Data analysis of assorted serum peptidome profiles

Josep Villanueva[1,2,4], John Philip[1,4], Lin DeNoyer[3] & Paul Tempst[1,2]

[1]Protein Center, [2]Molecular Biology Program, Memorial Sloan-Kettering Cancer Center, 1275 York Avenue, New York, New York 10021, USA. [3]Spectrum Square Associates Inc., 755 Snyder Hill Road, Ithaca, New York 14850, USA. [4]These authors contributed equally to this work. Correspondence should be addressed to P.T. (p-tempst@mskcc.org).

Published online 29 March 2007; doi:10.1038/nprot.2007.57

Discovery of biomarker patterns using proteomic techniques requires examination of large numbers of patient and control samples, followed by data mining of the molecular read-outs (e.g., mass spectra). Adequate signal processing and statistical analysis are critical for successful extraction of markers from these data sets. The protocol, specifically designed for use in conjunction with MALDI-TOF-MS-based serum peptide profiling, is a data analysis pipeline, starting with transfer of raw spectra that are interpreted using signal processing algorithms to define suitable features (i.e., peptides). We describe an algorithm for minimal entropy-based peak alignment across samples. Peak lists obtained in this way, and containing all samples, all peptide features and their normalized MS-ion intensities, can be evaluated, and results validated, using common statistical methods. We recommend visual inspection of the spectra to confirm all results, and have written freely available software for viewing and color-coding of spectral overlays.

INTRODUCTION

Polypeptide markers

There are alternative ways to go about mass spectrometry (MS)-based biomarker discovery in biological fluids. "Great-depth" analysis penetrates several orders of magnitude into the proteomic dynamic range but the level of sophistication and lengthy analysis time of the procedures, such as two-dimensional liquid chromatography (LC/LC)-coupled electrospray ionization (ESI)-MS/MS of trypsinized protein mixtures[1], preclude probing the large sample cohorts required for meaningful statistical analysis. "High-throughput" methods, on the other hand, relying for instance on MALDI-TOF-MS[2,3] and the more popular SELDI-TOF-MS[4,5], allow rapid profiling of pre-existing peptides in hundreds of serum samples but simply skim the top layer off complex peptidomes, thereby limiting discovery to fragments of abundant blood proteins[2,6,7]. Data analysis developments related to various LC- or LC/LC-front-ended discovery formats have already been widely reported in the peer-reviewed literature[8–15]. Here, we describe a data analysis workflow intended for use with the second of these approaches, that is, discovery of "biomarkers" in the form of distinct peptide signatures hidden in much larger serum peptidomes[2,3,16–19] revealed by high-throughput, low-resolution MS. More specifically, it has been developed for use in combination with a profiling approach whereby serum samples are processed automatically using solid-phase extraction in a liquid handler and a total peptide read-out is then obtained using MALDI-TOF-MS[20,21]. The first step in the data analysis pipeline (**Fig. 1**) is transfer of the raw spectra, which are then interpreted using signal processing algorithms to define suitable features (in this case peptides). These features will later be used to compare measured peptidomes and to derive differential "biomarker" panels, defining clinically meaningful differences between patient data sets, through appropriate statistical analysis.

Mass spectra processing

Signal processing of the spectra is performed through a series of steps: smoothing, baseline correction, normalization, calibration/alignment and peak labeling. The initial choice on how to process mass spectra as a prelude to biomarker discovery is whether to use all recorded *m/z* values as features or to use *m/z* "peaks" identified by a peak detection algorithm. Peak detection has the advantage of deriving a set of features by filtering out noise, but may also eliminate low-abundance features. On the other hand, the use of

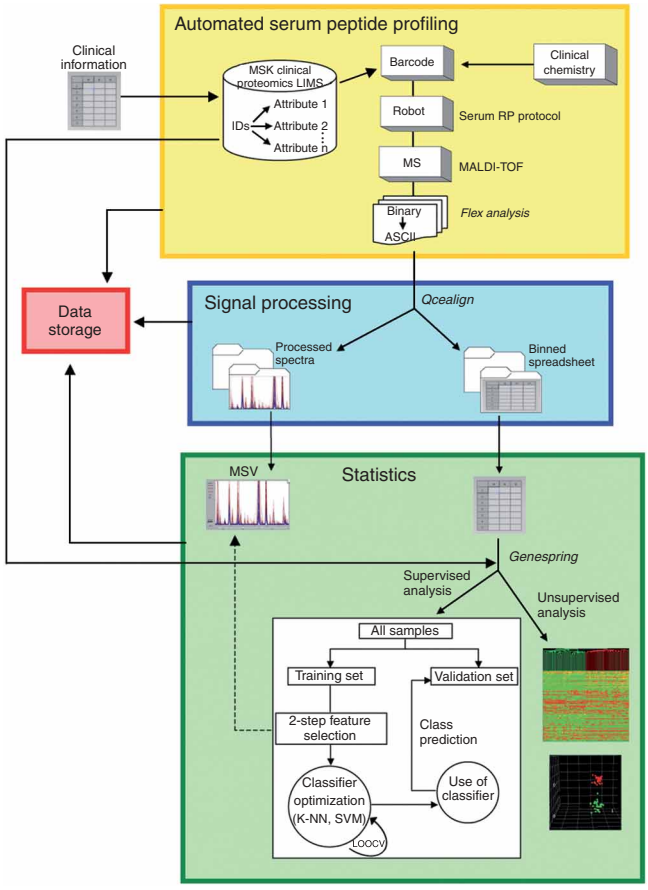


Figure 1 | MSKCC data analysis pipeline for serum peptidomics. The diagram shows the various data analysis steps for use in conjunction with the serum peptidomics protocol described by Villanueva *et al.*[21].

raw *m/z* values retains low-abundance features, but can also result in keeping several "features" deriving from noise. As our intention has always been to not only detect biomarker peptides but also to identify them (by MALDI TOF/TOF and MALDI Q/TOF MS/MS, as described[2]), we decided to only use peaks as features. The software described here, "Qpeaks," finds peaks in a spectrum and generates various output information, including a peak table, smoothed trace and an optional baseline.

The peak-finding function *rzrpic* is based on a Bayesian second derivative of the data. Smoothing is accomplished with the Maximum Entropy (Maxent) smoothing function *rzresm*. Both the Bayesian second derivative and the Maxent smoothing are optimal in that they produce the most probable results using only a single assumption: namely, a width for peaks in the data ("singletwidth"). They do not require any additional parameters, such as polynomial degree used by Savitsky–Golay algorithms. Bayesian and Maxent algorithms are self-adjusting to the noise, so no "degree of smoothing" parameter is required[22]. This handling of noise means that the Bayesian and Maxent functions cannot compromise resolution by oversmoothing. Thus, the use of primarily one parameter simplifies data processing without compromising data quality. Both the Bayesian second derivative and the Maximum Entropy smoothing begin by writing a probability function. For the Bayesian derivative, the probability statement is that the data consist of peaks with shape and width as specified, plus noise and thus calculate the most probable second derivative[22]. For the Maxent smoothing, the probability statement is that the data consist of peaks with shape and width as specified, plus noise and calculate the most probable noiseless spectrum[22]. Additionally, Qpeaks works well with low-resolution data such as the kind acquired in linear TOF-MS mode, unlike many other standard peak labeling algorithms that are optimized for isotopically resolved spectra. We prefer linear TOF-MS mode, as it gives better sensitivity than the reflectron mode; thus, it is imperative we can process lower resolution data and use the average isotopic value for *m/z* peaks instead of the monoisotopic ones.

Spectra are normalized to unit size by dividing each intensity value by the "total ion count." Once normalized, a scaling factor is applied by multiplying each intensity value by a user-selected number (e.g., 10^7). The scaling factor is constant within a data set and is used to convert the normalized spectrum to a "user friendly" scale, where most peak heights are greater than one. This normalization step is very conservative. Future developments will consider other methods of normalization, such as the use of added calibrants to the samples with subsequent scaling to these standards.

Mass spectra alignment

Once processed, the spectra must be aligned to compare peptides across samples, an operation that is perhaps the single most difficult task in peptide profiling studies. A common approach is to perform external calibration; that is, peptides of known molecular mass are analyzed alongside the samples. A calibration curve is then calculated to adjust the *x* axis of the calibrant spectrum so that its known peaks fit their known values. However, despite the best possible external calibration, *m/z* peaks representing identical peptides in different samples deviate to various extents from the theoretical molecular mass. They are slightly shifted to the left or right in the spectra, which makes strict "numerical" alignment (e.g., a spectrum divided in 1 Da, consecutive segments) all but impossible. Deciding automatically what to consider as the "same peak" between different sample spectra is a difficult task, and different methods have been proposed. A straightforward approach is window-based peak binning, whereby all peaks within a given *m/z* window across spectra are considered to be the "same" peak[17]. One may also use a genetic algorithm to group peaks trying to maximize the peak number in a group from different samples, and minimize the number of peaks in a group from the same sample[18]. Finally, hierarchical clustering and time warping have also been proposed for peak binning[19,23].

We have previously developed a new approach for alignment, applying a function, termed "Entropycal," which aligns sample data files to a reference file using a minimum entropy algorithm, and by taking unsmoothed ("raw"), baseline-corrected data[24]. Taking raw spectra for the alignment allows all the statistical information in the data to be used. The alignment is performed in three steps: reference spectrum creation, applying "Entropycal" and binning. First, a reference spectrum is created by summing all intensities of all the calibrated samples. Calibration of the spectra yields alignment of peaks within a 1,500-p.p.m. (0.15%) window. At this resolution, erroneous summing of different peaks (i.e., merging different peaks into each other) is generally not a problem. This results in a composite spectrum that contains the average of the peak information from all the data sets. The *x* axis of the reference spectrum is the *x* axis of the first calibrated sample. Next, "Entropycal" slides each data file by *n* data points to the right or left along the *x* axis of the reference file. At each relative position *n*, the Shannon entropy of the sum of the two files is computed. The optimal alignment occurs at the shift that produces the minimum Shannon entropy[25]. Third, the aligned peak lists are then binned by using the resolution of the peaks: all peaks in rows within $\Delta(m/z)$ of the strongest peak at a given value of *m/z* are binned together, and a spreadsheet is created for further statistical analysis. This approach appears to complement the signal processing of "Qpeaks." Again, no *a priori* information is assumed and no expert fine-tuning is required.

Feature selection

Statistical analysis to evaluate peptidomic data can readily be performed using commercially available software such as "GeneSpring" (Agilent). Unsupervised analysis is first performed to get a visual representation of the signal strength of the data. The spreadsheet containing all the peptides and their normalized intensities found in the data set (created during the signal processing) is used to create a hierarchical cluster and to do principal component analysis (PCA). Biomarker discovery requires identification of features that distinguish between classes of interest. This is typically accomplished by carrying out supervised analysis. It is advisable to use a "training set" to optimize feature selection and class prediction, and then a separate "validation set" to assess error rate of the final models generated using the training set. This strategy will help to avoid data overfitting leading to artificially low error rates. In practice, the available samples are randomized and 75% (or less) is assigned to the training set and 25% (or more) as validation set, depending on the total number. Alternatively, training and validation sets can be collected independently over time.

Although various methods such as self-organizing maps, genetic algorithms, neural networks have been used, for simplicity reasons, we prefer a two-step feature selection using the training set. First, a Mann–Whitney *U*-test is used to rank each feature based on its

ability to separate the samples into appropriate clinical groups using multiple hypothesis testing correction. Then, a second step is carried out to filter out masses by peak height aimed at removing peptides that would be extremely difficult to sequence-identify and that probably are not reproducible among different analyses and methodologies anyway. The peak height cutoff is arbitrary and is ascertained experimentally, and the threshold may be different for each mass spectrometer (type/manufacturer) as well as depend on expertise and confidence of the investigator(s). The output of feature selection containing the most significant peptides arising from feature ranking and the peak height cutoff filter are then used as input to class prediction statistical tools to find a biomarker panel (**Fig. 1**). Here, we use two different machine learning for class prediction, k-NN (K-nearest neighbor) or SVM (support vector machine). Different models are generated using leave-one-out-crossvalidation on the training set using the classification error rate for parameter optimization. Then, these classifiers are tested on the validation set.

Regardless of the signal processing routines and statistical methods used, it is prudent to visually inspect and confirm all results. *m/z* peaks obtained after feature selection should be examined using the MassSpectraViewer (MSV) that has been written in Matlab for this purpose. Peptides with low *P*-values (to be determined by the researcher, but typically $P<0.05$) should show clear overall differences in the viewer between clinical groups. We have often observed that visual inspection of the peaks surviving feature selection may reveal stronger differences than the statistics suggested. Conversely, the viewer can also serve as an error-check for signal processing. If a peak with low *P*-value shows negligible difference in the spectral overlays, there might have been an error in processing, most likely the result of poor binning (i.e., too many bins). Only in those cases where the outcome of the statistical analysis is visually confirmed, one can be sufficiently confident about the results.

Note concerning the procedure

All the steps described below are manufacturer-specific but the functionality can be reproduced elsewhere. For instance, all the statistical analysis tools (ANOVA, k-NN, SVM, PCA) are available in "R" or "SAS" languages and environment, and most are available in "Matlab" language for technical computing as well, but implementation in these languages is beyond the scope of this protocol. The only manufacturer-specific step is the conversion of Bruker MALDI-TOF MS raw data to ASCII text files using the FlexAnalysis macro.

MATERIALS

REAGENT

Spreadsheet with patient data and various clinical parameters
Unprocessed MALDI-TOF-MS data in binary format, generated using a Bruker AutoFlex or UltraFlex type instrument

EQUIPMENT

- Apple personal computer; PC; or Linux system
- Server for data storage
- FlexAnalysis software (Bruker)
- Matlab software (Mathworks)
- QPeaks software (Spectrum Square Associates)
- Entropycal software (Spectrum Square Associates)
- GeneSpring software (Agilent)
- MSV software (Memorial Sloan-Kettering Cancer Center); this software is available upon request from the authors

EQUIPMENT SETUP

Matlab software Use instructions available from the Mathworks website to install Matlab. The setup instructions are for an Apple computer, but the setup for a PC or Linux system is similar.

FlexAnalyis software Follow the instructions that come with the Bruker CD to install FlexAnalysis. Download the macro from the supplemental section of this protocol (**Supplementary Method 2**). On the PC where FlexAnalysis has been installed, copy the macro into the "FlexAnalysisMacroModules" folder. Typically, this folder is located in "C:\Methods\FlexAnalysisMacroModules." Launch FlexAnalysis. The Macro should now appear as "Convert Directory of Spectra to ASCII" in the last section of the Tool Menu.

Qpeaks, Qcealignf, Entropycal, calcMedian and MSV software Download information is available at http://cbio.mskcc.org/tempst. Place all files related to these programs in one folder, called "Data Process": docal.m; docaleffT.m; docalt0.m; docalt1.m; docalt2.m; docalt3.m; docalt6.m; docalt7.m; docaltof.m; dqpeak.mexmac; entropycal.mexmac; getacaldcal.m; getacaldcalt0.m; getacaldcalt1.m; getacaldcalt2.m; getacaldcalt3.m; getacaldcalt6.m; getacaldcalt7.m; getacaldcaltof.m; getaligncals.m; getcalibrants.m; ProcessCalFiles.m; ProcessDataFiles.m; qcealignf.m; qpeaks.m; SumDataFiles.m; ChgDirectory.m; GroupLegend.m; map.mat; marklegend3.m; masspectraviewer.m; calcMedians.fig; calcMedians.m; load_definitionFile.m; load_outFile.m; loadInfo.m; SimpleDescriptiveStats.m. Launch Matlab. Select "File" from the Menu. Then select "Set Path" from the "File" Menu. A window will appear. Push the "Add Folder" button and a dialog window will appear asking you to select a folder. Navigate the dialog to "Data Process" folder and select it. The folder will now appear at the top of the Set Path window. Press the save button and then the close button.

GeneSpring software Use the instructions available from Agilent's website (http://www.sigenetics.com/) to install GeneSpring.

Personal computer The preferred operation system is MacOS X (Apple). Other operating systems should also work fine as long as they can run Matlab. For larger data sets, the computer should have at least 1 GB of RAM and a high-end video card with 256 MB of video RAM.

Server for data storage This could be any computer with a large hard drive (>500 GB). It should be networked to allow multiple computers to access it. It should also have a fast network card. An example of such a computer would be the Apple Xserve (http://www.apple.com/xserve).

PROCEDURE

Converting raw data to ASCII

1| MS instrument manufacturers typically provide a procedure to convert data from their binary format to standard tab delimited ASCII files. We provide instructions for Bruker MALDI-TOF instrumentation here. Create a folder called "Raw Spectra_files" (see **Fig. 2a**). You may wish to append the date to the name of the folder as well; for example, Raw Spectra _files_051506. Place all the raw spectra to be processed in this folder. Sample names must be followed by "_1" or "_2," for example, "00ZG70005DVA_1," "00ZG70005DVA_2." The "_1" designates data acquired in the 700 Da to 4 kDa range whereas "_2" refers to data acquired in the 4–15 kDa range. Calibrant files use the same naming convention in addition to "-Cal" to designate the data file as a calibrant; for example, "00ZG7-Cal_1." The first five characters in the sample and calibrant name will be used to determine which calibrant file is used to calibrate the sample. Sample files will be calibrated using the calibrant file with the matching five digits (see **Fig. 2b**).

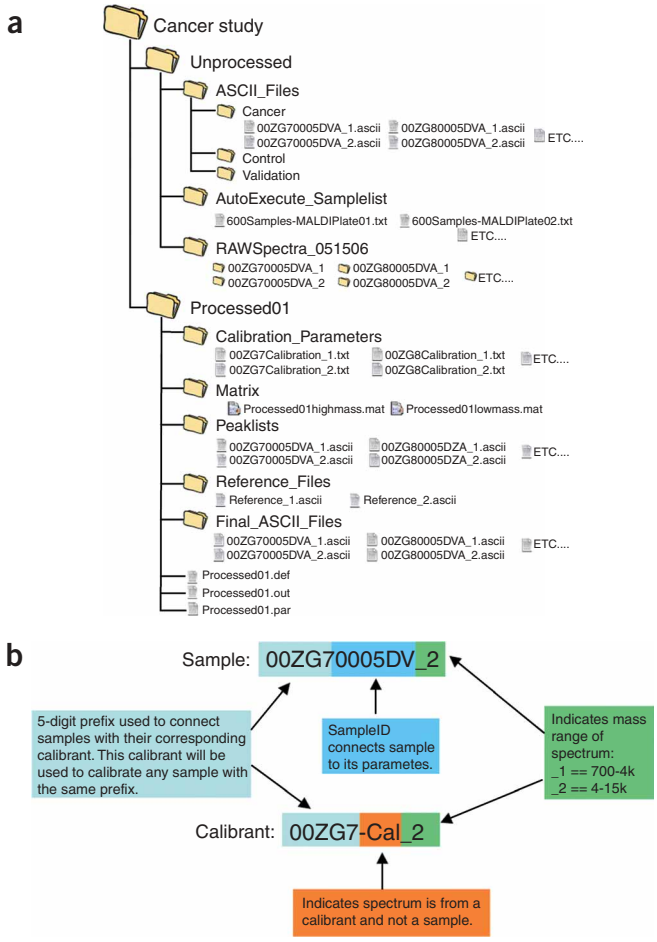


Figure 2 | Data folder structure and naming convention used by the MSKCC serum proteomics data analysis. (**a**) Directory structure for a typical serum peptidomics project. The parent folder structure is divided between unprocessed data, containing the raw spectra, and their ascii version. (**b**) Serum samples and calibrant files are linked using a defined naming convention. The first five digits in both sample and calibrant file names determine which calibrant file is used to calibrate a sample.

▲ **CRITICAL STEP** Bruker MALDI-TOF-MS instruments store the data for each sample inside a folder with many subfolders. The folder structure must not be changed in any way. Each data folder is named with the sample name. Simply copy (or move) the data folder to the "Raw Spectra_files" folder.

▲ **CRITICAL STEP** Sample naming convention must be followed or the script will treat each file as a sample file, and it will not be able to calibrate the data. Sample IDs must be unique.

2| Launch FlexAnalysis. Deselect all the items (Analysis List, Mass Spectrum, Mass List) in the Window Menu. Then select the "Tools" Menu item. Select "Convert Directory of Spectra to ASCII" in the last section of the Tool Menu. A dialog will appear. Press Browse to select the folder where the Raw Spectra data is stored. Navigate to the location of the "Raw Spectra_files" folder and press "OK." Wait as all the data files are loaded into the dialog window. Once this is performed, press the "Convert Spectra" button. This process takes about 300 s for 250 spectra if all the data are stored on a local SCSI hard drive (40 GB, 10,000 r.p.m.) on a Dual 2 GHz G5 Mac with 4 GB of RAM.

? TROUBLESHOOTING

3| Wait while the data are being converted. A dialog box will appear as the data are converted. Once the task is performed, press "Exit" to leave the macro. Now you will have an ASCII text file for each raw spectrum you placed in the "Raw Spectra_files" folder.

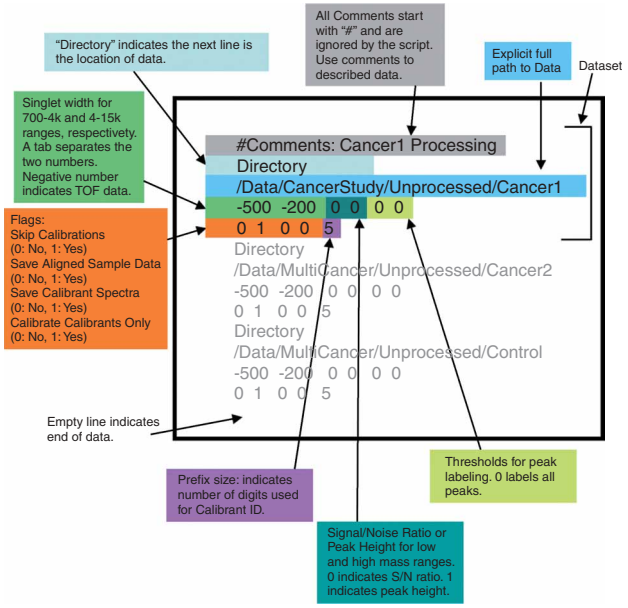


Figure 3 | Parameter file for signal processing. This file contains the processing parameters for the signal processing software (Qcealignf). The four lines of text describe the location of the unprocessed spectra and the signal processing parameters used by Qcealignf.

Setting up spectral processing

4| Create an empty folder. Name the folder with the project name (e.g., "CancerStudy"). Create two folders called "Processed01" and "Unprocessed" and put them in the folder "CancerStudy." These two folders are now subfolders. Navigate to the Unprocessed Folder. Inside this folder, place all the ASCII files of the samples (see **Fig. 2a**).

5| Inside the Processed01 folder, create a text file. Call this file "Processed01.par" (**Fig. 3**). This file will contain the processing parameters for the signal processing software (Qcealignf). The first line is "Directory," which tells the script that the next line in the file refers to the exact location of the data. The second line is the explicit full path of the data. In this case, we enter the full path as "/Data/CancerStudy/Unprocessed/Cancer1." The third line contains the signal processing parameters, separated by tabs. The first parameter refers to the singlet width for the low-mass range (see **Box 1**). The second to the singlet width for the high-mass range. We will enter "-400〈tab〉-200." Note: 〈tab〉 refers to the tab separator. The next two numbers refer to the criteria for minimum peak determination for each mass range. Use 1 to use signal cutoff or peak height threshold. Use 0 to indicate signal-to-noise ratio.

BOX 1 | HOW AN OPTIMAL SINGLET WIDTH IS SET

Singlet width is used for baseline subtraction, peak labeling, smoothing and noise statistics. The "width" is actually the full-width at the half-maximum point of typical isolated peaks in a mass spectrum. This width may be the same as the instrumental resolution; it may be the natural width of peaks. Choose whichever width matches the widths of peaks as seen in the data (see **Fig. 9**). Units are the same as the units of the *x*_axis of the data (usually *m/z*). The negative number indicates to Qpeaks that the resolution of the peaks decreases further along the mass axis, as we are using a time-of-flight mass spectrometer. Choose the width of an isotopic envelope as the peak width. Our default value for singlet width in the low-mass range is "−400," and "−200" for the upper mass range. Place an unprocessed ASCII data file in an empty folder. Although it is not necessary, the corresponding calibrant file can be placed there as well. The data file should be representative of a majority of the data files. Navigate to the folder where the raw file is. In MATLAB's command window, type the following commands:

(a) "rawData = figure"
(b) "h = load ('〈insertfilename〉');"
(c) rawData = plot (h(:,1),h(:,2));

This will display the raw file as a plot. Create a parameter file and run Qcealignf as described in Steps 5–7. Use the default singlet width of "−400" for the low-mass range and "−200" for the upper mass range in the parameter file. Navigate to the folder "Final_ASCII_Spectra." In MATLAB's command window, type the following commands:

(d) "processedData = figure"
(e) "h = load ('〈insertfilename〉');"
(f) processedData = plot (h(:,1),h(:,2));

Compare the two plots to each other. If the baseline subtraction and smoothing are fine, navigate to the Peaklists folder and see if all the peaks are labeled correctly. If the signal processing is off, change the singlet width in the parameter file and rerun Qcealignf. If the peaks in the data are very sharp, use a larger number such as 500. Remember to add a negative sign in the front of the number to indicate TOF data. Thus, the "−400" would change to "−500." For broader peaks, the "−400" could change to "−300." Smaller increments than a hundred can be used as well. Use Steps (a–f) (see above) to examine the results. Repeat until the singlet width parameter is optimized for each mass range

We will enter "0〈tab〉0." This means we are using signal to noise ratio not peak height to determine our peaks. The next two numbers refer to the peak height threshold. As we want everything labeled, we will use a signal-to-noise ratio greater than 0 in each mass range. Thus, we will enter "0〈tab〉0." If we were using signal cutoff, these numbers would refer to the minimum peak height desired for peak labeling. The last line is used to turn on/off various aspects of signal processing: calibration, alignment, process and save calibrant spectra. The first number is to skip doing calibrations altogether. The second is to save aligned data files. The third is to save the calibrated calibrant spectra. The fourth is to only calibrate the calibrant spectra. The last is to indicate the prefix size. Prefix the number preceding the sample name. This prefix allows the script to associate the sample spectrum with its corresponding calibrant spectrum. Prefix size refers to the number of digits in the prefix. Thus, we enter "0〈tab〉1〈tab〉0〈tab〉0〈tab〉5." Repeat for other data sets. Close the file and save.

6| Take the entire Parent folder called "CancerStudy" and move it to the computer where Matlab is installed.

Spectra signal processing: Qcealignf

7| Launch Matlab. Click on the command window and type "h=waitbar(0,'Processing Spectra');" and then press enter. Next, type "qcealignf('','',h);" and then press enter. A dialog box will ask where the parameter file is. Navigate to the location of the parameter file and select it. Qcealignf will calculate the other needed parameters automatically and process the data sets, resulting in the generation of a spreadsheet termed "Processed01.out." This spreadsheet will contain samples in the columns and *m/z* peaks as rows with the corresponding intensities. See **Figure 4** for a diagram on how Qcealignf works and **Box 2** for

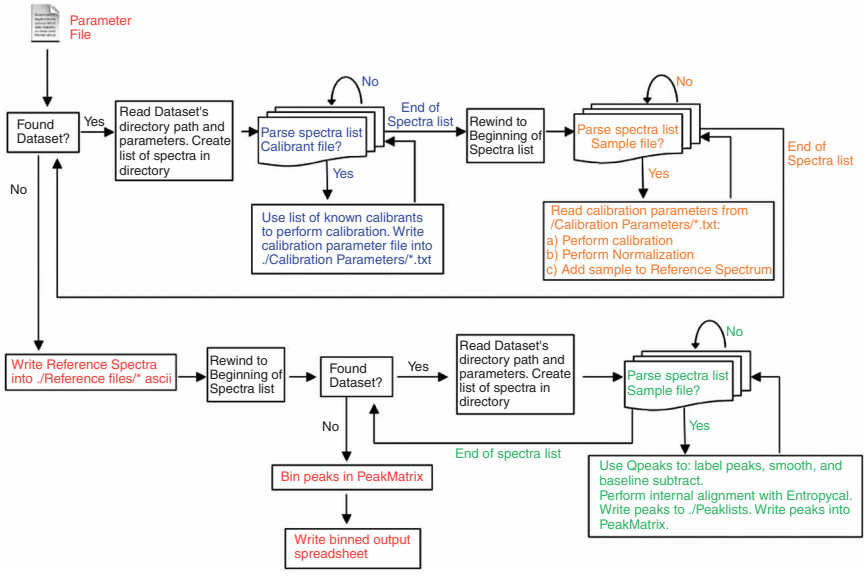


Figure 4 | Qcealignf workflow. Qcealignf automatically performs signal processing of raw spectra (containing samples and calibrant files). Its output is a spreadsheet termed "Processed01.out". This spreadsheet contains samples as columns and *m/z* peaks as rows, with the corresponding normalized intensities, and it is used as the input for statistical analysis. *Qcealignf* will also generate processed spectra traces that can be visualized using the MSV software.

BOX 2 | HOW QCEALIGNF WORKS

Open the parameter file. This file will be read twice. The first time through, each named data directory is processed to create reference files that will be subsequently used for internal mass alignment.
(1) Find the key word "Directory" to locate a data set. Read the directory name and associated parameters.
(2) LOOP #1 on "Directory" keyword: Create a list of all spectra in the named directory. When a calibrant spectrum is found in the directory, perform LOOP2 to create calibration curves. Repeat until all calibration spectra are processed for all data sets (designated by the keyword "Directory").

Use calibrant files to create calibration parameter files
(3) LOOP #2 ProcessCalFiles: Open the calibrant file and read the data. Select the appropriate peak width (low-mass or high-mass single width) for qpeaks processing of this file. Set signal to noise ratio (S/N) = 10. Set switches to perform baseline removal (baseline width = 3), omit Levenberg–Marquardt peak fitting (stopPercent = 50) and find all peaks with S/N $\geqslant$ 10 using peak finder #4. Use Qpeaks function rzrqpk, which (a) removes a baseline, (b) performs maximum entropy smoothing on the data, (c) finds all peaks which meet the S/N criterion, using Bayesian second derivative and (d) measures the peak positions, amplitudes, widths and areas. The results are returned to Qcealign in a peak table. The smoothed and baseline traces are also available for use. The list of the true mass positions of known calibrants is compared to the peak table to find the measured position of each of the known calibrants. Calibrant lists were True_masses = [782.46; 1,047.2; 1,297.51; 1,620.88; 2,094.46; 2,466.73; 3,149.61; 3,883.59] for low-mass files; True_masses = [4,282.945; 5,734.56; 6,181.048; 8,565.89; 12,361.088] for high-mass files. Vectors of Measured_masses and True_masses are created. Calibration consists of fitting the parameters of one of the following equations:

Equation 0: $\sqrt{\text{Calibrated\_mass}} = \sqrt{\text{Measured\_mass}} + b.$
Equation 1: $\sqrt{\text{Calibrated\_mass}} = a\sqrt{\text{Measured\_mass}}.$
Equation 2: $\sqrt{\text{Calibrated\_mass}} = \sqrt{\text{Measured\_mass}} + a\sqrt{\text{Measured\_mass}} + b.$

These equations fit the normal modes of a TOF spectrometer. Finding best-fit values for a and b is equivalent to finding C and t_0 in the TOF transform between acquisition time t and mass, $\sqrt{m} = C(t - t_o)$. Parameter a stretches the mass axis; b is a shift. The purpose of this external calibration is to align the files well enough that our subsequent internal calibration method (alignment with Entropycal) would be able to converge. Any one of the three equations seems good enough for this first-pass external calibration. Select calibration Equation 0. Use matrix methods to obtain $b \pm \Delta b$. Write b, Δb, Measured_masses, Calibrated_masses and True_masses of calibrants to the calibration parameters *.txt file, stored in the subfolder/Calibration_parameters. Repeat LOOP2 ProcessCalFiles until no more calibrant files are found in the current data set
(4) Go back to the first data set. Find a sample spectrum and go to LOOP3 to create the reference spectra needed for alignment.

Create reference files for internal calibration
(5) LOOP #3 SumDataFiles. Open the sample file and read the data. Open the appropriate file containing calibration parameters written out during LOOP2 ProcessCalFiles. Use the calibration parameters to compute a bin-shift that best fits the calibration parameters. Shift the mass axis of the sample file. Normalize the sample file by dividing it by the total ion current and scale it to 10^7, which is a user-defined setting. Add the rescaled, calibrated sample file into an accumulator. Go to the next sample in the filename list in the current data set. Repeat LOOP3 until all data sets are processed. Scale the outputs of the sample file accumulators by dividing by the number of sample files accumulated into each. Write out the scaled sum of all sample files in the low-mass range as a file named Reference_1; write the scaled sum of all sample files in the high-mass range as Reference_2. These files, stored in the subfolder/Reference_files, will be used for internal (self) calibration. Use current values of singletwidth_1 and singletwidth_2 as peak widths for Qpeaks processing of the Reference_1 and Reference_2 files. Use Qpeaks to calculate a baseline for the reference spectrum (baselinewidth = 3). Subtract the baseline from the Reference_1 and Reference_2 summed-data arrays. Store the baseline-subtracted Reference spectra in memory for later use by Entropycal.
(6) Go back to beginning of the Parameter file to find the first data set. Perform LOOP #4 on all data sets.

Perform calibration, smoothing, baseline subtraction, internal calibration and peak finding on sample files
(7) LOOP #4 ProcessDataFiles. Open the first sample file and read the data.
External calibration
(a) Open the appropriate file containing calibration parameters created during LOOP2 ProcessCalFiles. Use the calibration parameters to compute a bin-shift that best fits the calibration parameters. Shift the mass axis of the sample file.
Smoothing, baseline subtraction, peak finding with Qpeaks
(b) Select the appropriate peak width (depends on the mass range of the spectrum) for Qpeaks processing of this file. Use the signal-to-noise (S/N) value read from parameter file. Set switches to perform baseline removal (baselinewidth = 3), omit Levenberg–Marquardt peak fitting (stopPercent = 50) and find peaks using peak finder #4. Call the Qpeaks function rzrqpk to (i) remove a baseline, (ii) perform maximum entropy smoothing on the data, (iii) find all peaks that meet the S/N criterion, using a Bayesian second derivative and (iv) measure the peak positions, amplitudes, widths and areas. The results are returned to Qcealign in a peak table. The smoothed and baseline traces are also available for use. Subtract the baseline from the sample data.

Internal calibration using reference spectra and Entropycal
(c) Using the baseline-subtracted sample data array, and the baseline-subtracted Reference_1 or Reference_2, call Entropycal. Entropycal shifts the time axis of the data array relative to the reference array, to find the relative shift (alignment) that produces a minimum entropy result. This entropy calibration, added to the previous external calibration, then creates the internal mass calibration for the sample data. The sample data array is returned from Entropycal with the internal calibration applied. No additional steps are required.

BOX 2 | CONTINUED

Compute correction for peak table positions

(d) The positions listed in the peak table returned from Qpeaks were based on the original, external calibration of the sample data files. The additional calibration, performed by Entropycal alignment to the reference spectrum, must be added to all positions in the peak table. Apply this correction.

Write peaklists

(e) Using the peak table, write out a list of peaks. The list contains the calibrated centroid position of each peak (column 33 of Qpeaks peak table plus Entropycal correction), and the height (i.e., intensity) of the peak in the smoothed data trace, measured at the apex position taken from column 26 of Qpeaks peak table. The height is adjusted so that the total area of the original data file is 10^7, which is a user-defined setting. The peaklist is written into the subfolder /Peaklists.

Write out processed sample files

(f) Write the name of the processed sample file into a log file. Write the processed, smoothed, baseline-subtracted, rescaled data into the subfolder /Final_ASCII_Spectra.

Bin the peaks and write into a spreadsheet

(g) Write the sample filename into the first row of a spreadsheet named XL. Write the internally-calibrated peak positions (peak table column 33 plus Entropycal correction) and scaled amplitudes into the XL spreadsheet. Amplitudes are measured on the smoothed, baseline-subtracted traces, at the positions given in column 26 of the peak table. Amplitudes are scaled according to the MyScalingFactor as above. Excel contains one row for each dalton in mass. Peak amplitudes are written into the row closest to the calibrated mass position.

(h) Go to the next sample file. Repeat LOOP4 for all data sets.

Write output spreadsheet

The spreadsheet contains one column for each processed sample file, plus one column for the summed peak amplitudes. The number of rows is proportional to the *m/z* range of all the data, namely $m/z = 400–10{,}000$ will give $\sim 10{,}000$ rows. Compress the spreadsheet by deleting rows in which there are no peaks. In addition, when peaks from different sample files are clearly similar in position (i.e., their apexes are spaced not farther apart than the full-width at half-maximum), then collapse the spreadsheet into a single row for these peaks, using the following logic. Find rows of peak maxima in the summed-amplitudes column; find rows of the minima between peaks in the summed-amplitudes column. Note that a plot of the amplitude column should look very similar to the summed reference spectrum created in **Box 2**, Step 5. All peaks in other columns that fall into rows between the minima rows below and above a summed-amplitude maximum row are collapsed into the maximum row. Write out the final spreadsheet. This output file is given the same base name as the input *.par parameter file obtained from the user; the extension is changed to *.out. The final output file is written into the same directory as the input *.par parameter file.

its detailed description. ● **TIMING** This process takes about 120 min for 250 samples (four spectra per sample) using the type of computer described in Step 2.

? TROUBLESHOOTING

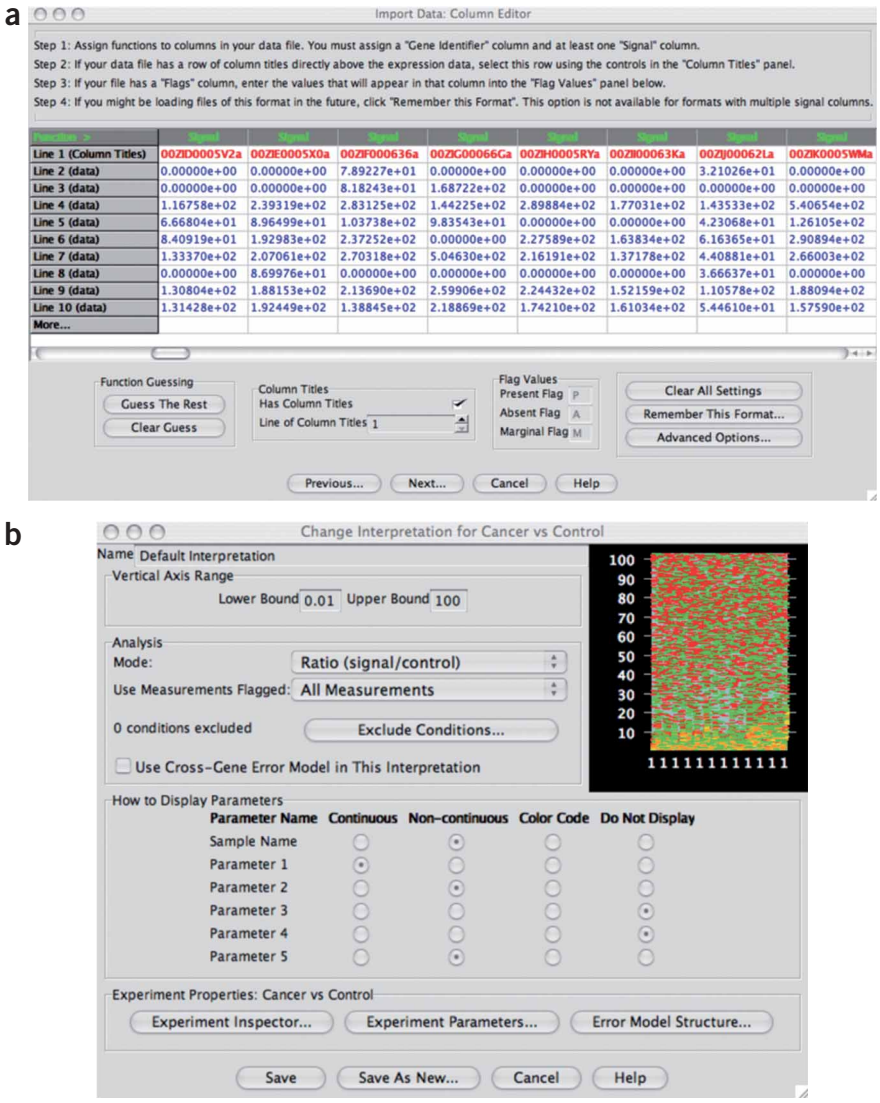
Matching clinical information to the processed spectra

8| Clinical information is stored in a dedicated database. Owing to privacy concerns and HIPAA (Health Insurance Portability and Accountability Act) regulations, all identifying information is removed. We typically receive information in a spreadsheet with a patient ID number and corresponding clinical parameters. As each sample enters the laboratory, it is assigned a sample ID, and its patient ID and corresponding clinical parameters are saved in a LIMS database system[21]. To perform statistical analysis, generate a custom report to match each spectrum and peaklist (identified by sample ID) with its appropriate clinical parameters. To do so, save this report as a spreadsheet where each sample ID is listed in rows and the corresponding clinical data in columns. This spreadsheet is called the clinical definition file and saved as "Processed01.def."

Importing data and creating experiments in GeneSpring

9| In the same location as the parameter file, Qcealignf will have created a "processed01.out" file. This is the binned, aligned peaklist for all samples.

10| Launch GeneSpring. From the File Menu, select "Import Data." In the resulting dialog window, navigate to and select the "processed01.out" file. This is the binned, aligned peaklist for all samples of the project; it is in the same location as the parameter file and it is created by Qcealignf (see **Fig. 2a**). A dialog will appear asking for the name. Click "Create a New Genome" and enter a name for the data set. For this example, we will use "CancerStudy," which is the name of the parent folder where the spectra are stored. Then click "Next." Another window appears. The first column should be set to "Gene Identifier." The remaining columns should be set to Signal. Also, select the checkbox "Has Column Titles" and make sure the "First Line of Column Titles" is set to 1 (see **Fig. 5a**). Then, click next to dismiss this window. Another window will appear asking to import more data. As we have none, click next again. A window will appear, saying how many samples are created. Click "Yes" to continue. Then, GeneSpring will create an experiment. When you hit "Next" to continue to create an experiment, a warning may appear saying that a column lacks a title. This means that there is an empty column. Scroll through and locate the empty column (usually the last column) and change the column designation from "Signal" to "Unused."



Line	Signal	Signal	Signal	Signal	Signal	Signal	Signal	Signal
Line 1 (Column Titles)	00ZID0005V2a	00ZIE0005X0a	00ZIF000636a	00ZIG00066Ga	00ZIH0005RYa	00ZII00063Ka	00ZIJ00062La	00ZIK0005WMa
Line 2 (data)	0.00000e+00	0.00000e+00	7.89227e+01	0.00000e+00	0.00000e+00	0.00000e+00	3.21026e+01	0.00000e+00
Line 3 (data)	0.00000e+00	0.00000e+00	8.18243e+01	1.68722e+02	0.00000e+00	0.00000e+00	0.00000e+00	0.00000e+00
Line 4 (data)	1.16758e+02	2.39319e+02	2.83125e+02	1.44225e+02	2.89884e+02	1.77031e+02	1.43533e+02	5.40654e+02
Line 5 (data)	6.66804e+01	8.96499e+01	1.03738e+02	9.83543e+01	0.00000e+00	0.00000e+00	4.23068e+01	1.26105e+02
Line 6 (data)	8.40919e+01	1.92983e+02	2.37252e+02	0.00000e+00	2.27589e+02	1.63834e+02	6.16365e+01	2.90894e+02
Line 7 (data)	1.33370e+02	2.07061e+02	2.70318e+02	5.04630e+02	2.16191e+02	1.37178e+02	4.40881e+01	2.66003e+02
Line 8 (data)	0.00000e+00	8.69976e+01	0.00000e+00	0.00000e+00	0.00000e+00	0.00000e+00	3.66637e+01	0.00000e+00
Line 9 (data)	1.30804e+02	1.88153e+02	2.13690e+02	2.59906e+02	2.24432e+02	1.52159e+02	1.10578e+02	1.88094e+02
Line 10 (data)	1.31428e+02	1.92449e+02	1.38845e+02	2.18869e+02	1.74210e+02	1.61034e+02	5.44610e+01	1.57590e+02
More...								

Figure 5 | Data import and interpretation for *GeneSpring*. (**a**) Screenshot of the data import function in GeneSpring. The imported data are contained in the ".out" spreadsheet generated by Qcealignf. (**b**) Screenshot of the experiment interpretation function in GeneSpring.

11| To create an experiment, first type a name for the experiment and hit "Save." We will call it "CancerStudy" for our data set. Next, a window appears to set the different statistical properties of the experiment. Click on "Normalizations." As Qcealignf already normalized the data, we need no normalizations (see **Box 2**). Remove all default normalizations using the "delete" button and press "OK." Then, click on the "Parameters" button. In the appearing window, enter the parameters you want to study. Click on "New Parameter" and enter the clinical information for each sample. Enter "Parameter1" as the name for the column, accepting the default column properties. If the Parameter is Numeric (i.e. body weight), then set "Numeric" to "Yes," otherwise leave it as "No." Similarly, set the value for "Logarithmic" to Yes or No, depending on whether the data are in log scale. One can enter many parameters with which to analyze the data. Copy the parameter information from the spreadsheet obtained in Step 8. Press "Save" when done.

12| Next, click on "Experiment Interpretation" (see **Fig. 5b**). In the resulting dialog, for the Interpretation named "Default Interpretation," set Mode to "Ratio (signal/control)." Select how to display the Clinical Parameters. Parameters such as body weight will be continuous. Parameters such as Gender will be non-continuous. We typically display the parameter of interest (in this cone, Parameter 1 as non-continuous) and the Sample Name (as non-continuous). Make sure "Use Cross-Gene Error Model in this Interpretation" is unchecked. No conditions should be excluded. Press "Save" to continue. As we do not use the Error Model, skip the "Error Model" button and press "Close."

Unsupervised analysis: hierarchical clustering

13| From the Tools Menu, select "Clustering" and select "Gene Tree."

14| The settings should be the following:

Choose "All Genes"
Choose the Default Interpretation of the Training Experiment
Use "Pearson Correlation" and "Average Linkage"
Click on "Calculate Confidence Intervals," if desired

Then, press "Start" and save the Results.

15| Then go back to the Tools Menu, and this time select "Clustering" and select "Condition Tree" in the Tab of the Clustering dialog box. Keep "Pearson Correlation" and "Average Linking" and select "Calculate Confidence Levels," if desired. Then, press

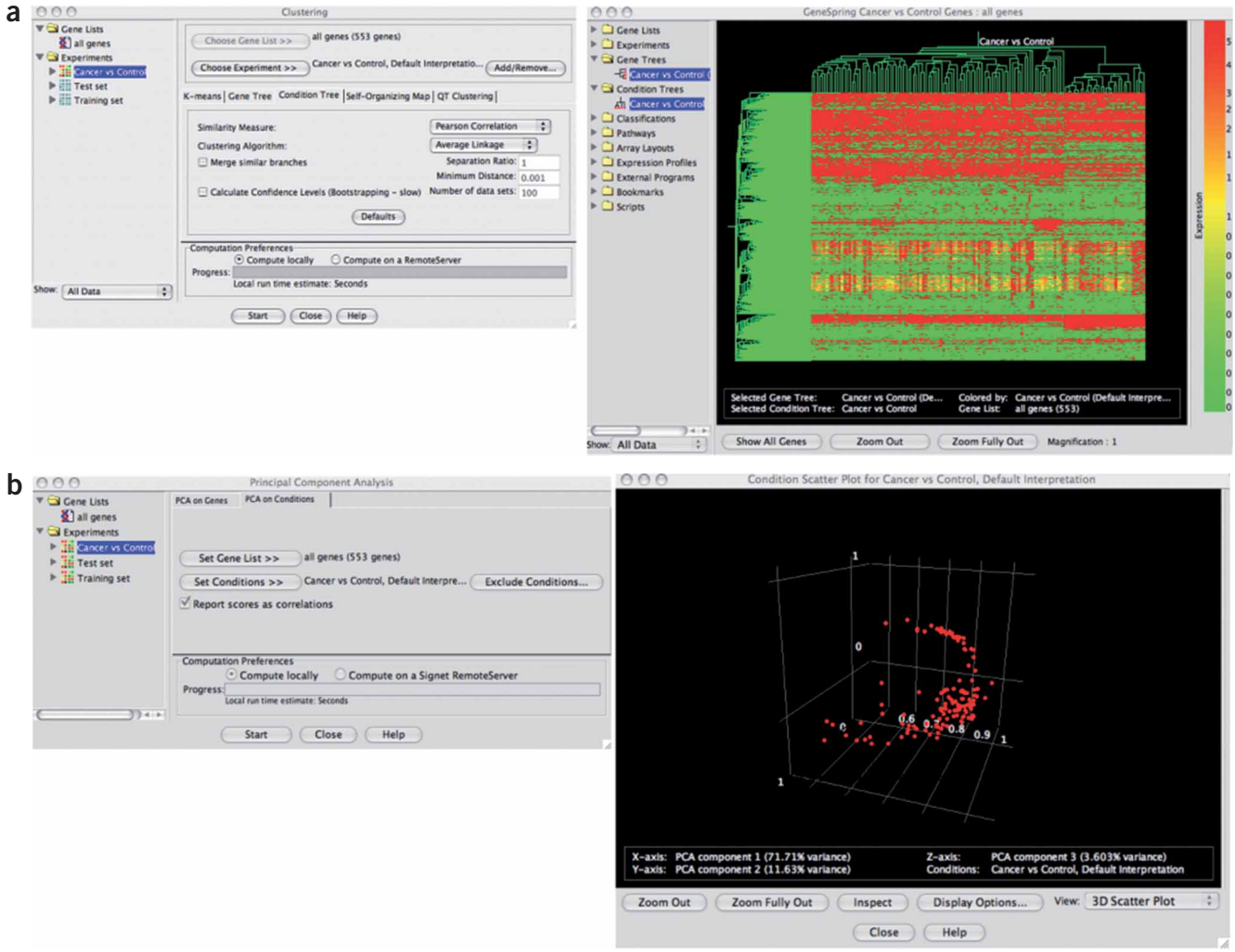


Figure 6 | Unsupervised statistical analysis. (**a**) Hierarchical clustering analysis. (**b**) PCA.

"Start" and save the Results (see **Fig. 6a**). ● TIMING This step takes 15 s to perform using the type of computer described in Step 2.

Unsupervised analysis: PCA

16| From the Tools Menu, Select "Principal Component Analysis." Then select "PCA on Conditions" from the tabs. Set the gene list to "All Genes." Set the conditions of the experiment and press "Start."

17| The next window will have the main components listed. Save Scores and Profiles, if desired. Then press "Close." The next window will have the main components mapped on an XYZ grid as a scatterplot. Press Display Options. Select the "Coloring" tab. Set parameter to be "Parameter1." Press "OK." The resulting scatterplot will show the samples colored by the chosen parameter. Zoom in and out using the buttons given below. Use Option-click to rotate the screen (see **Fig. 6b**). ● TIMING This step takes 30 s to perform using the type of computer described in Step 2.

Setting up training and test sets

18| From the Experiment Menu, select "Create New Experiments." Select the "Filter on Parameter" tab. One of the parameters added in Step 11 contained information about which samples were part of training or test set. In that parameter, the training set samples were labeled as "Training," whereas the additional test set samples were marked as "test."

19| From the "Filter On Parameter," we select the parameter value "Training" and click on "Add All." This adds all the samples from the training set to a new experiment. Click "Next." A window appears. Click "Import Parameter." Select the previous experiment. All the parameters from that experiment will appear. Click "Select All" and then "OK." The values for the training samples

will be moved into this experiment. Press "Next." A window asking about normalizations appears. Remove all normalizations as before and press "Next." A window asking for a name for the experiment appears. We name this set "Training." All dialogs disappear and the main window reappears.

20| With the new "Training" experiment selected, from the Experiment window, select "Experiment Interpretation" (see **Fig. 5b**). Change mode to "Ratio (signal/control)."

21| We repeat Steps 18–20 to create a new experiment called "Test" where the samples labeled "Test" are moved in with their corresponding parameter information.

Feature selection

22| From the Tools Menu, select "Statistical Analysis (ANOVA)." Make sure it indicates "all genes" next to the "Choose Gene list" button. Choose the default Interpretation for the Training experiment just created. Then click "Choose Experiment" and the name should appear next to the button. Mode should be "Ratio (signal/control)." Make sure that Cross-Gene Error Model is inactive (click next to dismiss this window). Click on the "Parameter to Test" pull-down menu and select the desired parameter. The settings should be the following (see **Fig. 7a**):

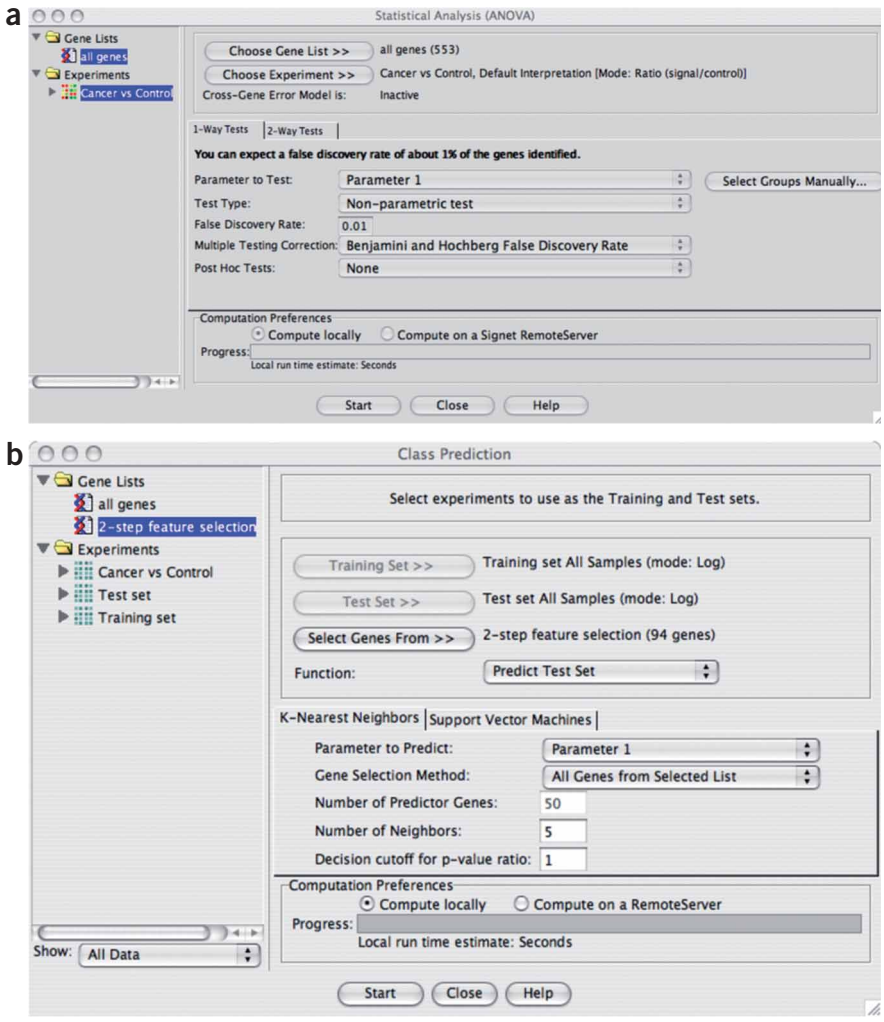


Figure 7 | Supervised statistical analysis. (**a**) Screenshot showing the Mann–Whitney *U*-test setup in GeneSpring. (**b**) Screenshot showing a k-NN class prediction setup.

Test Type should be set to "Non-parametric" test
False Discovery Rate should be set to 0.05
Multiple Testing Correction should be set to "Benjamini and Hochberg False Discovery Rate"
There should be no *post hoc* tests

▲ **CRITICAL STEP** If the Experiment and Cross-Gene Error Model values are not correct, click "Close." Then go back to the Experiment Interpretation window and set the proper values. One can do this by selecting "Experiment Interpretation" from the "Experiment" Menu. Then return to Step 22.

23| Click "Start" and Save your results (called a gene list) with an appropriate name. We will call our "CancerStudy_p05" and hit "Save" to record the results. The *P*-value can be changed to a more stringent value as needed. We generally save a gene list from the results of $P < 0.00001$. ● **TIMING** This step takes 30 s to perform using the type of computer described in Step 2.

Feature selection: peak height cutoff

24| Once Processed01.def and Processed01.out files are created, perform a second feature selection step based on normalized *m/z* peak height (i.e., peptide-ion intensity). This cutoff value was ascertained experimentally as follows. MS/MS was performed on a variety of peaks ranging from 100 to 2,000 normalized ion intensity units. Peaks taller than 500 units were robust and generally gave good MS/MS results whereas peaks smaller than 500 units were rather unreliable and gave mostly poor MS/MS

spectra. This cutoff was used to filter peaks from "Processed01.out." We expect that this threshold will be different for each mass spectrometer and it must, therefore, be experimentally determined in each case. To remove masses below the threshold, a median height must be calculated for each mass across all the samples within a group. Thus, if Gender is the parameter of interest, then a median intensity is calculated for each mass for all men and then for all women. If the calculated intensity is not higher than 500 in any of the clinical groups in the parameter of interest, then the mass is removed. This results in a mass list filtered for intensity.

25| To generate this list, make sure that the Processed01.def and Processed01.out are in the same folder. Following our example, they are in the Processed01 folder. Then, type "calcMedians" in the command window in Matlab. This should bring up a window. Press 'Load Data' in the window to select the processed01.def file. A list of clinical parameters will appear. Select the parameter that the median value should be calculated for and enter a cutoff filter (default is 500). Then press Run. A new file will be created in the same location as the processed01.def file. This file will contain all masses whose median value across a data set is greater than or equal to the cutoff. Additional information such as count, standard deviation, max and min are also listed for each mass. This tab-delimited file can be opened in any text editor or Microsoft Excel, which can be used to sort the masses.

26| To merge the "ion intensity"-filtered list with the "*P*-value"-filtered list, select the gene list in GeneSpring saved in Step 23. When this is done, select 'Copy Gene List' from the 'Edit' menu. Then, select 'Paste Gene List' from the "Edit" menu. A window will appear with a list of masses. Remove any mass that does not appear in the list generated in Step 25. The resulting list will be filtered for *P*-value and minimum peak height. In most cases, the filtering by *P*-value is very stringent, thus making it simple to manually delete masses eliminated by ion intensity. If so desired, save this list as "two-step feature selection."

Class prediction

27| In the 'From the Tools' Menu, select "Class Prediction." Set the Training Experiment to be the Training Set. Working in the K-Nearest Neighbors tab, select the "Parameter to Predict." We select "Parameter 1." Set Gene Selection Method to be "All Genes from Selected List." Select the Gene List saved from Step 26. Set the decision cutoff for *P*-value ratio to 1 (see **Fig. 7b**). Press "Start." To save the results from the pop-up window, copy and paste into an Excel spreadsheet. To optimize the results with the training set, repeat by varying the number of neighbor from 3 to 9. Once optimal conditions (as judged by the lowest prediction errors) are found through iterative crossvalidation using the training set, select the "Test" experiment and set it as the "Test Set." Change function from "Crossvalidate Training Set" to "Predict Test Set." Keeping the values for optimal conditions found from the Training Set, press "Start." Save the Results in an Excel spreadsheet. ● **TIMING** Each k-NN run (for 250 samples and 600 masses) takes about 1 min using the type of computer described in Step 2.

28| For Support Vector Machine, select the second tab in the "Class Prediction" window. Keeping the same training set, change function to "Crossvalidate Training set." Set "Parameter to Predict" to the appropriate parameter. We select "Parameter 1." The Gene list created in Step 26 should be the one used as before. Gene Selection Method remains "All genes from Selected List" unchanged. Press "Start." Save the results in an Excel spreadsheet by copying and pasting. Optimize the conditions for the training set by varying the kernel function. Occasionally, it might be necessary to change the scaling factor from 0 to 1 or 2 (this depends on how balanced the number of samples are in the groups tested). Once optimal values are found, then set the Test to the "Test" Experiment. Change the Function to "Predict Test Set." Save the results in an Excel spreadsheet. Press "Close" to exit Class Prediction. ● **TIMING** Each SVM run (for 250 samples and 600 masses) takes about 1 h using the type of computer described in Step 2.

Processed spectra overlay (MSV)

29| Create a viewer definition file. This tells the viewer which samples are in which clinical group (see **Supplementary Method 1**). The definition file needs three columns and looks like the following example. The second column does not use "_1" or "_2"—just the basefile name. The viewer automatically adds the suffix when looking for the ASCII files. No spaces are to be used anywhere. The columns are tab-separated:

1	000ZG60005DVA	Cancer1	Parameter 1	Parameter 2
2	000ZG70007DCK	Cancer1	Parameter 1	Parameter 2
3	000ZG80003VKD	Control	Parameter 1	Parameter 2
4	000ZG90002EGO	Control	Parameter 1	Parameter 2

30| Go to Matlab and type "MSV" in the command window (see **Fig. 8**). Click on "Folder Setting." A window appears. Press "Browse" next to the "Folder of ASCII files" and select the directory holding the processed ASCII data files. It should be called "Final_ASCII_Spectra" according to our data structure and located inside the Processed01 folder. Press the second "Browse" button and select the "Matrix" folder, also in the Processed01 folder. Press the last "Browse" button and select the viewer

definition file that was created in the previous step. Then press "Save and Return." Then press "Make Matrix." Once that is done, press "Group Color." A dialog box will appear. The clinical subgroups will appear in the left box. Select each item. Pick a color by scrolling up and down the color bar. Once the appropriate color is picked, press the "Select" button to assign that color to that subgroup. Repeat for each subgroup. Once all the groups are given a color, press "Save and Return." Once the main screen returns, press Update. Refer to the manual for additional instructions (see **Supplementary Method 1**). Zoom in on peaks selected for the clinical parameter and verify the results.
? TROUBLESHOOTING

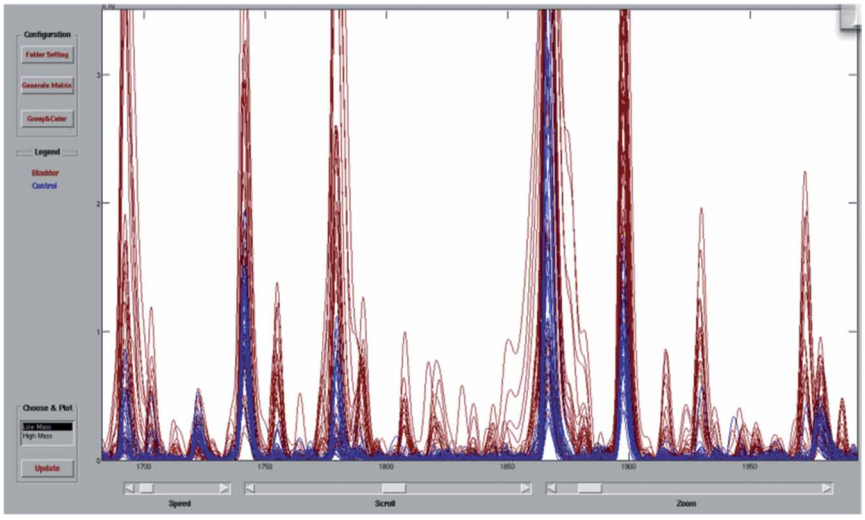


Figure 8 | MSV. This software allows to display and color-code overlays of MALDI-TOF mass spectra processed using Qcealignf.

31| Repeat Step 30 for each additional clinical parameter that was used during the statistical analysis (Steps 9–28).

? TROUBLESHOOTING
Troubleshooting advice can be found in **Table 1**.

TABLE 1 | Troubleshooting table.

Step	Problem	Possible reason	Solution
Converting raw data to ASCII	Macro does not appear in the tools menu	Macro is installed in the wrong folder	Find where the other macros are installed. Typically they are in "C:\Methods\ FlexAnalysisMacroModules". But setup may vary. But if other items are appearing under the Tools menu, this macro should be placed in the same location as the other items
Spectra signal processing: Qcealignf	Qcealignf does not run	There is an empty line at the beginning of the parameter file	Empty line tells Qcealignf that there are no more data to be processed. Thus, each line of the parameter file should have text—either a command or a comment. Remove all empty lines until the end of the file
		Matlab does not know where Qcealignf is	Follow directions in the setup to reinstall Qcealignf and associated files
		Parameter file is using spaces to separate items instead of tabs	All processing parameters in the parameter file are separated by tabs and not spaces. Do a search to remove unnecessary spaces in the parameter file. Refer to **Figure 3** to see the format of the parameter file
		Comments do not have the "#" at the beginning of the line	When Qcealignf reads a "#" at the beginning of a line, it ignores the rest of the line. If this "#" is missing, Qcealignf will not know how to interpret the line and may mistake it for a data path, causing the script to fail
Processed spectra overlay (MSV)	Typing "masspectraviewer" does not launch the viewer	Matlab does not know where the viewer is installed	Follow directions in the setup to reinstall the MSV and its associated files
	The Folder field shows "0" instead of the path to the folder in the "Folder Setting" dialog	The original folder was moved No folder was selected. The "cancel" button in the Browse button was pressed	Reselect the folder by pressing the "Browse" button

TABLE 1 | Troubleshooting table (continued).

Step	Problem	Possible reason	Solution
	"Group Color" does not show the clinical subgroups	The viewer definition file could not be found or is the wrong format	Create a viewer definition file as described in Step 29 (see also **Supplementary Method 1**)
		The viewer definition file is the wrong format	Check the definition file to make sure that tabs are used instead of spaces
			Make sure that all comments are preceded by the "#" character
			Make sure that there is no extra carriage return until the last data set is defined
	Clinical subgroup is of the wrong color or not displayed in the legend	Subgroup has not been selected	Be sure to click on the subgroup and set the color. Then hit the "Select" button

ANTICIPATED RESULTS

A critical aspect of serum proteomics data analysis as described here is the selection of the singlet width parameter. When the singlet width is properly optimized, the smoothing, baseline subtraction, peak labeling and alignment will also be optimized. In many uses, one would normally pick a singlet width that is the width of an isotopic peak. This would properly preserve the isotopic structure of the data, which is needed in many mass spectrometric techniques. However, for the methodology described here, the serum spectra contain peaks with full isotopic resolution (for peptides with m/z below 2,000 amu) whereas the rest of the peaks ($m/z > 2{,}000$ amu) do not have that high resolution. Therefore, we purposely pick a singlet width that represents an isotopic envelope rather than an isotopic singlet (**Fig. 9**). This causes the isotopes to merge by aggressive smoothing, reducing the number of peaks and thus the complexity of the data. Data analysis is unaffected however as Qpeaks and Entropycal still work with the unsmoothed raw spectra to perform their functions.

In choosing the proper singlet width, a few issues need to be considered. The ideal singlet width results in the construction of a proper baseline that will show the bottom of major peaks being connected from one to another with no major gaps. In general, if the singlet width is too wide (e.g., –200 instead of –400), the baseline drops down too low. If the singlet width is too narrow (e.g., –800 instead of –400), the baseline forces itself up under the peaks (see **Fig. 10**). One can also consider how much smoothing is needed. **Figure 10** also shows the effect of the singlet width on the ion intensity and resolution of the resulting processed data: an incorrect singlet width will cause a decrease of the ion intensity and a lower peak resolution (measured here using the peak width). To do this, it is best to pick peaks that are not near other peaks, thus minimizing confusion resulting from overlapping isotopic

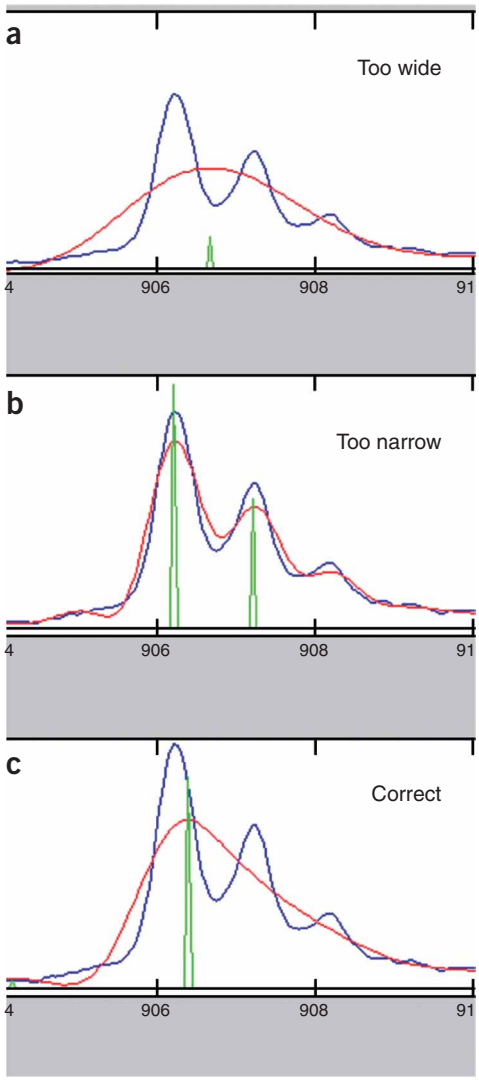
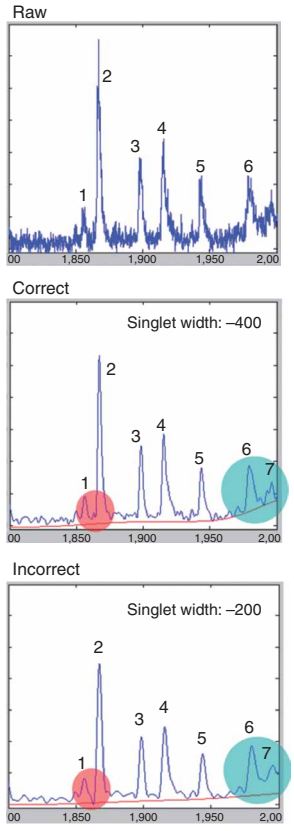


Figure 9 | Optimization of the singlet width. (**a**) When the selected singlet width is too wide, the peaks are flattened and the base of the smoothed peak is broader than that of the isotopic envelope. (**b**) When the singlet width is too narrow, multiple peaks in the isotopic envelope will be labeled. (**c**) When the singlet width is correct, the smoothed peak has almost the same width as the isotopic envelope and only one peak is found and labeled. The blue trace represents the raw spectra, the red trace represents the smoothed spectra and the green trace represents the peaks found by the peak labeling algorithm.

envelopes. Once an isolated peak is found, adjust the singlet width such that the tops of the isotopic peaks are merged but the base of the envelope does not broaden (**Figs. 9** and **10**). One can do this visually using the smoothed data, but it is best to use the peaklists instead. With a narrow singlet width, all the isotopes will be labeled. Slowly increase the value of the singlet width such that the isotopes are no longer labeled and replaced by a single labeled peak (**Fig. 9**). Although it is important to pick an optimal singlet width, Qpeaks and Entropycal are somewhat forgiving if the singlet width is near the optimal value (±20%).

The use of Entropycal is necessary to accommodate the use of external samples for calibration of the sample spectra. Calibration with peaks from within the sample would result in much better calibration and could preclude the need for Entropycal. But it is not possible to guarantee the presence of a given peak in every sample due to the complexity of the serum peptidome. Adding calibrant peptides to serum is problematic because this calibrants may affect the signal of the serum peptides in the sample. Entropycal resolves the issue of poor calibration by aligning the spectra to a reference spectrum. The reference spectrum is created by summing all the sample spectra of a project. As each sample contributes to the reference, Entropycal can align by finding common features between the reference spectrum and the sample spectrum. Once the commonality is established, the sample spectrum is adjusted to minimize the differences between it and the reference. Thus, in essence, Entropycal acts as internal calibration, which in turn results in better binning. With just external calibration, the binning of adjacent peaks becomes more difficult and results often in many bins. With better calibration (in other words, alignment), many of the adjacent peaks move closer to each other, allowing for an easier placement of peaks in bins (see **Fig. 11**). As a result, fewer bins are created with more accurate peak information in each bin. Better bins, in turn, lead to better statistical results as well.



	Mass	Norm. intensity	Peak width
1	1,856.2009	894.098	3.1694
2	1,867.5742	5,222.81	3.17899
3	1,898.7730	2,366.41	3.71898
4	1,916.1888	2,722.08	3.73565
5	1,944.2415	1,658.36	3.93105
6	1,980.5238	1,420.06	5.17008
7	1,996.2808	611.91	2.36432

	Mass	Norm. intensity	Peak width
1	1,856.2019	792.158	4.20867
2	1,867.4984	4,410.61	3.92765
3	1,898.6674	2,028.34	3.98914
4	1,916.5581	2,324.23	5.03787
5	1,944.2835	1,425.32	4.46687
6	1,980.6297	1,573.39	4.478
7	1,995.7735	904.367	4.49764

Figure 10 | Effects of the singlet-width parameter on ion intensity and resolution. Incorrect singlet-width selection will cause a decrease in the ion intensity and a lower peak resolution (lower panel).

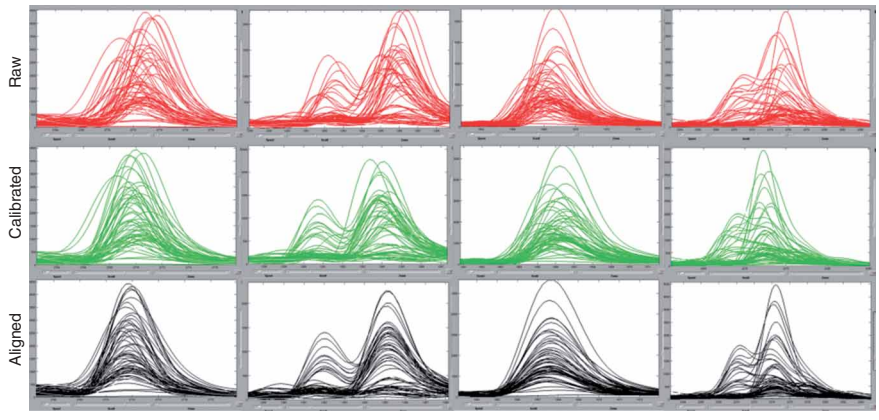


Figure 11 | Effects of mass calibration and 'Entropycal'-based alignment on mass spectral overlays. Mass spectra of 59 serum samples obtained from healthy control individuals and from thyroid carcinoma patients are shown in overlay. All spectra have been smoothed and baseline subtracted using the signal processing described in this protocol. Different mass calibrations were applied. Four regions of the spectra were selected and displayed using the MSV. Each spectral region is shown in a raw version (Raw), after external calibration (Calibrated) and after external calibration plus computer 'Entropycal' alignment (Aligned). External calibration and 'Entropycal'-based alignment are described in the protocol. Reprinted with permission from ref. 24. Copyright © 2005 American Chemical Society.

Regardless of the signal processing routines and improved statistical methods used, it is important to visually inspect and confirm the results. For instance, once a list of *m/z* peaks is obtained after feature selection, they should always be examined using the MSV. In case of a list of peptides with good (i.e., low) *P*-values, the overlays in the viewer should show clear differences between clinical groups; that is, selected peaks in the spectra from one subgroup will overall be either higher or lower than those of another subgroup (**Fig. 12**). We use conservative nonparametric statistics to calculate *P*-values for each peak, which reduce intensity information to ranks. Whereas this transformation reduces the effect of very large peaks and other such outliers, it may also minimize the actual differences between groups in the clinical parameter

being analyzed. Thus, visual inspection of the peaks surviving feature selection may reveal stronger differences than the statistics suggested. Conversely, the viewer can also serve as an error-check for signal processing. If a peak with low *P*-value shows negligible difference in the spectral overlays, there might have been an error in processing. Peaks will often give low *P*-values (and good hierarchical clustering) if calibration/alignment is bad or binning was poor, resulting in an inflated number of bins, a random few of which will inevitably show considerable differences between groups. In these cases, the selected singlet width was typically too narrow and the data will need to be reprocessed. If, on the other hand, only a few peaks survive feature selection but clear differences are visible in the overlays, the singlet width was too wide and needs to be decreased. Only when the outcome of the statistical analysis is visually confirmed, one can be sufficiently confident about the results.

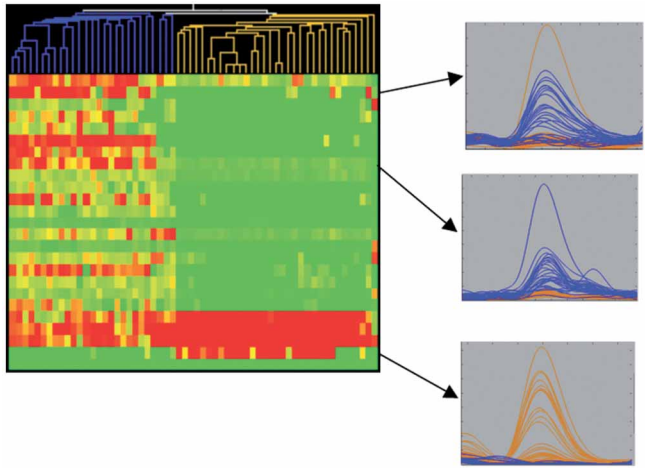


Figure 12 | Confirmation of statistical results using color-coded mass spectra overlays.

Note: Supplementary information is available via the HTML version of this article.

COMPETING INTERESTS STATEMENT The authors declare that they have no competing financial interests.

Published online at http://www.natureprotocols.com
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions

1. Qian, W.J., Jacobs, J.M., Liu, T., Camp, D.G. & Smith, R.D., 2nd. Advances and challenges in liquid chromatography–mass spectrometry-based proteomics profiling for clinical applications. *Mol. Cell Proteomics* **5**, 1727–1744 (2006).
2. Villanueva, J. *et al.* Differential exoprotease activities confer tumor-specific serum peptidome patterns. *J. Clin. Invest.* **116**, 271–284 (2006).
3. Villanueva, J. *et al.* Serum peptidome patterns that distinguish metastatic thyroid carcinoma from cancer-free controls are unbiased by gender and age. *Mol. Cell Proteomics* **5**, 1840–1852 (2006).
4. Issaq, H.J., Conrads, T.P., Prieto, D.A., Tirumalai, R. & Veenstra, T.D. SELDI-TOF MS for diagnostic proteomics. *Anal. Chem.* **75**, 148A–155A (2003).
5. Petricoin, E.F. & Liotta, L.A. SELDI-TOF-based serum proteomic pattern diagnostics for early detection of cancer. *Curr. Opin. Biotechnol.* **15**, 24–30 (2004).
6. Koomen, J.M. *et al.* Direct tandem mass spectrometry reveals limitations in protein profiling experiments for plasma biomarker discovery. *J. Proteome Res.* **4**, 972–981 (2005).
7. Richter, R. *et al.* Composition of the peptide fraction in human blood plasma: database of circulating human peptides. *J. Chromatogr. B* **726**, 25–35 (1999).
8. Gao, J., Opiteck, G.J., Friedrichs, M.S., Dongre, A.R. & Hefta, S.A. Changes in the protein expression of yeast as a function of carbon source. *J. Proteome Res.* **2**, 643–649 (2003).
9. Fach, E.M. *et al. In vitro* biomarker discovery for atherosclerosis by proteomics. *Mol. Cell Proteomics* **3**, 1200–1210 (2004).
10. Wang, W. *et al.* Quantification of proteins and metabolites by mass spectrometry without isotopic labeling or spiked standards. *Anal. Chem.* **75**, 4818–4826 (2003).
11. Li, X.J. *et al.* A tool to visualize and evaluate data obtained by liquid chromatography-electrospray ionization-mass spectrometry. *Anal. Chem.* **76**, 3856–3860 (2004).
12. Chen, S.S. *et al.* Improving mass and liquid chromatography based identification of proteins using bayesian scoring. *J. Proteome Res.* **4**, 2174–2184 (2005).
13. Silva, J.C. *et al.* Quantitative proteomic analysis by accurate mass retention time pairs. *Anal. Chem.* **77**, 2187–2200 (2005).
14. Jaitly, N. *et al.* Robust algorithm for alignment of liquid chromatography–mass spectrometry analyses in an accurate mass and time tag data analysis pipeline. *Anal. Chem.* **78**, 7397–7409 (2006).
15. Wang, P. *et al.* A statistical method for chromatographic alignment of LC–MS data. *Biostatistics* (2006).
16. Gillette, M.A., Mani, D.R. & Carr, S.A. Place of pattern in proteomic biomarker discovery. *J. Proteome Res.* **4**, 1143–1154 (2005).
17. Adam, B.L. *et al.* Serum protein fingerprinting coupled with a pattern-matching algorithm distinguishes prostate cancer from benign prostate hyperplasia and healthy men. *Cancer Res.* **62**, 3609–3614 (2002).
18. Yanagisawa, K. *et al.* Proteomic patterns of tumour subsets in non-small-cell lung cancer. *Lancet* **362**, 433–439 (2003).
19. Tibshirani, R. *et al.* Sample classification from protein mass spectrometry, by "peak probability contrasts". *Bioinformatics* **20**, 3034–3044 (2004).
20. Villanueva, J. *et al.* Serum peptide profiling by magnetic particle-assisted, automated sample processing and MALDI-TOF mass spectrometry. *Anal. Chem.* **76**, 1560–1570 (2004).
21. Villanueva, J., Lawlor, K., Toledo-Crow, R. & Tempst, P. Automated serum peptide profiling. *Nat. Prot.* **1**, 880–891 (2006).
22. DeNoyer, L. & Dodd, J. *Smoothing and Derivatives in Spectroscopy*. Vol. 3 (John Wiley and Sons, Chichester, UK, 2002).
23. Bylund, D., Danielsson, R., Malmquist, G. & Markides, K.E. Chromatographic alignment by warping and dynamic programming as a pre-processing tool for PARAFAC modelling of liquid chromatography–mass spectrometry data. *J. Chromatogr. A* **961**, 237–244 (2002).
24. Villanueva, J. *et al.* Correcting common errors in identifying cancer-specific serum peptide signatures. *J. Proteome Res.* **4**, 1060–1072 (2005).
25. Shannon, C.E. A mathematical theory of communication. *Bell System Tech. J.* **27**, 379–423 (1948).