molecules

MDPI

Article

Semi-Quantitative MALDI Measurements of Blood-Based Samples for Molecular Diagnostics

Matthew A. Koc, Senait Asmellash, Patrick Norman, Steven Rightmyer, Joanna Roder, Robert W. Georgantas III and Heinrich Roder *

Biodesix Inc., Boulder, Colorado 80301, USA; matthew.koc@biodesix.com (M.A.K.); senait.asmellash@biodesix.com (S.A.); patrick.norman@biodesix.com (P.N.); steven.rightmyer@biodesix.com (S.R.); joanna.roder@biodesix.com (J.R.); robert.georgantas@biodesix.com (R.W.G.III)

* Correspondence: heinrich.roder@biodesix.com

Abstract: Accurate and precise measurement of the relative protein content of blood-based samples using mass spectrometry is challenging due to the large number of circulating proteins and the dynamic range of their abundances. Traditional spectral processing methods often struggle with accurately detecting overlapping peaks that are observed in these samples. In this work, we develop a novel spectral processing algorithm that effectively detects over 1650 peaks with over 3.5 orders of magnitude in intensity in the 3 to 30 kD m/z range. The algorithm utilizes a convolution of the peak shape to enhance peak detection, and accurate peak fitting to provide highly reproducible relative abundance estimates for both isolated peaks and overlapping peaks. We demonstrate a substantial increase in the reproducibility of the measurements of relative protein abundance when comparing this processing method to a traditional processing method for sample sets run on multiple matrix-assisted laser desorption/ionization-time of flight (MALDI-TOF) instruments. By utilizing protein set enrichment analysis, we find a sizable increase in the number of features associated with biological processes compared to previously reported results. The new processing method could be very beneficial when developing high-performance molecular diagnostic tests in disease indications.

Keywords: proteomics; mass spectrometry; spectral processing; set enrichment analysis; peak detection

Citation: Koc, M.A.; Asmellash, S.; Norman, P.; Rightmyer, S.; Roder, J.; Georgantas III, R.W.; Roder, H. Semi-Quantitative MALDI Measurements of Blood-Based Samples for Molecular Diagnostics. *Molecules* **2022**, *27*, 997. https://doi.org/10.3390/molecules27030997

Academic Editors: Fulvio Magni, Marco Gaspari and Isabella Piga

Received: 1 December 2021
Accepted: 22 December 2021
Published: 1 February 2022

1. Introduction

Protein abundance in blood is related to outcomes in many systemic diseases and cancer. Standard measurements of known (pre-defined) proteins via enzyme-linked immunoassays (ELISAs) used in medical diagnostics typically measure small numbers of proteins [1–4], sometimes in combination with clinical attributes [5]. Due to the complexity of pathway interactions, it is likely that a multiplexed measurement of many proteins will allow for more accurate characterization of a patient cohort in a particular disease. Indeed, such diagnostic tests have been developed based on highly sensitive high-throughput matrix-assisted laser desorption/ionization (MALDI) profiling, Deep MALDI® [6] analysis, which enables the simultaneous measurement of proteins varying in abundance by four orders of magnitude. These highly multiplexed data can be combined into diagnostic tests using machine learning techniques designed to work well in the clinical setting where we generally have more attributes than samples, without overfitting [7,8]. Multiple tests in the area of oncology were developed using this approach [9–21].

One challenge with using MALDI profiling is the reliable definition and characterization of many hundreds to thousands of Deep MALDI peaks with a dynamic range of peak intensity varying over four orders of magnitude and with overlapping peaks in the

presence of background and noise. Reliable and reproducible peak intensity estimates are necessary as input into machine learning algorithms. Typical peak picking approaches [22–26] often miss many peaks. They often rely on simply finding candidate peaks, either through local intensity maxima or by finding minima in the second derivatives of the intensity, and then using intensity thresholding to select real peaks from the selection of candidate peaks. Although this method is computationally fast, it can fail to detect peaks when they overlap and may struggle to work well when there are large changes in peak intensity [27,28]. Improved peak detection algorithms using a continuous wavelet transform exhibit improved peak detection, but they often may not be accurate in the case of overlapping peaks or highly asymmetric peaks [25,26]. We propose an improved peak detection approach based on characteristics of Deep MALDI spectra. We separate well-defined (using the measured m/z, mass-charge ratio, dependent peak half-width) individual peaks from broad structures. These well-defined peaks are then fitted using a predefined peak shape function either individually, when isolated, or in a multi-peak fit algorithm, when overlapping. Finally, the intensity of the broad structures is added back to the intensity of the previously estimated well-defined peaks to give an expression value for a peak.

In this paper, we start by describing the data arising from a typical Deep MALDI application, and the associated peak shape function. Then, we show how to separate well-defined peaks from broad structures, and how to fit the list of peaks. We evaluate our approach on spectra from two different MALDI-TOF instruments using coefficients of variation (CVs) as the primary metric. Furthermore, we compare these results with those obtained from a common peak-fitting software package. Finally, we examine the association of the peak intensities with biological processes using set enrichment techniques [6,29] to evaluate the amount of useful biological information extracted from the spectra.

2. Results

Deep MALDI spectra were collected on two different MALDI-TOF instruments: the Bruker RapifleX (Bruker, Billerica, MA, USA) and the SimulTOF100 (SimulTOF Systems, Marlborough, MA, USA). Unless stated otherwise, the results shown in the main text are from the RapifleX and the corresponding results on the SimulTOF100 are shown in the Supplementary Materials.

In the Deep MALDI process, for each sample preparation, multiple 800 laser shot ("raster") spectra are collected. Individual raster spectra have significant noise, and only the strongest peaks can be accurately resolved, as shown in Figure 1. To improve the measurement sensitivity and to decrease the noise one averages 500 aligned raster spectra to create a single 400,000 shot-averaged spectrum. The 400,000 shot Deep MALDI averaged spectrum shows a greatly improved signal-to-noise ratio (SNR) and well-defined peaks are now visible that were previously hidden within the noise of a single raster spectrum. Although the sensitivity of our Deep MALDI spectra could be improved further by averaging more individual rasters, we determined that the 400,000 shot-averaged spectra result is a good compromise between sensitivity and instrument run time. We refer readers to Tsypin et al. [6] for further information on this technique.

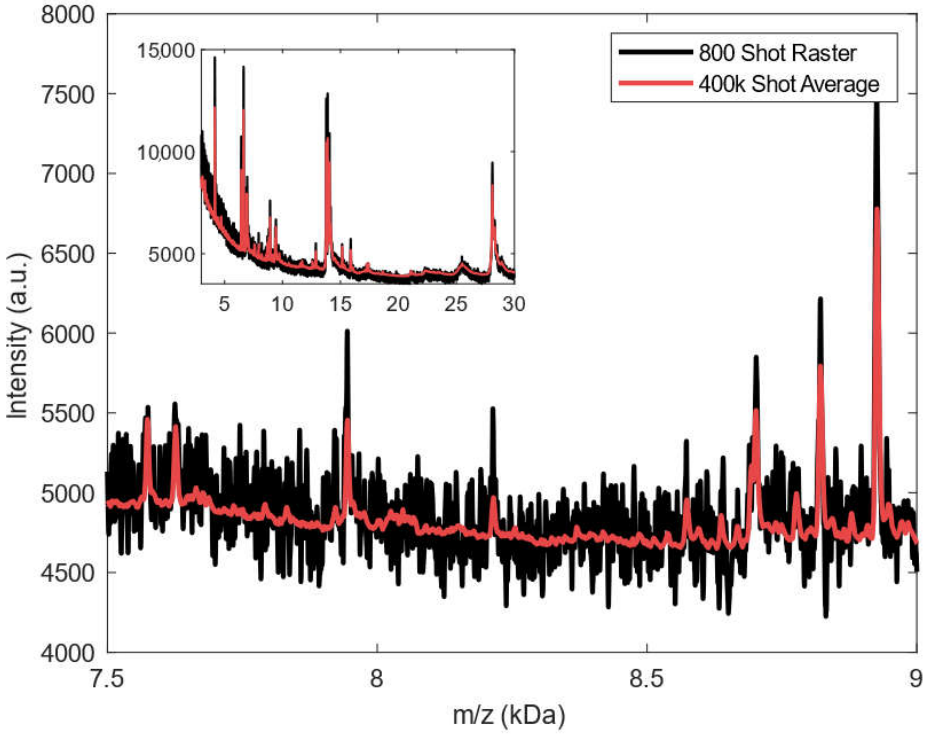
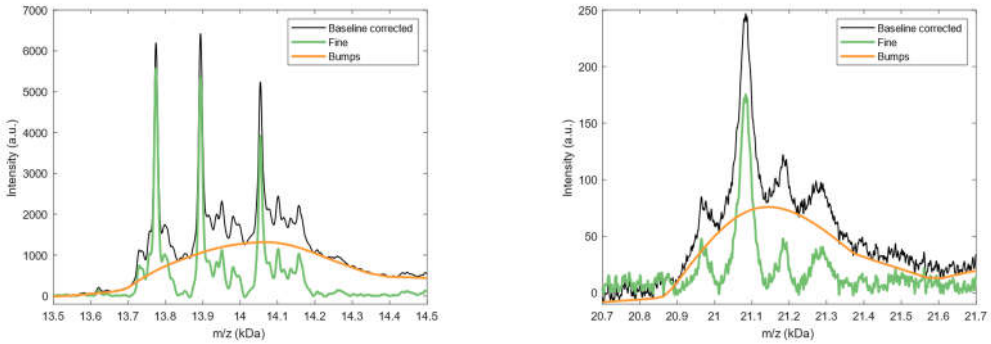


Figure 1. Example spectra collected on the RapifleX of an individual raster spectrum (black) and a 400,000 shot Deep MALDI averaged spectrum (red) from 7.5 to 9 kDa m/z range. The inset shows the same spectra over the full 3 to 30 kDa range analyzed in this work.

Because of the large number of proteins and peptides in serum samples, we observe complex structure in the baseline-corrected spectra. We often find that a baseline-corrected spectrum contains sharp features (peaks) sitting atop broad, wide features ("bumps"), as shown in Figure 2. The origin of the peaks is easy to understand as coming from proteins or peptides of a given mass. The bumps can be attributed to unresolved peaks, e.g., those arising from clusters of highly overlapping mixtures of prominent and less prominent peaks, or from multiply charged, higher-mass proteins (see Figure 2b). Because the bumps originate from biological content in the sample and are not purely an artifact of the measurement process, including the background, removing the bumps during the baselining process will reduce the potential information content available in a single spectrum. To address this problem, we separate and analyze these two components of the spectrum: the peaks (or "fine structure") and the bumps ("bumps"). As detailed below, we find better reproducibility when we include information from both the fine structure and the bumps when determining the feature values for each peak. To maintain a consistent naming convention used in previously published literature [6,7,11–15], we will use the general term "feature" to refer to the peaks and "feature value" to be the semi-quantitative numerical value we calculate to represent the relative abundance of that feature (protein or peptide) within the sample. High-resolution images of a representative unprocessed and processed MALDI-TOF spectrum across the entire acquisition range (m/z = 3 to 30 kDa) is shown in the Supplementary Materials Figure S1.



(**a**) (**b**)

Figure 2. Spectral component analysis showing the baseline corrected Deep MALDI spectrum (black), fine structure (green), and bumps (orange) for peak clusters around (**a**) 14 kDa and (**b**) 21 kDa.

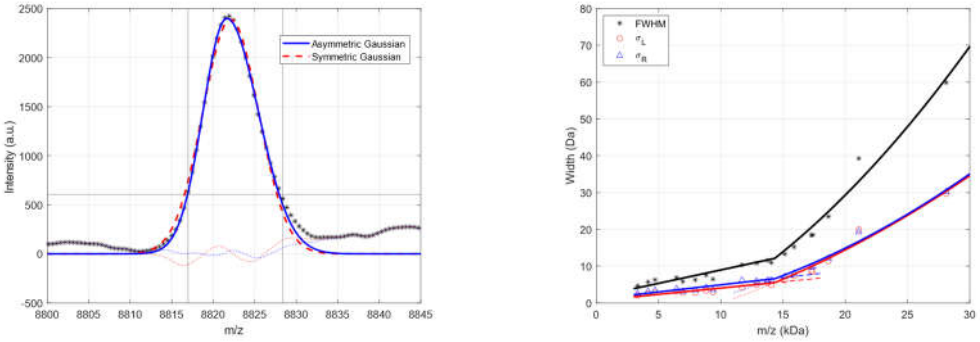
2.1. MALDI Peak Shape Analysis

To improve upon the accuracy of our peak detection algorithm, particularly for overlapping peaks, we utilize a convolution approach whereby the spectrum is convoluted with the peak shape function of the instrument. Previous work has successfully used Gaussian functions to describe the peak shape of MALDI-TOF mass spectra [28,30], but we find this simpler approach insufficient, especially at higher masses. Individual peaks that we have observed in typical spectra are asymmetrically broadened, with the right side (high-mass side) being wider than the left side (low-mass side). This asymmetric broadening comes from a convolution of the instrument broadening and the isotope distribution, which are m/z and mass dependent, respectively. The overall peak shape of the peaks in our spectra were empirically determined to fit well to an asymmetric Gaussian of the form

$$s(m) = \begin{cases} A_0 e^{-(m-m_0)^2 ln(2)/\sigma_L^2}, & m < m_0 \\ A_0 e^{-(m-m_0)^2 ln(2)/\sigma_R^2}, & m \geq m_0, \end{cases} \quad (1)$$

where A_0 is the amplitude, m_0 is the peak center, and σ_L and σ_R are the left and right half widths at half max (HWHM), respectively.

Figure 3a shows a typical, isolated peak in a Deep MALDI averaged spectrum and the symmetric (red-dashed) and asymmetric (blue-solid) Gaussian fits. Only data points with an intensity greater than 0.25 times the maximum intensity were used in the fit. The dotted lines show the calculated error between the raw data and the fitted peak. The sum of the absolute error in the fitting range is 1158.4 a.u. for the symmetric Gaussian fit and 150.6 a.u. for the asymmetric Gaussian fit. The asymmetric Gaussian fit shows a consistent improvement over the symmetric Gaussian fit across the entire m/z range of the peak.



(**a**) (**b**)

Figure 3. Peak shape determination of Bruker RapifleX MALDI-TOF spectral peaks. (**a**) Raw sample data (black stars) and peak fit to an asymmetric (blue-solid) and symmetric (red-dashed) Gaussian. Fit error is shown by the dotted lines. (**b**) Peak shape parameters as a function of m/z. Overall fitted trend is shown with solid lines and the linear (dashed) and quadratic (dotted) piecewise portions for σ_L and $\sigma_R$of the fits are extended past the trend range for reader visibility.

A selection of 19 prominent and isolated peaks, selected to span the acquisition range, was fitted to asymmetric Gaussians to determine the m/z dependence of the peak width. Figure 3b shows the full width at half max (FWHM), σ_L , and σ_R as a function of the m/z range. The right HWHM is consistently larger than the left HWHM across the range; although, at higher masses, the difference is less pronounced.

The trends of left and right HWHM across the m/z range were empirically found to fit well to a piecewise function of the m/z coordinate, m, as

$$\sigma_i(m) = \begin{cases} a_0 + a_1 m, & m < m_{int} \\ c_0 + c_1 m + c_2 m^2, & m \geq m_{int} \end{cases}, \quad (2)$$

where m_{int} is the intersection of the linear and quadratic portions of the piecewise function. The FWHM is simply

$$FWHM(m) = \sigma_L(m) + \sigma_R(m). \quad (3)$$

The average results of 12 different reference sample preparations were used to generate the peak width parameters for the RapifleX shown in Table 1. These parameters are stable over the course of multiple months and run hundreds of samples on the instrument, as shown in the Supplementary Materials, Figure S4. The corresponding peak shape parameters for the SimulTOF100 are shown in the Supplementary Materials, Table S1, but it is worth noting that, due to the difference in instrumentation, we found that the peak shape trends fitted well to a single quadratic for the SimulTOF100 and, therefore, we set $m_{int} = 0$ in Equation (1).

Table 1. The average peak width parameters for the FWHM, and the left and right HWHM for the RapifleX.

	a_0	a_1	c_0	c_1	c_2	m_{int}
FWHM	2.878	5.89E-04	5.764	-1.13E-03	1.05E-07	
σ_L	1.374	2.55E-04	-3.143	-9.53E-06	3.99E-08	14471
σ_R	1.504	3.33E-04	8.907	-1.12E-03	6.54E-08	

2.2. Spectral Analysis of Deep MALDI Spectra

2.2.1. Background Estimation

The 400,000 shot average has a pronounced background that varies across the acquisition range, as shown in Figure 4. The background was estimated using Eilers' estimation [31,32]. This method for background estimation allows for an elastic background that is

penalized differently for errors above and below the background line. For our spectra, we found Eilers' parameters of $\lambda_1 = 10^{11}$, $\lambda_2 = 10^4$, and $p = 0.001$ provided a good background estimation, (BG_1), without over-fitting to the spectral peaks. For further description of the Eilers' method and selection of its parameters, we direct the reader to Boelens et al. [32].

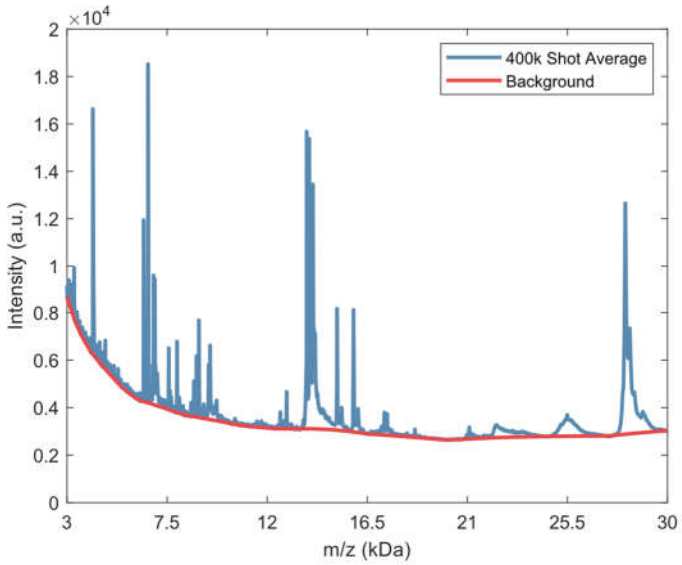


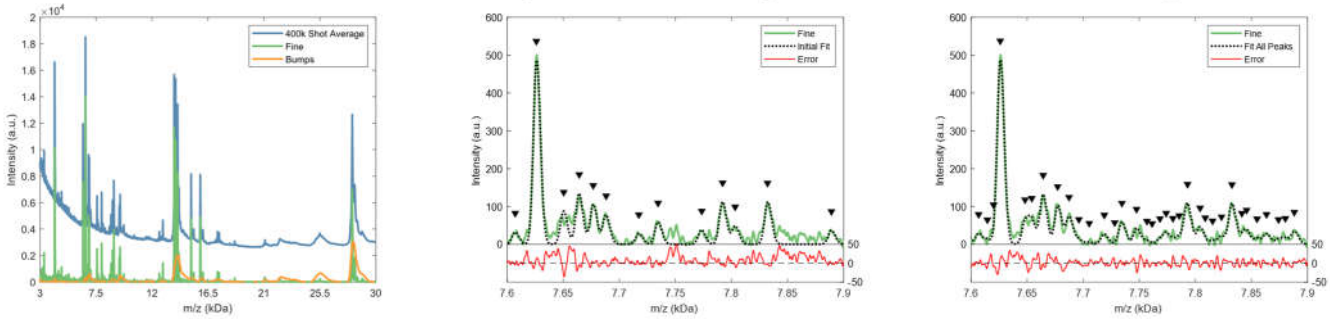
Figure 4. Example of a 400,000 shot-averaged Deep MALDI spectrum collected on the RapifleX and the associated Eilers' background estimation.

2.2.2. Fine Structure Determination and Peak Fitting

As described above, the fine structure is defined to be only the component of the MALDI spectra that contains the sharp features on a flat background. This was calculated by subtracting a relaxed Eilers' background ($\lambda_1 = 10^6$, $\lambda_2 = 10^2$, and $p = 0.001$) (BG_2), from the Deep MALDI spectra. The bumps were calculated as the difference between the relaxed and stiff backgrounds:

$$Bumps(m) = BG_2(m) - BG_1(m). \quad (4)$$

The visual representation of the spectral components is shown in Figure 5a.



(**a**) (**b**) (**c**)

Figure 5. Visual representation of peak fitting and feature value determination. (**a**) A single-processed Deep MALDI spectrum showing the fine structure (green) and bumps (yellow) components. (**b**) Initial peak finding and result of applying the fitting algorithm to the fine structure of a single spectrum in the range 7.5–7.9 kDa. (**c**) The complete fitting of the same range using the master list of all peaks. The triangles indicate locations of fitted peaks and the red trace at the bottom shows the error in the peak fit.

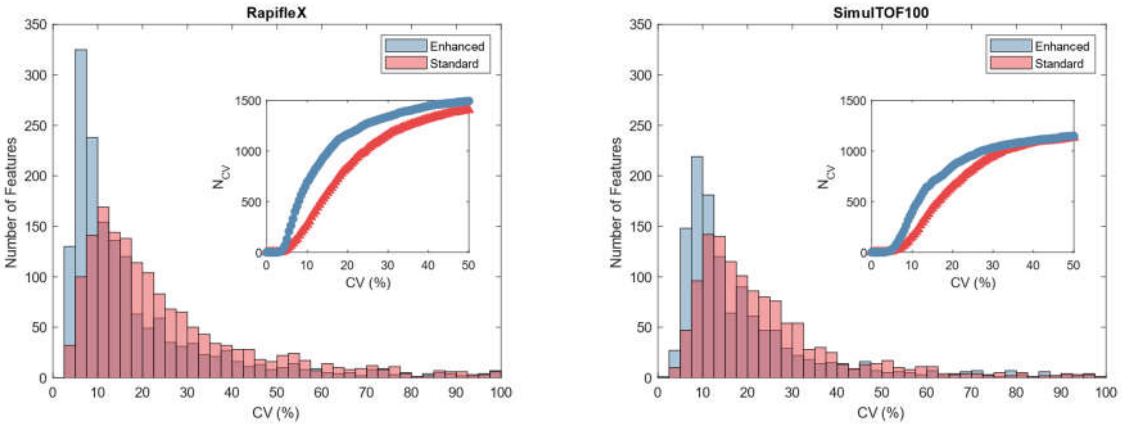
A peak-finding algorithm (see Section 4.4.4 for details) was used to determine the largest peaks that could be used to align spectra to a common m/z axis and to generate a master peak list from multiple samples. The master peak list is a collection of all unique peaks found across all samples and is used to accurately fit the entire spectrum, even peaks that are only sporadically present. Briefly, we first calculated the convolution of the spectrum fine structure with the peak shape function to differentiate true peaks from artifact structures, such as noise. The algorithm then searched for peaks that had SNR > 10 and whose centers were more than one FWHM away from adjacent peaks. Figure 5b shows a sub-range of the entire spectrum that has the fit of the peaks found by this algorithm. Typically, between 700 and 800 peaks with SNRs in this range are found for an individual 400,000 shot Deep MALDI spectrum acquired from 3–30 kDa on the RapifleX. This peak list was used to align the sample to a common m/z axis to allow direct comparison across different samples.

Although the peak fitting clearly does a good job of fitting the strongest features, there are some peaks (such as near 7.75 kDa or 7.85 kDa shown in Figure 5b) that are not fit, because they were not identified by the peak fitting algorithm with the specified parameters. Because each sample will have a different relative abundance of proteins, the intensity of individual peaks will vary across samples. This means that, if a peak is not detected by the peak finding algorithm for a given sample, either because the peak is low intensity, too close to a more prominent peak, or not present in the sample, it may be detected in a different sample. We assume that most proteins are present across different samples albeit at very different concentrations. To determine the set of potential peaks, in blood-based samples, we could detect the lists of peaks from the qualification set of 40 different samples and from the reference sample (measured with 220 unique preparations and acquisitions as described in Section 4.2) which were merged into a master list of unique peaks, resulting in a total of 1657 peaks for the RapifleX and 1256 peaks for the SimulTOF100 instruments.

Accurate peak intensities can be calculated by fitting the pre-defined peak shape function, described in Section 2.1, to each peak in the master peak list, yielding a semi-quantitative feature value for each peak ("Standard" feature value). Here, we utilize the fitted peak amplitude, A_0, as the standard feature value, but we note that other choices of the feature value, such as the area under the fitted peak, could also be used. The result of the fit of all peaks is shown in Figure 5c for the same acquisition and m/z range, as was shown in Figure 5b.

2.3. Reproducibility

To determine the reproducibility of the spectral processing and the resulting feature values, Deep MALDI spectra obtained from 20 replicate preparations of the reference sample were processed and the variations in the calculated feature values were compared. The coefficient of variation (CV) was measured for each feature using the standard feature values of the fitted peaks, as shown in red in Figure 6. We found that the reproducibility could be further improved by including the information in the bumps. For each peak, the "Enhanced" feature value was calculated as the sum of the fitted fine structure peak amplitude and the intensity of the bumps spectrum at the same m/z location. The CV distribution for the enhanced feature values for all peaks in the master peak list is shown in blue in Figure 6.



(**a**) (**b**)

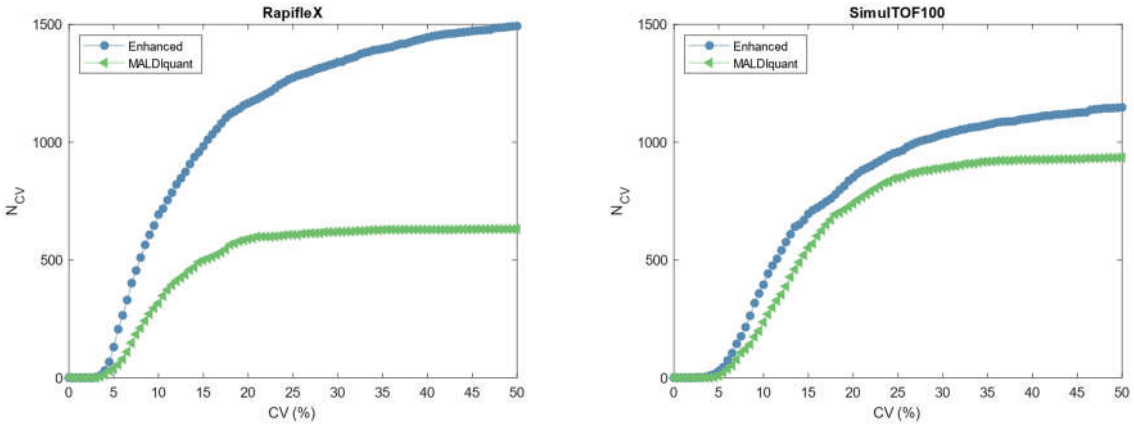
Figure 6. Reproducibility of feature values for a single sample over 20 preparations and acquisitions collected on the (**a**) RapifleX and (**b**) SimulTOF100. Histograms of CVs for standard (red) and enhanced (blue) feature values are shown in the main plot. The inset shows the cumulative CV distribution, N_{CV}, for the standard (red, triangles) and enhanced feature values (blue, circles) (only CVs up to 50% are shown for clarity).

The improvement in reproducibility is further shown by the traces in the insets of Figure 6 which show the cumulative CV distribution, N_{CV}, for the two methods of analysis, where

$$N_{CV}(x) = P(CV \leq x), \tag{5}$$

and $P(CV \leq x)$ is the probability that the CV is less than or equal to x. The enhanced feature value trace shows substantially more features with CVs that were lower than the standard features for the entire range. For example, for the RapifleX spectra, using enhanced feature values, there are 1000 feature with CV < 15.26%. However, while using standard feature values, there are only 594 features. In the following analysis, we restrict ourselves to the feature values calculated using the enhanced approach.

To compare the reproducibility of our processing to that of a commonly used software package, we analyzed our Deep MALDI spectra using the MALDIquant software package [24]. To maintain consistency with our method of creating a master peak list, we utilized the built-in functionality of MALDIquant to bin peaks over the identical 220 spectrum set to generate its own master peak list. The total number of peaks to fit was determined using the median absolute deviation (MAD) method with a SNR cutoff of 2 [22]. We intentionally chose a low SNR to select as many features as possible. Due to slight variations in peak positions not being numerically identical, we binned the peaks with a 0.002 Da tolerance, which resulted in 635 features for the RapifleX acquisition and 947 features for the SimulTOF100. CVs were calculated for the 20 replicate measurements of the reference sample and the comparison with the results from our processing methods, as shown in Figure 7.



(**a**) (**b**)

Figure 7. Comparison of reproducibility of our approach to a published method, MALDIquant. Cumulative CV distribution, N_{CV}, for the same Deep MALDI spectra analyzed with the presented methods using enhanced feature values (blue, circles) and with MALDIquant processing (green, triangles) for: (**a**) RapifleX and (**b**) SimulTOF100 acquisitions. Only CVs up to 50% are shown for clarity.

2.4. Association with Biological Processes

It is important that the features represented here not only present reproducible data, but also are also biologically relevant. We calculated the association of each feature with 23 biological processes using protein set enrichment analysis (PSEA) [6,29]. This commonly used bioinformatics tool determines the association between a measured quantity, and, in our case, a mass spectral feature and a biological process by assessing the correlation between the measured quantity and the abundances of a set of known proteins related to the biological process. Table 2 shows the total number of features that were determined to be associated with each biological process with p-value of association < 0.01 and a false discovery rate (FDR) of < 5%.

Table 2. Number of features associated with each biological process with FDR of < 5% and a p-value of association < 0.01 for 400,000 shot spectra collected on the RapifleX and SimulTOF100 mass spectrometers using the processing and feature definitions presented in this paper. A comparison is made with the number of associated features obtained with the same 400,000 shot spectra collected on the SimulTOF100 mass spectrometer using an alternative processing and feature definition method [6]. The percentages show the proportion of all analyzed features that show an association with the biological process. Note the substantial increase in the number of associated features identified when the new processing feature definition method is used.

Biological Process	Associated Features		
	RapifleX	**SimulTOF**	**Ref. [6]**
Acute phase response	655 (43.2%)	460 (40.6%)	122 (40.9%)
Complement activation	619 (40.8%)	383 (33.7%)	70 (23.5%)
Acute inflammatory response	434 (28.6%)	317 (27.9%)	109 (36.6%)
IFN γ signaling/response	266 (17.5%)	147 (12.9%)	25 (8.4%)
Immune tolerance and suppression	227 (15.0%)	189 (16.6%)	31 (10.4%)
Wound healing	202 (13.3%)	160 (14.1%)	100 (33.6%)
IFN type 1 signaling/response	195 (12.9%)	82 (7.2%)	33 (11.1%)
Type 17 immune response	73 (4.8%)	82 (7.2%)	2 (0.7%)
Angiogenesis	54 (3.6%)	28 (2.5%)	2 (0.7%)
Response to hypoxia	42 (2.8%)	54 (4.7%)	7 (2.3%)
Cellular component morphogenesis	27 (1.8%)	6 (0.5%)	4 (1.3%)

Cytokine production involved in immune response	21 (1.4%)	31 (2.7%)	3 (1.0%)
Glycolysis	21 (1.4%)	36 (3.2%)	2 (0.7%)
Chronic inflammatory response	19 (1.3%)	23 (2.0%)	2 (0.7%)
Innate immune response	18 (1.2%)	30 (2.6%)	8 (2.7%)
Extracellular matrix organization	16 (1.1%)	4 (0.4%)	0 (0.0%)
Behavior	7 (0.5%)	2 (0.2%)	8 (2.7%)
Epithelial-mesenchymal transition	6 (0.4%)	4 (0.4%)	0 (0.0%)
NK cell mediated immunity	4 (0.3%)	11 (1.0%)	N/A*
T-cell mediated immunity	4 (0.3%)	5 (0.4%)	N/A*
Type 2 immune response	2 (0.1%)	11 (1.0%)	N/A*
B-cell mediated immunity	2 (0.1%)	6 (0.5%)	N/A*
Type 1 immune response	1 (0.1%)	1 (0.1%)	N/A*

*N/A indicates the number of features associated with the biological process was not published in Ref. [6].

3. Discussion

The goal of our processing methods is to better characterize complex MALDI-TOF spectra by improving peak detection and quantification. Because common peak detection approaches often perform poorly for clustered peaks [27,28], we utilized the method of spectral convolution to select peaks. Quantitation of peak intensity is also difficult to accurately determine when the peak of interest is part of a clusters of peaks. One cannot simply take the maximum intensity at the peak location because the tails of adjacent and overlapping peaks will add to the overall intensity at that location. Due to this complication, previous work would define the entire cluster as a single feature that spanned a spectral range instead of decomposing the cluster into individual peaks [6,12]. By accurately fitting the clusters, each individual peak intensity can be accurately determined without the influence of adjacent peaks. By implementing these ideas and utilizing the often-neglected information in the component of the spectra that varies more slowly with m/z (the bumps), we can detect more peaks with improved reproducibility than the traditional processing method tested [24]. The improved characterization can also lead to a better understanding of the direct biological implications of different features in our spectra.

3.1. MALDI Peak Shape Analysis

The m/z dependence on the peak shape is due to inherent protein properties (isotopic distribution) and the instrument response function (IRF). For any individual protein, the peak shape that we observe in a spectrum is a convolution of the isotopic distribution of the protein with the instrument response function. As proteins get larger, we expect to see a wider isotopic distribution [33]. An estimation of the peak width change with mass is shown in the Supplementary Materials, Figure S5, based on proteins composed of the fictional amino acid known as averagine [33]. Similarly, the mass spectrometer is known to have a variable IRF over wide mass ranges that results in wider features further from the optimal (tuned) mass range [28,34,35]. The change in trend from linear to quadratic shown in Figure 3b is likely due to a change in the IRF.

The IRF is a difficult parameter to determine [28,34–36]; thus, for this work, we opted to use an empirical fit. We note that, if the IRF could be carefully measured, it would be possible to get higher-resolution spectra by deconvoluting the observed spectra with the IRF and the isotope distribution [28]. Such information could be useful in better determining component parts of the bumps or perhaps eliminate the bumps altogether.

3.2. Peak Detection and Feature Value Determination

Traditional fitting methods often rely on simply removing the broad structures (bumps) during background subtraction, resulting in only sharp features (fine structure) [37,38]. These broad features originate from real biological content that is unresolved (see Figure 2) and, thus, potentially valuable information is lost when the bumps are overlooked during background estimation and subtraction. During the spectral analysis presented here, an individual spectrum was decomposed into three separate components: the background, the fine structure, and the bumps. By maintaining a slowly varying background, we are able to extract and analyze the bumps spectrum which was shown to improve quantitative reproducibility.

Although our methods show improved quantification of highly reproducible features, we do note that our method of processing is computationally slower (~ 3 min/spectrum) than the MALDIquant software (several seconds/spectrum) [24]. This is, in part, due to the high level MATLAB language, which could be sped up with a faster language, but it is also due to the differences in our peak detection algorithms. In the present work, we enhance our peak detection by convoluting the spectrum with the asymmetric peak shape that we defined for each instrument. Although this is a computationally intensive task, the convolution sharpens features and effectively filters the noise, which allows for more accurate detection of low intensity peaks. The MAD peak detection algorithm that we utilized in the MALDIquant analysis is one of the most commonly used peak-finding algorithms [22–24]. It simply finds local maxima and only selects those that are above the SNR cutoff. Because of the convolution we used, for the RapifleX data, we were able to find a total of 1657 peaks while using a SNR cutoff of 10, while the traditional MAD method used in the MALDIquant processing only found 635 features with a SNR of 2.

To further evaluate our algorithms, we processed an additional 220 sample preparations on another mass spectrometer (SimulTOF100). The spectra acquired on the SimulTOF100 were processed using the presented methods, and we found 1256 unique features (see Supplementary Materials Figure S3 and Table S1 for peak shape analysis on the SimulTOF100). The Deep MALDI spectra collected on the SimulTOF100 were also analyzed using MALDIquant, which found 947 features. Although the MALDIquant processing appeared to do better with this instrument, we still find that our processing methods produce a greater number of highly reproducible features.

3.3. Reproducibility

Our methods show an improvement over current traditional processing techniques, as tested by the MALDIquant software package [24]. For both sample sets run on the RapifleX and SimulTOF100 instruments, we defined and characterized a greater number of features (with enhanced feature values) with smaller CVs with the presented processing than with the MALDIquant software package. Because the spectral processing methods presented here show improvements across multiple instruments, we expect a similar performance using alternative sample preparation methods, such as utilizing depletion methods to improve the detection of low-abundance proteins [39,40].

Due to the large number of highly reproducible features, diagnostic tests could be created to stratify and classify patients into different groups to predict patient outcome based on this processing method. Multiple tests are currently being developed using the methods presented here which we believe will provide useful and actionable results for physicians and patients.

3.4. PSEA

Gene set enrichment analysis (GSEA) is a commonly used tool in bioinformatics that associates a measured quantity (for example, gene expression) with a biological process by finding patterns of association across a set of genes known to be related to that process

[41]. Using a similar approach, we can associate our Deep MALDI features with biological processes in a protein set enrichment analysis [29,42].

In our previous work on the development of Deep MALDI [6], we demonstrated an improvement in the number of features associated with biological processes with ab increasing number of laser shots. The direct comparison (using the same number of laser shots generated by the Deep MALDI spectra, as well as the same p-value cutoff and FDR) to the previous work is shown in Table 2. The work by Tsypin et al. [6] was run on a SimulTOF100, so we can see a clear improvement in the number of associated features with our processing. The methods and procedures presented here show a substantial increase in the number of associated features in nearly all the biological processes investigated.

4. Materials and Methods

4.1. Serum Samples

A total of 40 serum samples ("qualification set"), derived from the blood of lung cancer and colorectal cancer patients, were purchased from Discovery Life Sciences, Inc. (Huntsville, AL, USA). A reference sample was created by pooling equal volumes of serum obtained from ten healthy individuals, also purchased from Discovery Life Sciences, Inc. The 100 serum samples collected from patients with non-small cell lung cancer used for association with biological processes via protein set enrichment analysis ("PSEA set") were purchased from Oncology Metrics, LLC (Fort Worth, TX, USA) and Discovery Life Sciences Inc (Huntsville, AL, USA). All samples were collected under ethics-approved protocols, according to the requirements of Discovery Life Sciences Inc and Oncology Metrics LLC, and were stored at −80 °C.

4.2. Sample Preparation

Serum samples were thawed and 3-μL aliquots of each sample were spotted onto a serum card (GE HealthCare, Chicago, IL, USA). The spots were allowed to dry for 1 h at ambient temperature after which the whole serum spot was punched out from the underside with a 6-mm skin biopsy punch (Acuderm, Fort Lauderdale, FL, USA). Each punch was placed in a centrifugal filter with a 0.45-μm nylon membrane (VWR, Randor, PA, USA). In cases where the serum spots had spread outside the 6-mm diameter, the section where serum was visible was excised and added to the tube containing the 6-mm punch. To the centrifugal filter containing the punch, 100 μL of HPLC grade water (VWR, Randor, PA, USA was added. The punches were vortexed gently for 10 min and then spun down at 14,000 rcf for two minutes. The flow-through was removed and transferred back on to the punch for a second round of extraction consisting of vortexing gently for three minutes and spinning down at 14,000 rcf for two minutes. Finally, 20 μL of the filtrate from each sample was then transferred to a 0.5-mL Eppendorf tube. All subsequent sample preparation steps were carried out in a custom designed humidity and temperature control chamber (Coy Laboratory). The temperature was set to 30 °C and the relative humidity at 10%.

An equal volume of freshly prepared matrix (25 mg of sinapinic acid per 1 mL of 50% acetonitrile mixed with 50% water plus 0.1% TFA) was added to each 20-μL serum extract and the mix was vortexed for 30 s. The first three aliquots (3 × 2 μL, for SimulTOF100) or five aliquots (5 × 2 μL, for Rapiflex) of sample–matrix mix were discarded into the tube cap. Eight aliquots of 2-μL sample–matrix mix were then spotted onto a stainless steel MALDI target plate (Bruker, Billerica, MA, USA and SimulTOF Systems, Marlborough, MA, USA for spectra acquisition on the RapifleX and SimulTOF 100, respectively). The MALDI plate was allowed to dry in the chamber before placement in the MALDI mass spectrometer.

For the work on generating a master peak list and reproducibility a total of five replicate batches of the qualification set (40 serum samples) were analyzed for each mass

spectrometer. Each batch consisted of the 40 serum samples with an additional four preparations of reference sample used as controls, with two preparations spotted at the start and two at the end of the batch. This resulted in a total of 220 spectra per mass spectrometer (5 × 40 samples in the qualification set and 20x of the reference sample).

Samples for the PSEA set were run in batches of up to 44 samples with an additional four preparations of the reference sample used as controls, with two preparations spotted at the start and two at the end of the batch for each mass spectrometer.

4.3. Mass Spectra Acquisition

4.3.1. RapifleX

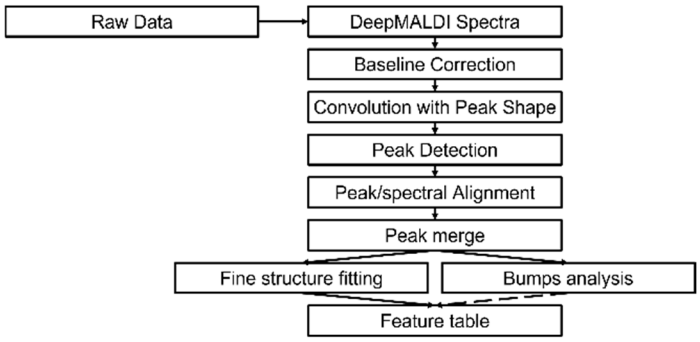
MALDI spectra were obtained using a RapifleX MALDI-TOF mass spectrometer (Bruker, Billerica, MA, USA). The instrument was operated in a positive ion mode, with ions generated using a frequency-tripled Nd:YAG laser emitting at 355 nm with a laser repetition rate of 5 kHz. Spectra were acquired in the 3 kDa to 30 kDa m/z range with a sampling rate of 0.63 Gs/s. External calibration was performed using the following peaks in the spectra generated from the reference samples included on every target plate (or batch): m/z = 3320, 4158.7338, 6636.7971, 9429.302, 13890.4398, 15877.5801, and 28093.951. From each spot, 100 raster spectra were collected, totaling 800 raster spectra per sample. A raster spectrum is an average over 800 laser shots measured across a single spot.

4.3.2. SimulTOF100

MALDI spectra obtained using a SimulTOF100 MALDI-TOF mass spectrometer (SimulTOF Systems, Marlborough, MA, USA). The instrument operated in the positive ion mode with ions generated using a 349-nm, diode-pumped, frequency-tripled Nd:YLF laser operated at a laser repetition rate of 0.5 kHz. Raster spectra were acquired in the 3 to 75 kDa m/z range (only the range from 3 to 30 kDa was used in this analysis) and were 'hardware averaged' to contain 800 laser shots as the laser fires continuously across the spot while the stage is moving at a speed of 0.25 mm/s. External calibration was performed using the following peaks generated in the reference sample included on every target plate: m/z =3320, 4158.7338, 6636.7971, 9429.302, 13890.4398, 15877.5801, and 28093.951.

4.4. Spectral Analysis

The spectral analysis workflow is shown in Scheme 1 for processing of raw data through generation of a feature table or matrix (a list of feature values for each feature for each sample). Post-processing, such as normalization or corrections, can be performed on the table of feature values.



Scheme 1. Spectral analysis workflow for mass spectrometer data using presented preprocessing methods.

4.4.1. Raster Averaging for Deep MALDI Spectra

To increase the number of observable peaks and to improve the SNR in the MALDI-TOF spectra, we employed the Deep MALDI raster averaging technique. For a complete

description of this technique, we refer the reader to Tsypin et al. [6]. Briefly, each raster spectrum of 800 shots was processed through an alignment workflow to align peaks to a set of internal alignment points (Tables S2, S3). Peaks were detected in each raster spectrum with a SNR cut-off > 3.0. The identified peaks for a raster spectrum were then used together with the set of predefined alignment peak positions to establish the coefficients in a second-order polynomial (in m/z) that was used to transform the m/z values of this raster spectrum. For successful alignment, we required a minimum of 20 detected peaks, with at least 13 peaks useable for alignment, i.e., an un-aligned peak position within a fixed alignment tolerance (1500 ppm) of the alignment peak. A maximum shift of 15 Da was allowed at the lowest m/z alignment point.

Averages were created from the pool of aligned raster spectra that satisfied the alignment quality criteria. A random selection of 500 raster spectra, without replacement, were averaged to create a final analysis spectrum of 400,000 shots for each sample.

4.4.2. Background Estimation

The spectrum background was calculated using a stiff Eilers' estimation ($\lambda_1 = 10^{11}$, $\lambda_2 = 10^4$, $p = 0.001$) [31,32]. Briefly, the Eilers' estimation is an asymmetric least squares fitting that penalizes positive and negative deviations separately, allowing for accurate estimation of the background.

4.4.3. Fine Structure and Bumps Determination

The stiff background does not overfit the spectra and subtraction from the spectra results in sharp features that still are sitting on top of broad features, as seen in Figures 2 and 4. Both the bumps and the sharp peaks may contain useful information as to total protein content in the sample but need to be treated separately to ensure accurate estimation of feature values.

A second, aggressive background was calculated to fit both the background and bumps with Eilers' parameters $\lambda_1 = 10^6$, $\lambda_2 = 10^2$, and $p = 0.001$. The difference between the spectrum and the aggressive background results in the fine structure, which contains the information of the sharp peaks on a flat background. The bumps were defined as the difference between the aggressive background and the stiff background.

4.4.4. Peak Detection

Peak candidates to be fit were estimated using a peak finding algorithm based on the convolution of the fine structure with the peak-shape defined in Section 2.1. Peak candidate locations were estimated using the MATLAB function *islocalmin* on the second derivative of the fine structure, with a prominence window equal to the width of the FWHM of a peak and a minimum separation of peaks equal to 1/4 of the peak FWHM at the m/z location. These candidates were only fit as peaks if the SNR was greater than 10 and if the candidate was not being influenced by adjacent peak candidates. The signal was simply the intensity of the signal at the m/z point, while the noise was measured as the deviation in the signal from the average, as estimated by a Gaussian-smoothing window the size of the peak width.

A given peak was determined to be influenced by an adjacent peak if the peak centers were within half a peak width of each other or if either peak intersected the other at more than 10% of the maximum amplitude. Peak candidates with SNR > 10, which were not found to be influenced by an adjacent peak, were fitted to a single asymmetric Gaussian to get precise peak position and amplitude. Peak candidates with SNR > 10, but which were determined to be influenced by adjacent peak candidates, were assigned to be part of a cluster. The multiplicity of a cluster, *N*, is defined as the number of peak candidates that are influenced by at least one other member of the cluster. Clusters were fitted simultaneously by *N* asymmetric Gaussians (i.e., a doublet would be fit to *N* = 2 asymmetric Gaussians and a triplet would be fit to *N* = 3 asymmetric Gaussians). This method of fitting

allows for accurate determination of the m/z position as well as peak intensity of all N peaks in the cluster. Peaks with a SNR < 10 were not considered for alignment purposes or merging into the master list.

A list of all peak locations for each sample was later merged into a master list of all measurable peaks from a wide range of multiple samples, as described in Section 4.6 below.

4.4.5. Spectral Alignment

Spectra were aligned to a common m/z axis to ensure accurate feature (peak) intensities across samples. Alignment was performed by minimizing the variation in peak positions (m/z value from the peak fitting) for the sample with respect to the pre-specified alignment points (Tables S4 and S5). The m/z axis was rescaled using a second-order polynomial in m/z. Only peaks in the 80th percentile of SNR were used for alignment and were weighted inversely to their location in m/z, to account for the greater weights a simple linear regression gives to instances at higher m/z. A peak was determined to be alignable if its spectral position was within half of a peak width (as defined at the m/z position) of the nearest alignment point. To account for variations over the large range in m/z, we split the alignment range into four sub-ranges, as shown in Table S6.

To ensure the high-quality data we require, only fits with at least five alignable features in each region were used. Spectra that failed to align were not used in further determination.

4.4.6. Feature Value Determination

The fine structure was fitted to 1657 (1256) asymmetric Gaussians at the specified m/z positions (see Section 4.6) to extract the peak intensity for the RapifleX (SimulTOF100). Isolated peaks were simply fitted to a single asymmetric Gaussian, while peaks that were part of a cluster were simultaneously fitted to N asymmetric Gaussians, where N is the multiplicity of the cluster. By fitting the entire cluster simultaneously, we ensured accurate peak amplitude measurements for peaks with significant overlap. A subtle, yet important, point to note is that we only fitted the intensity of each peak while keeping the m/z position and width parameters fixed (unlike in Section 4.4.5 above, where the m/z position was also fitted). Because the spectra were aligned in the previous step, we did not need to fit the m/z position and we could fit all features, even including those whose SNR was under 10. This results in our ability to accurately fit peaks whose intensity ranges over 3.8 orders of magnitude. A preliminary "Standard" feature value, characterizing the magnitude of a peak, was defined as the fitted peak amplitude. The preliminary feature value was further modified by adding the bump intensity at the m/z location to determine the "Enhanced" feature value. Mathematically,

$$FV_E(m) = FV_S(m) + Bumps(m), \quad (6)$$

where m is the m/z location, $FV_i(m)$ for $i = S, E$ is the feature value for "Standard" or "Enhanced", respectively, and $Bumps(m)$ is defined above in Equation (3).

4.4.7. MALDIquant Analysis

The same set of 220 Deep MALDI spectra (as described for the qualification set in Section 4.2 above), used to determine our master peak list (see Section 4.6), were analyzed using the MALDIquant mass spectra analysis software as a methods comparison [24]. Deep MALDI spectra were transformed using the square root transformation method to minimize any variance from the mean. Spectra were then smoothed with the Savitzky-Golay-Filter (halfWindowSize = 10) [43] and the baseline was corrected with the Statistics-sensitive Non-linear Iterative Peak-clipping (SNIP) algorithm (iterations = 100) [44]. Spectra were total ion current (TIC) normalized and spectra were aligned using the "lowess" warping method (halfWindowSize = 20, SNR = 2, and tolerance = 0.002) [45].

Peaks were determined using the MAD method (halfWindowSize = 20, SNR = 20) [22] and similar peaks were binned with a tolerance of 0.002. Feature values and CVs were calculated using the 635 unique peaks determined by this method for RapifleX Deep MALDI spectra and 947 unique peaks for the SimulTOF100 Deep MALDI spectra.

4.5. Peak Shape Fitting

A total of 19 isolated peaks were fitted to asymmetric gaussians (Equation (1)). The peaks were selected as isolated peaks that spanned the m/z range (3 to 30 kDa). Only the top 75% intensity was fitted for peak width determination. The trends of the left- and right-HWHM were then fit to a linear trend in the low mass region (3 to 17 kDa) and a quadratic fit for the high-mass region (13 to 30 kDa), as described by Equation (2). The linear and quadratic fits were intentionally made to overlap to accurately determine the intersection of the two curves (m_{int}). The average peak shape trend parameters came from the average of 12 different preparations of the same reference serum sample measured over three batches.

To ensure there were no discontinuities while fitting the entire spectra the final m_{int} for FWHM, σ_L and σ_R were calculated solely from the FWHM trend.

4.6. Merge Peak Lists

Two-hundred and twenty peak lists, determined in Section 4.4.4, were generated for five batches of the qualification set as described in Section 4.2. All the peaks were merged into a single list resulting in 1657 unique peaks for the RapifleX and 1256 unique peaks for the SimulTOF100 (Supplementary Materials Tables S7 and S8). The merged peak list was created by iteratively comparing the merged peak list (initially empty) with an un-merged list. Peaks from the un-merged list that had a peak center greater than 0.5x peak width away from adjacent peak centers in the merge list were added. Peaks with centers less than 0.5x peak width away from adjacent peaks in the merge list had their location averaged with the existing merge peak.

4.7. Reproducibility Analysis

A total of 20 replicate measurements, including sample preparation and spectra acquisition, of the reference sample were collected over five batches. Spectra were processed as described above and standard and enhanced feature values were calculated for each replicate. Feature values were normalized to the total feature value intensity for the sample. For each feature, the average feature value ($\bar{x}$), standard deviation (σ_x), and coefficient of variation ($CV = \sigma_x/\bar{x}$) were calculated.

4.8. Association with Biological Processes

We followed the procedure outlined in Grigorieva et al. [29] to determine the association of the identified mass spectral features with 23 biological processes using protein set enrichment analysis. The biological processes investigated included both those expected to be assessable in circulation of patients with cancer (e.g., acute phase response, acute inflammatory response, wound healing) and some processes designed as controls (behavior, cellular components of morphogenesis). Briefly, protein abundance for 1305 known proteins was obtained for the PSEA set of 100 serum samples using the aptamer-based 1.3k SOMAscan assay (SomaLogic, Boulder, CO) [46,47]. The subsets of the 1305 proteins known to be associated with each of the 23 biological processes were identified using database searches, as has been previously described in detail [29]. Deep MALDI spectra were acquired from the PSEA set using both the RapifleX and the SimulTOF100 mass spectrometers using the methods of sections 4.2–4.3. The spectra were processed, and the feature values for each sample are described in sections 4.4–4.6. The Spearman correlation was calculated between each feature and each of the 1305 proteins across the 100 different samples. An enrichment score was generated for each of the 23 biological

processes for each mass spectral feature using the method of Roder et al. [42] with 25 splits of the sample set to provide increased power to detect association with biological processes compared with the standard GSEA enrichment score [41]. We calculated the p-values of association between each feature and the biological processes by comparing the enrichment score to a null distribution generated by a random permutation of feature values across the sample set. Features with a p-value of association < 0.01 and a false discovery rate of 5% or less, as estimated by the method of Benjamini-Hochberg [48] for multiple comparisons across the 23 biological processes, were determined to be associated with a given biological process. A subset of 1516 features were used from the Rapiflex processing and 1138 features for the SimulTOF100. For comparison, Ref [6] used 298 features. These reduced feature sets were determined by removing features that are known to depend strongly on sample collection and processing details, i.e., features that are related to hemoglobin (and its multiply charged analogs) or to fibrinogen (whose spectral intensity often vary).

5. Conclusions

Here, we developed a novel method for analyzing MALDI-TOF spectra over a wide spectral range. We used our method to analyze spectra from multiple samples to find 1657 unique peaks with over 3.5 orders of magnitude intensity, compared to only 635 for the traditional processing methods [24]. Our use of a well-defined peak shape function for our instrumentation allows us to accurately detect a greater number of peaks, particularly among overlapping peaks. The use of peak shape also allows for accurate fitting of overlapping peaks for accurate peak amplitude measurements. When compared to a traditional processing method, we found a substantial increase in the number of highly reproducible features with low CVs. We further validated our processing by performing the same analysis on spectra collected on a mass spectrometer from a different manufacturer and showing improved detection and reproducibility. Finally, we analyzed a set of 100 samples with known protein variation to determine the number of features associated with biological processes. We found an increase in the number of features associated with biological processes compared to analysis of the same sample set with a different spectral processing method [6].

Supplementary Materials: The following are available online, Figure S1: Representative, high-resolution unprocessed and processed RapifleX spectrum, Figure S2: SimulTOF100 Deep MALDI comparison, Figure S3: SimulTOF100 peak shape fitting, Table S1: SimulTOF100 peak shape parameters, Figure S4: Peak shape parameter stability, Figure S5: Isotopic contribution to peak shape broadening, Table S2: Raster alignment peaks for RapifleX, Table S3: Raster alignment peaks for SimulTOF100, Table S4: Deep MALDI Alignment points for RapifleX, Table S5: Deep MALDI Alignment points for SimulTOF100, Table S6: Spectral alignment ranges, Table S7: Peak list for RapifleX, and Table S8: Peak list for SimulTOF100. A .csv file of an example spectrum measured on the RapifleX and the calculated background, fine structure, and bumps spectral components is also available.

Author Contributions: Conceptualization, H.R.; methodology, M.A.K., S.A., J.R., and H.R.; software, M.A.K. and H.R.; validation, M.A.K.; formal analysis, M.A.K.; investigation, S.A., S.R., and P.N.; resources, S.A.; data curation, M.A.K.; writing—original draft preparation, M.A.K. and H.R.; writing—review and editing, M.A.K., S.A., P.N., S.R., J.R., R.W.G.III, and H.R.; visualization, M.A.K.; supervision, R.W.G.III and H.R.; project administration, R.W.G.III and H.R.; funding acquisition, R.W.G. All authors have read and agreed to the published version of the manuscript.

Funding: All authors are employees of Biodesix, Inc. The authors received no other funding for this work. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Institutional Review Board Statement: All samples were collected under ethics-approved protocols according to the requirements of Discovery Life Sciences Inc and Oncology Metrics LLC.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in the study are available in the supplementary materials. MATLAB code used for analysis is available on request from the corresponding author

Acknowledgments: The authors would like to thank A. Wiles and T. Campbell for helpful discussions and for the critical reading of the manuscript.

Conflicts of Interest: All authors are current employees of and have or had stock options in Biodesix, Inc. H.R., J.R., and S.A. are inventors on patents describing Deep MALDI TOF mass spectrometry of complex biological samples, assigned to Biodesix, Inc.

Sample Availability: Samples of the compounds are not available from the authors.

References

1. Bast, R., Jr.; Xu, F.-J.; Yu, Y.-H.; Barnhill, S.; Zhang, Z.; Mills, G. CA 125: The Past and the Future. *Int. J. Biol. Markers* **1998**, *13*, 179–187, doi:10.1177/172460089801300402.
2. Scarà, S.; Bottoni, P.; Scatena, R. CA 19-9: Biochemical and Clinical Aspects. *Adv. Cancer Biomark.* **2015**,867 247–260, doi:10.1007/978-94-017-7215-0_15.
3. Dorcely, B.; Katz, K.; Jagannathan, R.; Chiang, S.S.; Oluwadare, B.; Goldberg, I.J.; Bergman, M. Novel Biomarkers for Prediabetes, Diabetes, and Associated Complications. *Diabetes Metab. Syndr. Obes. Targets Ther.* **2017**, *10*, 345, doi:10.2147/dmso.s100074.
4. Patel, S.P.; Kurzrock, R. PD-L1 Expression as a Predictive Biomarker in Cancer Immunotherapy. *Mol. Cancer Ther.* **2015**, *14*, 847–856, doi:10.1158/1535-7163.mct-14-0983.
5. Johnson, P.J.; Pirrie, S.J.; Cox, T.F.; Berhane, S.; Teng, M.; Palmer, D.; Morse, J.; Hull, D.; Patman, G.; Kagebayashi, C.; et al. The Detection of Hepatocellular Carcinoma Using a Prospectively Developed and Validated Model Based on Serological Biomarkers. *Cancer Epidemiol. Prev. Biomark.* **2014**, *23*, 144–153.
6. Tsypin, M.; Asmellash, S.; Meyer, K.; Touchet, B.; Roder, H. Extending the Information Content of the MALDI Analysis of Biological Fluids via Multi-Million Shot Analysis. *PLoS ONE* **2019**, *14*, e0226012, doi:10.1371/journal.pone.0226012.
7. Roder, J.; Oliveira, C.; Net, L.; Tsypin, M.; Linstid, B.; Roder, H. A Dropout-Regularized Classifier Development Approach Optimized for Precision Medicine Test Discovery from Omics Data. *BMC Bioinform.* **2019**, *20*, 325, doi:10.1186/s12859-019-2922-2.
8. Roder, H.; Oliveira, C.; Net, L.; Linstid, B.; Tsypin, M.; Roder, J. Robust Identification of Molecular Phenotypes Using Semi-Supervised Learning. *BMC Bioinform.* **2019**, *20*, 273, doi:10.1186/s12859-019-2885-3.
9. Taguchi, F.; Solomon, B.; Gregorc, V.; Roder, H.; Gray, R.; Kasahara, K.; Nishio, M.; Brahmer, J.; Spreafico, A.; Ludovini, V.; et al. Mass Spectrometry to Classify Non–Small-Cell Lung Cancer Patients for Clinical Outcome after Treatment with Epidermal Growth Factor Receptor Tyrosine Kinase Inhibitors: A Multicohort Cross-Institutional Study. *J. Natl. Cancer Inst.* **2007**, *99*, 838–846, doi:10.1093/jnci/djk195.
10. Weber, J.S.; Sznol, M.; Sullivan, R.J.; Blackmon, S.; Boland, G.; Kluger, H.M.; Halaban, R.; Bacchiocchi, A.; Ascierto, P.A.; Capone, M.; et al. A Serum Protein Signature Associated with Outcome after Anti–PD-1 Therapy in Metastatic Melanoma. *Cancer Immunol. Res.* **2018**, *6*, 79–86, doi:10.1158/2326-6066.CIR-17-0412.
11. Mahalingam, D.; Chelis, L.; Nizamuddin, I.; Lee, S.S.; Kakolyris, S.; Halff, G.; Washburn, K.; Attwood, K.; Fahad, I.; Grigorieva, J.; et al. Detection of Hepatocellular Carcinoma in a High-Risk Population by a Mass Spectrometry-Based Test. *Cancers* **2021**, *13*, 3109, doi:10.3390/cancers13133109.
12. Grigorieva, J.; Asmellash, S.; Net, L.; Tsypin, M.; Roder, H.; Roder, J. Mass Spectrometry-Based Multivariate Proteomic Tests for Prediction of Outcomes on Immune Checkpoint Blockade Therapy: The Modern Analytical Approach. *Int. J. Mol. Sci.* **2020**, *21*, 838, doi:10.3390/ijms21030838.
13. Kasimir-Bauer, S.; Roder, J.; Obermayr, E.; Mahner, S.; Vergote, I.; Loverix, L.; Braicu, E.; Sehouli, J.; Concin, N.; Kimmig, R.; et al. Definition and Independent Validation of a Proteomic-Classifier in Ovarian Cancer. *Cancers* **2020**, *12*, 2519, doi:10.3390/cancers12092519.
14. Ascierto, P.A.; Capone, M.; Grimaldi, A.M.; Mallardo, D.; Simeone, E.; Madonna, G.; Roder, H.; Meyer, K.; Asmellash, S.; Oliveira, C.; et al. Proteomic Test for Anti-PD-1 Checkpoint Blockade Treatment of Metastatic Melanoma with and without BRAF Mutations. *J. Immunother. Cancer* **2019**, *7*, 91, doi:10.1186/s40425-019-0569-1.
15. Carbone, D.P.; Salmon, J.S.; Billheimer, D.; Chen, H.; Sandler, A.; Roder, H.; Roder, J.; Tsypin, M.; Herbst, R.S.; Tsao, A.S.; et al. VeriStrat® Classifier for Survival and Time to Progression in Non-Small Cell Lung Cancer (NSCLC) Patients Treated with Erlotinib and Bevacizumab. *Lung Cancer* **2010**, *69*, 337–340, doi:10.1016/j.lungcan.2009.11.019.
16. Carbone, D.P.; Ding, K.; Roder, H.; Grigorieva, J.; Roder, J.; Tsao, M.-S.; Seymour, L.; Shepherd, F.A. Prognostic and Predictive Role of the VeriStrat Plasma Test in Patients with Advanced Non–Small-Cell Lung Cancer Treated with Erlotinib or Placebo in the NCIC Clinical Trials Group BR. 21 Trial. *J. Thorac. Oncol.* **2012**, *7*, 1653–1660, doi:10.1097/jto.0b013e31826c1155.
17. Kuiper, J.; Lind, J.; Groen, H.; Roder, J.; Grigorieva, J.; Roder, H.; Dingemans, A.; Smit, E. VeriStrat® Has Prognostic Value in Advanced Stage NSCLC Patients Treated with Erlotinib and Sorafenib. *Br. J. Cancer* **2012**, *107*, 1820–1825, doi:10.1038/bjc.2012.470.
18. Gautschi, O.; Dingemans, A.-M.; Crowe, S.; Peters, S.; Roder, H.; Grigorieva, J.; Roder, J.; Zappa, F.; Pless, M.; Brutsche, M.; et al. VeriStrat® Has a Prognostic Value for Patients with Advanced Non-Small Cell Lung Cancer Treated with Erlotinib and

Bevacizumab in the First Line: Pooled Analysis of SAKK19/05 and NTR528. *Lung Cancer* **2013**, *79*, 59–64, doi:10.1016/j.lungcan.2012.10.006.
19. Stinchcombe, T.E.; Roder, J.; Peterman, A.H.; Grigorieva, J.; Lee, C.B.; Moore, D.T.; Socinski, M.A. A Retrospective Analysis of VeriStrat Status on Outcome of a Randomized Phase II Trial of First-Line Therapy with Gemcitabine, Erlotinib, or the Combination in Elderly Patients (Age 70 Years or Older) with Stage IIIB/IV Non–Small-Cell Lung Cancer. *J. Thorac. Oncol.* **2013**, *8*, 443–451, doi:10.1097/jto.0b013e3182835577.
20. Grossi, F.; Genova, C.; Rijavec, E.; Barletta, G.; Biello, F.; Dal Bello, M.G.; Meyer, K.; Roder, J.; Roder, H.; Grigorieva, J. Prognostic Role of the VeriStrat Test in First Line Patients with Non-Small Cell Lung Cancer Treated with Platinum-Based Chemotherapy. *Lung Cancer* **2018**, *117*, 64–69, doi:10.1016/j.lungcan.2017.12.007.
21. Fidler, M.J.; Fhied, C.L.; Roder, J.; Basu, S.; Sayidine, S.; Fughhi, I.; Pool, M.; Batus, M.; Bonomi, P.; Borgia, J.A. The Serum-Based VeriStrat® Test Is Associated with Proinflammatory Reactants and Clinical Outcome in Non-Small Cell Lung Cancer Patients. *BMC Cancer* **2018**, *18*, 310, doi:10.1186/s12885-018-4193-0.
22. Yasui, Y.; McLerran, D.; Adam, B.-L.; Winget, M.; Thornquist, M.; Feng, Z. An Automated Peak Identification/Calibration Procedure for High-Dimensional Protein Measures from Mass Spectrometers. *J. Biomed. Biotechnol.* **2003**, *2003*, 242, doi:10.1155/s111072430320927x.
23. Morris, J.S.; Coombes, K.R.; Koomen, J.; Baggerly, K.A.; Kobayashi, R. Feature Extraction and Quantification for Mass Spectrometry in Biomedical Applications Using the Mean Spectrum. *Bioinformatics* **2005**, *21*, 1764–1775, doi:10.1093/bioinformatics/bti254.
24. Gibb, S.; Strimmer, K. MALDIquant: A Versatile R Package for the Analysis of Mass Spectrometry Data. *Bioinformatics* **2012**, *28*, 2270–2271, doi:10.1093/bioinformatics/bts447.
25. Lange, E.; Gröpl, C.; Reinert, K.; Kohlbacher, O.; Hildebrandt, A. High-Accuracy Peak Picking of Proteomics Data Using Wavelet Techniques. In *Biocomputing 2006*; World Scientific: Singapore; 2006; pp. 243–254.
26. Du, P.; Kibbe, W.A.; Lin, S.M. Improved Peak Detection in Mass Spectrum by Incorporating Continuous Wavelet Transform-Based Pattern Matching. *Bioinformatics* **2006**, *22*, 2059–2065, doi:10.1093/bioinformatics/btl355.
27. Dubrovkin, J. Evaluation of the Peak Location Uncertainty in Second-Order Derivative Spectra. Case Study: Symmetrical Lines. *Int. J. Emerg. Technol. Comput. Appl. Sci.* **2014**, *7*, 45–53.
28. Picaud, V.; Giovannelli, J.-F.; Truntzer, C.; Charrier, J.-P.; Giremus, A.; Grangeat, P.; Mercier, C. Linear MALDI-ToF Simultaneous Spectrum Deconvolution and Baseline Removal. *BMC Bioinformatics* **2018**, *19*, 123, doi:10.1186/s12859-018-2116-3.
29. Grigorieva, J.; Asmellash, S.; Oliveira, C.; Roder, H.; Net, L.; Roder, J. Application of Protein Set Enrichment Analysis to Correlation of Protein Functional Sets with Mass Spectral Features and Multivariate Proteomic Tests. *Clin. Mass Spectrom.* **2020**, *15*, 44–53, doi:10.1016/j.clinms.2019.09.001.
30. Trede, D.; Kobarg, J.H.; Oetjen, J.; Thiele, H.; Maass, P.; Alexandrov, T. On the Importance of Mathematical Methods for Analysis of MALDI-Imaging Mass Spectrometry Data. *J. Integr. Bioinforma. JIB* **2012**, *9*, 189.
31. Eilers, P.H.; Boelens, H.F. Baseline Correction with Asymmetric Least Squares Smoothing. *Leiden Univ. Med. Cent. Rep.* **2005**, *1*, 5.
32. Boelens, H.F.; Dijkstra, R.J.; Eilers, P.H.; Fitzpatrick, F.; Westerhuis, J.A. New Background Correction Method for Liquid Chromatography with Diode Array Detection, Infrared Spectroscopic Detection and Raman Spectroscopic Detection. *J. Chromatogr. A* **2004**, *1057*, 21–30, doi:10.1016/j.chroma.2004.09.035.
33. Senko, M.W.; Beu, S.C.; McLaffertycor, F.W. Determination of Monoisotopic Masses and Ion Populations for Large Biomolecules from Resolved Isotopic Distributions. *J. Am. Soc. Mass Spectrom.* **1995**, *6*, 229–233, doi:10.1016/1044-0305(95)00017-8.
34. Blaum, K.; Geppert, C.; Müller, P.; Nörtershäuser, W.; Wendt, K.; Bushaw, B. Peak Shape for a Quadrupole Mass Spectrometer: Comparison of Computer Simulation and Experiment. *Int. J. Mass Spectrom.* **2000**, *202*, 81–89, doi:10.1016/s1387-3806(00)00237-2.
35. Foxon, C.; Joyce, B.; Holloway, S. Instrument Response Function of a Quadrupole Mass Spectrometer Used in Time-of-Flight Measurements. *Int. J. Mass Spectrom. Ion Phys.* **1976**, *21*, 241–255, doi:10.1016/0020-7381(76)80125-8.
36. Osborn, D.L.; Zou, P.; Johnsen, H.; Hayden, C.C.; Taatjes, C.A.; Knyazev, V.D.; North, S.W.; Peterka, D.S.; Ahmed, M.; Leone, S.R. The Multiplexed Chemical Kinetic Photoionization Mass Spectrometer: A New Approach to Isomer-Resolved Chemical Kinetics. *Rev. Sci. Instrum.* **2008**, *79*, 104103, doi:10.1063/1.3000004.
37. Yang, C.; He, Z.; Yu, W. Comparison of Public Peak Detection Algorithms for MALDI Mass Spectrometry Data Analysis. *BMC Bioinformatics* **2009**, *10*, 4, doi:10.1186/1471-2105-10-4.
38. Danielsson, R.; Bylund, D.; Markides, K.E. Matched Filtering with Background Suppression for Improved Quality of Base Peak Chromatograms and Mass Spectra in Liquid Chromatography–Mass Spectrometry. *Anal. Chim. Acta* **2002**, *454*, 167–184, doi:10.1016/s0003-2670(01)01574-4.
39. Ahn, S.-M.; Simpson, R.J. Body Fluid Proteomics: Prospects for Biomarker Discovery. *PROTEOMICS–Clinical Appl.* **2007**, *1*, 1004–1015, doi:10.1002/prca.200700217.
40. Müller, A.C.; Breitwieser, F.P.; Fischer, H.; Schuster, C.; Brandt, O.; Colinge, J.; Superti-Furga, G.; Stingl, G.; Elbe-Bürger, A.; Bennett, K.L. A Comparative Proteomic Study of Human Skin Suction Blister Fluid from Healthy Individuals Using Immunodepletion and ITRAQ Labeling. *J. Proteome Res.* **2012**, *11*, 3715–3727, doi:10.1021/pr3002035.

41. Subramanian, A.; Tamayo, P.; Mootha, V.K.; Mukherjee, S.; Ebert, B.L.; Gillette, M.A.; Paulovich, A.; Pomeroy, S.L.; Golub, T.R.; Lander, E.S.; et al. Gene Set Enrichment Analysis: A Knowledge-Based Approach for Interpreting Genome-Wide Expression Profiles. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 15545–15550, doi:10.1073/pnas.0506580102.
42. Roder, J.; Linstid, B.; Oliveira, C. Improving the Power of Gene Set Enrichment Analyses. *BMC Bioinformatics* **2019**, *20*, 257, doi:10.1186/s12859-019-2850-1.
43. Savitzky, A.; Golay, M.J.E. Smoothing and Differentiation of Data by Simplified Least Squares Procedures. *Anal. Chem.* **1964**, *36*, 1627–1639, doi:10.1021/ac60214a047.
44. Ryan, C.G.; Clayton, E.; Griffin, W.L.; Sie, S.H.; Cousens, D.R. SNIP, a Statistics-Sensitive Background Treatment for the Quantitative Analysis of PIXE Spectra in Geoscience Applications. *Nucl. Instrum. Methods Phys. Res. Sect. B Beam Interact. Mater. At.* **1988**, *34*, 396–402, doi:10.1016/0168-583x(88)90063-8.
45. Tibshirani, R.; Hastie, T.; Narasimhan, B.; Soltys, S.; Shi, G.; Koong, A.; Le, Q.-T. Sample Classification from Protein Mass Spectrometry, by "Peak Probability Contrasts. *Bioinformatics* **2004**, *20*, 3034–3044, doi:10.1093/bioinformatics/bth357.
46. Gold, L.; Ayers, D.; Bertino, J.; Bock, C.; Bock, A.; Brody, E.; Carter, J.; Cunningham, V.; Dalby, A.; Eaton, B.; et al. Aptamer-Based Multiplexed Proteomic Technology for Biomarker Discovery.*PLoS ONE* **2010**, 5, 12, e16004 doi: 10.1371/journal.pone.0015004. .
47. Kraemer, S.; Vaught, J.D.; Bock, C.; Gold, L.; Katilius, E.; Keeney, T.R.; Kim, N.; Saccomano, N.A.; Wilcox, S.K.; Zichi, D.; et al. From SOMAmer-Based Biomarker Discovery to Diagnostic and Clinical Applications: A SOMAmer-Based, Streamlined Multiplex Proteomic Assay. *PLoS ONE* **2011**, *6*, e26332, doi:10.1371/journal.pone.0026332.
48. Benjamini, Y.; Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B Methodol.* **1995**, *57*, 289–300, doi:10.1111/j.2517-6161.1995.tb02031.x.