Published in final edited form as:

J Proteome Res. 2020 July 02; 19(7): 2664–2675. doi:10.1021/acs.jproteome.9b00892.

Gain-Scanning for Protein Microarray Assays

Feng Feng,

Boston University School of Medicine, Boston, Massachusetts

Sila Toksoz Ataca,

Boston University School of Medicine, Boston, Massachusetts

Mingxuan Ran,

Boston University School of Medicine, Boston, Massachusetts

Yumei Wang,

Boston University School of Medicine, Boston, Massachusetts

Michael Breen,

Boston University School of Medicine, Boston, Massachusetts

Thomas B. Kepler

Boston University School of Medicine, Boston, Massachusetts, and Boston University, Boston, Massachusetts

Abstract

Protein microarrays consist of known proteins spotted onto solid substrates and are used to perform highly multivariate assessments of protein-binding interactions. Human protein arrays are routinely applied to pathogen detection, immune response biomarker profiling, and antibody specificity profiling. Here, we describe and demonstrate a new data processing procedure, gain-scan, in which data were acquired under multiple photomultiplier tube (PMT) settings, followed by data fitting with a power function model to estimate the incident light signals of the array spots. Data acquisition under multiple PMT settings solves the difficulty of determining the single optimal PMT gain setting and allows us to maximize the detection of low-intensity signals while avoiding the saturation of high-intensity ones at the same time. The gain-scan data acquisition and fitting also significantly lower the variances over the detectable range of signals and improve the linear data normalization. The performance of the proposed procedure was verified by analyzing the profiling data of both the human polyclonal serum samples and the monoclonal antibody

Corresponding Authors: Feng Feng – Phone: 617 638 4123; ffeng@bu.edu; Fax: 617 638 4286, Thomas B. Kepler – Phone: 617 638 4124; tbkepler@bu.edu; Fax: 617 638 4286.

Author Contributions

T.B.K. and F.F. conceived of the idea. F.F., Y.W., S.A., and M.B. carried out the experiments. F.F., T.B.K., S.T.A., and M.R. ran the data analyses and wrote the manuscript. All authors have given approval to the final version of the manuscript.

The authors declare no competing financial interest.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at <https://pubs.acs.org/doi/10.1021/acs.jproteome.9b00892>.

Figures of signal intensities, raw signal strengths, volcano plots, model fittings and inter-array variances and tables of dynamic range, discontinued commercial microarray scanners, sample demographics, and multiple sclerosis (MS) serum biomarkers (PDF)

Complete contact information is available at: <https://pubs.acs.org/10.1021/acs.jproteome.9b00892>

samples with both technical replicates and biological replicates. We showed that the multigain power function was an appropriate model for describing data acquired under multiple PMT settings. The gain-scan fitting alone or in combination with the linear normalization could effectively reduce the technical variability of the array data and lead to better sample separability and more sensitive differential analysis.

Graphical Abstract

Keywords

protein microarray; ProtoArray; data acquisition; data analysis; photomultiplier gain; intra-array and inter-array variability; feature identification

INTRODUCTION

Protein microarray technology provides the means to determine the protein-binding specificity of one or more analytes in solution. Each microarray consists of a large number of distinct proteins immobilized on a solid substrate in well-localized discrete spots. Binding is assessed by incubating the primary and secondary probes sequentially with the microarray and by using imaging to detect the locations and hence the identities of the bound spots on the array. This technology was first studied and developed over 25 years ago.^{1–3} It is now used commonly for screening the binding activities of thousands of proteins simultaneously and efficiently.⁴ There are mainly three types of protein microarrays: analytical microarrays, functional microarrays, and reverse-phase microarrays.¹ Among different types of protein microarrays, ProtoArray Human Protein Microarray was once the most popular human functional protein array adopted by many researchers.^{5–7} It was the first commercial high-density protein microarray and contains over 9000 unique full-length human proteins, including kinases, phosphatases, G-protein coupled receptors (GPCRs), nuclear receptors, and proteases, spotted in duplicate on a 1 in. × 3 in. thin nitrocellulose coated glass slide.⁸ It also includes control proteins in each subarray for quality assessment and normalization.⁹ It enables the rapid elucidation of protein interactions on a proteome scale and has been

extensively applied to pathogen detection,^{5,6,10,11} immune response biomarker profiling,^{12–16} and antibody specificity profiling.^{11,17,18}

In spite of similarities to DNA/oligonucleotide microarrays, the software for the analysis of DNA microarray data is not well-suited to the analysis of protein microarrays. The statistical distribution of the image intensities differs greatly between DNA and protein arrays. For example, protein arrays show a bimodal distribution of signal intensities and exhibit large technical variability (~25% coefficient of variance (CV) for the intra-array variability).⁹ Efforts have been devoted to meet the needs specific for protein array data.^{19,20} Sboner et al. developed a robust linear model (RLM) to normalize protein array data as an important preprocessing step to reduce technical variability.⁹ Turewicz et al.²¹ further formalized and implemented this method into an R package, ProtoArray Analyzer (PAA). PAA also incorporates other useful tools for protein microarray analyses, such as the empirical Bayesian method to adjust the batch effects among samples.²² Protein Microarray Analyzer (PMA) is another open-source software that allows for neighborhood background correction, net intensity correction, and “pin-to-pin” “array-to-array” normalization as well as user-defined intensity and coefficient of variation (CV) thresholds to exclude the background spots.²³

Our purpose in undertaking this study is to offer further improvement to the reliability of protein microarray analysis through the introduction of an alternative procedure and the required statistical method to implement it effectively. The key component for imaging microarrays is the photomultiplier tube (PMT), which uses the photoelectric effect to convert incident light into an electron current, amplifies this current through a multiple dynode cascade, and produces an output current many times larger than that produced by the incident photons. This gain is adjustable by the operator and is controlled by the voltage applied across the dynodes. The contemporary practice is to find the optimal gain setting and fix it for data acquisition. If the gain is set too high, features with strong analyte binding, including control features, will have an intensity greater than the maximum that is reportable by the instrument; i.e., they will become saturated. As the gain is decreased, the sensitivity is lost.²⁴

In this work, we provide an alternative approach to optimizing the gain. In this alternative approach, one images the protein array multiple times with a different gain setting for each image and then combines these data in the estimation procedure. Although multiple scanning with different PMTs to achieve an extended dynamic range is not something totally new, it has not been implemented widely. The reason is obvious: The implementation of an extended dynamics range needs extra hardware support²⁵ and the integration of the data needs an appropriate mathematical model to represent the effect of the changing gains. To our knowledge, only two commercially available array scanners have such a function (Supporting Information Tables S1 and S2). Moreover, the available data integration model in these implementations is crude and involves removing data points below an arbitrarily defined signal threshold and a linear regression model.^{26,27} In this work, we develop a more appropriate model based on the principle of the PMT operation with small modifications. Our version of such a model fits the raw data very well and allows us to expand the dynamic range of the instrument while reducing the inter- and intra-array variance substantially. We

demonstrate the method with assays performed with both polyclonal sera and monoclonal antibodies, with both technical replicates and biological replicates, and with acquisition using a large (nine) or small (two) number of PMT gain settings. We also have implemented the method in an R package and made it freely available.

METHOD AND MATERIAL

Protein Microarrays

ProtoArrays (version 5.0, PAH0525101) were purchased from Thermo Fisher Scientific (Waltham, MA).

Antibodies and Sera

Two human antibodies were used. The first one was the human IgG isotype control purchased from Sigma (St. Louis, MO). The other one was a recombinant human monoclonal antibody, anti-histone H2B IgG (ab213288), from Abcam (Cambridge, MA).

Independent serum samples from multiple sclerosis (MS) patients and age- and gender-matched normal controls ($n = 5$ for each group) were obtained from *BioIVT* (Westbury, NY). All samples were handled using the vendor suggested standard procedures and stored at -80°C until use. The demographic information for the sera is listed in Supporting Information Table S3.

Array Hybridization and Data Acquisition

The standard manufacturer's protocol was used with minor modifications as previously reported.²⁸ Briefly, the protein microarrays were blocked in a prechilled incubation tray and incubated for 1 h at 4°C with gentle shaking. The blocking buffer was aspirated, and the primary probe, either $1.0\ \mu\text{g/mL}$ of human antibody or 1:500 diluted serum sample, was incubated for 90 min at 4°C without shaking. The primary probe was aspirated, and the microarray was washed for 5 min in 0.1% Tween-20 and $1\times$ synthetic block PA017 (Thermo Fisher Scientific, Waltham, MA) in PBS with gentle shaking. The secondary Alexa-Fluor 647 antihuman IgG antibody (Thermo Fisher Scientific, Waltham, MA) was added at $1.0\ \mu\text{g/mL}$ and incubated for 90 min at 4°C with gentle shaking. Microarrays were washed for 5 min with gentle shaking and dried by centrifugation at $200g$ for 1 min at room temperature.

Microarrays were scanned using a GenePix 4000B scanner (Molecular Devices, San Jose, CA) at a wavelength of 635 nm, with $10\ \mu\text{m}$ resolutions, using 100% power. The arrays probed with the isotype control IgG were scanned with nine PMT gain voltages, 250, 300, 350, 400, 450, 500, 550, 600, and 650, and the ones with the anti-histone H2B antibody or the sera were scanned with five PMT gains, 350, 450, 550, 600, 650, and 700. Fluorescence intensities were quantified with the GenePix Pro 5.0 software (Molecular Devices, San Jose, CA) using the lot-specific protein location information provided by the manufacturer (of the protein array).

After the acquisition, the protein array data were exported as .GPR files, which were loaded into the software package R for preprocessing, logarithm-transformation, and gain-scan

power function model fitting (see below for the model and fitting). The raw and processed data were also normalized with the RLM in order to compare the performance.^{9,29}

Serum Protein Array Data from the Literature

The microarray data sets from the work by Sboner et al.⁹ were downloaded from <http://archive.gersteinlab.org/proj/proteinchipRLM/>. In their work, two sera were utilized, normal serum and positive control sera. The former is a sample from a subject with no known pathology and the latter is an equal mixture of six independent sera qualified for reactivity to different human proteins. Each serum was probed by ProtoArrays six times (they are technical replicates). The data were acquired with two PMT voltages, 550 and 700. Detailed information about the data set is described in the original reference.⁹ The data sets were loaded into R and similarly preprocessed as mentioned above.

PMT Gain-Scan Power Function Modeling and Data Fitting

To acquire data, microarray scanners rely on a PMT to collect the incident light signals and turn them into amplified electrical signals.³⁰ The signal amplification, termed as the PMT gain, is determined by the voltage applied to the PMT dynodes.^{30,31} In the simplest model, the PMT output signal y depends linearly on the incident light flux ϕ , with the constant of proportionality an adjustable parameter known as the gain G . The PMT theory models the gain as a function of the adjustable dynode voltage v : $g(v) = \gamma(v/v_0)^\delta$, where δ is the number of dynodes in the amplification cascade and γ and v_0 are parameters related to the design of the PMT. The light flux itself is taken to be proportional to the quantity of analyte bound to the given spot. We expect nonlinearity in the relationship between the bound analyte and the light flux, but in our assays, such nonlinearities are undetectable. Finally, note that we are not concerned with the absolute value of the light flux but only with the ratios of fluxes. For that reason, we may absorb both γ and v_0 into the light flux measurement units and write simply

$$y = \phi v^\delta \quad (1)$$

In practice, however, this ideal model must be revised to account for several additional components. The first is a background signal, comprising the so-called dark current as well as other sources of signal independent of the light flux.^{30–33} Then a term for random fluctuations, or measurement error, and an envelope function enforcing the signal saturation at high intensities make the model complete. The background signal is accommodated by adding a constant term to eq 1. The error term appears as a normal random variable ϵ with a mean zero and unknown variance σ^2 added to the logarithm of the “true” signal. The choice of using the logarithm and adding an error gives a reasonable and empirically validated relationship between the signal mean and variance. The envelope function is a simple piecewise function equal to the identity function below a fixed level Λ and having a constant value Λ above that level. This envelope function is expected in the analogue-to-digital transformation (AD) and is implemented in the hardware. In our specific case, for example, the signal is stored as a 16-bit integer. Values above $\Lambda = 2^{16} - 1$ are simply reported as Λ . There are potentially other sources of signal saturation besides that induced by the AD

conversion, but we do not find any discernible nonlinearity below the cutoff. Therefore, the model is written as

$$\log(Y) = \begin{cases} \log(\beta + \phi v^\delta) + \varepsilon & \text{for } (\beta + \phi v^\delta)e^\varepsilon \leq \Lambda \\ \log \Lambda & \text{for } (\beta + \phi v^\delta)e^\varepsilon > \Lambda \end{cases} \quad (2)$$

where the measured signal Y is treated as a random variable. The way this model is used in practical data analysis depends on the experimental design and choices made about the sources of variability among the parameters. Λ , for example, is a fixed and known constant across all experiments and therefore does not need to be estimated. The voltage v is controlled by the analyst and so is not estimated. β and δ are treated as independent of the spot, constant within an array, and variable among arrays. The gain exponent δ has a simple interpretation as the number of dynodes in the amplification cascade (and so could be treated as known with no need to estimate it). It is known that there are small corrections that make the effective exponent slightly different from the theoretical value. Indeed, we do see slight differences among arrays. The analyses shown below are not sensitive to whether we take δ as fixed or estimated for each array, but we err on the side of caution and estimate δ for each array. For a given analyte solution, the binding to each spot is different and so is the resulting light flux, which we take to be characteristic of the spot/solution combination independent of the array. On the basis of all these considerations, we have the statistical model as

$$\log(Y_{ij}) = \begin{cases} \log(\beta_j + \phi_i v^\delta) + \varepsilon_{ij} & \text{for } (\beta_j + \phi_i v^\delta)e^{\varepsilon_{ij}} \leq \Lambda \\ \log \Lambda & \text{for } (\beta_j + \phi_i v^\delta)e^{\varepsilon_{ij}} > \Lambda \end{cases} \quad (3)$$

$i = 1, 2, \dots, n; j = 1, 2, \dots, m$

where n is the total number of proteins on each array and m the total number of arrays.

The key characteristic of this model for integrating data across voltages is its scaling property. The PMT gain function is

$$f(v, \phi) = \phi v^\delta \quad (4)$$

The scaling property of interest is

$$f(v, s\phi) = f(s^{1/\delta}v, \phi) \quad (5)$$

The importance of this property is that there is essentially a single reference function and that the intensity of each feature on the array may be expressed in terms of a standard intensity and translated back to a standard curve.

The fitting is carried out using the Levenberg–Marquardt nonlinear least-squares algorithm (R MINPACK package). We will show in the Results and Discussion section that the above model is appropriate for ProtoArray data and the assumptions are satisfactorily met.

Data Variability and Differential Analysis

Following Sboner et al.,⁹ we assess the performance of the proposed gain-scan acquisition and data fitting method by computing the intra- and inter-array variations for data variability before and after processing. To evaluate intra-array variation, we computed the coefficient of variation as

$$CV = \frac{s}{\bar{x}} \quad (6)$$

where $\bar{x}$ and s are respectively the sample mean and standard deviation for the log signal intensities for each control protein spotted in different subarrays (blocks) within the same array. To assess the inter-array variability, CV was computed similarly but for the same protein spot across multiple arrays. To evaluate the separability of the profiling data of two group of samples, again following Sboner et al.,⁹ we computed Fisher's signal-to-noise ratio

$$S = \frac{(\bar{x}_1 - \bar{x}_2)^2}{s_1^2 + s_2^2} \quad (7)$$

where $\bar{x}_1$ and $\bar{x}_2$ are the sample mean intensities of a given protein feature within group 1 and group 2, respectively, and s_1 and s_2 are the sample standard deviations. The ratios were computed on either raw protein array data or preprocessed data generated with different methods (RLM normalized, gain-scan fitted, or both). For calculations done on raw and RLM-normalized data, we chose the data set acquired at one single PMT gain setting, the PMT voltage of 650 for the monoclonal antibody and MS serum experiments, and the voltage of 700 for the serum data by Sboner et al.,⁹ since those are recommended by the scanner manufacturer. For the proposed gain-scan method, all the data acquired at different PMT settings were fitted to estimate the original incident light signals, which were then used for the analyses.

We also carried out the differential analysis comparing the anti-H2B antibody with the isotype control to identify the possible targets of the anti-H2B antibody. Each group had three microarrays probed with the antibody or control. The array data were preprocessed with the RLM or the gain-scan fitting or both. Then, individual t tests were carried out for each feature point comparing the two groups and the obtained p -values were adjusted by the false discovery rate (FDR) q -values to correct the multiple comparison error.^{34–36} To identify the targets of the anti-H2B antibody, the features were sorted on the basis of the FDR q -values in increasing order. The top ones with the smallest q -values were treated as the potential targets. The performance was assessed by the true discovery rate among the top N selected features. The true discovery was defined according to the signal responses of Histone H2B isoforms on ProtoArray. There are nine available Histone H2B isoforms on the ProtoArray, HIST1H2BA, HIST1H2BB, HIST1H2BD, HIST1H2BH, HIST1H2BM, HIST1H2BN, HIST1H2BO, HIST3H2BB, and HISTONE, which all showed much stronger signals on arrays probed with the anti-H2B antibody than those of the IgG isotype control group. The true positive discovery rates were plotted and the areas under the curve (AUC)³⁷ were also computed for evaluation.

For the MS sera data, the differential analysis was done similarly. The MS serum group and the control group ($n = 5$ for each group) were compared, the t tests were carried out, and the FDR q -values were obtained. The positive hits were determined by the following criteria: a minimum signal intensity of 1500 RU, a minimum 1.4-fold difference between two groups, and a p -value lower than 0.1 for the t test. The true serum biomarkers for MS were defined on the basis of the work done by DeMarshall et al.,¹⁴ and a total of 61 markers available on ProtoArray were chosen in this study (Supporting Information Table S4). The true positive discoveries were plotted, and the AUCs were computed.

All analyses were carried out within the R software environment with the following packages: limma, minpak.lm, ggplot2, and qvalue, as well as our gain-scan fitting package (available to the public at <https://github.com/BULQI/gainscan>).

RESULTS AND DISCUSSION

Signal Intensity Distribution in Protein Microarray

The human IgG isotype control (Sigma) and monoclonal anti-histone H2B IgG antibody (Abcam) were hybridized to ProtoArrays according to the manufacturer's instruction. Then, arrays were scanned with nine different PMT gain settings ranging from 250 up to 650. The first blocks of the same array scanned at the voltages of 250, 450, and 650 are shown in Figure 1. Each circular spot shown on the array is for one of the duplicate protein features. High intensity indicates avid binding between the analyte (IgG) and the protein printed at that position. The image scanned at the lowest PMT setting (voltage of 250) shows blank spots at many positions, while increasing the gain voltage enhances the intensity for many spots. At the highest PMT setting, however, four features are saturated as indicated by their whiteness. A saturated pixel is not an accurate measurement of the signal from the pixel.²⁴ As suggested by the manufacturer, on one hand the PMT gain needs to be large enough to obtain a good sensitivity, while on the other hand they need to remain low enough to avoid saturation of the most avidly bound features.

Next, we studied the distribution and variation of signal intensities in the protein arrays. We first examined the signal distributions of arrays hybridized with the monoclonal anti-H2B antibody and the isotype control. Parts A and B of Figure 2 show the distributions of raw and RLM normalized signals. In the case of the monoclonal anti-histone H2B antibody, the empirical density function appears unimodal and the variation between the technical replicates is low, while in the case of the isotype IgG control, a substantial technical variation among the replicates is evident. In this case, the RLM normalization did not lower these technical variations. This negative result is likely because the control proteins used for normalization were high in signal intensity and some of them were saturated with intensities close to or equal to the AD cutoff $2^{16} - 1$ (Figure S1). Such saturation resulted in biased and noisy normalization and therefore poor data processing, and this saturation effect can be partly corrected by acquiring data at lower PMTs (Figure S2).

Following Sboner et al.,⁹ we examined the signal distributions of arrays probed with the polyclonal sera, either a positive control serum or a normal one (Figure 2C,D). In this case, bimodal distributions were observed. Variations among technical replicates were obvious;

the RLM normalization was effective in reducing the technical variation. Notably, the control proteins contained no saturated signals (Figure S3).

To reduce the technical variation and improve the data quality, we propose to (1) scan each array multiple times with different PMT voltages/settings, (2) fit the acquired data with a power function model to estimate the original incident light signal for each spot, and (3) (optionally) normalize the estimated light signals with the RLM.

Multigain Power Function Data Fitting

In this section, we examine the quality of the processed data obtained using our gain-scan method. Two data sets were used here. First, we used the IgG isotype control protein array data (Figure 3A,B). For each antibody-probed array, nine PMT settings were used to scan and acquire the data, which were then fitted to the power function. Figure 3A shows the processed signals of different features over PMT voltages. We further used the scaling property of the intensity function to shift and overlay all the data onto a single reference curve illustrating the effectiveness of the method for covering a wide range of voltages (Figure 3B). For the second example, we used the data from the polyclonal sera generated and made public by Sboner et al.⁹ In this project, arrays were probed with either the polyclonal positive control or the normal sera and then data were acquired with only two PMT voltages, 550 and 700. The gain-scan power function model fitted these data very well too (Figure 3C,D).

Intra-array and Inter-array Variability of Protein Microarray Data

Figure 4 shows that gain-scan reduced the intra-array variances substantially across features with both high- and low-intensity signals without changing their mean values. In contrast, RLM normalization had little or no effect on the intra-array variances of data. Similar results were also observed using data from the polyclonal sera. The distributions of the intra-array variances are summarized in Figure 4D,E, and the overall mean intra-array variances were largely reduced for both isotype IgG treated ProtoArrays and the sera array data from Sboner et al.⁹

The RLM alone can only slightly lower the inter-array variances of the data with technical replicates (Figure 5A), while the variance-reducing effect of the gain-scan is much bigger (Figure 5B,D,E). Moreover, we observed that RLM and gain-scan synergized to achieve a better performance in processing the polyclonal serum data (Figure 5E) but not so in the IgG control data (Figure 5D).

Sample Separability and Differential Analysis

As shown above, the gain-scan fitting effectively reduces both inter- and intra-array variations (Figures 4 and 5). In this section, we further demonstrate that the proposed preprocessing procedure reduces variances without introducing bias. Again, following Sboner et al., we used Fisher's signal-to-noise ratio.⁹ The array data were either unprocessed or processed by the RLM, gain-scan, or both, and then, the signal-to-noise ratios were computed for each case.

The first analysis was carried out on the protein array data generated by Sboner et al.⁹ In their work, there were two groups of polyclonal human sera with technical replicates, the positive control sera and the normal sera, which were assumed to have distinct proteomic profiles (see details in Sboner et al.). In agreement with Sboner's results, we find (Figure 6A) that normalization with RLM increases the signal-to-noise ratio as does the gain-scan processing.

Next, we compared the monoclonal anti-histone H2B antibody with the IgG isotype control ($n = 3$ for each group). In this case, similar results were obtained. As shown in Figure 6B, the raw data still show some level of separability between the anti-histone H2B antibody and the control with the nonzero mean Fisher's signal-to-noise ratio. Processing with either RLM or the gain-scan method increases the signal-to-noise ratio, and no obvious synergistic effect is observed

One major application of protein arrays is to compare the binding profiles among different analyte solutions to identify the binding partners that differ between them. To this end, we ran a differential analysis on the array data of the monoclonal anti-histone H2B antibody against the IgG isotype control. First, individual t tests were carried out for features comparing the two groups ($n = 3$ for each group). Then, FDR q -values were computed on the basis of the p -values of the t tests to control for the multiple comparison error.^{34–36} Those features with the smallest q -values were selected as the potential targets of the tested antibody. Then, the true positive discovery was obtained to assess the performance of the analysis. On ProtoArray, there are nine histone H2B isoforms. The monoclonal anti-histone H2B antibody showed good reactivity and specificity to all nine histone H2B isoforms (Table 1 and Supporting Information Figures S4 and S5). Therefore, the signal responses of Histone H2B isoforms on ProtoArray were used to define the criteria of the “true” targets of the tested antibody, log intensity ratio greater than 5.0, and a FDR q -value smaller than 0.003 (Table 1, Figure 6C, and Supporting Information Figures S4, S5, and S6). As shown in Figure 6D, the gain-scan acquisition and fitting (alone or in combination with the RLM) can help to improve the differential analysis and result in selecting more true discoveries. The true discovery curves have higher AUC values than do the ones prior to the application of multigain processing. Furthermore, volcano plots (Figure 6C and Supporting Information Figure S6) provide another way to visualize the salutary changes brought about by the gain-scan processing. Overall, it is obvious to see that the gain-scan acquisition and fitting could lower the probability (therefore smaller FDR q -values) of t tests of comparison (Figure 6C, Supporting Information Figure S6, and Table 1) and can slightly increase the fold changes between two groups (in Figure 6D). As a result, the processed data allowed for sensitive identification. Most likely the RLM and gain-scan achieve their effects by reducing the data variation and correcting the data skewness caused by technical factors. The former can be seen in Figures 4 and 5 in the previous sections and the latter in Figure 6C and Figure S6 in this section.

We further test the gain-scan method under a disease setting and with biological replicates. Independent sera from multiple sclerosis (MS) patients and normal controls ($n = 5$ for each group) were obtained and profiled using the gain-scan method with six different PMTs (between 350 and 700). First, we can see that the data fits the gain-scan power function

model well (Figure 7A). Next, the performance of gain-scan was compared. As a result, the gain-scan method can effectively lower both the intra-array and inter-array variances (Figure 7B,C). It can also increase Fisher's signal-to-noise ratios comparing the MS sera data with the normal control (Figure 7D). Lastly, we also tested the effect of the gain-scan processing on the identification of the serum MS biomarkers. We compared the MS sera with the control sera. The positive hits are defined with the criteria as a minimum signal intensity higher than 1500 RFU,¹⁴ a p -value of the t test smaller than 0.1, and a log fold difference larger than 1.4. The true MS serum biomarkers were determined on the basis of the previous work done by DeMarshall et al. using 51 cases and 31 controls (Supporting Information Table S4).¹⁴ We can see from the true discovery plot (Figure 7E) that the gain-scan data integration increases the true discovery of the MS biomarkers compared with the unprocessed raw data. One thing worth noting is that in this experiment we did not apply the RLM normalization. The reason is that the RLM assumes that the control features in different arrays are saturated and reflect only the technical variations.⁹ In this experiment, however, the control protein (human Igs) cannot be assumed to a good reference, since lower/depleted blood immunoglobulins could be one of the manifestations of MS.¹⁴

Minimum Number of PMT Scans for Gain-Scan to Achieve the Variance Reduction Effect

A full scan of microarray at high resolutions takes time. For example, on a GenePix 4000B scanner it takes more than 7 min to scan a ProtoArray microarray. Therefore, the gain-scan method with six PMTs takes ~50 min for the data acquisition of each array. In order to shorten the experimental time, we tried to determine the minimum number of scans that is necessary to achieve the model fitting and the variance reduction. We carried out experiments and data analyses on both the monoclonal antibody and MS serum data. As a result, we found out that as few as two PMT scans could reach the goal (Figure 3C,D and Supporting Information Figure S8).

CONCLUSION

The power function with baseline is an appropriate model describing the protein array data acquired under different PMT settings/voltages. Instead of choosing one single optimal PMT gain setting for data acquisition, the multigain data acquisition and the following power function data fitting can effectively avoid saturation of signals and reduce the technical data variations (intra- and inter-array variations). Most importantly, the gain-scan processing likely achieves the above-mentioned impacts without distorting the real biological difference, leads to better sample separability, and improves the differential analysis.

We note the unfortunate circumstance that while our work was being completed, Thermo-Fisher announced the discontinuance of the ProtoArray brand. We believe our method will provide reliable inferences, after a re-estimation of the global parameters, with all similar devices. This re-estimation is essential for the unbiased use of the gain-scan method.

SOFTWARE AVAILABILITY

An R package, protein array software tool (PAST), implementing the gain-scan fitting with the power function model and the RLM normalization, can be downloaded from <https://github.com/BULQI/gainscan>

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Funding

This work was supported by NIH/NIAID research contract 1 U19 AI117892-01 (TBK, PI) and U19AI117905 (JEC, PI).

ABBREVIATIONS

PMT	photomultiplier tubes
RLM	robust linear model
CV	coefficient of variation

REFERENCES

- (1). Hall DA; Ptacek J; Snyder M Protein microarray technology. *Mech. Ageing Dev* 2007, 128 (1), 161–167. [PubMed: 17126887]
- (2). Macbeath G; Schreiber SL Printing Proteins as Microarrays for High-Throughput Function Determination. *Science* 2000, 289 (5485), 1760–1763. [PubMed: 10976071]
- (3). Templin MF; Stoll D; Schrenk M; Traub PC; Vöhringer CF; Joos TO Protein microarray technology. *Trends Biotechnol.* 2002, 20 (4), 160–6. [PubMed: 11906748]
- (4). Stoevesandt O; Taussig MJ; He M Protein microarrays: high-throughput tools for proteomics. *Expert Rev. Proteomics* 2009, 6 (2), 145–157. [PubMed: 19385942]
- (5). Auger I; Balandraud N; Rak J; Lambert N; Martin M; Roudier J New autoantigens in rheumatoid arthritis (RA): screening 8268 protein arrays with sera from patients with RA. *Ann. Rheum. Dis* 2009, 68 (4), 591. [PubMed: 18957483]
- (6). Tribollet L; Cantacessi C; Pickering DA; Navarro S; Doolan DL; Trieu A; Fei H; Chao Y; Hofmann A; Gasser RB; Giacomini PR; Loukas A Probing of a human proteome microarray with a recombinant pathogen protein reveals a novel mechanism by which hookworms suppress B-cell receptor signaling. *J. Infect. Dis* 2015, 211 (3), 416–25. [PubMed: 25139017]
- (7). Satoh J; Obayashi S; Misawa T; Sumiyoshi K; Oosumi K; Tabunoki H Protein microarray analysis identifies human cellular prion protein interactors. *Neuropathol. Appl. Neurobiol* 2009, 35 (1), 16–35. [PubMed: 18482256]
- (8). Thermo Fisher Scientific, ProtoArray™ Human Protein Microarray. <https://www.thermofisher.com/order/catalog/product/PAH0525101?SID=srch-hj-PAH0525101> (3 27).
- (9). Sboner A; Karpikov A; Chen G; Smith M; Dawn M; Freeman-Cook L; Schweitzer B; Gerstein MB Robust-linear-model normalization to reduce technical variability in functional protein microarrays. *J. Proteome Res* 2009, 8, 5451–5464. [PubMed: 19817483]
- (10). Fernandez S; Cisney ED; Tikhonov AP; Schweitzer B; Putnak RJ; Simmons M; Ulrich RG Antibody recognition of the dengue virus proteome and implications for development of vaccines. *Clin. Vaccine Immunol* 2011, 18 (4), 523–32. [PubMed: 21270280]

- (11). Yang G; Holl TM; Liu Y; Li Y; Lu X; Nicely NI; Kepler TB; Alam SM; Liao HX; Cain DW; Spicer L; VandeBerg JL; Haynes BF; Kelsoe G Identification of autoantigens recognized by the 2F5 and 4E10 broadly neutralizing HIV-1 antibodies. *J. Exp. Med* 2013, 210 (2), 241–56. [PubMed: 23359068]
- (12). Sigdel TK; Li L; Tran TQ; Khatri P; Naesens M; Sansanwal P; Dai H; Hsieh SC; Sarwal MM Non-HLA antibodies to immunogenic epitopes predict the evolution of chronic renal allograft injury. *J. Am. Soc. Nephrol* 2012, 23 (4), 750–63. [PubMed: 22302197]
- (13). DeMarshall C; Oh E; Kheirkhah R; Sieber F; Zetterberg H; Blennow K; Nagele RG Detection of early-stage Alzheimer's pathology using blood-based autoantibody biomarkers in elderly hip fracture repair patients. *PLoS One* 2019, 14 (11), No. e0225178. [PubMed: 31730624]
- (14). DeMarshall C; Goldwaser EL; Sarkar A; Godsey GA; Acharya NK; Thayasivam U; Belinka BA; Nagele RG Autoantibodies as diagnostic biomarkers for the detection and subtyping of multiple sclerosis. *J. Neuroimmunol* 2017, 309, 51–57. [PubMed: 28601288]
- (15). Furukawa H; Oka S; Shimada K; Masuo K; Nakajima F; Funano S; Tanaka Y; Komiya A; Fukui N; Sawasaki T; Tadokoro K; Nose M; Tsuchiya N; Tohma S Autoantibody Profiles in Collagen Disease Patients with Interstitial Lung Disease (ILD): Antibodies to Major Histocompatibility Complex Class I-Related Chain A (MICA) as Markers of ILD. *Biomarker Insights* 2015, 10, 63–73. [PubMed: 26327779]
- (16). Li Y; Nakka M; Kelly AJ; Lau CC; Krailo M; Barkauskas DA; Hicks JM; Man TK p27 Is a Candidate Prognostic Biomarker and Metastatic Promoter in Osteosarcoma. *Cancer Res.* 2016, 76 (13), 4002–11. [PubMed: 27197201]
- (17). Stafford P; Halperin R; Legutki JB; Magee DM; Galgiani J; Johnston SA Physical characterization of the "immunosignaturing effect". *Mol. Cell. Proteomics* 2012, 11 (4), M111.011593.
- (18). Brandle SM; Obermeier B; Senel M; Bruder J; Mentele R; Khademi M; Olsson T; Tumani H; Kristoferitsch W; Lottspeich F; Wekerle H; Hohlfeld R; Dornmair K Distinct oligoclonal band antibodies in multiple sclerosis recognize ubiquitous self-proteins. *Proc. Natl. Acad. Sci. U. S. A* 2016, 113 (28), 7864–9. [PubMed: 27325759]
- (19). Peschl P; Ramberger M; Hofberger R; Johrer K; Baumann M; Rostasy K; Reindl M Methodological Challenges in Protein Microarray and Immunohistochemistry for the Discovery of Novel Autoantibodies in Paediatric Acute Disseminated Encephalomyelitis. *Int. J. Mol. Sci* 2017, 18 (3), 679–89.
- (20). Sundaresh S; Doolan DL; Hirst S; Mu Y; Unal B; Davies DH; Felgner PL; Baldi P Identification of humoral immune responses in protein microarrays using DNA microarray data analysis techniques. *Bioinformatics* 2006, 22 (14), 1760–6. [PubMed: 16644788]
- (21). Turewicz M; Ahrens M; May C; Marcus K; Eisenacher M PAA: an R/bioconductor package for biomarker discovery with protein microarrays. *Bioinformatics* 2016, 32 (10), 1577–1579. [PubMed: 26803161]
- (22). Johnson WE; Li C; Rabinovic A Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* 2007, 8 (1), 118–27. [PubMed: 16632515]
- (23). Da Gama Duarte J; Goosen RW; Lawry PJ; Blackburn JM PMA: Protein Microarray Analyser, a user-friendly tool for data processing and normalization. *BMC Res. Notes* 2018, 11 (1), 156–61. [PubMed: 29482592]
- (24). Gu Q; Sivanandam TM Optimizing scan parameters for antibody microarray experiments: accelerating robust systems diagnostics for life sciences. *OMICS* 2014, 18 (6), 385. [PubMed: 24754828]
- (25). Corson JF; Curry B; Troup C; Zhou X Extended dynamic range reading of chemical arrays. *US* 2007/0178480 A1, 8 2, 2007.
- (26). Agilent G5761A SureScan Dx Microarray Scanner System. Agilent Technologies, Incorporated: Santa Clara, CA, 2018; Vol. G5761–90000, p 358.
- (27). Agilent, SureScan Dx Microarray Scanner. <https://www.agilent.com/en/product/gene-expression-microarray-platform/gene-expression-microarray-scanners-equipment/surescan-dx-microarray-scanner-228492#productdetails> (8 30).

- (28). Liu M; Yang G; Wiehe K; Nicely NI; Vandergrift N. a.; Rountree W; Bonsignori M; Alam SM; Gao J; Haynes BF; Kelsoe G Polyreactivity and Autoreactivity among HIV-1 Antibodies. *J. Virol* 2015, 89, 784–798. [PubMed: 25355869]
- (29). Ritchie ME; Phipson B; Wu D; Hu Y; Law CW; Shi W; Smyth GK Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* 2015, 43, e47. [PubMed: 25605792]
- (30). MacLeod A Measuring the gain of a photomultiplier tube. Master's thesis, McGill University, Montreal, Quebec, Canada, 2007.
- (31). Flyckt S-O; Marmonier C Photomultiplier Tubes Principles and applications; Photonis: Brive, France, 2002; p 311.
- (32). Photomultiplier Tubes: Basics and Applications, 3a ed.; Hamamatsu Photonics K.K. Electron Tube Division: 2007; p 308.
- (33). Lung K; Arisaka K; Bargetzi A; Beltrame P; Cahill A; Genma T; Ghag C; Gordon D; Sainz J; Teymourian A; Yoshizawa Y Characterization of the Hamamatsu R11410–10 3-in. photomultiplier tube for liquid xenon dark matter direct detection experiments. *Nucl. Instrum. Methods Phys. Res., Sect. A* 2012, 696, 32–39.
- (34). Storey J A direct approach to false discovery rates. *Journal of the Royal Statistical Society Series B-Statistical Methodology* 2002, 64, 479–498.
- (35). Storey J; Taylor J; Siegmund D Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *Journal of the Royal Statistical Society Series B-Statistical Methodology* 2004, 66, 187–205.
- (36). Storey J; Dai J; Leek J The optimal discovery procedure for large-scale significance testing, with applications to comparative microarray experiments. *Biostatistics* 2007, 8 (2), 414–432. [PubMed: 16928955]
- (37). Robin X; Turck N; Hainard A; Tiberti N; Lisacek F; Sanchez JC; Müller M pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinf.* 2011, 12, 77.

Figure 1.

Representative images of protein array probed with the human IgG isotype control scanned at different PMT gain settings. Human IgG isotype control was hybridized to ProtoArrays microarrays according to the manufacturer's instruction. Arrays were scanned at nine different PMT gain settings from 250 up to 650 at an interval of 50. The first blocks of one array scanned at the PMT voltages of 250, 450, and 650 are shown here. Each circular spot indicates the signal generated between one single feature protein and the antibody. A black spot means no detectable interaction and a white dot is the saturated signal.

Figure 2.

Distribution of signal intensities of noncontrol proteins in protein microarray. (A) and (B) The signal distributions of noncontrol proteins in arrays probed by the monoclonal anti-histone H2B antibody and the IgG isotype control. The array data were acquired with the PMT voltage of 600. Then the raw (A) or RLM normalized (B) data were plotted to show the signal distributions of noncontrol proteins. (C) and (D) The signal distributions of noncontrol proteins in microarrays probed by the autoreactive and control sera. The data were obtained from the work done by Sboner et al.⁹ In their work, the arrays were scanned and acquired at two different PMT gain voltages (550 and 700). We used the data acquired at the 700 voltage for this figure. The raw (C) or RLM normalized data (D) were plotted to show the signal distributions of noncontrol proteins.

Figure 3.

Gain-scan power function fitting of protein microarray data. (A) and (B) Fitting of IgG isotype control probed array data. Data were acquired under nine different PMT settings, from 250 to 650, and then were fitted with the gain-scan power function model. Only blocks 1–8 of the control protein data are shown here. We show the individual fitting for each protein feature separately (A) and also the overlaid fitting (B) by shifting the curves along the x -axis toward the reference line (see explanations in the Method and Material section). (C) and (D) Fitting of polyclonal sera probed array data from the work done by Sboner et al.⁹ In their work, the arrays were acquired with two PMT voltages, 550 and 700. The data were fitted with the gain-scan power function model. Blocks 1–8 of the control protein data were plotted. We show the individual fitting (C) and overall fitting (D) by overlaying the curves on top of the reference line. The dots are the raw readings and the lines the fitted curves.

Figure 4.

Intra-array variances of microarray data before and after the gain-scan fitting and the RLM normalization. The IgG isotype control sample was hybridized in triplicate to protein arrays. The data were acquired at nine different PMT voltages, from 250 to 650, and then were processed using the gain-scan procedure, the RLM procedure, or both. The intra-array variance was calculated for each feature on arrays. (A), (B), and (C) show the decrease of intra-array variances after processing. The intra-array variances are connected by a line to identify the effect of processing. Red circles: intra-array variances of the raw data. Black dots: intra-array variances of the processed data (RLM normalized, gain-scan fitted, or both). (A) Intra-array variances before and after the RLM normalization. (B) Intra-array variances before and after the gain-scan fitting. (C) Intra-array variances before and after both procedures. (D) The distribution of intra-array variances of the isotype IgG control-probed arrays. (E) The distribution of intra-array variances of the arrays probed with the normal sera from Sboner et al.⁹

Figure 5.

Inter-array variances of microarray data before and after the gain-scan fitting and the RLM normalization. The IgG isotype control antibodies were hybridized to the protein array in triplicate. The protein array data were acquired at nine different PMT settings, from 250 to 650, and then were processed by the gain-scan fitting, the RLM normalization, or both. The inter-array variance was calculated for each feature protein across multiple arrays. The data sets from IgG isotype control antibodies and the polyclonal sera of Sboner et al. work⁹ are both shown here. (A), (B), and (C) show the decrease of inter-array variances after the gain-scan fitting, the RLM normalization, or both. The inter-array variances are connected by a line to identify the effect of processing. Red circles: inter-array variances of the raw data. Black dots: inter-array variances of the processed data. (A) Inter-array variances before and after the RLM normalization. (B) Inter-array variances before and after the gain-scan fitting. (C) Inter-array variances before and after both. (D) Distribution of inter-array variances of IgG isotype control data. (E) Distribution of inter-array variances of the polyclonal sera data from Sboner et al.⁹

Figure 6.

Effect of gain-scan on sample separability and differential analysis for the profiling data with technical replicates. Two data sets were analyzed. The first one was the microarray data profiling the polyclonal human sera from the work by Sboner et al.,⁹ in which positive control serum was compared with normal serum. The positive control was obtained by mixing six independent human sera reactive to different human antigens, and the normal serum was from a subject without known pathology. Six technical replicates for each sample were used in the experiment. The second data set was generated in this work profiling the monoclonal anti-histone H2B antibody and the IgG isotype control. Three technical replicates were used in the experiment. The data were acquired and then processed with the RLM, the gain-scan fitting, or both. Then, Fisher's signal-to-noise ratios, t tests, and FDR q -

values comparing different groups were computed. Results of different processing methods including unprocessed raw data are displayed to compare their performance. (A) The box plot for Fisher's signal-to-noise ratios comparing the polyclonal positive control sera with the normal serum by Sboner et al.⁹ (B) The box plot for Fisher's signal-to-noise ratios comparing the monoclonal anti-histone H2B with the IgG isotype control antibodies. (C) The volcano plot showing $-\log_{10}(q)$ versus fold change for the profiling data between the anti-histone H2B and IgG isotype control before and after processing. Each symbol on the figure indicates the values for one feature. The filled symbols are the data for the nine histone H2B isoform protein features (See the Supporting Information Figure S6 for separated volcano plots.) (D) The true discovery rate for different data processing methods. The true positive targets were defined on the basis of the signals of the nine histone H2B isoforms available on the ProtoArray (Table 1). The criteria for positive hits were a log fold change (FC) larger than 5.0 and a FDR q -value smaller than 0.003. The areas under the curve (AUC) were also obtained and used to assess the performance. For (C) and (D), the color of symbols indicates how the data is processed, either unprocessed raw, RLM normalized, gain-scan, or RLM and gain-scan processed.

Figure 7.

Effect of gain-scan on sample separability and differential analysis for serum profiling data with biological replicates. Five independent multiple sclerosis (MS) sera and five normal ones were obtained. Protein microarrays were probed with these sera at 1:500 dilution. The data were acquired at different PMTs and integrated with the gain-scan method. Finally, Fisher's signal-to-noise ratios, t tests, and FDR q -values were computed. The results using the profiling data acquired at one single PMT of 650 were also obtained and displayed as the raw results. (A) The example gain-scan power model fitting. The fitting using the data in the first eight blocks of one MS serum profiled array was displayed. The fittings for different features were scaled and overlaid to show the fitting. (B) The box plot for the intra-array variances. (C) The box plot for the inter-array variances. The variances were computed using either the unprocessed data acquired with a single PMT or the gain-scan integrated data. (D) Fisher's signal-to-noise ratios comparing the MS sera with the normal sera ($n = 5$ for each group). (E) The true discovery plot for identification of the MS biomarkers. The criteria for positive hits were a minimum signal intensity higher than 1500 RFU, a log fold change (FC) larger than 1.4, and a p -value smaller than 0.1. The true serum biomarkers were determined on the basis of the previous work done by DeMarshall et al.¹⁴ (Supporting Information Table

S4). The areas under the curve (AUC) were also obtained for each curve and used to assess the performance.

FDR q -Values and Fold Changes for Histone H2B Isoform Features Comparing the Anti-Histone H2B Antibody Probed Arrays with the Isotype Control Ones^{ab}

Table 1.

histone H2B isoforms	raw		RLM		multigain		multigain + RLM	
	$-\log_{10}(q)$	$\log_2(\text{FC})$	$-\log_{10}(q)$	$\log_2(\text{FC})$	$-\log_{10}(q)$	$\log_2(\text{FC})$	$-\log_{10}(q)$	$\log_2(\text{FC})$
HIST1H2BD	3.52	7.39	3.84	7.77	4.02	7.86	4.03	8.06
HIST1H2BN	2.54	7.96	3.07	8.33	3.36	9.29	3.36	9.48
HIST1H2BM	2.28	7.42	2.83	7.81	3.10	8.58	3.08	8.77
HIST1H2BH	2.16	8.22	2.76	8.60	3.10	9.73	3.04	9.92
HIST1H2BA	3.48	8.91	3.70	9.29	4.02	9.98	3.71	10.16
HIST1H2BO	2.39	7.08	2.92	7.47	3.50	8.88	3.38	9.07
HIST1H2BB	3.51	8.07	3.69	8.46	3.62	9.28	3.42	9.48
HIST3H2BB	3.57	7.87	3.75	8.25	3.62	9.14	3.42	9.33
HISTONE	3.23	4.61	3.36	5.00	3.61	5.57	3.39	5.76

^aThe abbreviations are listed here: HIST1H2BA, histone H2B type 1-A; HIST1H2BB, histone H2B type 1-B; HIST1H2BD, histone H2B type 1-D; HIST1H2BH, histone H2B type 1-H; HIST1H2BM, histone H2B type 1-M; HIST1H2BN, histone H2B type 1-N; HIST1H2BO, histone H2B type 1-O; HIST3H2BB, histone H2B type 3-B; FC, fold change; $-\log(q)$, the negative logarithm of the FDR q -value.

^bThe ProtoArrays were applied to profile the monoclonal anti-histone H2B and IgG isotype control antibodies ($n = 3$ for each group), and the data were acquired with multiple gain voltages (350, 450, 550, 600, 650, and 700). The raw data or processed data (RLM normalized, gain-scan fitted, or both) were used for further statistical analyses. Student's t -tests, FDR q -values, and fold changes comparing the two groups were computed for all feature proteins. These values of the nine histone H2B isoform features are listed.