Supplement: Details of data analysis

Identification of peaks

Each spectrum (range 500 to 25000 Da) was divided into nine tiles using the cut points 1500, 3000, 5000, 7000, 9000, 12000, 15000, and 20000 Da and peaks were identified separately for each tile (Figure 4).

A certain m/z value was defined as a peak if it exhibited the highest intensity within an interval of length δ around the m/z value. The N highest peaks in each spectrum were then selected. Finally, reproducible peaks across the spectra were identified as those which could be found in at least four of the nine spectra within an interval of length δ. The choice of δ and N is of course crucial, as δ must be large enough to cover the measurement error across the spectra with respect to the peak location, but small enough to allow separation of neighboring peaks, and N must be chosen close to a value separating peaks reflecting a signal from peaks reflecting noise. δ and N were chosen separately for each tile taking into consideration the number of common peaks identified as a function of δ for fixed N or as a function of N for fixed δ in each patient. An elbow criterion focusing on a sharp decrease of the slope was used to identify the most appropriate values for δ and N. An example is shown in Figure 5. The maximal values of δ and N over all patients were then chosen for each tile. This procedure resulted in the values of δ being 5, 6, 7, 8, 10, 15, 20, 30, and 50 Da for the nine tiles mentioned above. These values reflect the decrease in precision of peak locations with increasing mass values. The values for N were 80, 80, 100, 80, 60, 40, 15, 15, and 15 within the nine tiles, reflecting that most of the peaks can be found in the mass range up to 10000 Da. The appropriateness of these choices was confirmed by marking the reproducible peaks identified in plots combining the nine spectra from each subject. To each peak in the individual profile, i.e. the list of reproducible peaks, the mean m/z value and the mean intensity of the identified peak were assigned. The mean peak mass value and intensity was calculated over all the spectra in which the peak was identified.

All the individual profiles were joined into one list to identify common peaks across subjects. A visual inspection of these lists showed that groups of common peaks could easily be identified, as illustrated in Figure 6. To obtain an objective definition, we required a peak-free interval of a certain length between two neighboring groups. The interval lengths chosen to separate the groups for the nine tiles were 1, 1.1, 1.25, 1.5, 1.75, 2, 2.25, 2.5, and 3 Da. Only groups of peaks present in at least ten subjects were included in the overall proteomic profile considered in this paper. The peaks in this profile were characterized by the mean m/z values of the subjects contributing to the group. For each subject the intensity of the peak was set to 0, if the subject did not contribute to the group. Nine groups included several peaks from the same subject; here we made a random selection among the peaks from each subject. For each peak the Box-Cox transformation of the measured intensities resulting in a skewness of 0 was also determined. In subjects with no measured peak intensity, the transformed intensities were set to the minimal observed value minus 20% of the range.

Statistical analysis – Step 1

The Wilcoxon rank sum test was used to compare the intensities between cases and controls in order to identify peaks with discriminative power. P-values were corrected for multiple testing by determining the permutation test distribution of the minimal p-value over all peaks, and using this distribution as reference for all peaks.[45] The significance of the number of peaks with a correlation of the intensity to a certain prognostic factor or of the average correlation was assessed by determining the permutation test distribution of these statistics computed among the 67 peaks with smallest p-value among all peaks where the median intensities in cases was higher than in controls.

All permutation test results were based on 1000 random permutations.

Statistical analysis – Step 2

We started by subjecting the mass values with discriminative power ($p \leq 0.01$) to an average linkage clustering, using the absolute Spearman correlation of the intensities as similarity measure. We identified clusters of mass values with high correlation of the intensities using a correlation of 0.5 as cut point in the dendrogram, i.e. within each cluster we have an average correlation above 0.5 and between the clusters we have an average correlation below 0.5. All clusters with only one mass value were collected into one cluster. To investigate whether single subjects show high intensities across all clusters of mass values or whether they show high intensities only for some clusters, we computed for each individual and each cluster a score value. Here the 7 highest intensities for each mass values gave 1 point for the 7 corresponding subjects, and the next 7 highest intensities gave 0.5 points for the 7 corresponding subjects. Then the score value for each subject and each cluster was computed by summing up the points obtained for this subject by the mass value of the cluster. An overall score was computed by summing up the cluster specific scores.

To construct a diagnostic rule, we computed for each subject and each cluster the value of the first principal component score based on a principal component analysis using only the mass values of the cluster and the transformed intensities. So we have in our specific case with four clusters four variables reflecting for each subject the “average” intensity for the mass values of each of the four clusters. These four variables were then used as input to a forward logistic regression, where we fixed in advanced the size of the final model to 1, 2, 3 or 4 variables. As our main interest was to describe the potential diagnostic accuracy in combining the different mass values, we computed cross validated sensitivity and specificity for each of the four models obtained. Sensitivity and specificity refers here to the classification obtained by classifying all subjects with a predicted probability above 0.63 (48/76) according to the fitted logistic regression model as cases. Cross validation refers to a leave-one-out cross validation, and we repeated the complete modeling process for each subset with one subject left out. This means that we repeated for each subject the selection of peaks with a p-value less than 0.01, average linkage clustering of these mass values, selecting all non-singleton clusters at a cut point of 0.5, computing the first principal component score in each cluster and using these as input to a forward logistic regression.

All statistical computations of Step 1 and step 2 were performed with Stata 9.2.

Statistical Analysis – Step 3

The original data were additionally submitted to an independent analysis by using the support vector machines (SVM) for class prediction using the R package *e1071* with radial basis as kernel.

As a popular machine learning algorithm, the support vector machines (SVM) compute a maximal margin hyperplane or boundary that separates given samples into classes with maximal distance to the nearest data point.[46] The SVM is introduced with consideration of its good empirical performance and capacities in combating over-fitting and in approximating complex classification functions. Moreover, the probability estimate for the classification of each sample provides certainty measurement for its prediction. The cut point for classification probability was set to 0.63 as, of the 76 samples, 48 are tumors. Validation was performed again by leave-one-out cross validation. Features or peaks were selected by thresholding based on the absolute values of the t-statistic. That is, for a given threshold T, we selected the set of all peaks with $| t |> T$. Then performance of the selected features was examined by cross-validation as just described. The set of features that gave the best performance (lowest cross-validation error rate) was selected.

Supplement: Figures

Figure 4

Nine MALDI mass spectra (range 500 to 25000 Da) illustrating the nine tiles using cut points 1500, 3000, 5000, 7000, 9000, 12000, 15000, and 20000 Da. The asterisk illustrates an example of a perfectly reproducible peak.

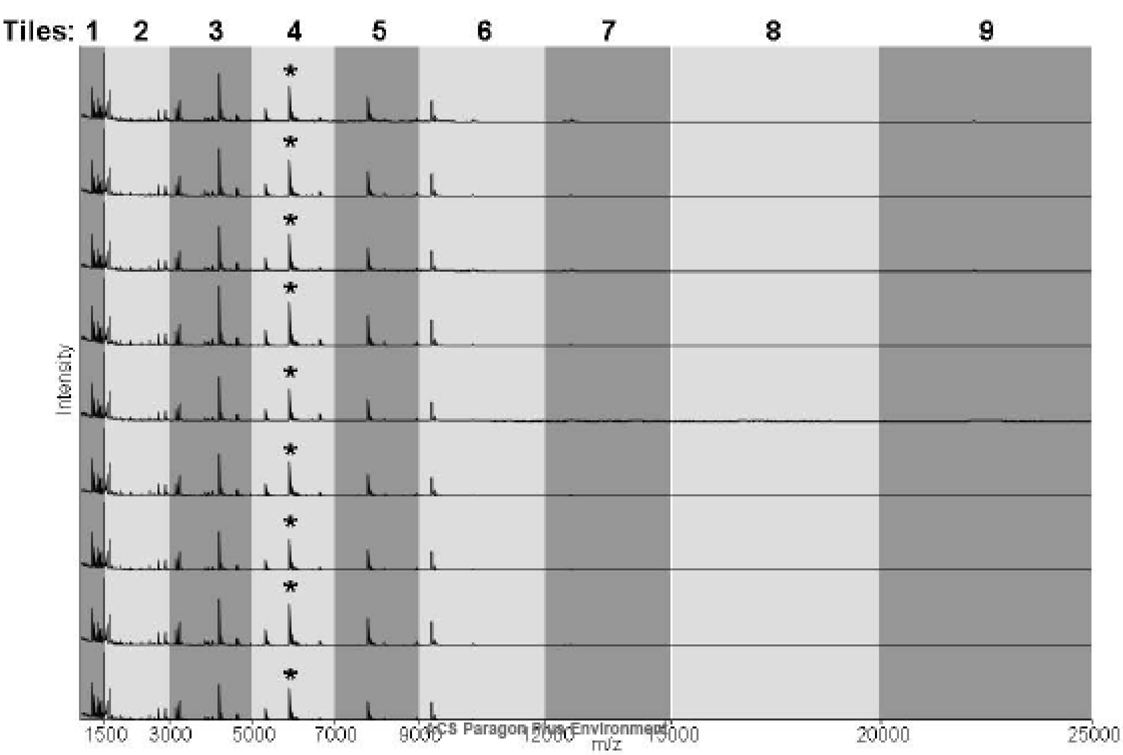
Figure 5

The number of perfectly reproducible peaks (found in all nine spectra) between 1500 and 3000 Da as a function of δ for 8 selected subjects and a preliminary choice of $N = 50$. The choice of δ for each patient using an elbow criterion is indicated by a dashed vertical line.

Figure 6

The union of all individual lists of reproducible peaks between 2720 and 2840 Da. Each point represents the m/z value of one reproducible peak for one subject. We can observe 8 groups of peaks, defining 8 peaks common across subjects.

Figure 4



Tiles: 1 2 3 4 5 6 7 8 9
*
Intensity
1500
3000
5000
7000
9000
12000
15000
20000
25000
m/z

Figure 5

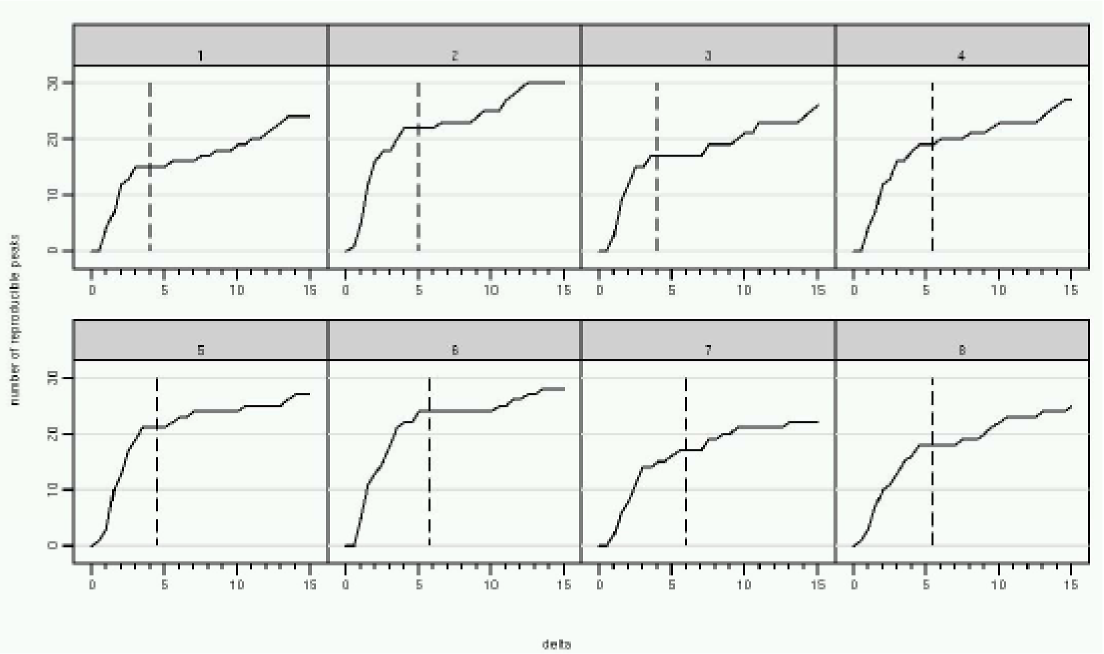
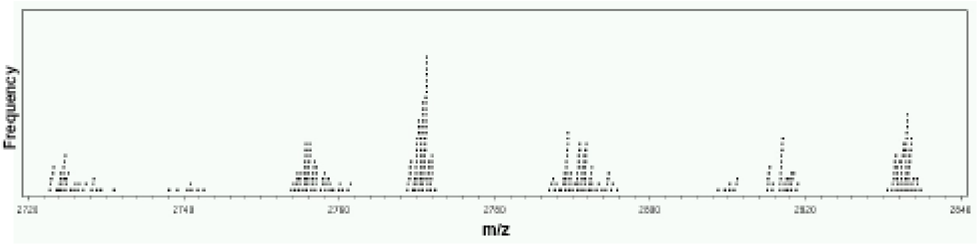

1
2
3
4
5
6
7
8
number of reproducible peaks
0
10
20
30
0
5
10
15
delta

Figure 6



Supplementary Table A

Identified m/z values discriminating cases and controls with a p-value less than 0.01.

Mean m/z value	p-value	Corrected p-value	Median intensity for controls (n=28)	Median intensity for early stage tumors (n=22)	Median intensity for late stage tumors (n=15)	Tendency
1045.7	.0084337	0.882	294.0	441.4	338.2	+ -
1693.0	.000031	0.006	0.0	198.4	283.7	+ +
1755.5	.0021463	0.454	0.0	132.9	140.6	+ +
1780.0	.0030637	0.582	296.4	535.9	485.3	+ -
1867.2	.0052874	0.760	619.5	958.9	1092.6	+ +
2024.0	.0024784	0.503	221.2	287.7	299.5	+ +
2106.5	4.26e-06	0.002	425.2	753.6	766.7	+ +
2136.5	.0024784	0.503	190.5	262.1	276.2	+ +
2282.1	.008664	0.890	110.5	131.8	122.0	+ -
2469.3	.0003188	0.092	120.2	183.0	222.6	+ +
2555.5	1.94e-06	0.001	419.0	779.6	1034.5	+ +
2670.8	.0001722	0.051	154.4	292.7	277.4	+ -
2770.7	.000021	0.005	1016.2	2290.9	2938.3	+ +
2832.7	.0010642	0.260	115.7	145.5	314.0	+ +
2933.5	.000336	0.096	1575.4	2888.8	3126.7	+ +
2953.3	.0000255	0.006	427.1	784.5	993.7	+ +
2985.0	.000026	0.006	160.6	268.0	323.5	+ +
2995.3	.0028061	0.557	250.2	354.1	422.6	+ +
3144.7	.0005047	0.139	96.4	130.6	227.8	+ +
3192.9	.0000355	0.007	1160.7	2682.3	3453.4	+ +
3209.7	.000202	0.062	101.6	203.2	314.6	+ +
3242.2	.0039678	0.670	217.7	399.2	591.5	+ +
3263.9	.000039	0.008	2297.0	5434.0	5565.5	+ +
3281.5	.0000848	0.019	203.9	387.5	471.2	+ +
3306.2	.0007214	0.178	63.1	147.7	262.5	+ +
3325.8	.000144	0.040	305.3	664.8	791.5	+ +
3366.7	.0067642	0.837	107.5	125.0	149.7	+ +
3378.1	.0039295	0.667	22.5	72.1	119.0	+ +
3427.4	.0015324	0.365	60.9	79.1	77.8	+ -
3448.9	.000039	0.008	148.4	263.6	252.1	+ -
3634.8	.0003495	0.097	86.9	103.4	111.6	+ +
3883.3	.0032849	0.607	862.9	1064.5	1104.7	+ +
3914.0	.0024791	0.504	220.9	265.4	314.8	+ +
4092.0	.0021481	0.454	139.2	200.4	238.5	+ +
4166.6	.0063684	0.824	0.0	0.0	61.9	+ +
4190.1	.000421	0.117	96.0	148.7	167.8	+ +
4210.2	.000212	0.066	6530.8	11452.8	11144.3	+ -
4231.3	.0000243	0.005	249.1	475.6	494.8	+ +
4240.2	.0000969	0.021	245.3	462.0	499.4	+ +
4249.5	.0003571	0.097	106.7	167.3	222.2	+ +
4271.5	.0002212	0.070	1236.7	1977.7	2143.7	+ +
4396.5	.0009469	0.228	199.5	229.2	351.7	+ +

4786.9	.0008489	0.204	35.1	52.6	73.6	+ +
4963.4	.0054659	0.773	316.9	404.0	406.2	+ +
5246.8	.0066348	0.835	32.9	44.0	76.6	+ +
5299.9	.000792	0.190	35.6	0.0	0.0	- =
5336.0	.0000469	0.009	723.4	1783.0	2618.0	+ +
5398.7	.000039	0.008	188.6	439.4	561.8	+ +
5461.2	.0000887	0.019	83.1	132.2	201.4	+ +
5522.1	.0001011	0.022	76.2	93.9	138.6	+ +
5804.1	.0004508	0.128	43.3	103.0	177.6	+ +
5864.1	.0033078	0.610	34.7	51.4	80.9	+ +
5885.6	.0014776	0.350	0.0	52.5	141.3	+ +
5903.3	.0000157	0.005	3029.3	6216.9	8613.7	+ +
5928.2	.0000454	0.008	62.8	217.1	412.9	+ +
5938.7	.0011441	0.275	72.2	181.1	361.1	+ +
5965.4	.000021	0.005	718.4	1746.2	2589.1	+ +
5981.7	.0015945	0.377	52.3	84.9	181.1	+ +
6027.9	.000212	0.066	231.7	660.2	964.0	+ +
6089.6	.0002032	0.062	143.0	369.2	590.7	+ +
6152.1	.0044713	0.712	67.9	98.5	164.7	+ +
6188.6	.0004177	0.116	39.5	0.0	0.0	- =
6239.1	.0013985	0.334	44.7	27.9	0.0	- -
6302.3	.0027573	0.548	119.0	209.9	211.2	+ +
6558.2	.0013663	0.326	22.3	38.1	66.4	+ +
7762.3	.0022268	0.469	4882.0	6039.1	6664.1	+ +
7790.1	.00998	0.917	105.8	160.5	168.6	+ +
8138.3	.0068751	0.841	596.5	757.5	821.4	+ +
8640.8	.0042223	0.690	11.2	0.0	0.0	- =
11966.5	.0011203	0.273	0.0	0.0	0.0	= =
12600.3	.0040377	0.676	153.3	253.0	265.8	+ +
12660.4	.0006412	0.159	36.4	58.1	60.8	+ +

Late stage tumors are defined as tumors over 20 mm and with positive lymph nodes, whereas early stage tumors are defined as tumors under 20 mm with negative lymph nodes.

Tendency refers to the difference in intensity of the m/z-value between 1) cases and controls and between 2) late and early stage tumors (+ corresponds to higher intensity of the m/z-value among cases/late stage tumors compared to controls; - corresponds to lower intensity of the m/z-value among cases/late stage tumors compared to controls)

Corrected p-values are based on a permutation test.