Supporting Information: Metabolic Profiling And The Metabolome-Wide Association Study: Significance Level For Biomarker Identification

Marc Chadeau-Hyam,[†,#] Timothy M D Ebbels,[‡,#] Ian J Brown,[†] Queenie Chan,[†] Jeremiah Stamler,[¶] Chiang Ching Huang,[¶] Martha L Daviglus,[¶] Hirotsugu Ueshima,[§] Liancheng Zhao,[‖] Elaine Holmes,[‡,⊥] Jeremy K Nicholson,[‡,⊥] Paul Elliott,[*,†,⊥] and Maria De Iorio[*,†]

Department of Epidemiology and Biostatistics, School of Public Health, Imperial College, London W2 1PG, UK, Biomolecular Medicine, Department of Surgery and Cancer, Faculty of Medicine, Imperial College, London SW7 2AZ, UK, Department of Preventive Medicine, Feinberg School of Medicine, Northwestern University, Chicago, Illinois 60611, US, Department of Health Science, Shiga University of Medical Science, Otsu, Japan, Department of Epidemiology, Fu Wai Hospital and Cardiovascular Institute, Chinese Academy of Medical Sciences, Beijing, People's Republic of China, and MRC-HPA Center for Environment and Health, Imperial College London UK

E-mail: p.elliott@imperial.ac.uk; m.deiorio@imperial.ac.uk

[*]To whom correspondence should be addressed

[†]Department of Epidemiology and Biostatistics, School of Public Health, Imperial College, London, UK

[‡]Biomolecular Medicine, Department of Surgery and Cancer, Faculty of Medicine, Imperial College, London, UK

[¶]Department of Preventive Medicine, Feinberg School of Medicine, Northwestern University, Chicago, US

[§]Department of Health Science, Shiga University of Medical Science, Otsu, Japan

[‖]Department of Epidemiology, Fu Wai Hospital and Cardiovascular Institute, Chinese Academy of Medical Sciences, Beijing, People's Republic of China

[⊥]MRC-HPA Center for Environment and Health, Imperial College London UK

[#]Contributed equally to this work

Table 1: Reference metabolites specifications. Figures are based on the mean spectrum calculated over the whole Chinese population, and the 'true positive range' for each metabolite is defined in accordance with the mean profile.

Metabolite	First Peak Position (ppm) $[Multiplicity]^{\dagger}$	True Positive range Δ (ppm)
Hippurate	7.846 $[d]^{*}$	[7.827 - 7.854]
	7.655 $[t]$	[7.625 - 7.660]
	7.569 $[t]$	[7.537 - 7.578]
	3.979 $[d]$	[3.966 - 3.984]
Alanine	1.49 $[d]^{*}$	[1.474 - 1.494]
	3.79 $[q]$	[3.77 - 3.81]
Formate	8.46 $[s]^{*}$	[8.456 - 8.465]

† $[s]$: single peak; $[d]$: doublet; $[t]$: triplet; $[q]$: quadruplet;

* standard location used in the disease model;

Table 2: Significance threshold α' ($\times 10^{-5}$) and effective number of tests ($\times 10^{3}$) based on Bonferroni correction (ENT_B). 95% confidence intervals are presented between brackets. Bold figures are the proportion of Effective/Actual number of tests. Figures are based on 50,000 re-samples of the disease indicator under the null hypothesis. Figures are based on the full INTERMAP population (4,630 spectra) at the medium resolution (7,100 variables).

Sample size #Cases/controls		Overall error rate $\alpha =$ 1%	5%	10%
50/50	α'	0.88 (0.81;0.96)	3.91 (3.79;4.03)	7.71 (7.50;7.93)
	ENT_B	1.14 (1.04;1.24) **16 %**	1.28 (1.24;1.32) **18 %**	1.30 (1.26;1.33) **18 %**
100/100	α'	0.73 (0.68;0.79)	3.23 (3.13;3.35)	6.44 (6.27;6.62)
	ENT_B	1.36 (1.27;1.46) **19 %**	1.55 (1.49;1.60) **22 %**	1.55 (1.51;1.59) **22 %**
200/200	α'	0.56 (0.51;0.60)	2.57 (2.47;2.66)	5.33 (5.18;5.49)
	ENT_B	1.80 (1.66;1.97) **25 %**	1.94 (1.88;2.03) **27 %**	1.88 (1.82;1.93) **26 %**
500/500	α'	0.43 (0.40;0.46)	2.11 (2.04;2.19)	4.38 (4.24;4.50)
	ENT_B	2.31 (2.16;2.50) **32 %**	2.36 (2.28;2.45) **33 %**	2.28 (2.22;2.36) **32 %**

Table 3: Mean number of false positive associations under the null hypothesis of no association. The results shown are averages over 50 replicates (minimum and maximum values over the replicates are given in parentheses). Both O2PLS approaches are based on a Bonferroni corrected threshold. T-test results are reported uncorrected, Bonferroni corrected, using the exact metabolome-wide significance level (MWSL), and using the general MWSL we estimated. Estimates of the MWSL are based on a FWER α of 5%.

# Cases/Controls	50/50	100/100	200/200
O2PLS - Bootstrap	71.1	81.6	45.5
	(11-250)	(14-602)	(6-211)
O2PLS - Permutation	16.1	15.9	14.9
	(1-65)	(0-49)	(0-48)
T-test uncorrected	686.9	846.8	762.3
	(292-1,439)	(349-2,328)	(328-1,740)
T-test Bonferroni	0.0	0.0	0.0
	(0-0)	(0-0)	(0-0)
T-test exact MWSL	0.0	0.1	0.0
	(0-1)	(0-2)	(0-1)
T-test general MWSL	0.0	0.0	0.0
$\alpha' = 9 \times 10^{-6}$	(0-0)	(0-2)	(0-0)

Table 4: Per-metabolite statistical power, calculated over the 50 replications of the disease model, for both the single (Table 4-a) and the multi-metabolite models (Table 4-b). FP rate is defined as the mean number of false positive associations as a proportion of the number of significant variables. The FWER is set to 5%.

Table 4-a One disease associated metabolite (hippurate).

	Prevalence	10%		30%		50%	
	Sample Size	Power	FP Rate	Power	FP Rate	Power	FP Rate
O2PLS	50/50	100%	41.8%	100%	37.7%	100%	39.4%
Bootstrap	100/100	N.A.	N.A.	100%	44.4%	100%	60.1%
	200/200	N.A.	N.A.	100%	58.0%	100%	73.8%
O2PLS	50/50	98%	9.3%	98%	9.2%	100%	3.5%
Permutation	100/100	N.A.	N.A.	100%	6.1%	100%	4.7%
	200/200	N.A.	N.A.	100%	4.9%	100%	6.0%
T-test	50/50	82%	0.9%	82%	0.8%	98%	1.0%
Bonferroni	100/100	N.A.	N.A.	100%	3.6%	100%	7.8%
	200/200	N.A.	N.A.	100%	15.6%	100%	36.9%
T-test	50/50	94%	4.2%	92%	5.9%	100%	2.5%
exact MWSL	100/100	N.A.	N.A.	100%	6.8%	100%	13.9%
	200/200	N.A.	N.A.	100%	22.7%	100%	45.3%
T-test	50/50	90%	3.8%	90%	3.1%	100%	1.7%
general MWSL	100/100	N.A.	N.A.	100%	5.5%	100%	11.3%
$\alpha' = 9 \times 10^{-6}$	200/200	N.A.	N.A.	100%	20.1%	100%	43.4%

Table 4-b Three disease associated metabolites (hippurate, alanine and formate)

		Power			FP rate
	Metabolite	Hippurate	Alanine	Formate	
	Logistic coefficient ($\|\beta\|$)	1.0	2.0	4.0	
O2PLS	50/50	20%	18%	98%	84.1%
Bootstrap	100/100	38%	38%	100%	78.6%
	200/200	70%	52%	100%	73.8%
O2PLS	50/50	14%	8%	92%	65.9%
Permutation	100/100	26%	22%	100%	49.7%
	200/200	34%	28%	100%	46.9%
T-test	50/50	0%	0%	66%	1.3%
Bonferroni	100/100	2%	4%	100%	2.5%
	200/200	8%	10%	100%	14.7%
T-test	50/50	0%	0%	78%	9.0%
exact MWSL	100/100	8%	6%	100%	6.9%
	200/200	18%	18%	100%	23.9%
T-test	50/50	0%	0%	74%	3.4%
general MWSL	100/100	8%	6%	100%	4.9%
$\alpha' = 9 \times 10^{-6}$	200/200	18%	16%	100%	20.8%

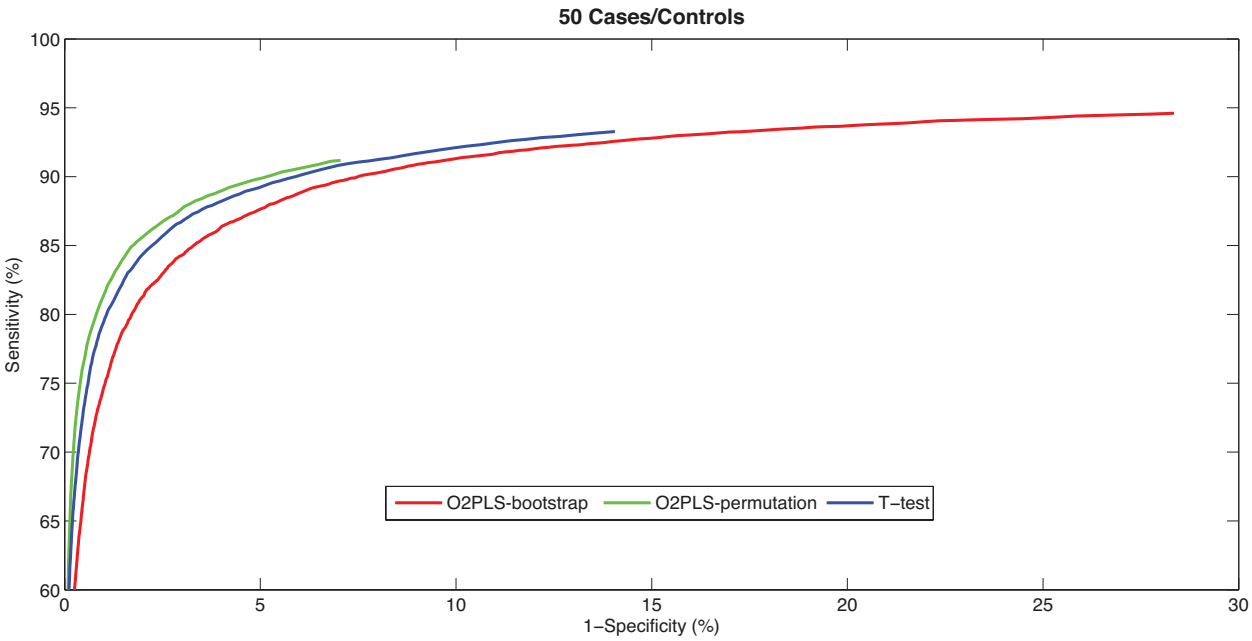

Figure 1: ROC curves for the single metabolite model, prevalence is set to 10%. Figures are based on 500 data points corresponding to $\alpha \in [10^{-10}; 10^{-1}]$. Note that the size of the reference population ($N = 836$) was not large enough for us to examine 100 and 200 cases/controls at this prevalence.

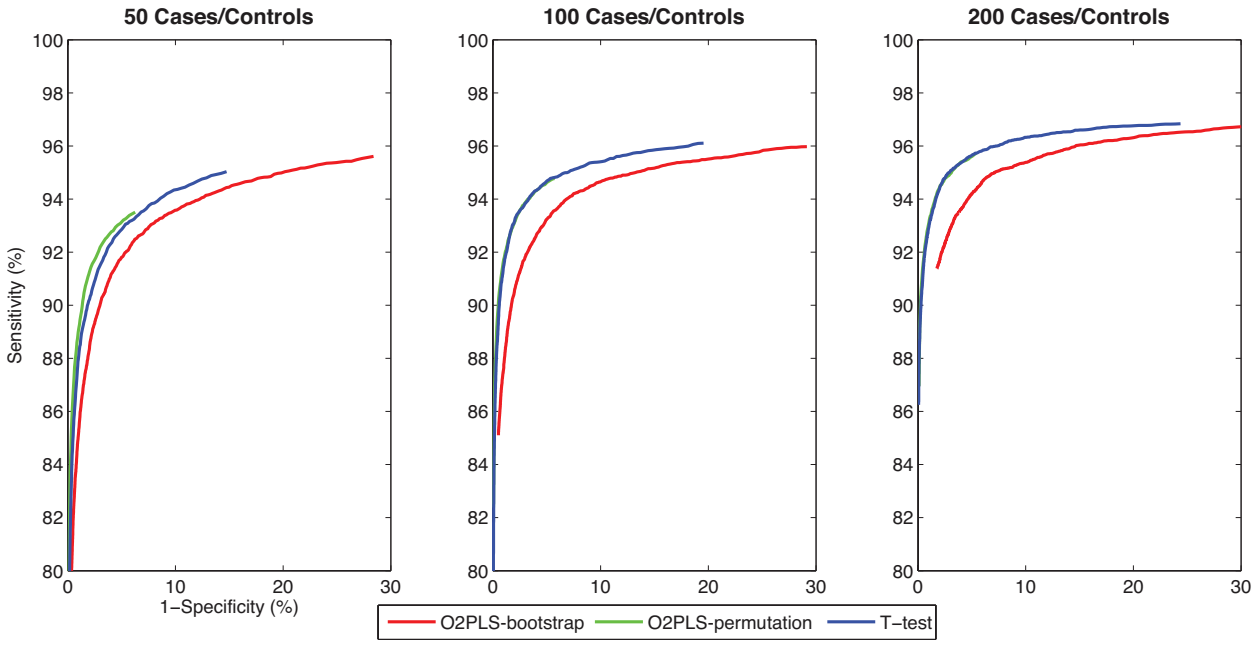

Figure 2: ROC curves for the single metabolite model, prevalence is set to 50%. Figures are based on 500 data points corresponding to $\alpha \in [10^{-10}; 10^{-1}]$.

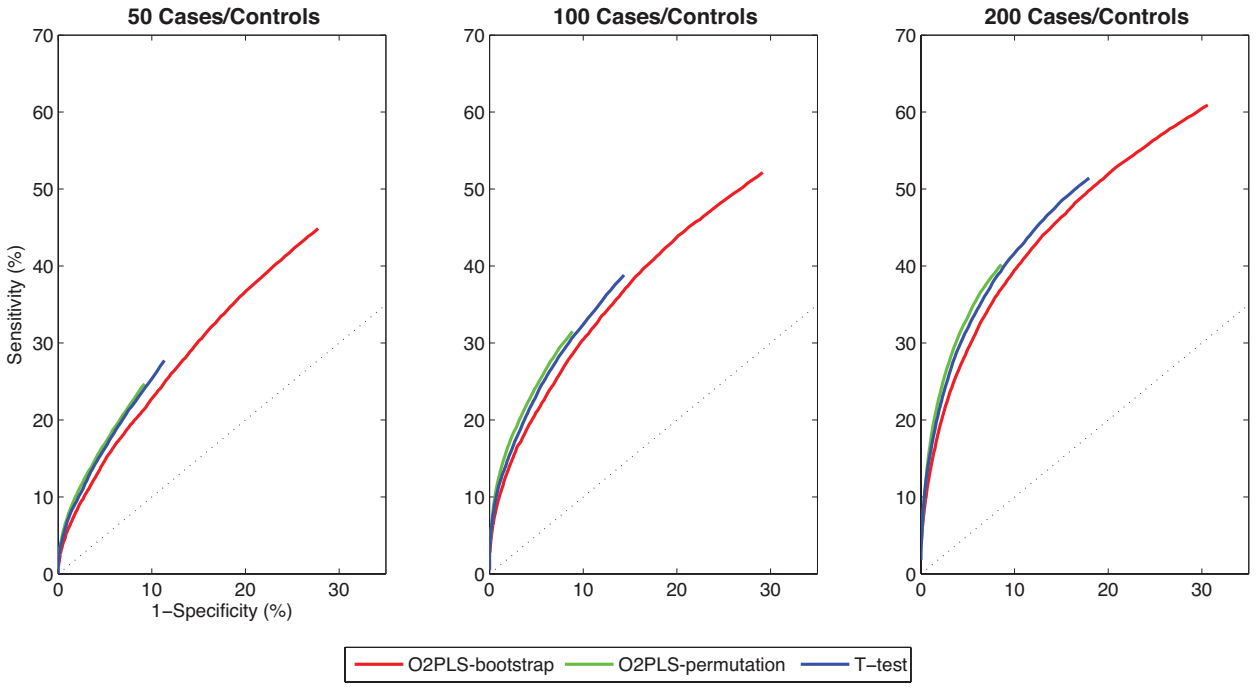

Figure 3: ROC curves for the multi-metabolite model, these were obtained for 500 data points corresponding to $\alpha \in [10^{-10}; 10^{-1}]$.

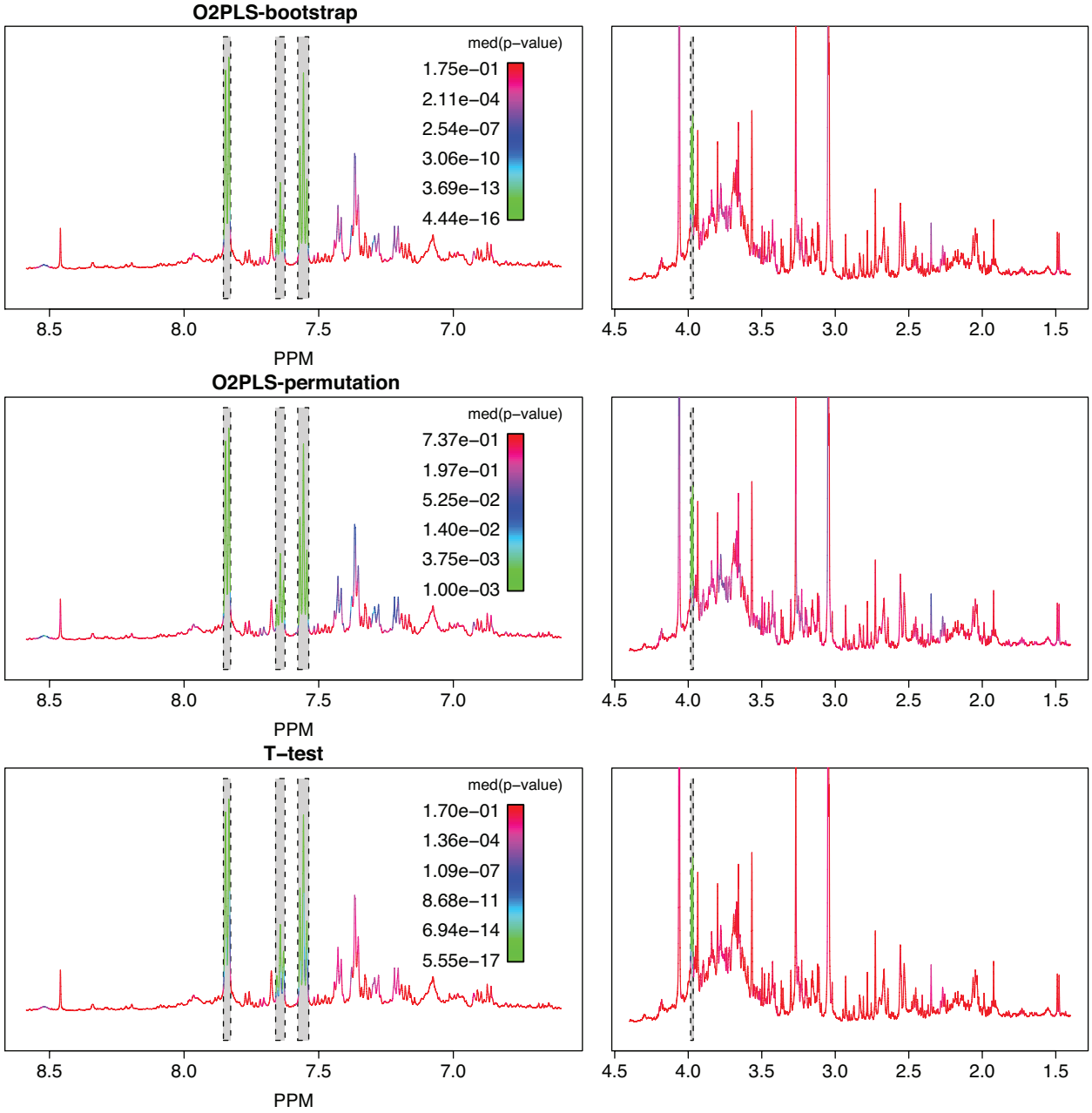

Figure 4: Mean spectrum colored according to median p-values calculated over 50 replications. The light grey area represents the 'true positive range'. This figure applies to the single metabolite model, for 200 cases/controls and a prevalence set to 30%. **Note that color scale is different for each method.**

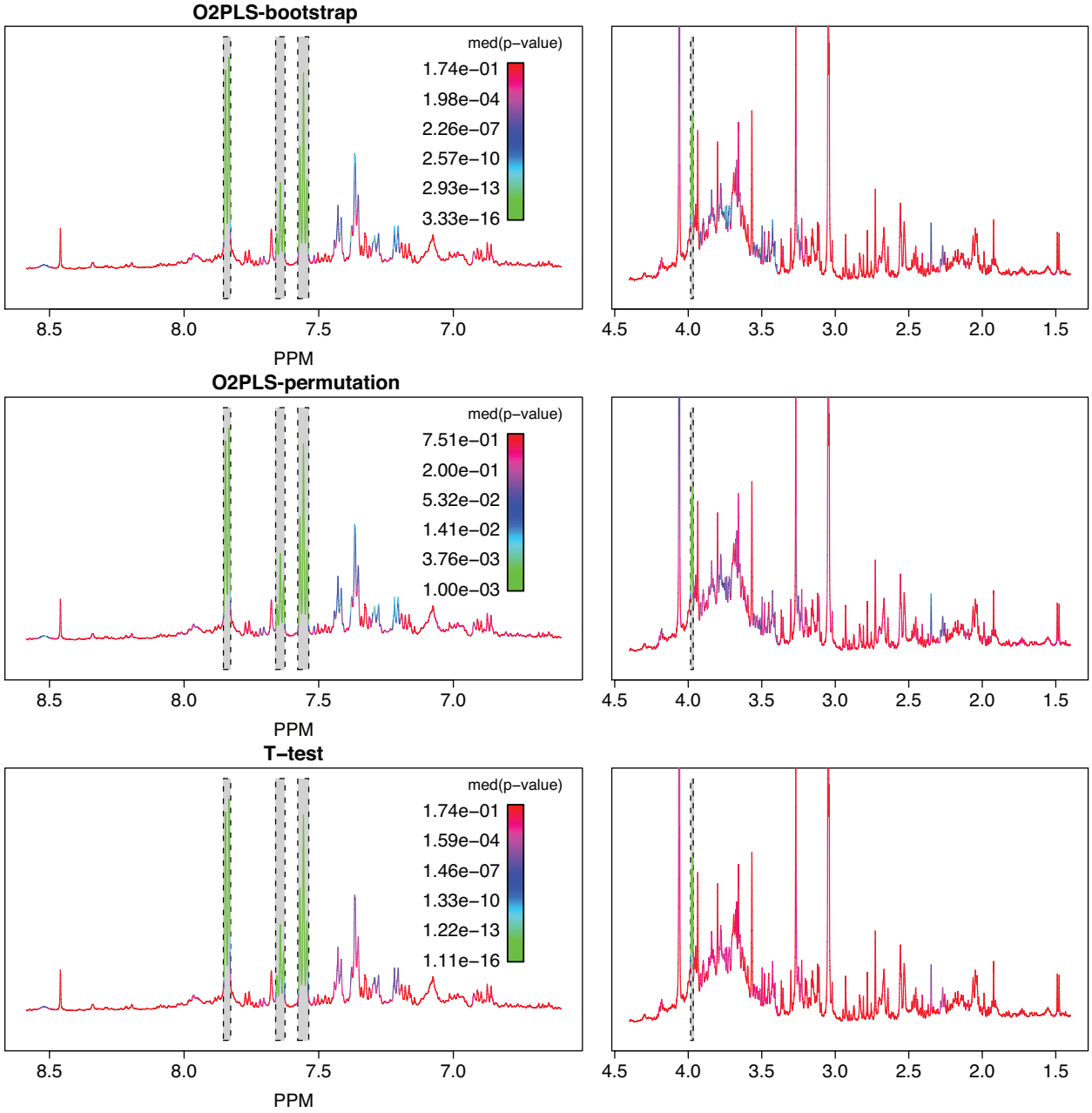

Figure 5: Mean spectrum colored according median p-values calculated over 50 replications. The light grey area represents the 'true positive range'. This figure applies to the single metabolite model, for 200 cases/controls and a prevalence set to 50%. **Note that color scale is different for each method**.

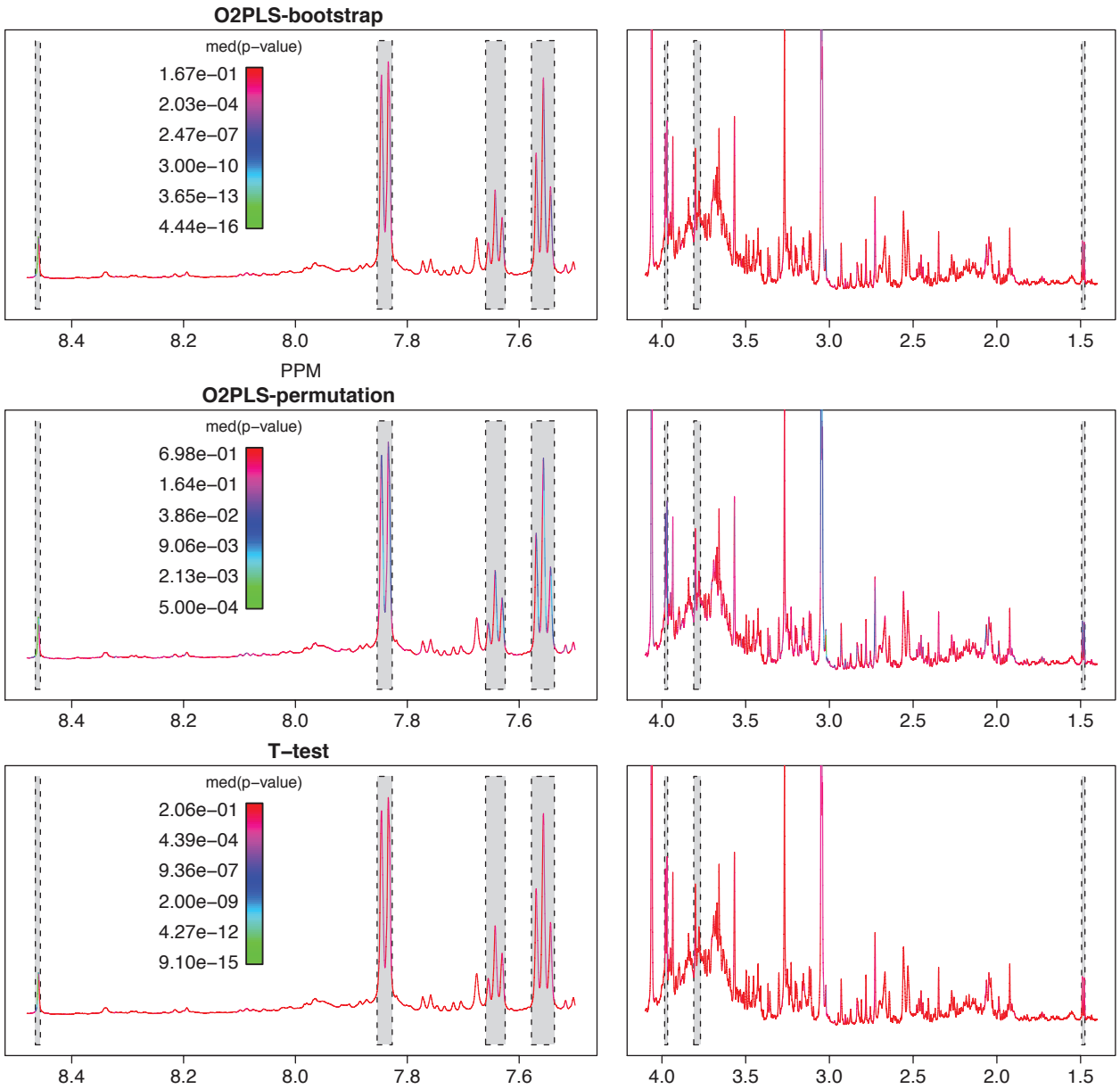

Figure 6: Mean spectrum colored according median p-values calculated over 50 replications. The light grey area represents the 'true positive range'. This figure applies to the multi metabolite model with 200 cases/controls. **Note that color scale is different for each method**.

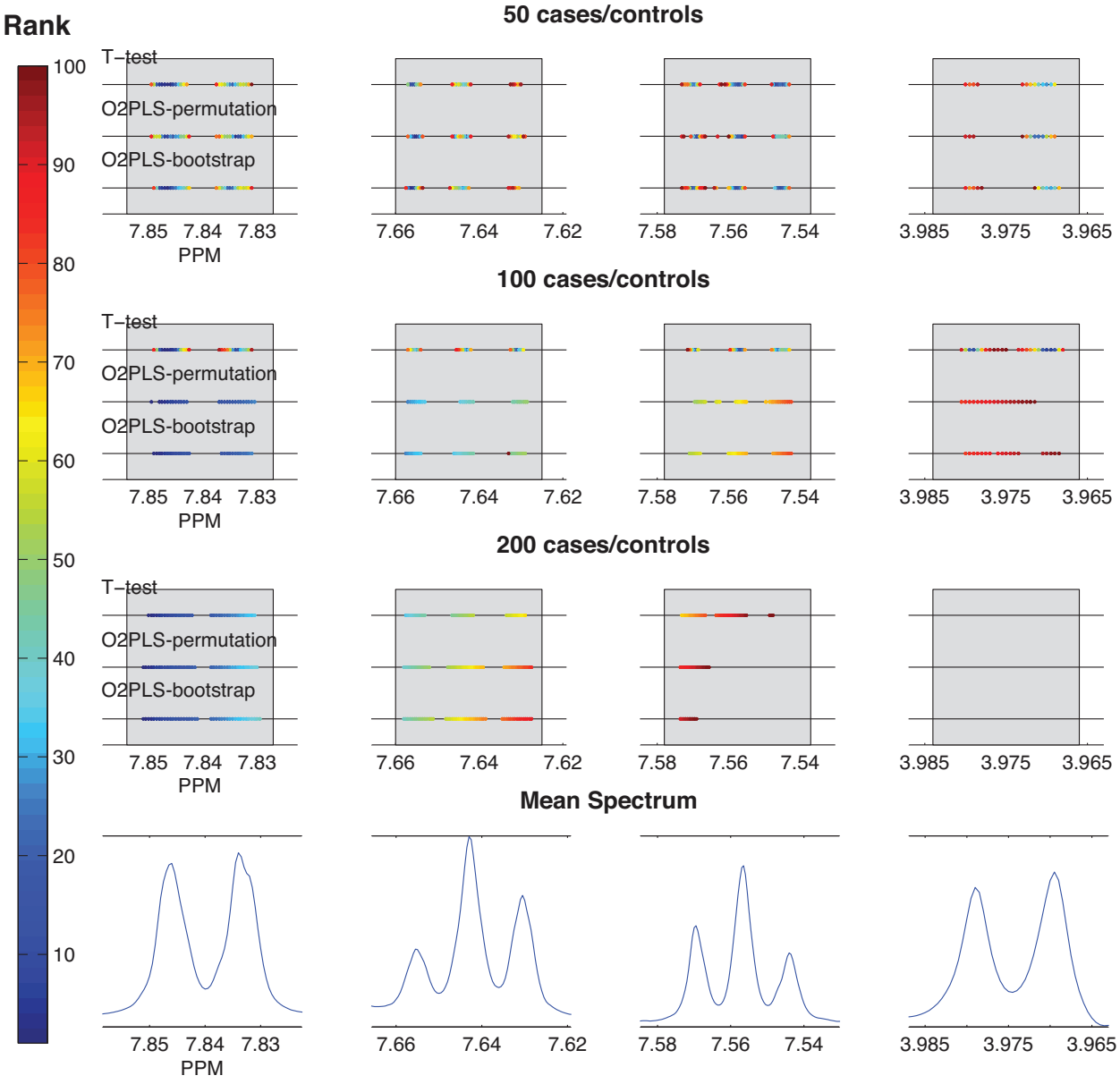

Figure 7: Location of the 100 metabolites with the lowest mean p-values using the three methods, single metabolite model (hippurate). For all simulations, none of the top 100 metabolites were found outside the 'true positive range' (represented in light grey in the figure). Points are colored according to their rank. Results are provided for a prevalence set to 50%.

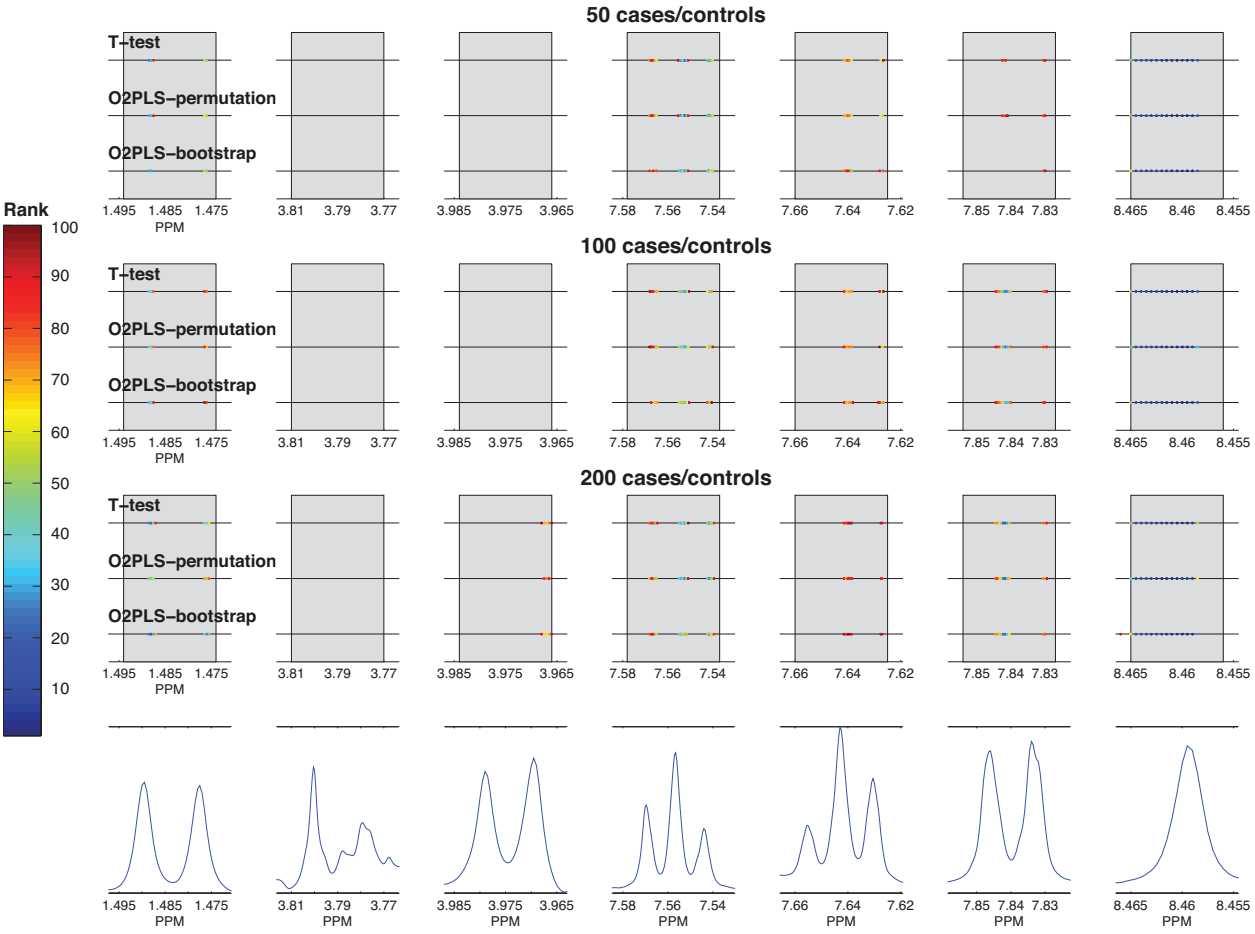

Figure 8: Location of the 100 metabolites with the lowest mean p-values using the three methods, multiple metabolites model (hippurate, alanine, formate). For all simulations, very few of the top 100 metabolites were found outside the 'true positive range' (represented in light grey in the figure). Points are colored according to their rank.