Supplementary Materials

Supplementary Methods

aROC curve analysis of winner's curse

We evaluated whether the estimated aAUC and aROC may be biased because of the "winner's curse" or using those metabolites most significantly associated with cancer.

We first performed stepwise selection to identify the peaks most strongly associated with liver cancer and, using those peaks, calculated the resulting aAUC and aROC. Note this is a slightly modified and simplified selection procedure and is intended only to evaluate the influence of the winner's curse.

We then estimated the bias in the above estimates of aAUC and aROC using the following steps. We split the data into training (90%) and test (10%) sets, selected metabolites and built the predictor using the training set and estimated the aAUC and aROC using the test set. We repeat this calculation 10x using 10 splits of the data and report the average aAUC and aROC (**Supplementary Figure 5**). Note, the difference between these averages and the overall estimates (based on the entire dataset) reflect the bias from potentially overfitting the model.

Supplementary Tables

Supplementary Table 1. Identification of trigonelline and 38 features associated with liver cancer, liver disease mortality and coffee intake.

						P-value		
Compound #	**ID (monoisotopic mass@RT)**	**Compound name**	**MSI identification confidence level**	***m/z***	**RT (min)**	**Coffee intake†**	**Liver cancer‡**	**Liver disease mortality‡**
1	449.3132@6.451681*	Glycochenodeoxycholic acid	1	450.3205	6.451681	2.69x10^{-14}	6.85X10^{-7}	3.32x10^{-10}
1	471.2943@6.4516096			472.3016	6.4516096	2.29x10^{-12}	9.98x10^{-7}	1.52x10^{-9}
2	465.3082@6.1825604*	Glycocholic acid	1	466.3155	6.1825604	2.23x10^{-8}	6.63x10^{-8}	1.57x10^{-7}
2	482.3352@6.181003			483.3425	6.181003	8.73x10^{-9}	1.31x10^{-6}	4.19x10^{-7}
2	518.2202@6.1818233			519.2275	6.1818233	1.15x10^{-7}	5.98x10^{-6}	1.87x10^{-7}
3	136.0382@0.86198366*	Hypoxanthine	1	137.0455	0.8620	1.05x10^{-7}	1.19x10^{-7}	1.98x10^{-6}
4	159.0675@1.5374656*	Serotonin	1	160.0748	1.5374656	8.44x10^{-8}	1.17x10^{-5}	1.25x10^{-7}
5	181.0767@1.275114*	Tyrosine	1	182.0840	1.2751	1.84x10^{-6}	5.75x10^{-7}	4.08x10^{-7}
5	181.0766@0.86087453			182.0839	0.8609	4.75x10^{-8}	3.91x10^{-8}	3.23x10^{-9}
5	135.0678@0.86145175			136.0751	0.8615	1.72x10^{-8}	7.10x10^{-8}	2.47x10^{-9}
5	118.042@0.8618372			119.0493	0.8618	3.01x10^{-7}	2.45x10^{-7}	2.11x10^{-9}
5	164.0467@0.86240983			165.0540	0.8624	1.25x10^{-9}	4.89x10^{-8}	8.78x10^{-10}
5	135.0673@1.2750853			136.0746	1.2751	5.63x10^{-6}	4.58x10^{-6}	4.77x10^{-7}
6	545.7893@6.9137807	LysoPC(18:2)	2	546.7966	6.9137807	9.08x10^{-11}	3.96x10^{-8}	4.40x10^{-9}
6	286.1276@6.9126344			287.1349	6.9126344	7.70x10^{-10}	1.90x10^{-7}	1.62x10^{-8}
6	538.8086@6.914731*			539.8159	6.914731	1.39x10^{-8}	4.06x10^{-7}	5.31x10^{-10}
6	806.4585@6.9148936			807.4658	6.9148936	1.45x10^{-7}	8.19x10^{-8}	4.10x10^{-9}
6	538.3067@6.915277			539.3140	6.915277	1.58x10^{-6}	1.24x10^{-6}	4.43x10^{-6}
6	805.9551@6.9153476			806.9624	6.9153476	4.72x10^{-10}	2.64x10^{-7}	4.39x10^{-9}
6	798.4699@6.916353			799.4772	6.916353	3.59x10^{-10}	5.30x10^{-8}	3.65x10^{-9}
6	797.9663@6.9167204			798.9736	6.9167204	6.44x10^{-11}	2.74x10^{-8}	2.75x10^{-8}
7	481.3229@6.8804502*	LysoPC(15:0)	2	482.3302	6.8804502	<1.11x10^{-16}	4.33x10^{-8}	5.96x10^{-10}
8	479.3426@7.109446*	LysoPC(P-16:0)	2	480.3499	7.109446	4.82x10^{-11}	7.63x10^{-8}	3.16x10^{-7}
9	230.1626@2.280263*	Dipeptide: Leu-Val or isomer	3	231.1699	2.280263	4.32x10^{-7}	1.15x10^{-5}	1.07x10^{-6}
10	481.3495@7.157514*	Unknown	4	482.3568	7.157514	5.77x10^{-15}	1.01x10^{-7}	1.57x10^{-7}
11	124.0638@2.238123*	Unknown	4	125.0711	2.238123	<1.11x10^{-16}	6.59x10^{-6}	1.10x10^{-6}
12	249.0052@2.7640297	Unknown	4	250.0125	2.7640297	<1.11x10^{-16}	1.15x10^{-5}	5.65x10^{-11}

12	202.0168@2.7641954*			203.0241	2.7641954	<1.11x10^{-16}	3.78x10^{-6}	8.74x10^{-11}
13	242.1067@4.0267076*	Unknown	4	243.1140	4.0267076	5.01x10^{-14}	1.05x10^{-5}	1.13x10^{-8}
14	278.076@4.280106*	Unknown	4	279.0833	4.280106	9.86x10^{-10}	1.65x10^{-6}	1.20x10^{-7}
14	238.0836@4.280605			239.0909	4.280605	9.15x10^{-10}	1.93x10^{-10}	1.04x10^{-9}
15	356.1964@5.496086*	Unknown	4	357.2037	5.496086	<1.11x10^{-16}	1.69x10^{-6}	2.47x10^{-7}
16	388.2573@6.339673*	Unknown	4	389.2646	6.339673	5.44x10^{-6}	3.05x10^{-6}	4.30x10^{-8}
17	504.3029@6.8187194*	Unknown	4	505.3102	6.8187194	<1.11x10^{-16}	5.46x10^{-6}	6.31x10^{-12}
18	793.9569@6.9292555*	Unknown	4	794.9642	6.9292555	1.23x10^{-11}	1.34x10^{-5}	3.98x10^{-6}
19	292.9962@0.6360238*	Unknown	4	294.0035	0.6360	2.76x10^{-10}	1.68x10^{-6}	7.99x10^{-8}
20	162.0534@0.64166015*	Unknown	4	163.0607	0.6417	6.09x10^{-7}	1.53x10^{-7}	4.83x10^{-7}
21	202.1313@0.8667429*	Unknown	4	203.1386	0.8667	2.76x10^{-10}	3.53x10^{-6}	2.30x10^{-6}
22	159.0292@0.6753366*	Trigonelline	1	160.0365	0.6753	<1.11x10^{-16}	4.68x10^{-5}	3.07x10^{-8}

* Indicates the main feature that was used to estimate associations with outcomes of interest.

†P-values were estimated from linear regression models treating a given spectral feature as the (log$_2$-transformed) continuous response variable and coffee consumption (continuous, g/day) as the exposure variable. Models were adjusted for age, smoking intensity (cigarettes/day), run order and the surrogate variables identified by SVA and observations from both case-control sets was used (N=940). Statistical tests were two-sided.

‡P-values were estimated from conditional logistic regression models treating case status, liver cancer (n=221 cases; n=221 controls) or liver disease death (n=242 cases; n=242 controls), as the response variable and a given spectral feature as the (log$_2$-transformed) continuous exposure variable. Models were adjusted for age, smoking intensity (cigarettes/day) and run order. Statistical tests were two-sided.

Abbreviations: MSI, Metabolomics Standards Initiative; RT, retention time

Supplementary Table 2. Odds ratios and 95% confidence intervals for incident liver cancer comparing men in the 90th and 10th percentiles, based on the distribution in controls, for top unknown metabolites, using conditional logistic regression

Direction of association	**Metabolite No. (mass@RT)**	**Unadjusted model**	**Model 1 †**	**>0 to 10 years of follow-up (model 1) † ‡**	**>10 years of follow-up (model 1) †§**	**Model 2 (diet adjusted) †‖**
Increased risk	Unknown 12 (202.0168@2.7641954)					
	OR (95% CI) *	3.71 (2.15-6.42)	3.61 (1.92-6.79)	7.02 (1.40-35.10)	3.26 (1.51-7.00)	3.05 (1.53-6.11)
	P-value¶	<0.001	<0.001	0.02	0.003	0.002
	Unknown 13 (242.1067@4.0267076)					
	OR (95% CI) *	3.23 (2.00-5.22)	2.93 (1.71-5.02)	3.58 (1.13-11.33)	3.66 (1.78-7.55)	2.55 (1.42-4.56)
	P-value¶	<0.001	<0.001	0.03	<0.001	0.002
	Unknown 16 (388.2573@6.339673)					
	OR (95% CI) *	5.51 (2.75-11.07)	4.93 (2.32-10.48)	4.83 (1.04-22.43)	6.36 (2.35-17.20)	4.64 (2.04-10.57)
	P-value¶	<0.001	<0.001	0.04	<0.001	<0.001
	Unknown 17 (504.3029@6.8187194)					
	OR (95% CI) *	3.07 (1.90-4.94)	3.31 (1.87-5.86)	3.01 (0.87-10.36)	3.99 (1.89-8.44)	3.29 (1.73-6.24)
	P-value¶	<0.001	<0.001	0.08	<0.001	<0.001
	Unknown 19 (292.9962@0.6360238)					
	OR (95% CI) *	5.18 (2.82-9.51)	4.21 (2.15-8.23)	9.15 (1.73-48.34)	4.31 (1.88-9.88)	5.21 (2.40-11.34)
	P-value¶	<0.001	<0.001	0.01	<0.001	<0.001
	Unknown 20 (162.0534@0.64166015)					
	OR (95% CI) *	8.32 (3.96-17.46)	6.77 (2.94-15.56)	8.60 (1.79-41.45)	9.00 (3.07-26.38)	7.03 (2.91-16.97)
	P-value¶	<0.001	<0.001	0.01	<0.001	<0.001
Decreased risk	Unknown 10 (481.3495@7.157514)					
	OR (95% CI) *	0.08 (0.03-0.18)	0.12 (0.05-0.29)	0.004 (<0.001-0.08)	0.22 (0.08-0.61)	0.12 (0.04-0.32)
	P-value¶	<0.001	<0.001	<0.001	<0.001	<0.001
	Unknown 11 (124.0638@2.238123)					
	OR (95% CI) *	0.31 (19-0.51)	0.30 (0.17-0.55)	0.07 (0.01-0.40)	0.42 (0.21-0.83)	0.31 (0.15-0.66)
	P-value¶	<0.001	<0.001	0.003	0.01	0.002
	Unknown 14 (278.076@4.280106)					
	OR (95% CI) *	0.41 (0.29-0.58)	0.39 (0.26-0.57)	0.45 (0.22-0.92)	0.35 (0.21-0.57)	0.30 (0.19-0.47)
	P-value¶	<0.001	<0.001	0.03	<0.001	<0.001
	Unknown 15 (356.1964@5.496086)					
	OR (95% CI) *	0.31 (0.19-0.50)	0.27 (0.16-0.48)	0.27 (0.09-0.80)	0.24 (0.11-0.50)	0.29 (0.16-0.55)
	P-value¶	<0.001	<0.001	0.02	<0.001	<0.001
	Unknown 18 (793.9569@6.9292555)					
	OR (95% CI) *	0.26 (0.15-0.44)	0.28 (0.15-0.52)	0.03 (0.004-0.22)	0.48 (0.24-0.96)	0.31 (0.16-0.61)
	P-value¶	<0.001	<0.001	<0.001	0.04	<0.001
	Unknown 21 (202.1313@0.8667429)					
	OR (95% CI) *	0.22 (0.12-0.42)	0.20 (0.10-0.41)	0.21 (0.05-0.90)	0.14 (0.06-0.36)	0.20 (0.09-0.42)
	P-value¶	<0.001	<0.001	0.04	<0.001	<0.001

* ORs for 221 liver cancer cases and 221 matched controls compare the 90th to the 10th percentile of metabolite values based on the distribution in the controls; letting X_{90} and X_{10}, denote the 90th percentile and 10th percentile in controls, and β denoted the log(OR) from the conditional logistic regression model, the OR is $e^{\beta(X90-X10)}$

† Models adjusted for entry age (years), body mass index (kg/m^2), smoking intensity (cigarettes/day), smoking duration (years), alcohol intake (none, <11.6 g/day, ≥11.6 g/day, or missing), self-reported diabetes status (yes or no), education (≤ or > elementary education), and run order

‡ n=146 (73 cases; 73 matched controls); missing alcohol assigned to highest frequency category owing to unstable risk estimates

§ n=296 (148 cases; 148 matched controls)

|| Models additionally adjusted for coffee intake (none, <1, 1 to <2, 2 to <3, or ≥3 cups (8 oz) per day), fruit and vegetable intake (g/1000 kcal), red meat intake (g/1000 kcal), white meat intake (g/1000 kcal), processed meat intake (g/1000 kcal), fish intake (g/1000 kcal), saturated fat intake (g/1000 kcal), energy intake (kcal); subjects with missing FFQ data were grouped using an indicator variable for missing

¶ P-value for X^2 test obtained from conditional logistic regression model for a given metabolite (modeled on a continuous basis); all tests were two-sided.

Abbreviations: CI, confidence interval; OR, odds ratio; RT, retention time

Supplementary Table 3. Odds ratios and 95% confidence intervals for liver disease death comparing men in the 90th and 10th percentiles, based on the distribution in controls, for top unknown metabolites, using conditional logistic regression

Direction of association	**Metabolite No. (mass@RT)**	**Unadjusted model**	**Model 1 †**	**>0 to 10 years of follow-up (model 1) †‡**	**>10 years of follow-up (model 1) †§**	**Model 2 (diet adjusted) †‖**
Increased risk	Unknown 12 (202.0168@2.7641954)					
	OR (95% CI) *	4.39 (2.82-6.81)	3.72 (2.34-5.93)	3.82 (1.95-7.50)	3.94 (1.91-8.13)	3.69 (2.18-6.26)
	P-value¶	<0.001	<0.001	<0.001	<0.001	<0.001
	Unknown 13 (242.1067@4.0267076)					
	OR (95% CI) *	3.69 (2.40-5.66)	3.41 (2.14-5.42)	3.04 (1.58-5.86)	3.84 (1.90-7.78)	3.04 (1.82-5.07)
	P-value¶	<0.001	<0.001	<0.001	<0.001	<0.001
	Unknown 16 (388.2573@6.339673)					
	OR (95% CI) *	7.92 (3.83-16.37)	10.01 (4.33-23.14)	59.98 (9.64-372.98)	4.85 (1.60-14.73)	9.91 (4.02-24.40)
	P-value¶	<0.001	<0.001	<0.001	0.01	<0.001
	Unknown 17 (504.3029@6.8187194)					
	OR (95% CI) *	5.43 (3.38-8.74)	4.61 (2.75-7.74)	2.83 (1.42-5.65)	8.41 (3.41-20.73)	4.03 (2.30-7.06)
	P-value¶	<0.001	<0.001	0.003	<0.001	<0.001
	Unknown 19 (292.9962@0.6360238)					
	OR (95% CI) *	5.26 (2.89-9.59)	5.06 (2.60-9.83)	5.84 (2.14-15.91)	5.73 (2.10-15.64)	5.04 (2.46-10.35)
	P-value¶	<0.001	<0.001	<0.001	<0.001	<0.001
	Unknown 20 (162.0534@0.64166015)					
	OR (95% CI) *	5.51 (2.90-10.47)	6.79 (3.24-14.22)	7.26 (2.58-20.46)	9.59 (2.64-34.82)	7.47 (3.28-17.06)
	P-value¶	<0.001	<0.001	<0.001	<0.001	<0.001
Decreased risk	Unknown 10 (481.3495@7.157514)					
	OR (95% CI) *	0.16 (0.08-0.31)	0.19 (0.09-0.40)	0.24 (0.08-0.69)	0.14 (0.05-0.43)	0.22 (0.10-0.50)
	P-value¶	<0.001	<0.001	0.01	<0.001	<0.001
	Unknown 11 (124.0638@2.238123)					
	OR (95% CI) *	0.31 (0.20-0.48)	0.37 (0.23-0.58)	0.55 (0.31-0.98)	0.22 (0.10-0.50)	0.40 (0.22-0.72)
	P-value¶	<0.001	<0.001	0.04	<0.001	0.003
	Unknown 14 (278.076@4.280106)					
	OR (95% CI) *	0.47 (0.36-0.61)	0.50 (0.38-0.66)	0.48 (0.32-0.71)	0.44 (0.27-0.73)	0.50 (0.37-0.69)
	P-value¶	<0.001	<0.001	0.003	0.001	<0.001
	Unknown 15 (356.1964@5.496086)					
	OR (95% CI) *	0.26 (0.16-0.43)	0.32 (0.19-0.54)	0.36 (0.15-0.84)	0.32 (.16-0.65)	0.41 (0.23-0.74)
	P-value¶	<0.001	<0.001	0.02	0.002	0.003
	Unknown 18 (793.9569@6.9292555)					
	OR (95% CI) *	0.34 (0.22-0.53)	0.38 (0.24-0.61)	0.29 (0.15-0.59)	0.49 (0.23-1.05)	0.42 (0.25-0.71)
	P-value¶	<0.001	<0.001	<0.001	0.07	0.001
	Unknown 21 (202.1313@0.8667429)					
	OR (95% CI) *	0.32 (0.20-0.51)	0.33 (0.20-0.56)	0.41 (0.21-0.79)	0.28 (0.12-0.65)	0.40 (0.23-0.70)
	P-value¶	<0.001	<0.001	0.01	0.003	0.001

* ORs for 242 fatal liver disease cases and 242 matched controls compare the 90th to the 10th percentile of metabolite values based on the distribution in the controls; letting X_{90} and X_{10}, denote the 90th percentile and 10th percentile in controls, and β denoted the log(OR) from the conditional logistic regression model, the OR is $e^{\beta(X90-X10)}$

† Models adjusted for entry age (years), body mass index (kg/m^2), smoking intensity (cigarettes/day), smoking duration (years), alcohol intake (none, <11.6 g/day, ≥11.6 g/day, or missing), self-reported diabetes status (yes or no), education (≤ or > elementary education), and run order

‡ n=228 (114 cases; 114 matched controls)

§ n=256 (128 cases; 128 matched controls)

|| Models additionally adjusted for coffee intake (none, <1, 1 to <2, 2 to <3, or ≥3 cups (8 oz) per day), fruit and vegetable intake (g/1000 kcal), red meat intake (g/1000 kcal), white meat intake (g/1000 kcal), processed meat intake (g/1000 kcal), fish intake (g/1000 kcal), saturated fat intake (g/1000 kcal), energy intake (kcal); subjects with missing FFQ data were grouped using an indicator variable for missing.

¶ P-value for X^2 test obtained from conditional logistic regression model for a given metabolite (modeled on a continuous basis); all tests were two-sided.

Abbreviations: CI, confidence interval; OR, odds ratio; RT, retention time

Supplementary Table 4. Odds ratios and 95% confidence intervals for sensitivity analyses of incident liver cancer comparing men in the 90th and 10th percentiles, based on the distribution in controls, for top metabolites, using conditional logistic regression, and excluding those with: 1) a self-reported history of diabetes or a baseline fasting glucose level ≥126 mg/dl; 2) a seropositive hepatitis B or C test; or ICD9 code 155.2

		Exclusion Factor		
Chemical Class	**Metabolite**	**Diabetes †**	**HCV or HBV ‡**	**ICD9:155.2 §**
Alkaloid	Trigonelline			
	OR (95% CI) *	0.27 (0.08-0.87)	0.17 (0.05-0.56)	0.36 (0.19-0.66)
	P-value ‖	0.01	0.004	0.001
Amino Acid	Tyrosine			
	OR (95% CI) *	2.92 (0.85-10.02)	7.18 (1.94-26.61)	4.37 (2.12-9.00)
	P-value ‖	0.09	0.003	<0.001
Indoleamine	Serotonin			
	OR (95% CI) *	0.17 (0.05-0.56)	0.34 (0.14-0.83)	0.33 (0.18-0.58)
	P-value ‖	0.003	0.02	<0.001
Dipeptide	Leucyl-valine			
	OR (95% CI) *	0.12 (0.03-0.45)	0.13 (0.04-0.42)	0.24 (0.13-0.45)
	P-value ‖	0.002	<0.001	<0.001
Bile Acid	Glycochenodeoxycholic acid			
	OR (95% CI) *	2.08 (0.86-5.04)	3.70 (1.49-9.20)	4.29 (2.30-8.00)
	P-value ‖	0.11	0.01	<0.001
	Glycocholic acid			
	OR (95% CI) *	3.36 (1.19-9.55)	6.72 (2.20-20.60)	5.70 (2.88-11.28)
	P-value ‖	0.02	<0.001	<0.001
Glycerophospholipid	LysoPC(15:0)			
	OR (95% CI) *	0.14 (0.03-0.56)	0.22 (0.07-0.67)	0.17 (0.09-0.35)
	P-value ‖	0.01	0.01	<0.001
	LysoPC(P-16:0)			
	OR (95% CI) *	0.29 (0.07-1.21)	0.16 (0.05-0.55)	0.21 (0.10-0.42)
	P-value ‖	0.09	0.004	<0.001
	LysoPC(18:2)			
	OR (95% CI) *	0.27 (0.08-0.85)	0.08 (0.02-0.31)	0.24 (0.13-0.46)
	P-value ‖	0.02	<0.001	<0.001
Purine derivative	Hypoxanthine			
	OR (95% CI) *	0.16 (0.04-0.63)	0.30 (0.11-0.86)	0.17 (0.08-0.36)
	P-value ‖	0.01	0.03	<0.001

* ORs for liver cancer cases and matched controls compare the 90th to the 10th percentile of metabolite values based on the distribution in the controls; letting X_{90} and X_{10}, denote the 90th percentile and 10th percentile in controls, and β denoted the log(OR) from the conditional logistic regression model, the OR is $e^{\beta(X90-X10)}$. Models adjusted for entry age (years), body mass index (kg/m^2), smoking intensity (cigarettes/day), smoking duration (years), alcohol intake (none, <11.6 g/day, ≥11.6 g/day, or missing), self-reported diabetes status (yes or no; *if applicable*), education (≤ or > elementary education), and run order

† Analysis includes n=184 (92 cases; 92 matched controls) without a self-reported history of diabetes or a baseline fasting glucose level ≥126 mg/dl; models are additionally adjusted for Homeostatic Model Assessment of Insulin Resistance (HOMA-IR)

‡ Analysis includes n=184 (92 cases; 92 matched controls) without a positive hepatitis B or C test

§ Analysis includes n=408 (204 cases; 204 matched controls) without a liver cancer diagnosis code of 155.2 (not specified as primary or secondary liver cancer)

‖ P-value for X^2 test obtained from conditional logistic regression model for a given metabolite (modeled on a continuous basis); all tests were two-sided.

Abbreviations: CI, confidence interval; ICD, International Classification of Diseases; OR, odds ratio; RT, retention time

Supplementary Table 5. Odds ratios and 95% confidence intervals for sensitivity analyses of incident liver disease mortality comparing men in the 90th and 10th percentiles, based on the distribution in controls, for top metabolites, using conditional logistic regression, and excluding those with: (1) a self-reported history of diabetes or a baseline fasting glucose level ≥126 mg/dl; (2) or a seropositive hepatitis B or C test

		Exclusion Factor	
Chemical Class	**Metabolite**	**Diabetes †**	**HCV or HBV ‡**
Alkaloid	Trigonelline		
	OR (95% CI) *	0.17 (0.07-0.39)	0.23 (0.11-0.48)
	P-value	0.001	<0.001
Amino Acid	Tyrosine		
	OR (95% CI) *	2.98 (1.25-7.10)	4.71 (2.26-9.84)
	P-value	0.003	<0.001
Indoleamine	Serotonin		
	OR (95% CI) *	0.31 (0.16-0.61)	0.35 (0.19-0.65)
	P-value	<0.001	<0.001
Dipeptide	Leucyl-valine		
	OR (95% CI) *	0.29 (0.15-0.57)	0.34 (0.20-0.59)
	P-value	<0.001	<0.001
Bile Acid	Glycochenodeoxycholic acid		
	OR (95% CI) *	4.83 (2.16-10.82)	7.25 (3.32-15.82)
	P-value	<0.001	<0.001
	Glycocholic acid		
	OR (95% CI) *	2.79 (1.54-5.00)	5.10 (2.63-9.90)
	P-value	<0.001	<0.001
Glycerophospholipid	LysoPC(15:0)		
	OR (95% CI) *	0.22 (0.10-0.49)	0.24 (0.12-0.48)
	P-value	<0.001	<0.001
	LysoPC(P-16:0)		
	OR (95% CI) *	0.44 (0.19-1.01)	0.32 (0.16-0.65)
	P-value	0.05	0.002
	LysoPC(18:2)		
	OR (95% CI) *	0.15 (0.05-0.39)	0.16 (0.07-0.35)
	P-value	<0.001	<0.001
Purine derivative	Hypoxanthine		
	OR (95% CI) *	0.38 (0.18-0.77)	0.38 (0.21-0.69)
	P-value	0.01	0.001

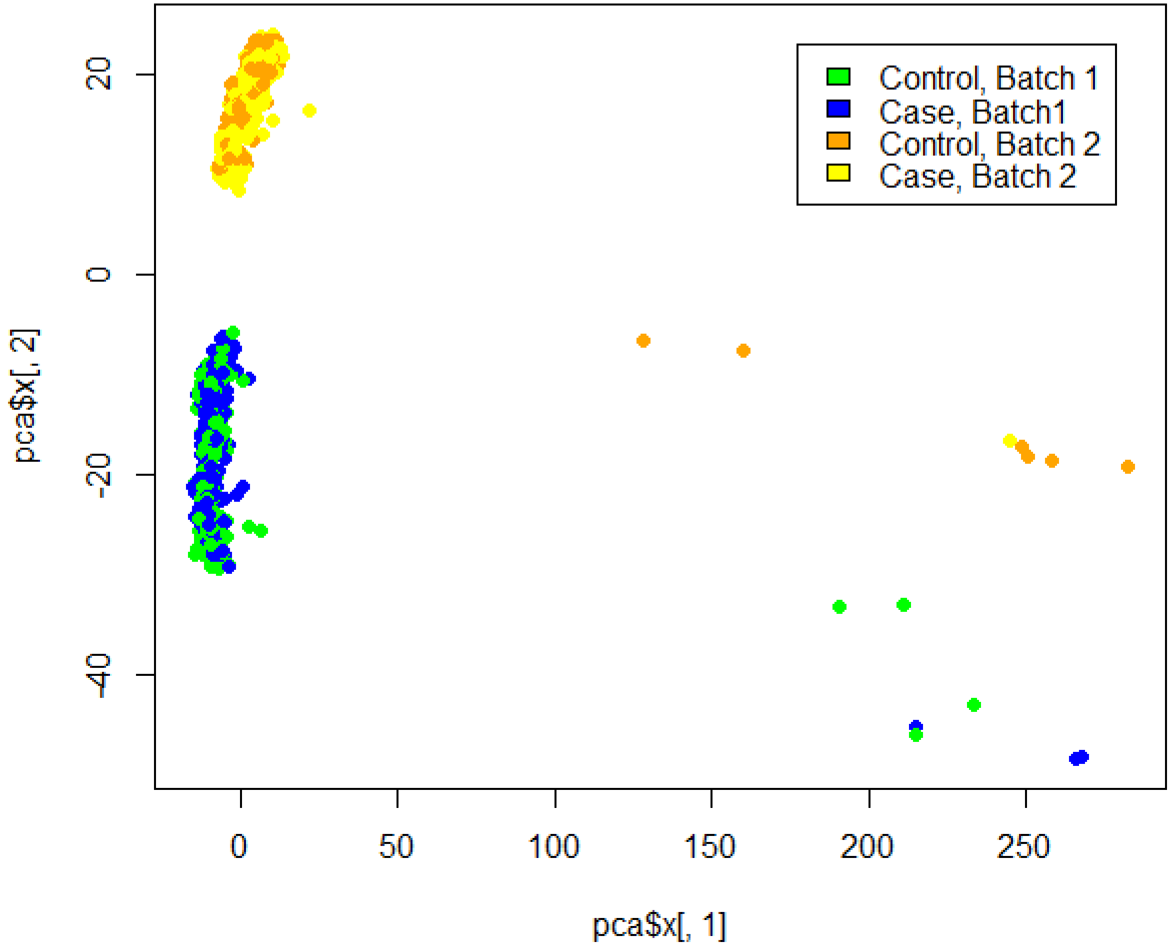
* ORs for fatal liver disease cases and matched controls compare the 90th to the 10th percentile of metabolite values based on the distribution in the controls; letting X_{90} and X_{10}, denote the 90th percentile and 10th percentile in controls, and β denoted the log(OR) from the conditional logistic regression model, the OR is $e^{\beta(X90-X10).}$ Models are adjusted for entry age (years), body mass index (kg/m^2), smoking intensity (cigarettes/day), smoking duration (years), alcohol intake (none, <11.6 g/day, ≥11.6 g/day, or missing), self-reported diabetes status (yes or no; *if applicable*), education (≤ or > elementary education), and run order

† Analysis includes n=330 (165 cases; 165 matched controls) without a self-reported history of diabetes or a baseline fasting glucose level ≥126 mg/dl; models are additionally adjusted for Homeostatic Model Assessment of Insulin Resistance (HOMA-IR)

‡ Analysis includes n=360 (180 cases; 180 matched controls) without a positive hepatitis B or C test

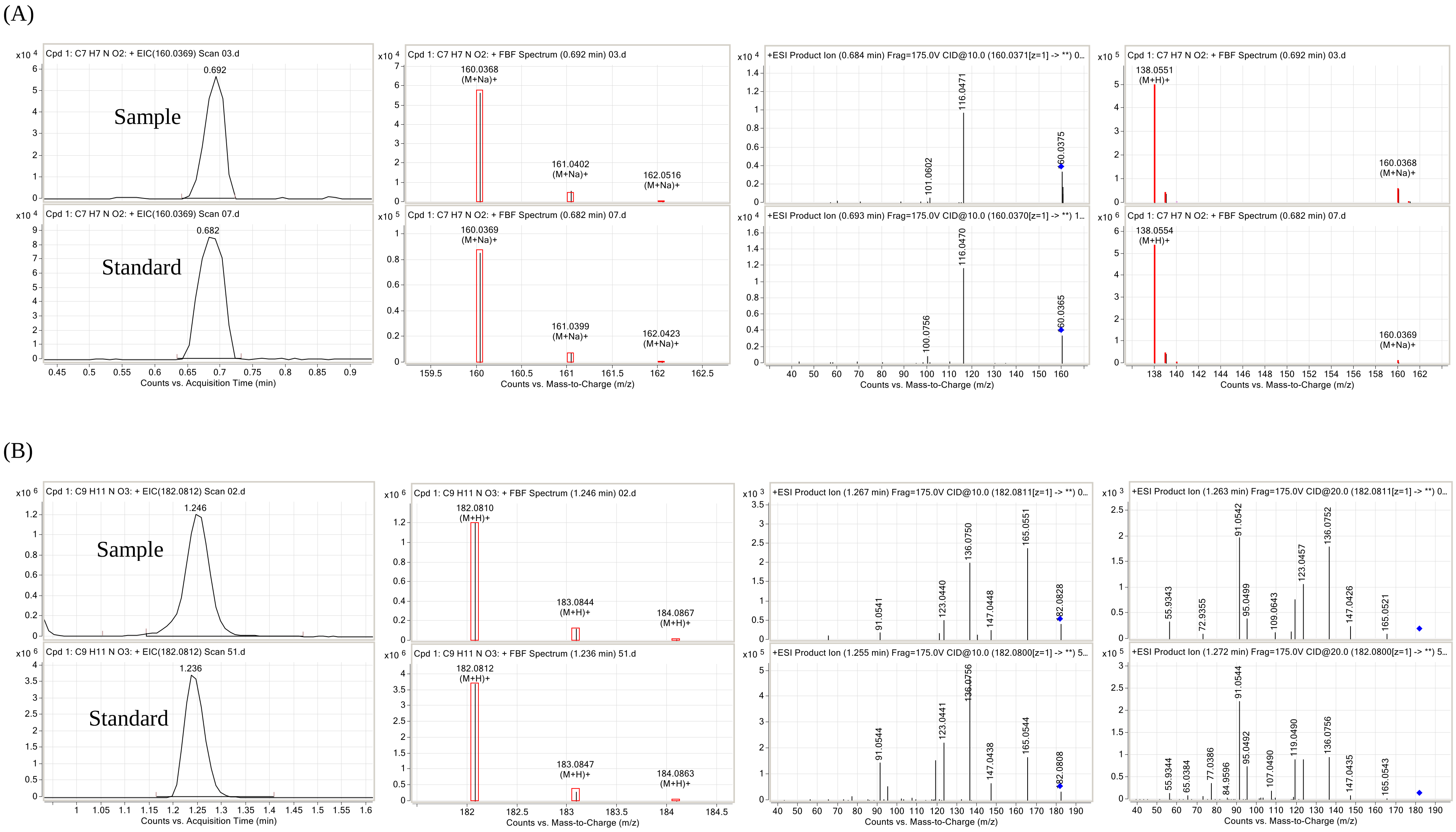
§ P-value for X^2 test obtained from conditional logistic regression model for a given metabolite (modeled on a continuous basis); all tests were two-sided.

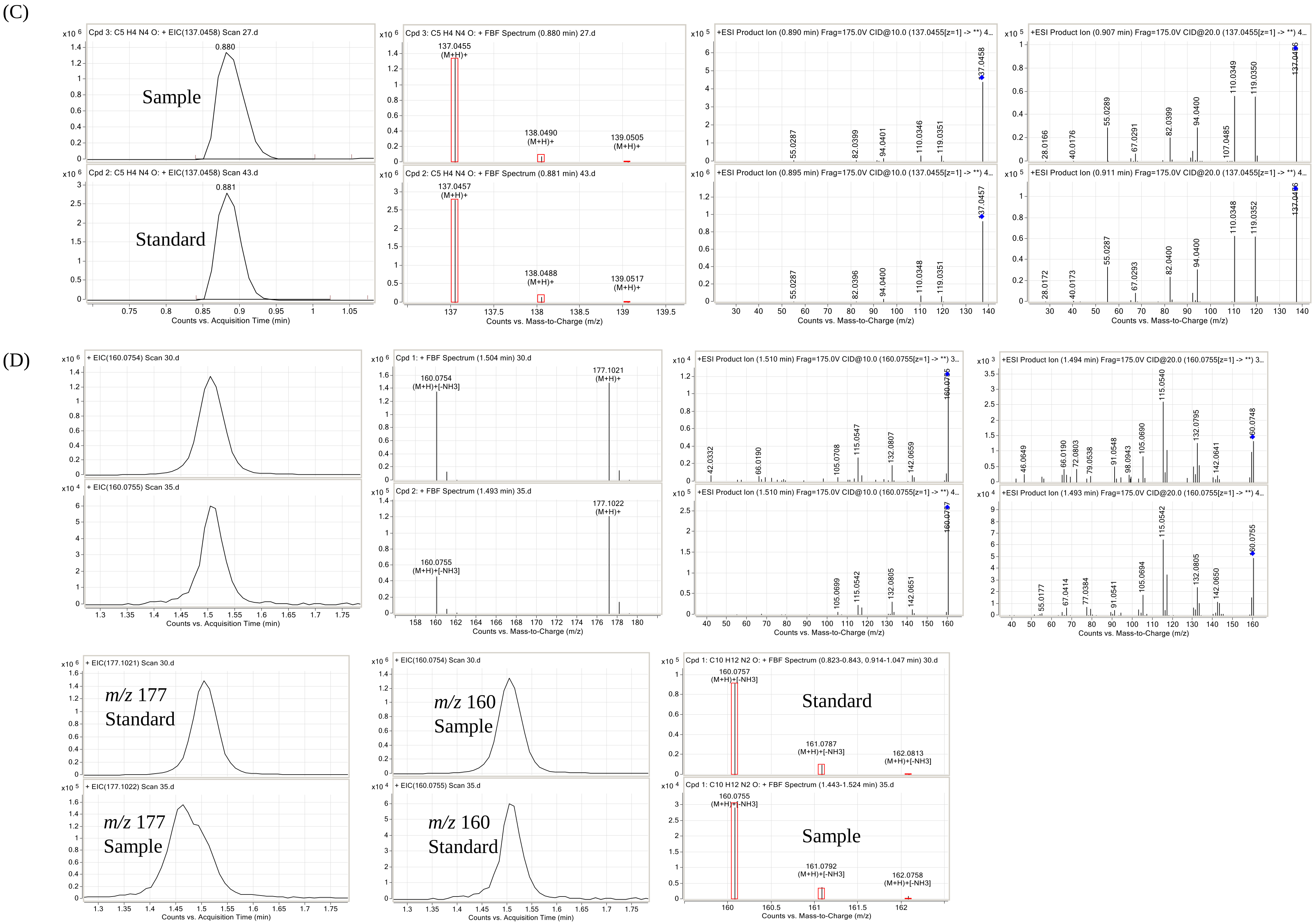
Abbreviations: CI, confidence interval; HBV, hepatitis B virus; HCV, hepatitis C virus; OR, odds ratio; RT, retention time



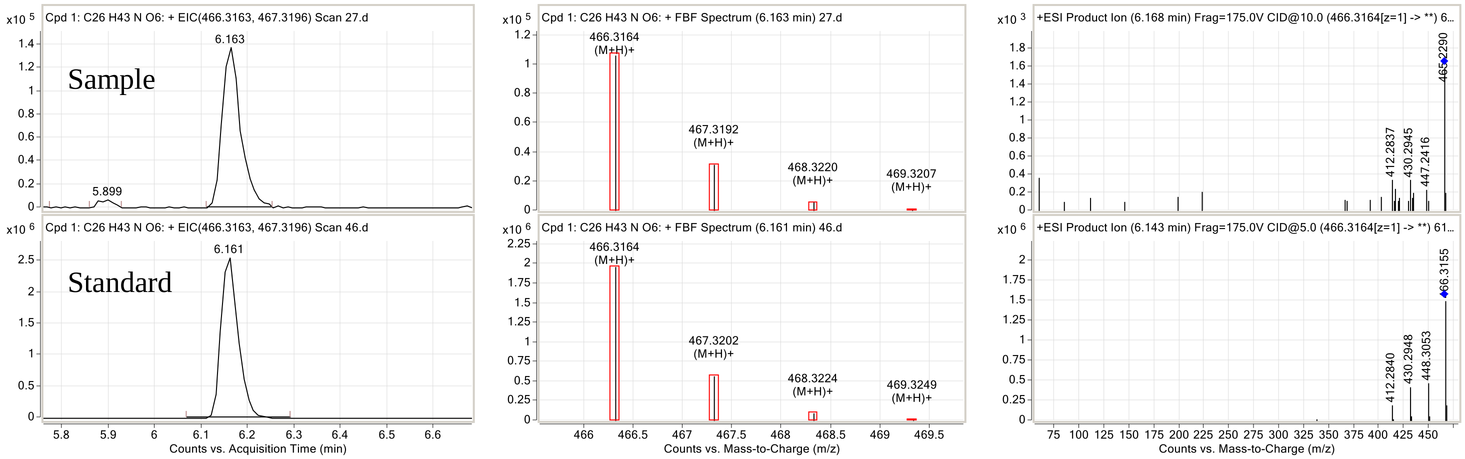
Supplementary Figure 1. PCA of 2,879 spectral features to identify outlying observations.

The PCA plot shows 14 observations that were excluded as extreme outliers (PC1>100) by case status and batch. Controls in batch 1 and 2 are highlighted in green and orange, respectively; cases in batch 1 and 2 are highlighted in blue and yellow, respectively. Abbreviations: PCA, principle components analysis.

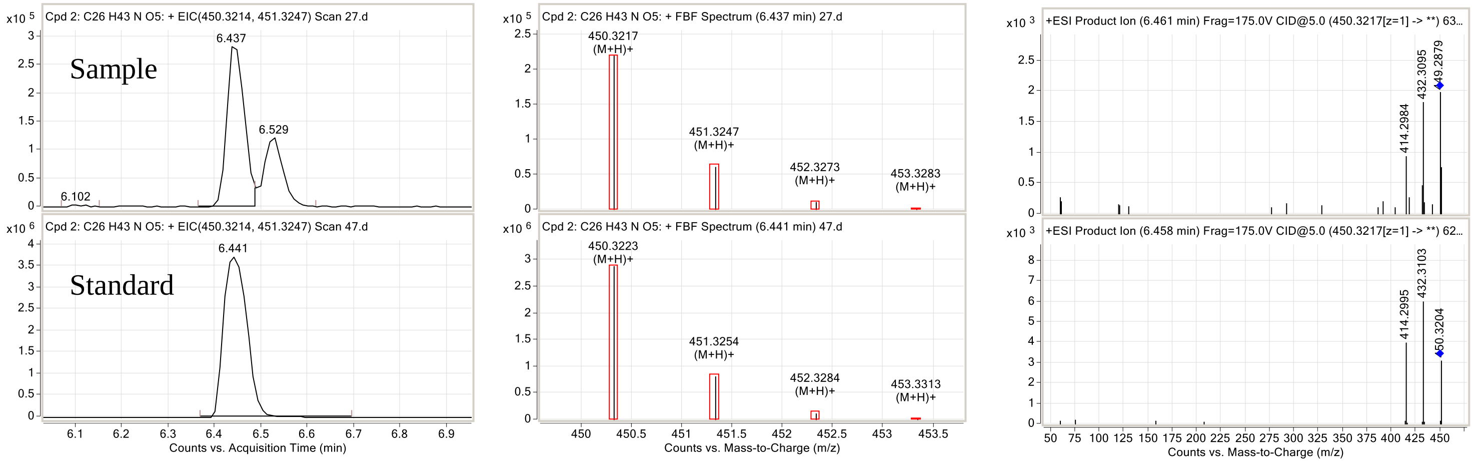

(A)
Cpd 1: C7 H7 N O2: + EIC(160.0369) Scan 03.d
0.692
Sample
Cpd 1: C7 H7 N O2: + EIC(160.0369) Scan 07.d
0.682
Standard
Counts vs. Acquisition Time (min)
Cpd 1: C7 H7 N O2: + FBF Spectrum (0.692 min) 03.d
160.0368 (M+Na)+
161.0402 (M+Na)+
162.0516 (M+Na)+
Cpd 1: C7 H7 N O2: + FBF Spectrum (0.682 min) 07.d
160.0369 (M+Na)+
161.0399 (M+Na)+
162.0423 (M+Na)+
Counts vs. Mass-to-Charge (m/z)
+ESI Product Ion (0.684 min) Frag=175.0V CID@10.0 (160.0371[z=1] -> **) 0...
116.0471
101.0602
160.0375
+ESI Product Ion (0.693 min) Frag=175.0V CID@10.0 (160.0370[z=1] -> **) 1...
116.0470
100.0756
160.0365
Counts vs. Mass-to-Charge (m/z)
Cpd 1: C7 H7 N O2: + FBF Spectrum (0.692 min) 03.d
138.0551 (M+H)+
160.0368 (M+Na)+
Cpd 1: C7 H7 N O2: + FBF Spectrum (0.682 min) 07.d
138.0554 (M+H)+
160.0369 (M+Na)+
Counts vs. Mass-to-Charge (m/z)
(B)
Cpd 1: C9 H11 N O3: + EIC(182.0812) Scan 02.d
1.246
Sample
Cpd 1: C9 H11 N O3: + EIC(182.0812) Scan 51.d
1.236
Standard
Counts vs. Acquisition Time (min)
Cpd 1: C9 H11 N O3: + FBF Spectrum (1.246 min) 02.d
182.0810 (M+H)+
183.0844 (M+H)+
184.0867 (M+H)+
Cpd 1: C9 H11 N O3: + FBF Spectrum (1.236 min) 51.d
182.0812 (M+H)+
183.0847 (M+H)+
184.0863 (M+H)+
Counts vs. Mass-to-Charge (m/z)
+ESI Product Ion (1.267 min) Frag=175.0V CID@10.0 (182.0811[z=1] -> **) 0...
91.0541
123.0440
136.0750
147.0448
165.0551
182.0828
+ESI Product Ion (1.255 min) Frag=175.0V CID@10.0 (182.0800[z=1] -> **) 5...
91.0544
123.0441
136.0756
147.0438
165.0544
182.0808
Counts vs. Mass-to-Charge (m/z)
+ESI Product Ion (1.263 min) Frag=175.0V CID@20.0 (182.0811[z=1] -> **) 0...
55.9343
72.9355
91.0542
95.0499
109.0643
123.0457
136.0752
147.0426
165.0521
+ESI Product Ion (1.272 min) Frag=175.0V CID@20.0 (182.0800[z=1] -> **) 5...
55.9344
65.0384
77.0386
84.9596
91.0544
95.0492
107.0490
119.0490
136.0756
147.0435
165.0543
Counts vs. Mass-to-Charge (m/z)

(C)
x10 6 Cpd 3: C5 H4 N4 O: + EIC(137.0458) Scan 27.d
0.880
Sample
x10 6 Cpd 2: C5 H4 N4 O: + EIC(137.0458) Scan 43.d
0.881
Standard
0.75 0.8 0.85 0.9 0.95 1 1.05
Counts vs. Acquisition Time (min)
x10 6 Cpd 3: C5 H4 N4 O: + FBF Spectrum (0.880 min) 27.d
137.0455 (M+H)+
138.0490 (M+H)+
139.0505 (M+H)+
x10 6 Cpd 2: C5 H4 N4 O: + FBF Spectrum (0.881 min) 43.d
137.0457 (M+H)+
138.0488 (M+H)+
139.0517 (M+H)+
137 137.5 138 138.5 139 139.5
Counts vs. Mass-to-Charge (m/z)
x10 5 +ESI Product Ion (0.890 min) Frag=175.0V CID@10.0 (137.0455[z=1] -> **) 4...
55.0287 82.0399 94.0401 110.0346 119.0351 137.0458
x10 6 +ESI Product Ion (0.895 min) Frag=175.0V CID@10.0 (137.0455[z=1] -> **) 4...
55.0287 82.0396 94.0400 110.0348 119.0351 137.0457
30 40 50 60 70 80 90 100 110 120 130 140
Counts vs. Mass-to-Charge (m/z)
x10 5 +ESI Product Ion (0.907 min) Frag=175.0V CID@20.0 (137.0455[z=1] -> **) 4...
28.0166 40.0176 55.0289 67.0291 82.0399 94.0400 107.0485 110.0349 119.0350 137.0456
x10 5 +ESI Product Ion (0.911 min) Frag=175.0V CID@20.0 (137.0455[z=1] -> **) 4...
28.0172 40.0173 55.0287 67.0293 82.0400 94.0400 110.0348 119.0352 137.0456
30 40 50 60 70 80 90 100 110 120 130 140
Counts vs. Mass-to-Charge (m/z)
(D)
x10 6 + EIC(160.0754) Scan 30.d
x10 4 + EIC(160.0755) Scan 35.d
1.3 1.35 1.4 1.45 1.5 1.55 1.6 1.65 1.7 1.75
Counts vs. Acquisition Time (min)
x10 6 Cpd 1: + FBF Spectrum (1.504 min) 30.d
160.0754 (M+H)+[-NH3]
177.1021 (M+H)+
x10 5 Cpd 2: + FBF Spectrum (1.493 min) 35.d
160.0755 (M+H)+[-NH3]
177.1022 (M+H)+
158 160 162 164 166 168 170 172 174 176 178 180
Counts vs. Mass-to-Charge (m/z)
x10 4 +ESI Product Ion (1.510 min) Frag=175.0V CID@10.0 (160.0755[z=1] -> **) 3...
42.0332 66.0190 105.0708 115.0547 132.0807 142.0659 160.0755
x10 5 +ESI Product Ion (1.510 min) Frag=175.0V CID@10.0 (160.0755[z=1] -> **) 4...
105.0699 115.0542 132.0805 142.0651 160.0757
40 50 60 70 80 90 100 110 120 130 140 150 160
Counts vs. Mass-to-Charge (m/z)
x10 3 +ESI Product Ion (1.494 min) Frag=175.0V CID@20.0 (160.0755[z=1] -> **) 3...
46.0649 66.0190 72.0803 79.0538 91.0548 98.0943 105.0690 115.0540 132.0795 142.0641 160.0748
x10 4 +ESI Product Ion (1.493 min) Frag=175.0V CID@20.0 (160.0755[z=1] -> **) 4...
55.0177 67.0414 77.0384 91.0541 105.0694 115.0542 132.0805 142.0650 160.0755
40 50 60 70 80 90 100 110 120 130 140 150 160
Counts vs. Mass-to-Charge (m/z)
x10 6 + EIC(177.1021) Scan 30.d
m/z 177 Standard
x10 5 + EIC(177.1022) Scan 35.d
m/z 177 Sample
1.3 1.35 1.4 1.45 1.5 1.55 1.6 1.65 1.7 1.75
Counts vs. Acquisition Time (min)
x10 6 + EIC(160.0754) Scan 30.d
m/z 160 Sample
x10 4 + EIC(160.0755) Scan 35.d
m/z 160 Standard
1.3 1.35 1.4 1.45 1.5 1.55 1.6 1.65 1.7 1.75
Counts vs. Acquisition Time (min)
x10 5 Cpd 1: C10 H12 N2 O: + FBF Spectrum (0.823-0.843, 0.914-1.047 min) 30.d
160.0757 (M+H)+[-NH3]
161.0787 (M+H)+[-NH3]
162.0813 (M+H)+[-NH3]
Standard
x10 4 Cpd 1: C10 H12 N2 O: + FBF Spectrum (1.443-1.524 min) 35.d
160.0755 (M+H)+[-NH3]
161.0792 (M+H)+[-NH3]
162.0758 (M+H)+[-NH3]
Sample
160 160.5 161 161.5 162
Counts vs. Mass-to-Charge (m/z)

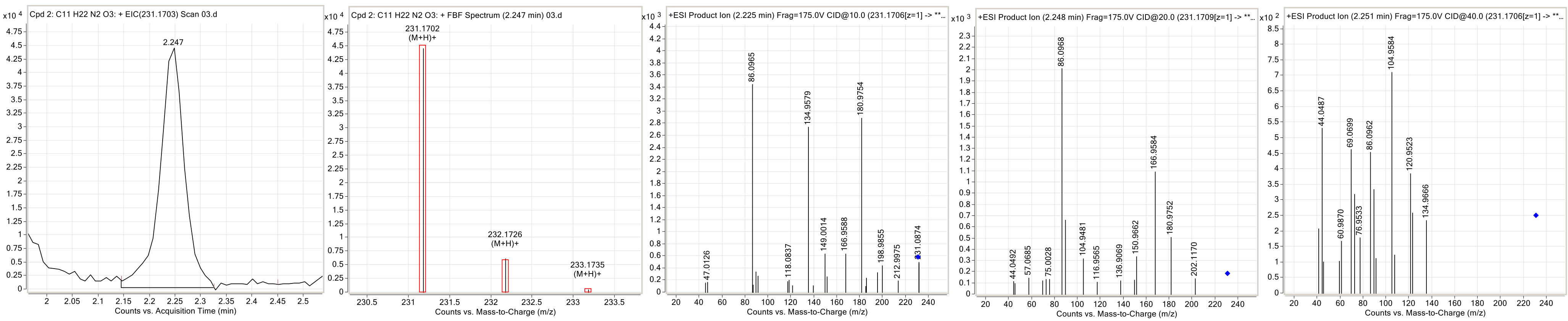
(E)


x10 5
Cpd 1: C26 H43 N O6: + EIC(466.3163, 467.3196) Scan 27.d
Sample
6.163
5.899
1.4
1.2
1
0.8
0.6
0.4
0.2
0
x10 6
Cpd 1: C26 H43 N O6: + EIC(466.3163, 467.3196) Scan 46.d
Standard
6.161
2.5
2
1.5
1
0.5
0
5.8
5.9
6
6.1
6.2
6.3
6.4
6.5
6.6
Counts vs. Acquisition Time (min)
x10 5
Cpd 1: C26 H43 N O6: + FBF Spectrum (6.163 min) 27.d
466.3164
(M+H)+
467.3192
(M+H)+
468.3220
(M+H)+
469.3207
(M+H)+
1.2
1
0.8
0.6
0.4
0.2
0
x10 6
Cpd 1: C26 H43 N O6: + FBF Spectrum (6.161 min) 46.d
466.3164
(M+H)+
467.3202
(M+H)+
468.3224
(M+H)+
469.3249
(M+H)+
2.25
2
1.75
1.5
1.25
1
0.75
0.5
0.25
0
466
466.5
467
467.5
468
468.5
469
469.5
Counts vs. Mass-to-Charge (m/z)
x10 3
+ESI Product Ion (6.168 min) Frag=175.0V CID@10.0 (466.3164[z=1] -> **) 6...
465.2290
412.2837
430.2945
447.2416
1.8
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
x10 6
+ESI Product Ion (6.143 min) Frag=175.0V CID@5.0 (466.3164[z=1] -> **) 61...
466.3155
412.2840
430.2948
448.3053
2
1.75
1.5
1.25
1
0.75
0.5
0.25
0
75
100
125
150
175
200
225
250
275
300
325
350
375
400
425
450
Counts vs. Mass-to-Charge (m/z)

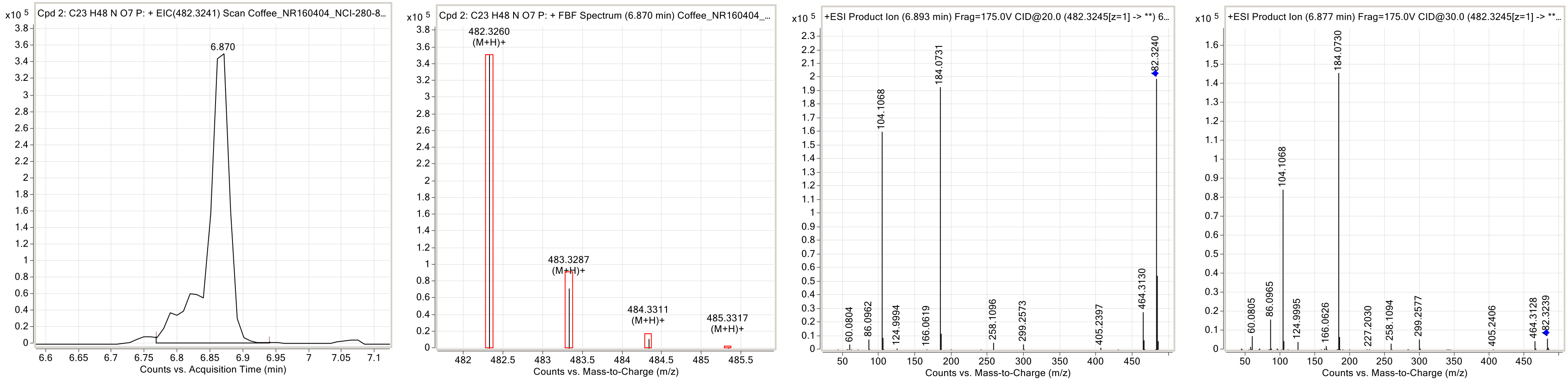
(F)

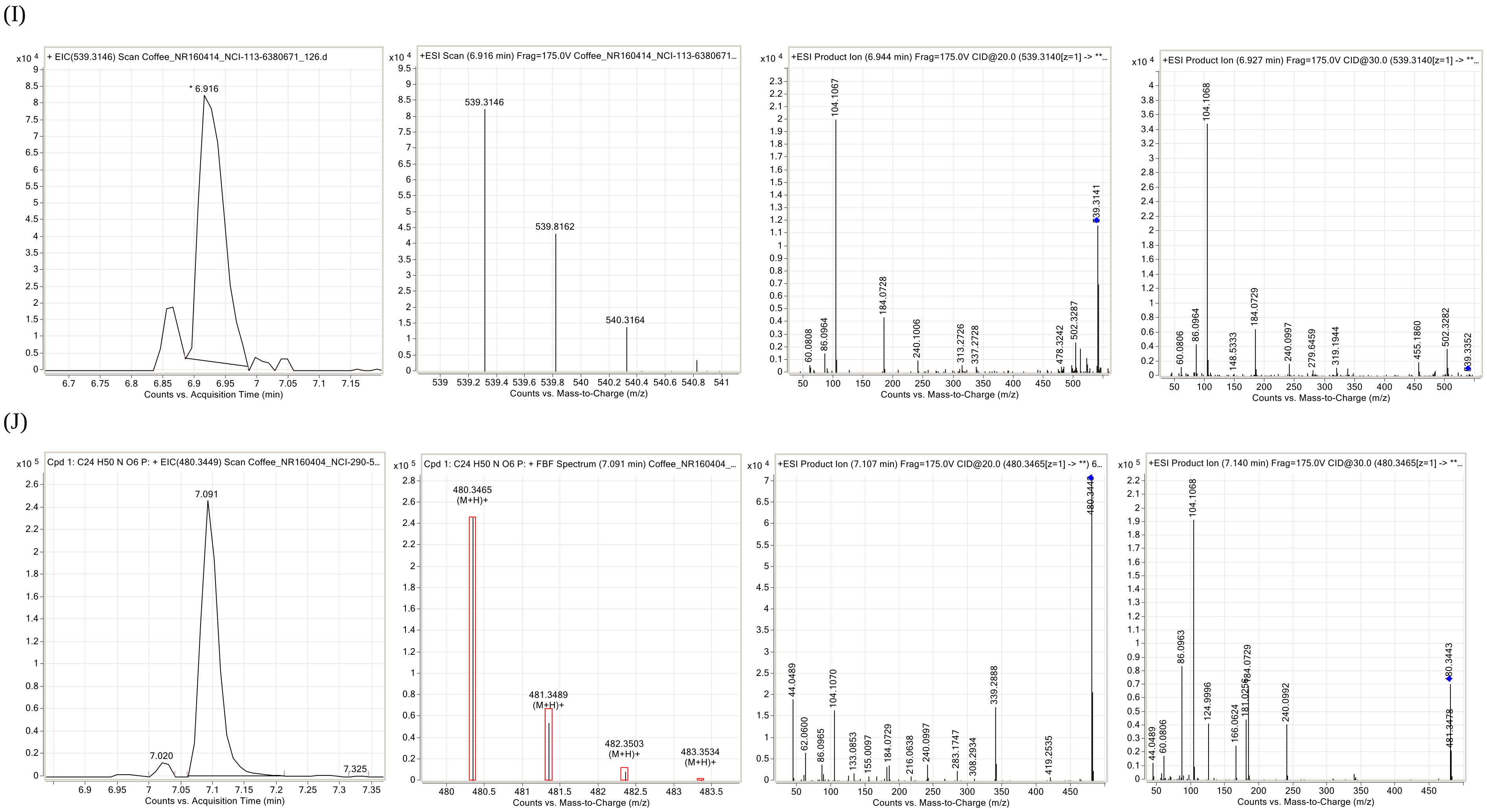

x10 5
Cpd 2: C26 H43 N O5: + EIC(450.3214, 451.3247) Scan 27.d
Sample
6.437
6.529
6.102
3
2.5
2
1.5
1
0.5
0
x10 6
Cpd 2: C26 H43 N O5: + EIC(450.3214, 451.3247) Scan 47.d
Standard
6.441
4
3.5
3
2.5
2
1.5
1
0.5
0
6.1
6.2
6.3
6.4
6.5
6.6
6.7
6.8
6.9
Counts vs. Acquisition Time (min)
x10 5
Cpd 2: C26 H43 N O5: + FBF Spectrum (6.437 min) 27.d
450.3217
(M+H)+
451.3247
(M+H)+
452.3273
(M+H)+
453.3283
(M+H)+
2.5
2
1.5
1
0.5
0
x10 6
Cpd 2: C26 H43 N O5: + FBF Spectrum (6.441 min) 47.d
450.3223
(M+H)+
451.3254
(M+H)+
452.3284
(M+H)+
453.3313
(M+H)+
3
2.5
2
1.5
1
0.5
0
450
450.5
451
451.5
452
452.5
453
453.5
Counts vs. Mass-to-Charge (m/z)
x10 3
+ESI Product Ion (6.461 min) Frag=175.0V CID@5.0 (450.3217[z=1] -> **) 63...
449.2879
432.3095
414.2984
2.5
2
1.5
1
0.5
0
x10 3
+ESI Product Ion (6.458 min) Frag=175.0V CID@5.0 (450.3217[z=1] -> **) 62...
432.3103
414.2995
450.3204
8
7
6
5
4
3
2
1
0
50
75
100
125
150
175
200
225
250
275
300
325
350
375
400
425
450
Counts vs. Mass-to-Charge (m/z)

(G)


Cpd 2: C11 H22 N2 O3: + EIC(231.1703) Scan 03.d
x10 4
2.247
Counts vs. Acquisition Time (min)
Cpd 2: C11 H22 N2 O3: + FBF Spectrum (2.247 min) 03.d
x10 4
231.1702 (M+H)+
232.1726 (M+H)+
233.1735 (M+H)+
Counts vs. Mass-to-Charge (m/z)
+ESI Product Ion (2.225 min) Frag=175.0V CID@10.0 (231.1706[z=1] -> **...
x10 3
47.0126
86.0965
118.0837
134.9579
149.0014
166.9588
180.9754
198.9855
212.9975
231.0874
Counts vs. Mass-to-Charge (m/z)
+ESI Product Ion (2.248 min) Frag=175.0V CID@20.0 (231.1709[z=1] -> **...
x10 3
44.0492
57.0685
75.0028
86.0968
104.9481
116.9565
136.9069
150.9662
166.9584
180.9752
202.1170
Counts vs. Mass-to-Charge (m/z)
+ESI Product Ion (2.251 min) Frag=175.0V CID@40.0 (231.1706[z=1] -> **...
x10 2
44.0487
60.9870
69.0699
76.9533
86.0962
104.9584
120.9523
134.9666
Counts vs. Mass-to-Charge (m/z)

(H)

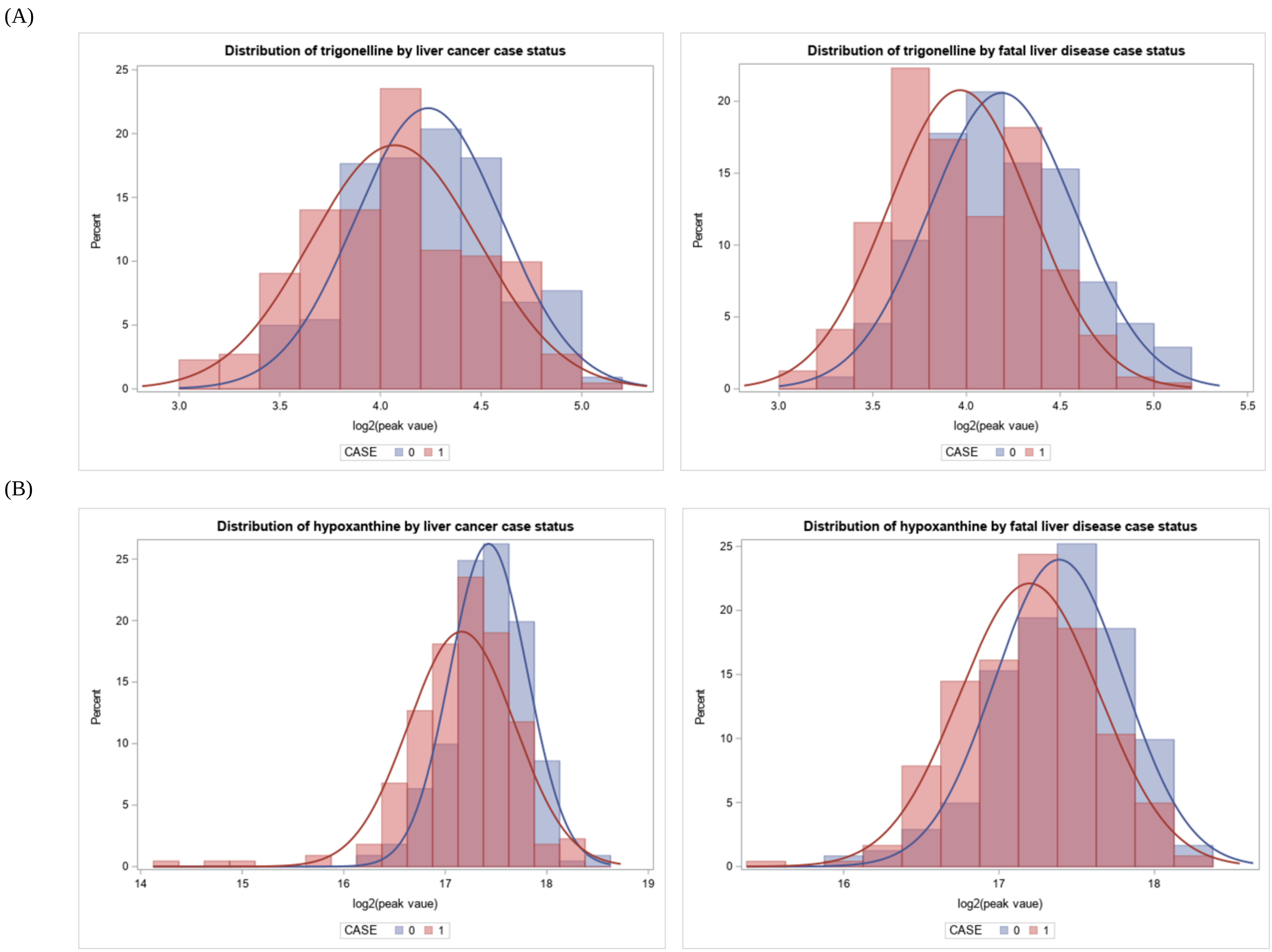

Cpd 2: C23 H48 N O7 P: + EIC(482.3241) Scan Coffee_NR160404_NCI-280-8...
x10 5
6.870
Counts vs. Acquisition Time (min)
Cpd 2: C23 H48 N O7 P: + FBF Spectrum (6.870 min) Coffee_NR160404_...
x10 5
482.3260 (M+H)+
483.3287 (M+H)+
484.3311 (M+H)+
485.3317 (M+H)+
Counts vs. Mass-to-Charge (m/z)
+ESI Product Ion (6.893 min) Frag=175.0V CID@20.0 (482.3245[z=1] -> **) 6...
x10 5
60.0804
86.0962
104.1068
124.9994
166.0619
184.0731
258.1096
299.2573
405.2397
464.3130
482.3240
Counts vs. Mass-to-Charge (m/z)
+ESI Product Ion (6.877 min) Frag=175.0V CID@30.0 (482.3245[z=1] -> **...
x10 5
60.0805
86.0965
104.1068
124.9995
166.0626
184.0730
227.2030
258.1094
299.2577
405.2406
464.3128
482.3239
Counts vs. Mass-to-Charge (m/z)



Supplementary Figure 2. Chromatograms and spectra from representative study samples and pure chemical standards from IROA Technologies' Mass Spectrometry Metabolite Library (when available).

Chromatograms and isotope patterns were generated with Find Compounds by Formula in Agilent MassHunter Qualitative Analysis B.06.00 SP1. Isotope patterns, highlighted with red rectangles, represent an isotope pattern calculated from the elemental composition (shown in the title of each spectra), and bars inside the rectangles show the observed isotope peaks. For MS/MS spectra, the precursor ion is indicated with a blue dot above the ion, with collision energy labeled on top of the spectra. **Panels A-J** correspond to the following compounds: **(A)** Trigonelline (HMDB0000875, trigonelline was detected as [M+Na]+ (m/z 160.037). Chromatographic specificity was confirmed by identification with a pure chemical standard and by ensuring separation from the major potentially interfering compounds of identical elemental composition p-Aminobenzoic acid (RT: 2.33 min), anthranilic acid (RT: 3.52 min), and salicylamide (RT: 3.45 min)); **(B)** L-

Tyrosine (HMDB0000158, L-Tyrosine appeared as 2 strongly correlated (Pearson's r: 0.93) chromatographic peaks); **(C)** Hypoxanthine (HMDB0000157); **(D)** Serotonin (HMDB0000259, serotonin was detected as an in-source fragment $[M-NH_3+H]^+$ (*m/z* 160). The $[M+H]^+$ ion of serotonin (*m/z* 177) was not chromatographically separated from an interfering ion of identical mass but produced fragments specific to serotonin. The specificity of the ion *m/z* 160 was confirmed by comparing its retention time and MS/MS fragments against those from a pure serotonin standard.); **(E)** Glycocholic acid (HMDB0000138); **(F)** Glycochenodeoxycholic acid (HMDB0000637); **(G)** Leu-Val or isomer; **(H)** LysoPC(15:0) (HMDB0010381); **(I)** LysoPC(18:2) (Several ions were detected, with [2M+H+K]2+ (*m/z* 539.3146) being the most intense, from which MS/MS spectra were acquired.); **(J)** LysoPC(P-16:0) (HMDB0010407).



(A)

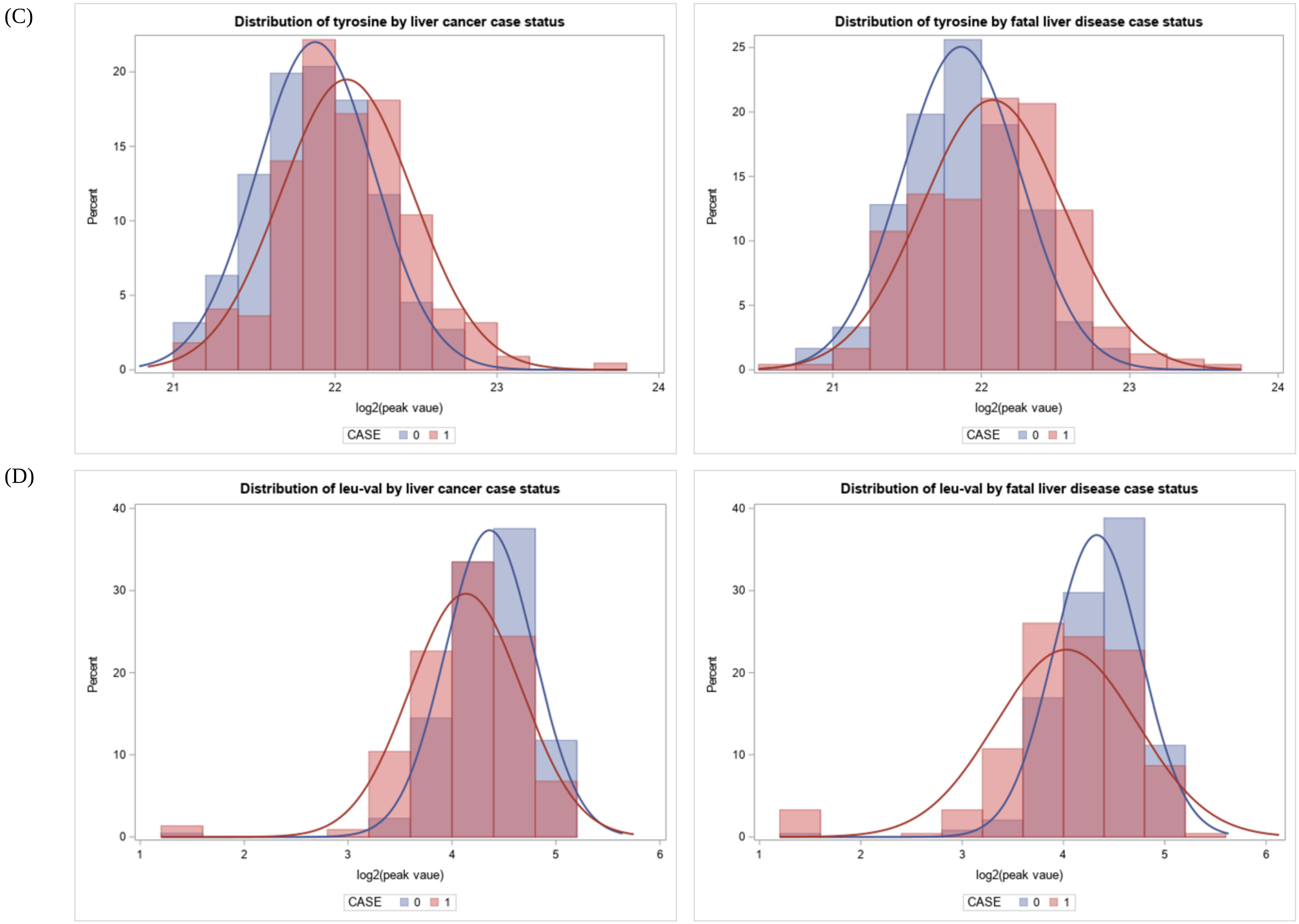
Distribution of trigonelline by liver cancer case status

Distribution of trigonelline by fatal liver disease case status

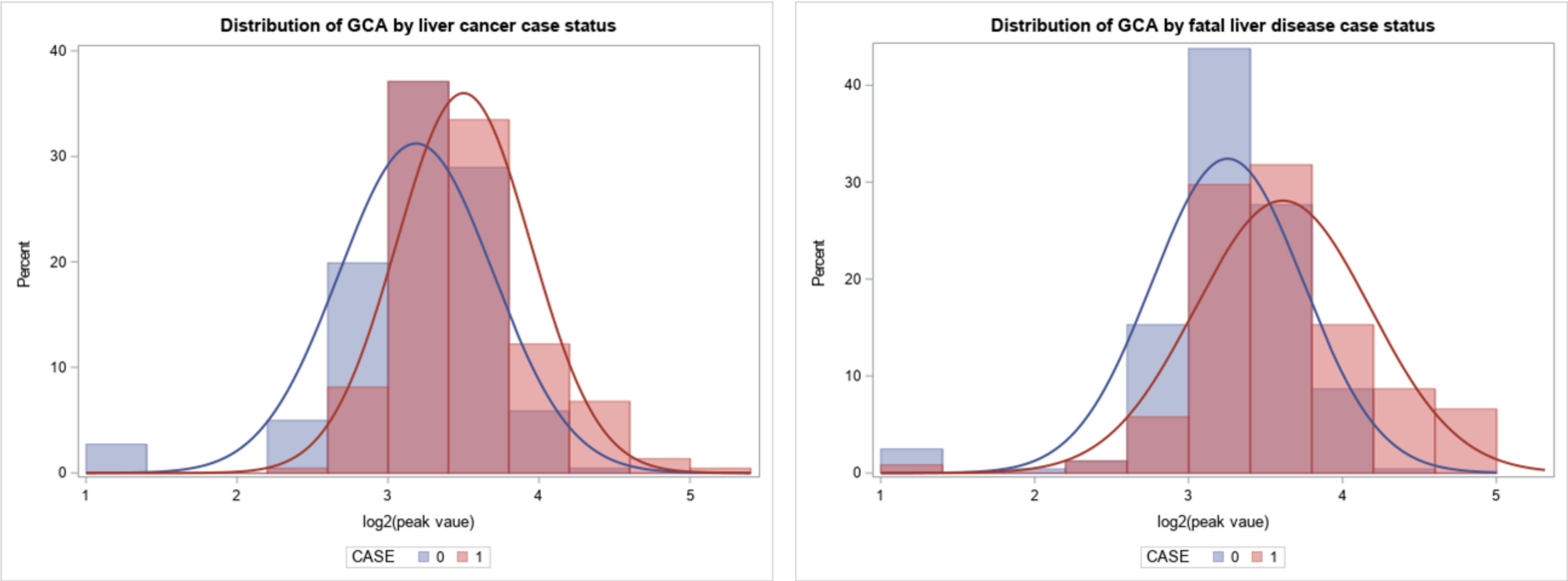
(B)

Distribution of hypoxanthine by liver cancer case status

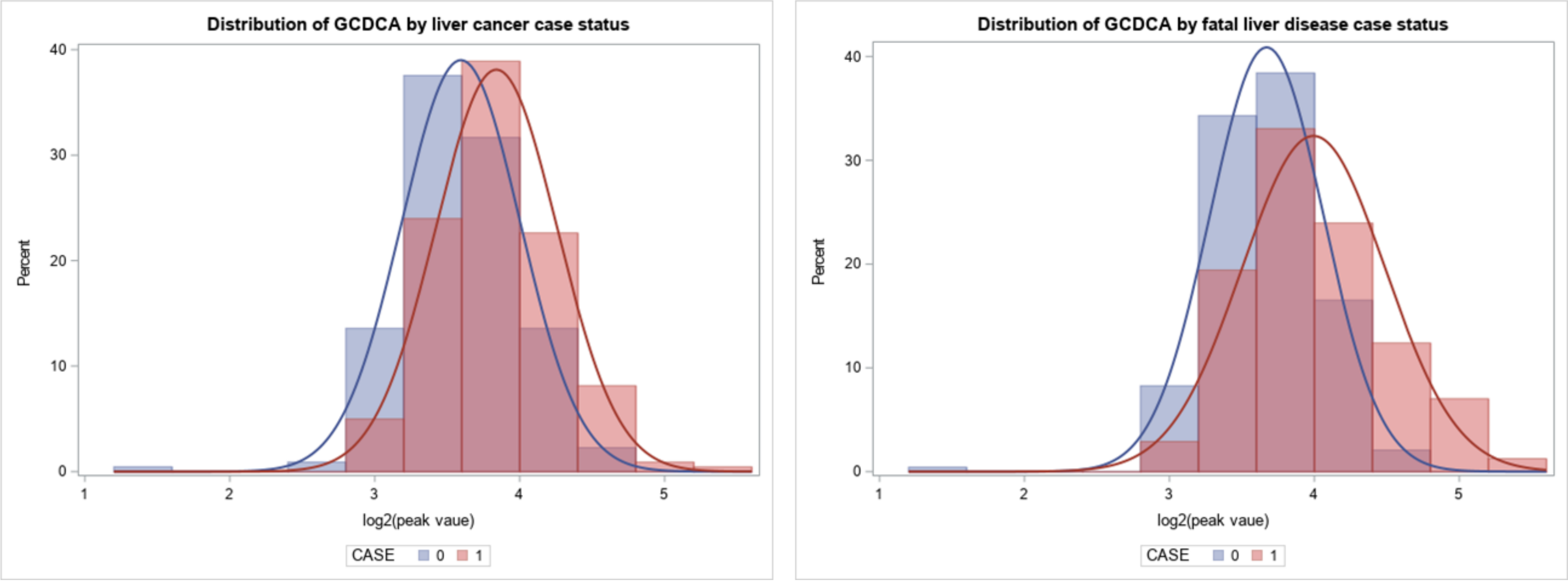
Distribution of hypoxanthine by fatal liver disease case status


(C)
Distribution of tyrosine by liver cancer case status
Percent
0
5
10
15
20
21
22
23
24
log2(peak vaue)
CASE ▨ 0 ▨ 1
Distribution of tyrosine by fatal liver disease case status
Percent
0
5
10
15
20
25
21
22
23
24
log2(peak vaue)
CASE ▨ 0 ▨ 1
(D)
Distribution of leu-val by liver cancer case status
Percent
0
10
20
30
40
1
2
3
4
5
6
log2(peak vaue)
CASE ▨ 0 ▨ 1
Distribution of leu-val by fatal liver disease case status
Percent
0
10
20
30
40
1
2
3
4
5
6
log2(peak vaue)
CASE ▨ 0 ▨ 1

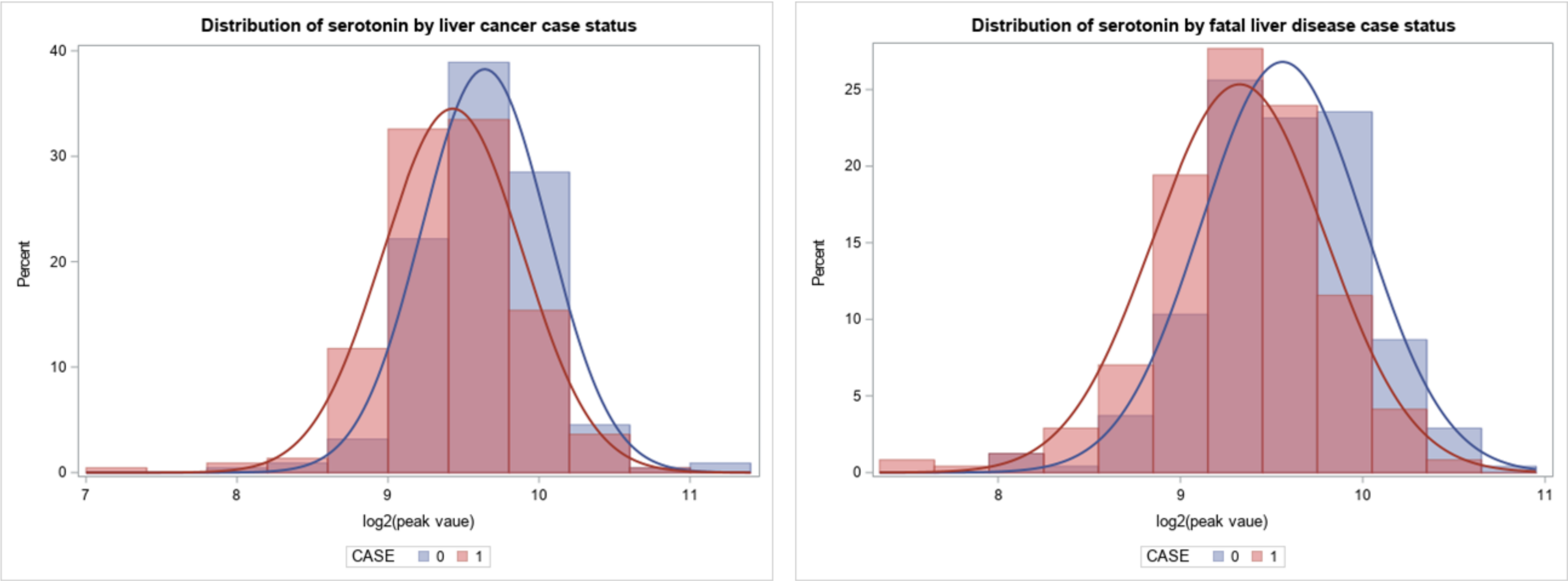
(E)


Distribution of GCA by liver cancer case status
Percent
0
10
20
30
40
1
2
3
4
5
log2(peak vaue)
CASE 🟦 0 🟥 1
Distribution of GCA by fatal liver disease case status
Percent
0
10
20
30
40
1
2
3
4
5
log2(peak vaue)
CASE 🟦 0 🟥 1

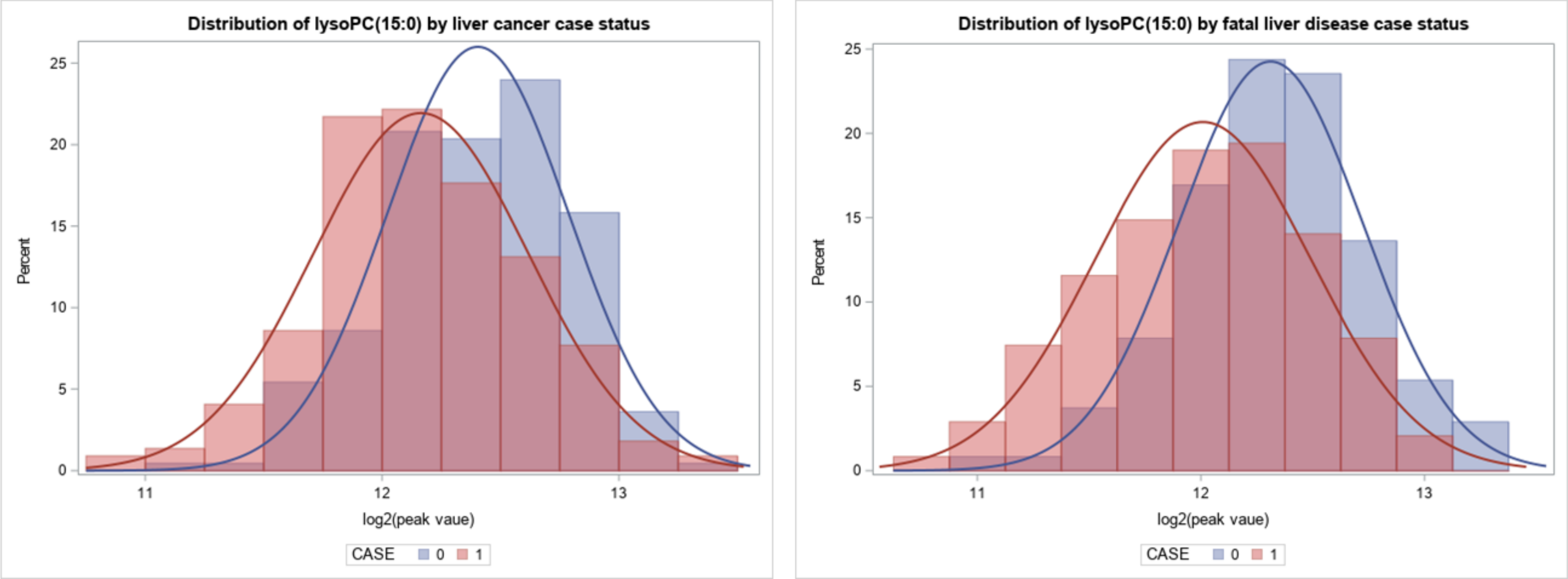
(F)

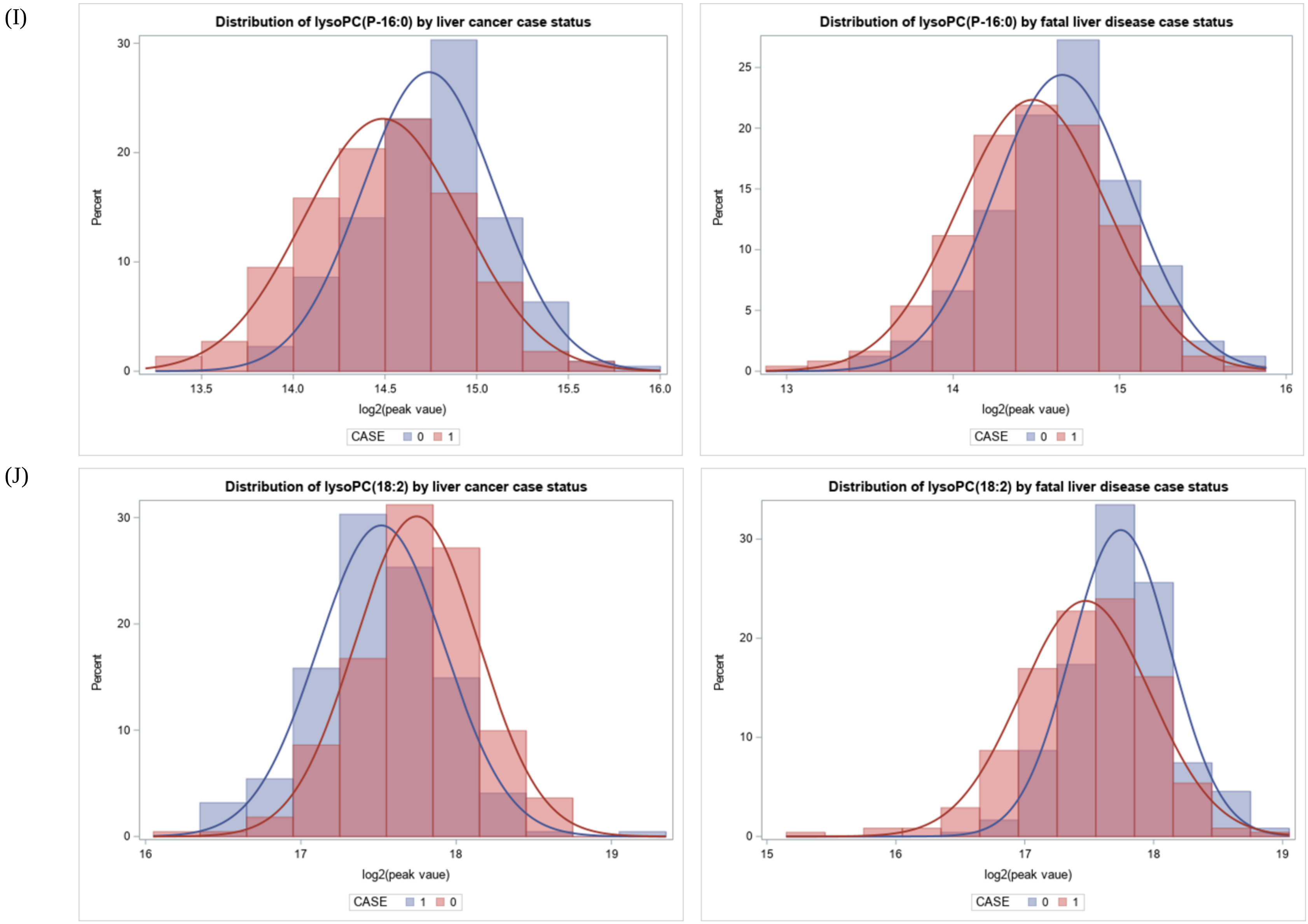

Distribution of GCDCA by liver cancer case status
Percent
0
10
20
30
40
1
2
3
4
5
log2(peak vaue)
CASE 🟦 0 🟥 1
Distribution of GCDCA by fatal liver disease case status
Percent
0
10
20
30
40
1
2
3
4
5
log2(peak vaue)
CASE 🟦 0 🟥 1

(G)


Distribution of serotonin by liver cancer case status
Percent
0
10
20
30
40
7
8
9
10
11
log2(peak vaue)
CASE 🟦 0 🟥 1
Distribution of serotonin by fatal liver disease case status
Percent
0
5
10
15
20
25
8
9
10
11
log2(peak vaue)
CASE 🟦 0 🟥 1

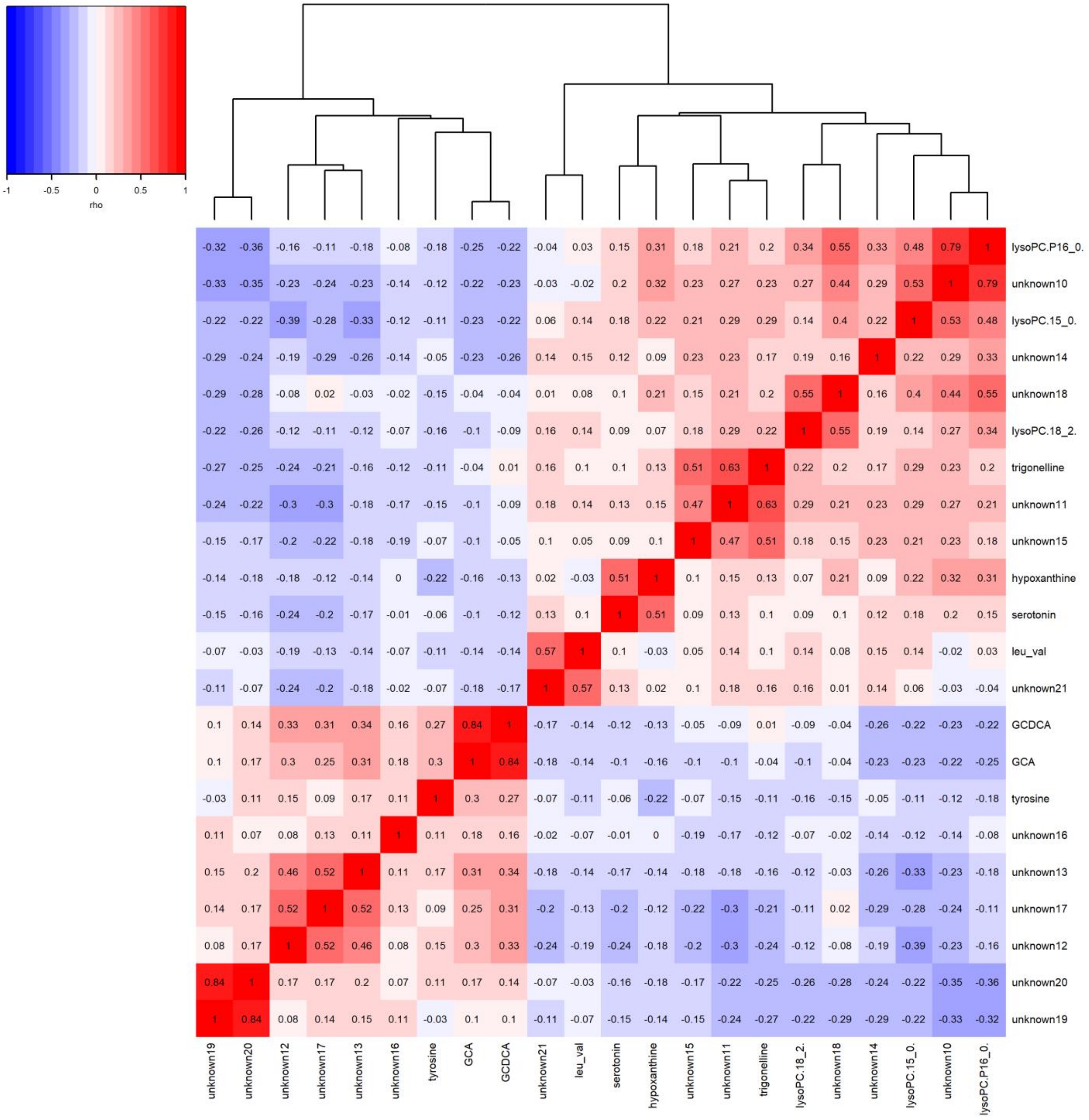
(H)


Distribution of lysoPC(15:0) by liver cancer case status
Percent
0
5
10
15
20
25
11
12
13
log2(peak vaue)
CASE 🟦 0 🟥 1
Distribution of lysoPC(15:0) by fatal liver disease case status
Percent
0
5
10
15
20
25
11
12
13
log2(peak vaue)
CASE 🟦 0 🟥 1



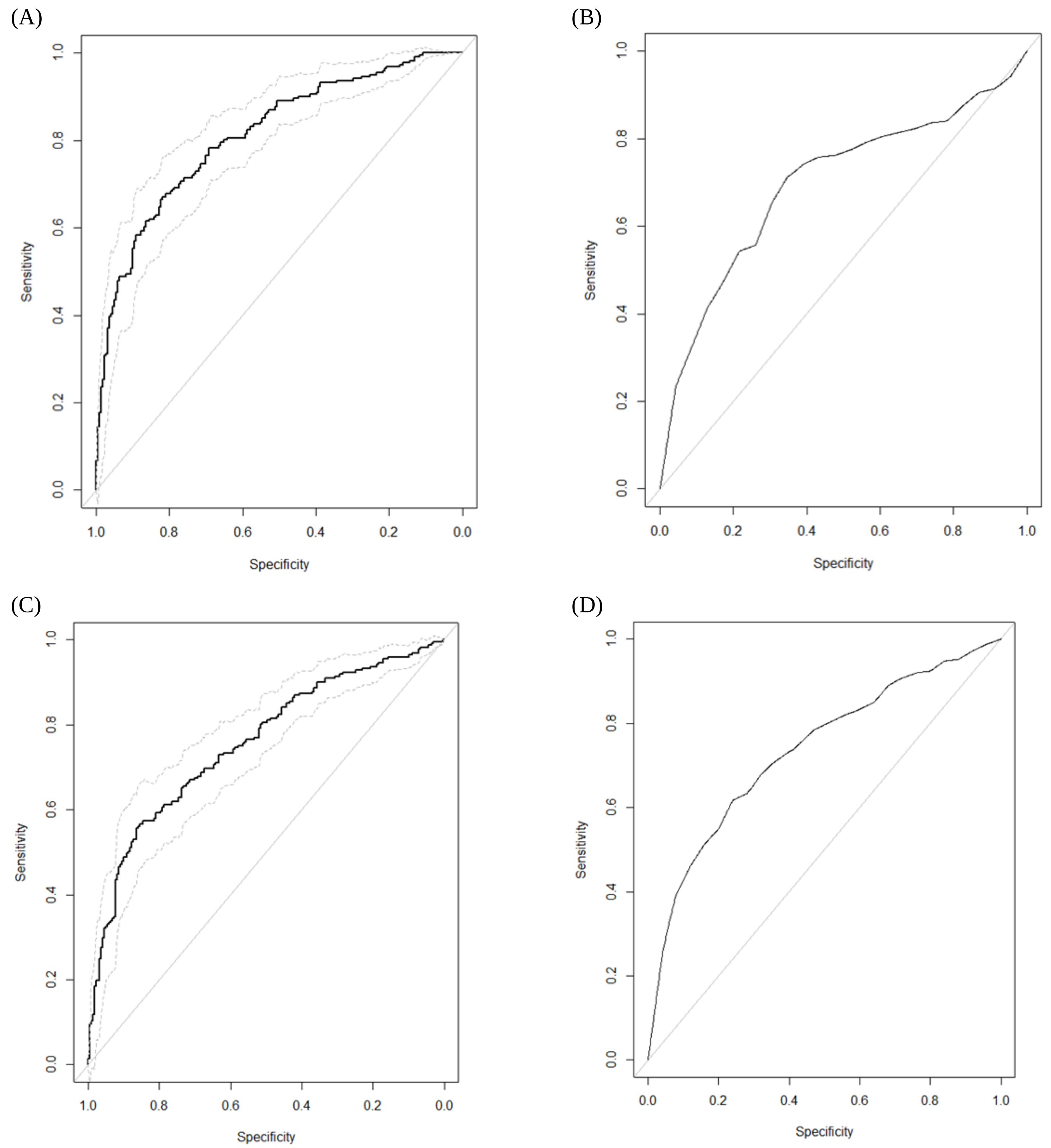
Supplementary Figure 3. Histograms of metabolites (log$_2$-transformed) by case type and status.

Histograms illustrate the distribution of each identified metabolite (log$_2$-transformed) among cases (red) and controls (blue) for the liver cancer and fatal liver disease case-control sets. **Panels A-J** correspond to the following compounds: **(A)** trigonelline, **(B)** hypoxanthine, **(C)** tyrosine, **(D)** leu-val, **(E)** glycocholic acid (GCA), **(F)** glycochenodeoxycholic acid (GCDCA), **(G)** serotonin, **(H)** lysoPC(15:0), **(I)** lysoPC(P-16:0), **(J)** lysoPC(18:2).



Supplementary Figure 4. Heatmap of metabolite correlations.

The heatmap was generated using hierarchical clustering of partial Spearman correlations, adjusted for case status and batch variables, between metabolite features that were statistically significantly associated with coffee drinking, liver cancer, and/or fatal liver disease.



Supplementary Figure 5. aROC curve of winner's curse.

We evaluated whether the estimated aAUC and aROC may be biased because of the "winner's curse" or using those metabolites most significantly associated with cancer. Panels A and B show the overall (aAUC = 0.81 (95% CI: 0.77-0.85) and unbiased estimate (aAUC = 0.72) for liver cancer, respectively. The liver cancer model includes glycocholic acid, leu-val, hypoxanthine, and lysoPC(18:2) Panels C and D show the overall (0.75 (95% CI: 0.71-0.80) and unbiased estimate (aAUC = 0.76) for liver disease mortality, respectively. The liver disease mortality model includes glycochenodeoxycholic acid, lysoPC(18:2), serotonin, and trigonelline.