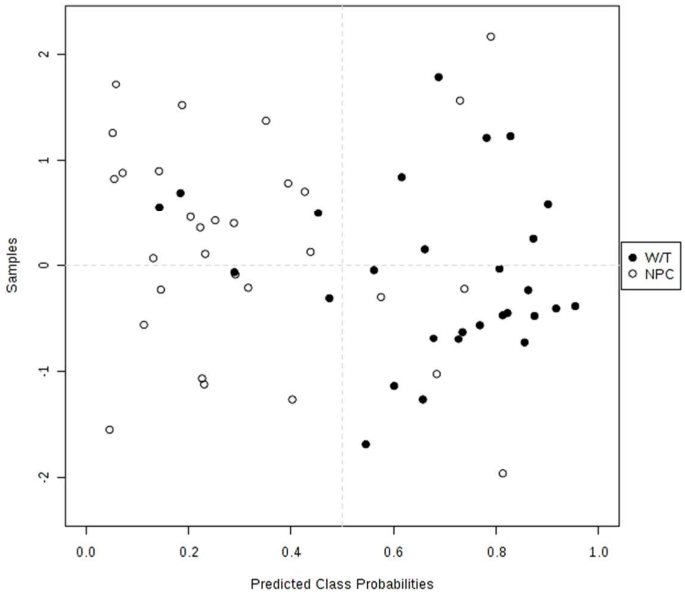
Table S1. ^{1}H-NMR Resonance assignments with chemical shifts, multiplicities and coupling constant values for signals identified in aqueous liver biopsy extracts.

Metabolites identified	Chemical shift (δ, ppm)	Multiplicity (*j*, Hz)
Unassigned	0.80	*t* (7.5)
2-Hydroxybutyrate-CH_3	0.89	*t* (7.5)
L-Valine-CH_3	0.98	*d* (7.1)
2-Aminobutyrate-CH_3	0.98	*t* (7.5)
Propylene glycol-CH_3	1.14	*d* (6.5)
Lactate-CH_3	1.33	*d* (7.1)
Alanine-CH_3	1.48	*d* (7.1)
Lysisne-C5-CH_2	1.73	*m*
Ornithine-C4-CH_2	1.73	*m*
Acetate-CH_3	1.93	*s*
Methionine-S-CH_3	2.13	*s*
Glutamate-C4-CH_2	2.35	*m*
Oxaloacetate-CH_2	2.38	*s*
Succinate-CH_2	2.41	*s*
Glutamine-C4-CH_2	2.45	*m*
Hypotaurine-C4-CH_2	2.65	*t* (7.1)
Cadaverine-C1/5-CH_2	2.68	*t* (7.3)
Aspartate-CH_{2a}	2.69	*dd* (8.9)
Sarcosine-CH_3	2.74	*s*
TMA-$CH_{3's}$	2.88	*s*
Dimethylglycine-$CH_{3's}$	2.90	*s*
Creatine-CH_3	3.08	*s*
Choline/Phosphocholine-$CH_{3's}$	3.21	*s*
Betaine-$CH_{3's}$	3.23	*s*
Taurine-S-CH_2	3.26	*t* (6.8)
GSSG-Cys-CH_{2a}	3.33	*m*
Methanol	3.36	*s*
Glycine	3.57	*s*
Glycerol-C1/3-CH_2	3.63	*m*
Threonine-CH_2	4.26	*m*
Ascorbate	4.50	*d* (2.1)
GSH-Cys-CH	4.58	*m*
β-Glucose-/Glucose-6P-C1-CH	4.66	*d* (8.0)
Phosphoenolpyruvate-CH_{2a}	5.18	*t* (1.1)
α-Glucose-/Glucose-6P-C1-CH	5.25	*d* (3.8)
Glycogen	5.42	*br.*
Uracil-CO-CH	5.81	*d* (7.7)

GTP-C1'-CH	5.98	*d* (6.8)
3-Hydroxyphenylacetate-CH_2	6.78	*m*
Tyr-C2/6-CH	6.90	*d* (8.2)
Phenylalanine	7.32	*m*
Guanosine-C8-CH	7.97	*s*
GTP/GMP-C8-CH	8.12	*s*
Hypoxanthine-C8-CH	8.21	*s*
Inosine-C2-CH	8.35	*s*
Niacinamide-C2-CH	8.91	*s*
Nicotinate-C2-CH	8.93	*s*

Abbreviations: Cys, cysteine; Glucose-6P, glucose-6-phosphate; GSH and GSSG, reduced and oxidised glutathione, respectively; GMP, guanosine monophosphate; GTP, guanosine triphosphate; TMA, trimethylamine; b*r.*, broad.

(a)



(b)

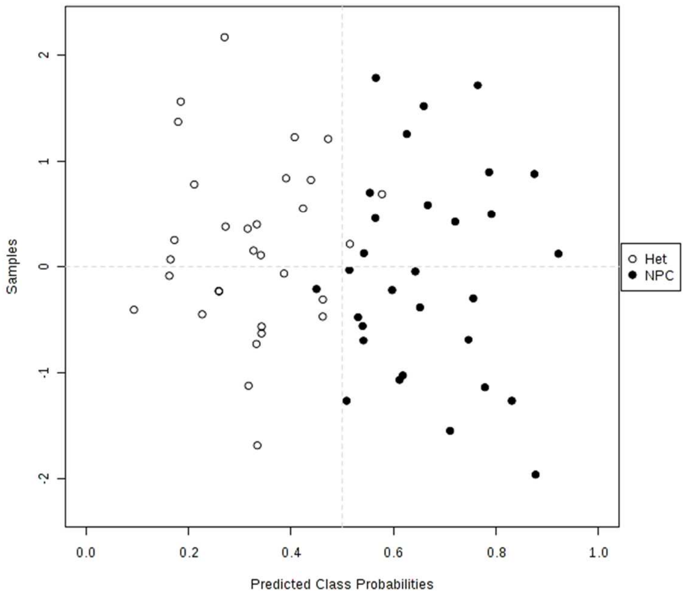


Figure S1. Receiver operating characteristic (ROC) curve exploration and testing probability view for direct comparisons of the NP-C1 disease classification with the (a) WT and (b) HET ones. These plots are derived from a balanced sub-sampling RFs model training strategy (predicted class probabilities for each sample employed the most significant 25 variables determined from the AUROC testing strategy).

Figure S2. 400 MHz 2D ^{1}H-^{1}H COSY NMR profiles of aqueous liver biopsy sample extracts

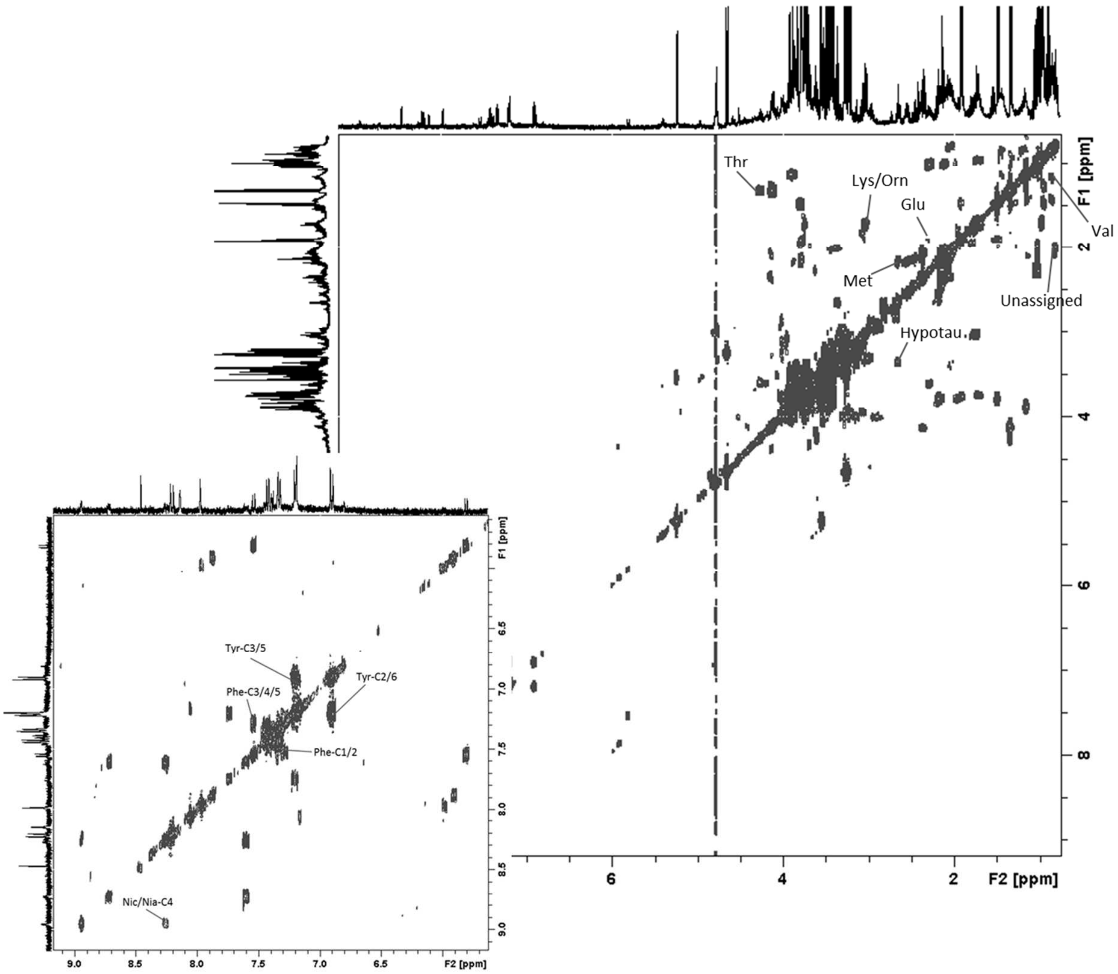


Figure S2. Full and expanded 5.8-9.2 ppm region of the 400 MHz ^{1}H-^{1}H COSY spectrum of an aqueous liver sample extract (prepared as outlined in section 2.2). Typical spectra are shown. Cross-peaks for selected metabolites are indicated in the spectrum. Abbreviations: Phe-C1/C2, Phenylalanine-C1/C2-CH; Tyr-C3/C5, Tyrosine-C3/C5-CH; Tyr-C2/C6, Tyrosine-C2/C6-CH; Lys/Orn, Lysine-C6-CH_2/Ornithine-C5-CH_2; Hypotau, Hypotaurine-C4-$CH_2SO_2^-$; Met, Methionine-C4-CH_2; Val, Valine-CH_3; Thr, Threonine-C3-CH; Nic/Nia-C4, Nicotinate/Niacinamide-C4-H; Phe-C3/4/5, Phenylalanine-C3/4/5-CH; Glu, Glutamate-C4-CH_{2a}.

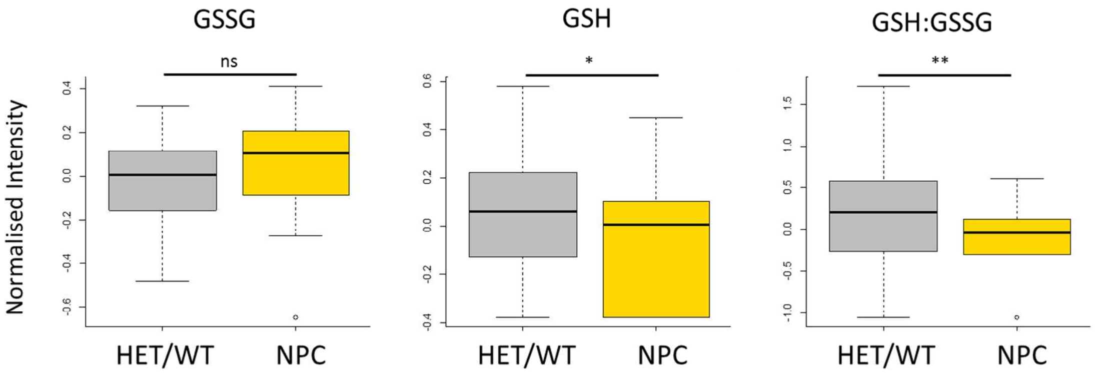


Figure S3: Box plots of TSP-normalised and Pareto-scaled intensities of GSSG and GSH ^{1}H NMR resonances (Table S1), and the GSH:GSSG molar ratio. *P*-values were determined from the ANCOVA model employed (equation 1). Abbreviations: ns, not significant; *, $p < 0.05$; **, $p < 0.01$.

Section S1.

Metabolomic advantages offered by the RFs multivariate analysis strategy employed: Several MV analysis strategies have been successfully utilised for the analysis of datasets in which the number of possible predictor variables exceeds the number of samples available, i.e. P > n situations; one approach is the RFs strategy, which has been successfully applied in a range of metabolomics studies.[1,2] Indeed, the RFs technique provides a prediction/classification model consisting of an ensemble of tree-structured classifiers.[3]

Our RFs model was primarily tuned in order to improve its performance, and this approach has been previously investigated by Diaz-Uriate *et al.,*[4] in which the investigators demonstrated that both *mtry* and *ntree* can be tuned seperately since they are independent parameters. Moreover, the computational time required proportionally increases with the number of trees, and selecting a very large number does not necessarily provide an improved performance, and this approach was further explored in this investigation. *Mtry* has been shown to be the most critical parameter to tune. The default value for the dataset analysed was 11.66 [$(136)^{1/2}$], and therefore the value employed was 11.

In order to improve the accuracy of the variable selection, an iterative cross-validated process was employed for the analysis of our dataset. In view of the random sub-sampling procedure conducted by RFs for classification purposes, the variables with selectively higher MDA values can, of course, vary when this classification technique is repetitively applied to

MV analysis of the same dataset; moreover, in order to explore as many combinations of samples in the training and test sets possible, an iterative process was implemented here, although the OOB error term computation already involves a cross-validation (CV) procedure.[5] An assessment of the performance of this iterative cross-validated process was also conducted in order to test the reliabilities of the selected metabolite variables, and this revealed that, with the employment of the MDA value of the 15th-ranked variable as a threshold, the top 5 ranked variables selected would not have been selected in 2, 3, 5, 10 and 12% respectively of the CV testing cases (Figure S4). As expected, this percentage increases for variables with lower rankings.

Further iterative processes have been proposed, such as the recursive elimination of features by Diaz-Uriarte *et al.,*[4] in which the variable importance parameters (VIPs) are computed during the RFs analysis, a process repeated following removal of 20% of the least important variables until the OOB error term decreases to a stable value. This was also investigated in detail in.[6] The MV analysis strategies described in these works have the final goal of reducing the number of variables to a minimum. However, in cases where the selected features may be related to further, albeit associated, health disorders arising from the classifiable disease process and not the disease *per se*, such reductions in the number of biomarker variables retained may limit the level of metabolomics information derived therefrom.

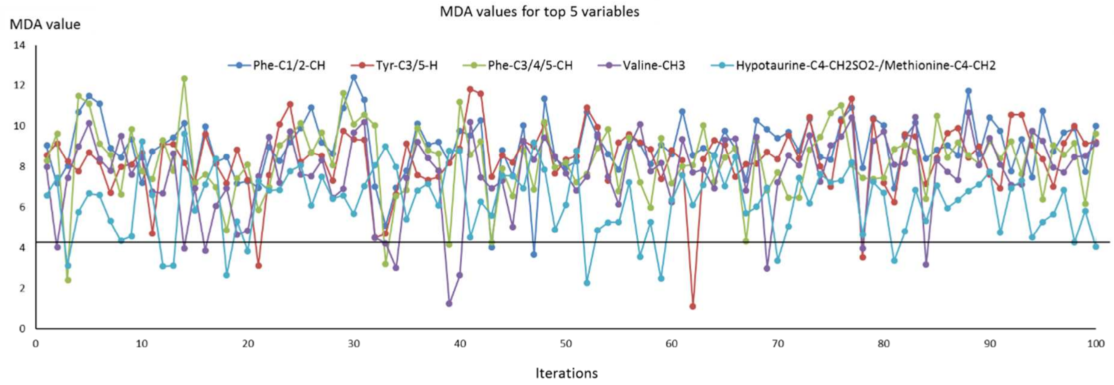


Figure S4. Mean decrease in accuracy (MDA) values computed for the 5 most effective discriminatory variables throughout 100 iterations. For this model, the black line represents the mean MDA value for the 15th most important selected ISB variable (nicotinate/niacinamide-C4-CH) acting as a threshold for variable selection.

Supplementary References

(1) Bertini, I.; Luchinat, C.; Miniati, M.; Monti, S.; Tenori, L. Phenotyping COPD by 1H NMR metabolomics of exhaled breath condensate. *Metabolomics* **2014***, 10*, 302-311.

(2) Smolinska, A.; Klaassen, E. M.; Dallinga, J. W.; van de Kant, K. D.; Jobsis, Q.; Moonen, E. J.; van Schayck, O. C.; Dompeling, E.; van Schooten, F. J. Profiling of volatile organic compounds in exhaled breath as a strategy to find early predictive signatures of asthma in children. *PLoS One* **2014***, 9*, e95668.

(3) Breiman, L. Random Forest. *Mach Learn* **2001***, 45*, 5-32.

(4) Diaz-Uriarte, R.; Alvarez de Andres, S. Gene selection and classification of microarray data using random forest. *BMC Bioinformatics* **2006***, 7*, 3.

(5) Bao, L.; Cui, Y. Prediction of the phenotypic effects of non-synonymous single nucleotide polymorphisms using structural and evolutionary information. *Bioinformatics* **2005***, 21*, 2185-2190.

(6) Genuer, R.; Poggi, J. M.; Tuleau-Malot, C. Variable selection using random forests. *Pattern Recogn. Lett.* **2010***, 31*, 2225-2236.